Uffe Kock Wiil

Editor

Counterterrorism and Open Source Intelligence

SpringerWien NewYork

Lecture Notes in Social Networks
(LNSN)

Uffe Kock Wiil
Editor

Counterterrorism and Open Source Intelligence

SpringerWienNewYork

Editor
Uffe Kock Wiil
The Maersk McKinney Moller Institute
University of Southern Denmark
Campusvej 55, 5230 Odense
Denmark
ukwiil@mmmi.sdu.dk

Contents

Part II   Tools and Techniques

Region-Based Geospatial Abduction with Counter-IED Applications **105**
Paulo Shakarian and V.S. Subrahmanian

**Finding Hidden Links in Terrorist Networks by Checking
Indirect Links of Different Sub-Networks** **143**
Alan Chen, Shang Gao, Panagiotis Karampelas, Reda Alhajj,
and Jon Rokne

**Counterterrorism Mining for Individuals Semantically-Similar
to Watchlist Members** **223**

James A. Danowski

Detection of Illegitimate Emails Using Boosting Algorithm **249**

Sarwat Nizamani, Nasrullah Memon, and Uffe Kock Wiil

Part III   Case Studies

Part IV   Alternative Perspective

Contributors

Sara Aghakhani Computer Science Department, University of Calgary, Calgary, Alberta, Canada

Reda Alhajj Computer Science Department, University of Calgary, Calgary, Alberta, Canada
and
Department of Computer Science, Global University, Beirut, Lebanon
and
Department of Information Technology, Hellenic American University, NH, USA;
email: alhajj@ucalgary.ca

Carlo Aliprandi Synthema Language & Semantic Intelligence, Pisa, Italy;
email: carlo.aliprandi@synthema.it

Martin Atkinson Joint Research Centre of the European Commission, Ispra, Italy;
email: martin.atkinson@jrc.europa.eu

Claus Atzenbeck Institute of Information Systems, Hof University, Hof, Germany;
email: claus.atzenbeck@iisys.de

Peter Borm Tilburg University, Tilburg, the Netherlands; email: p.e.m.borm@uvt.nl

M.A.J. Bourassa Royal Military College of Canada, Kingston, Ontario, Canada

Patricia L. Brantingham School of Criminology, Simon Fraser University, Burnaby, British Columbia, Canada; email: pbrantin@sfu.ca

Furio Camillo Department of Statistical Sciences, Faculty of Economics, University of Bologna, Bologna, Italy; email: furio.camillo@unibo.it

Ahmet Celik Diyarbakir A. Gaffar Okkan Vocational School, Turkish National Police, Diyarbakir, Turkey; email: acelik@rutgers.edu

Hsinchun Chen Artificial Intelligence Lab, Management Information Systems Department, The University of Arizona, Tucson, AZ, USA;
email: hchen@eller.arizona.edu

Alan Chen Computer Science Department, University of Calgary, Calgary, Alberta, Canada

James A. Danowski Department of Communication, University of Illinois at Chicago, Chicago, Illinois, USA; email: jimd@uic.edu

Khaled Dawoud Computer Science Department, University of Calgary, Calgary, Alberta, Canada

Douglas Demetrick Departments of Pathology, Oncology and Biochemistry & Molecular Biology, University of Calgary, Calgary, Alberta, Canada

Abdallah M. ElSheikh Computer Science Department, University of Calgary, Calgary, Alberta, Canada

Martin Ester School of Computing Science, Simon Fraser University, Vancouver, British Columbia, Canada; email: ester@cs.sfu.ca

Richard Frank School of Criminology, Simon Fraser University, Burnaby, British Columbia, Canada; email: rfrank@sfu.ca

Shang Gao Computer Science Department, University of Calgary, Calgary, Alberta, Canada

Nadir Gergin Diyarbakir Police Department, Turkish National Police, Diyarbakir, Turkey; email: nadirgergin@yahoo.com

Uwe Glässer School of Computing Science, Simon Fraser University, Vancouver, British Columbia, Canada; email: glaesser@cs.sfu.ca

Jolanta Gniadek Counterterrorism Research Lab, The Maersk Mc-Kinney Moller Institute, University of Southern Denmark, Odense, Denmark

Erik Van der Goot Joint Research Centre of the European Commission, Ispra, Italy; email: erik.van-der-goot@jrc.europa.eu

Alexander Gutfraind Center for Nonlinear Studies and T-5, Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM, USA; email: agutfraind.research@gmail.com

Herbert Hamers Tilburg University, the Netherlands; email: h.j.m.hamers@uvt.nl

Panagiotis Karampelas Department of Information Technology, Hellenic American University, Manchester, NH, USA and Hellenic Air Force Academy, Athens, Greece; email: pkarampelas@gmail.com

Keivan Kianmehr Computer Science Department, University of Calgary, Calgary, Alberta, Canada

Roy Lindelauf Netherlands Defence Academy, Breda, the Netherlands and Tilburg University, Tilburg, the Netherlands; email: rha.lindelauf.01@nlda.nl

Nasrullah Memon Counterterrorism Research Lab, The Maersk Mc-Kinney
Moller Institute, University of Southern Denmark, Odense, Denmark
and
Hellenic American University, Manchester, NH, USA;
email: memon@mmmi.sdu.dk

Mohamad Nagi Department of Computing, School of Computing Informatics &
Media, University of Bradford, Bradford, UK

Ghada Naji Department of Biology, Faculty of Sciences III, Lebanese University,
Tripoli, Lebanon

Federico Neri Synthema Language & Semantic Intelligence, Pisa, Italy;
email: federico.neri@synthema.it

Sarwat Nizamani Counterterrorism Research Lab, The Maersk Mc-Kinney
Moller Institute, University of Southern Denmark, Odense, Denmark
and
University of Sindh, Jamshoro, Pakistan; email: saniz@mmmi.sdu.dk

Fatih Ozgul Faculty of Computing, Engineering and Technology, University of
Sunderland, Sunderland, UK; email: fatih.ozgul@istanbul.com

Tansel Özyer Computer Science Department, University of Calgary, Calgary,
Alberta, Canada

Jakub Piskorski Research & Development Unit, Frontex, Warsaw, Poland;
email: jakub.piskorski@frontex.europa.eu

Pir Abdul Rasool Qureshi Counterterrorism Research Lab, The Maersk
Mc-Kinney Moller Institute, University of Southern Denmark, Odense, Denmark;
email: parq@mmmi.sdu.dk

Christopher J. Rhodes Networks and Complexity Programme, Institute of Math-
ematical Sciences and Imperial College Institute for Security Science and Technol-
ogy, Imperial College London, UK; email: c.rhodes@imperial.ac.uk

Mick Ridley Department of Computing, School of Computing Informatics &
Media, University of Bradford, Bradford, UK

Jon Rokne Computer Science Department, University of Calgary, Calgary,
Alberta, Canada

Paulo Shakarian Department of Computer Science, University of Maryland,
College Park, MD, USA; email: fpshak@cs.umd.edu

David Skillicorn School of Computing, Queen's University, Kingston, ON,
Canada; email: skill@cs.queensu.ca

Robert David Steele Earth Intelligence Network, Oakton, VA, USA;
email: robert.david.steele.vivas@gmail.com

V.S. Subrahmanian Department of Computer Science, University of Maryland, College Park, MD, USA; email: vsg@cs.umd.edu

Mohammad A. Tayebi School of Computing Science, Simon Fraser University, Vancouver, British Columbia, Canada; email: tayebi@cs.sfu.ca

Uffe Kock Wiil Counterterrorism Research Lab, The Maersk Mc-Kinney Moller Institute, University of Southern Denmark, Odense, Denmark; email: ukwiil@mmmi.sdu.dk

Gordon Woo Risk Management Solutions, London, UK; email: Gordon.Woo@rms.com

Roman Yangarber Department of Computer Science, University of Helsinki, Helsinki, Finland; email: roman.yangarber@cs.helsinki.fi

Counterterrorism and Open Source Intelligence: Models, Tools, Techniques, and Case Studies

Uffe Kock Wiil

1 Introduction

Terrorism is a serious threat to international peace and security. No nation can consider themselves immune from the dangers of terrorism. No nation can disengage themselves from the efforts to combat terrorism.

In order to combat terrorism, law enforcement and intelligence officials are faced with the overwhelming tasks of finding, integrating, and making sense of critical pieces of information from the vast amount of information collected and processed from classified and open sources.

Open source information is increasingly important to support intelligence tasks. The increasing ability to successfully mine data from large, incoherent sources, allows analysts to build up detailed composite overviews of their targets. Open source analysis is not a substitute for traditional classified work. Analysts can use the generated open source overview in combination with information from other (classified) sources to build a comprehensive all source overview of a particular target.

As the amount of available information rapidly increases (particularly from open sources), counterterrorism officials increasingly depend upon advanced software tools to collect and process the information in an effective and efficient manner.

An increasing number of researchers are engaging in the fight against terrorism:

- By developing new ways to model terrorism information
- By developing new techniques to collect, filter, process, store, mine, structure, relate, analyze, interpret, and visualize terrorism information
- By developing software tools that can assist law enforcement and intelligence officials in supporting (automating) critical and time-consuming terrorism information and knowledge management tasks

U.K. Wiil (✉)

Counterterrorism Research Lab, The Maersk Mc-Kinney Moller Institute, University of Southern Denmark, Odense, Denmark
e-mail: ukwiil@mmmi.sdu.dk

- By demonstrating the usefulness of new approaches in case studies – often jointly performed with law enforcement and intelligence officials to ensure the validity of the approaches on real data sets etc.

Overall, this type of research provides valuable contributions to the next generation of counterterrorism tools.

This book includes 20 chapters from some of the leading research groups engaged in terrorism informatics research. The chapters provide a state-of-the-art overview of models, tools, and techniques for counterterrorism and open source intelligence supported by several prominent case studies. The chapters demonstrate the applicability of terrorism informatics research to a number of practical application areas:

- Modeling transnational terrorist groups
- Knowledge discovery for intelligence analysts
- Interdiction of plots
- Understanding terrorist network topologies
- Crime data analysis and mining
- Countering Improvised Explosive Devices (IEDs)
- Finding hidden links in networks
- Prediction of network structure
- Measuring link importance in networks
- Measuring organizational strength of networks
- Identification of potential new terrorist watchlist members
- Detection of suspicious emails
- Studying biological agents in relation to bioterrorism
- Dark Web information collection and processing
- Monitoring border security
- Using telephone Call Detail Records (CDRs) in investigations
- Web sentiment mining
- Detecting new trends in terrorist organizations

2 Organization

The book is organized into four parts. The first part (Models) includes five chapters that are primarily related to proposing new models of terrorism information. Part II (Tools and Techniques) includes nine chapters that primarily focus on presenting state-of-the-art tools and techniques for counterterrorism and open source intelligence. The next part (Case Studies) include five chapters that focus on demonstrating the applicability of terrorism informatics research through prominent case studies. Finally, Part IV (Alternative Perspective) includes one chapter that provides a different perspective on how to view and cope with terrorism in a broader context.

2.1 Models

Terrorist groups often operate across international borders where different countries host different stages of terrorist operations. Stopping such attacks is difficult because intervention in any region or route might merely shift the terrorists elsewhere. In the chapter "Targeting by Transnational Terrorist Groups," Gutfraind proposes a model of transnational terrorism based on the theory of activity networks. The model represents attacks on different countries as paths in a network. The group is assumed to prefer paths of lowest cost (or risk) and maximal yield from attacks.

In the chapter "A Framework for Analyst Focus from Computed Significance," Skillicorn and Bourassa argue that attention is the critical resource for intelligence analysts and that tools should help provide focus. One way to determine focus is by computing significance. The authors propose a framework for understanding significance and explore its impact on the knowledge discovery process for intelligence analysts.

Elaborate computer models for terrorism threat prediction have been criticized for preconceived false notions of how terrorist networks behave. In the chapter "Interdiction of Plots with Multiple Operatives," Woo presents a basic quantitative model of plot interdiction which is grounded on terrorist network behavior. The model captures the key elements of the interdiction process and can be used to support risk-informed public policy decisions.

The chapter by Lindelauf, Borm, and Hamers ("Understanding Terrorist Network Topologies and their Resilience against Disruption") investigates the structural nature of covert (terrorist or criminal) networks. Using the secrecy versus information tradeoff characterization of covert networks, Lindeluaf, Borm, and Hamer show that network structures are generally not small-worlds, in contradistinction to many overt social networks.

The chapter "Co-offending Network Mining" by Brantingham et al. propose a computational framework for co-offending network mining defined in terms of a process that combines formal data modeling with data mining of large crime and terrorism data sets as gathered and maintained by law enforcement and intelligence agencies. Crime data analysis aims at exploring relevant properties of criminal networks in arrest-data. The framework is evaluated based on 5 years of real-world crime data that was made available for research purposes. This data was retrieved from a large database system with several million data records keeping information for the regions of the Province of British Columbia (Canada).

2.2 Tools and Techniques

Criminologists use geographic profiling and crime pattern theory to determine locations of criminals based on where the criminal activity occurred. Military analysts try to find terrorist safe houses and weapons locations by studying where the attacks occurred. In the chapter "Region-based Geospatial Abduction with

Counter-IED Applications," Shakarian and Subrahmanian study how the theory of geospatial abduction problems can be used to locate weapons caches in Baghdad based on IED attack locations.

Modeling and analyzing criminal and terrorists networks is a challenging problem that has attracted considerable attention from academia, industry, and government institutions. In the chapter "Finding Hidden Links in Terrorist Networks by Checking Indirect Links of Different Sub-Networks," Chen et al. propose a method to identify hidden links between nodes in a network. The method was evaluated on multiple small terrorism data sets.

Open source intelligence is playing an increasing role in helping agencies responsible for national security to determine the characteristics, motivations, and intentions of adversary groups that threaten the stability of civil society. In the chapter "The Use of Open Source Intelligence in the Construction of Covert Social Networks," Rhodes presents a statistical inference method to maximise the insight that can be gained into the structure of covert social networks from the limited and fragmentary data gathered from intelligence operations or open sources.

In the chapter "A Novel Method to Analyze the Importance of Links in Terrorist Networks," Wiil, Gniadek, and Memon propose a new terrorist network analysis method inspired by previous work from social network analysis, graph theory, and transportation networks. The proposed link importance measure is implemented in CrimeFighter Assistant and evaluated based on known terrorist networks harvested from open sources.

The organizational structure and critical members of a group are key indicators in determining its strengths and weaknesses. In the chapter "A Global Measure for Estimating the Degree of Organization and Effectiveness of Individual Actors with Application to Terrorist Networks," Aghakhani et al. presents a novel approach for extracting structural patterns of terrorist networks based on social network analysis measurements and techniques. A global organization measure is proposed in order to estimate the degree of organization of a social network.

A key counterterrorism problem is how to identify people that should be added to a watchlist even though they have no direct communication with its members. In the chapter "Counterterrorism Mining for Individuals Semantically-Similar to Watchlist Members," Danowski presents a method for locating individuals in discussion forums who have highly similar semantic networks to some reference network. A Pakistani discussion forum is used to test the method.

In the following two chapters "Detection of Illegitimate Emails using Boosting Algorithm" and "Cluster Based Text Classification Model," Nizamani, Memon, and Wiil demonstrate how various data mining techniques (classification and clustering) can be used to detect suspicious emails. Based on various data sets, it is demonstrated that the proposed algorithms and models performs better than existing algorithms and models.

Bioterrorism involves the aggressive and planned release of biological agents and is a serious concern for humanity. In the chapter "Effectiveness of Social Networks for Studying Biological Agents and Identifying Cancer Biomarkers," Naji et al. propose a model to study social networks of genes within the body leading to the

identification of disease biomarkers. The model is used to identify potential cancer diagnostic biomarkers. The methodology can also be applied to identify potential biomarkers of other diseases (including outbreaks related to bioterrorism).

2.3   Case Studies

In the chapter "From Terrorism Informatics to Dark Web Research," Chen provides an overview of terrorism informatics including several critical books that lay the foundation for studying terrorism in the new Internet era and important terrorism research centers and resources that are of relevance to the Dark Web project conducted at The University of Arizona. The Dark Web project is a long-term scientific research program that aims to study and understand the international terrorism (Jihadist) phenomena via a computational, data-centric approach. Information from various sources are collected and processed – including web sites, forums, chat rooms, blogs, social networking sites, videos, virtual worlds, etc.

Telephone CDR are commonly used by police and intelligence services all over the world. In the chapter "Investigating Terrorist Attacks using CDR Data: A Case Study," Ozgul et al. presents a case study that shows how mining CDR data helped in the investigation of a terrorist attack that happened in Istanbul, Turkey in 2007. A truck was put on fire by a terrorist organization to protest against the conditions of a terrorist leader in prison. Arsonists were identified and arrested after the interrogation of suspects.

The chapter "Multilingual Real-Time Event Extraction for Border Security Intelligence Gathering" by Atkinson et al. gives an overview of tools developed for Frontex, the European Agency for the Management of Operational Cooperation at the External Borders of the Member States of the European Union, to facilitate the process of extracting structured information on events related to border security from on-line news articles, with a particular focus on incidents and developments in the context of illegal migration, cross-border crime, and related crisis situations at the EU external borders and in third countries.

Web sources are more accessible, ubiquitous, and valuable than ever before. But the most valuable information is often hidden and encoded in blog posts or pages, which are neither structured, nor classified, being free textual. In the chapter "Mining the Web to Monitor the Political Consensus," Neri, Aliprandi, and Camillo describes a sentiment mining case study performed on over 1,000 news articles and forum/blog posts, concerning the Italian Prime Minister Silvio Berlusconi's involvement in an escort scandal.

The chapter "Exploring the Evolution of Terrorist Networks" by Memon et al. discusses new trends in terrorist networks. A case study regarding a recent terror plan against a target in Denmark is investigated. Based on open source information, a part of the terrorist network centered on David Headley is mapped. Despite its deficiencies, the network provides insight into new trends in terrorist organizations and people involved in terrorist plots.

2.4 Alternative Perspective

This part provides an alternative perspective on counterterrorism and open source intelligence given by one of the open source intelligence pioneers, Robert Steele.

Steele argues in his chapter "The Ultimate Hack: Re-Inventing Intelligence to Re-Engineer Earth" that terrorism is a symptom and not a root cause or threat. Steele lists the ten high-level threats to humanity within which terrorism only comes in as number nine. Steele then goes on to outline a strategy to address all ten high-level threats to humanity.

3 Conclusion and Acknowledgments

The terrorism informatics research field is an active field of research dedicated towards innovating, developing, deploying, and evaluating new models, software tools, and techniques to help in the fight against terrorism. This book provides a comprehensive overview of state-of-the-art models, tool, techniques, and case studies related to counterterrorism and open source intelligence.

I would like to acknowledge the efforts of all those who helped in the creation of this book. Firstly, it would not have been possible for a book such as this one to provide such a broad and extensive overview of the latest research without the important contributions made by those expert researchers and practitioners who have authored and contributed chapters. Secondly, the support, guidance, and encouragement from the LNSN book series editors (Nasrullah Memon and Reda Alhajj) and the people at Springer (in particular Stephen Soehnlen and Wolfgang Dollhäubl) is greatly acknowledged. Finally, Asadullah Shaikh provided invaluable support with the layout, typesetting, and formatting of this book. Thanks Asad!

Part I
Models

Targeting by Transnational Terrorist Groups

Alexander Gutfraind

Abstract Many successful terrorist groups operate across international borders where different countries host different stages of terrorist operations. Often the recruits for the group come from one country or countries, while the targets of the operations are in another. Stopping such attacks is difficult because intervention in any region or route might merely shift the terrorists elsewhere. Here, we propose a model of transnational terrorism based on the theory of activity networks. The model represents attacks on different countries as paths in a network. The group is assumed to prefer paths of lowest cost (or risk) and maximal yield from attacks. The parameters of the model are computed for the Islamist–Salafi terrorist movement based on open source data and then used for estimation of risks of future attacks. The central finding is that the United States (US) has an enduring appeal as a target, due to lack of other nations of matching geopolitical weight or openness. It is also shown that countries in Africa and Asia that have been overlooked as terrorist bases may become highly significant threats in the future. The model quantifies the dilemmas facing countries in the effort to cut terror networks, and points to a limitation of deterrence against transnational terrorists.

1 Introduction

Despite vast investments in counter-terrorism, victory in the global war on terror remains elusive. In part, this is because terrorist groups are highly adaptive in their tactics and strategy. When airport scanners were installed to detect weapons and explosives, terrorists switched to explosives that cannot be detected using the

A. Gutfraind (✉)

Center for Nonlinear Studies and T-5, Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM, USA
e-mail: agutfraind.research@gmail.com

Fig. 1 Terrorist attacks by
Islamist groups over
1990–2008. The fraction of
all transnational plots that
originated at country i and
targeted country j is
proportional to the area of a
circle at coordinates (i, j).
The source countries on the
horizontal axis account for
>99% of all attacks against
developed countries, the
countries in the Organisation
for Economic Cooperation
and Development, the OECD
(vertical). The dataset is very
small: only 82 incidents fit
the criteria

Countries of origin
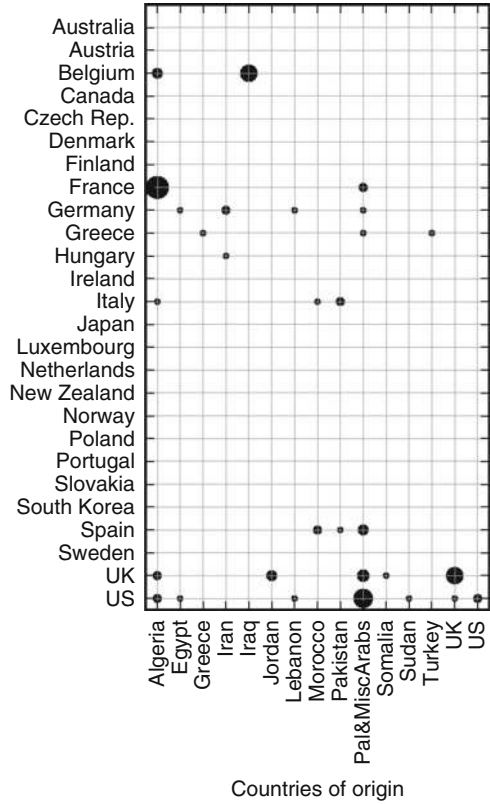
scanners and to other modes of attack [12, 27]. When it became harder to reach US soil or attack US embassies, groups shifted to attacks against other countries or less fortified installations [11, 43]. Like international businesses, globalized terrorist groups are vast international enterprises that tap into the most successful business practices and cost-efficient solutions [3]. If a country erects high barriers to entry or develops an effective domestic counter-terrorism response then terrorists switch their targeting to a safer and more accessible place. If a country no longer provides a haven for recruitment, training and planning of operations, those will shift elsewhere [24].

Adaptability makes risk estimation challenging. One possible basis for risk assessment is extrapolation of historical data, such as the ITERATE dataset of transnational attacks [33]. Figure 1 shows all ITERATE attacks carried out by Islamist groups on developed countries in which the national origins of the attackers are known. Many of the incidents in the matrix are due ethno-nationalist conflicts, such as the GIA attacks in France or due to attacks by home-grown cells inspired by Salafis. While a substantial fraction of attacks were against the US, many attacks also targeted Germany, the UK and other countries.

The rest of the paper will introduce another method for risk assessment: a quantitative network-based model. The model takes demographic and economic information pertaining to violent Islamist–Salafi groups – the most probable source of future attacks – and estimates the risk of various transnational terror plots. The model suggests that the future of transnational terrorism may be substantially different from the past:

1. Several regions will become large new sources of transnational terrorism (Sect. 4)
2. The US will rise as the terrorists' preferred attack destination (Sect. 4)
3. Successes in stopping foreign-based plots against the US will increase the threat to other countries (Sect. 4.1)
4. Deterrence will be hard to achieve (Sect. 4.2)

The model is based on an activity network for stages of terrorist attacks. The network represents decisions required for terrorist operations on the global scale, such as which country to attack. No distinction is made between "transnational" and "international" terrorism, both referring to terrorist groups that operate using foreign bases, support or inspiration. This coarsened scale of analysis exposes the strategic picture and can guide counter-terrorism decision making at the national and international levels. It also quantifies a kind of unintended effect from counter-terrorism measures known as "transboundary externality" [36, 37]: the redirection of terrorists from one country to another, because the latter is less protected.

To the author's knowledge, this is the first model in the open literature that models transnational networks and estimates which countries might be selected for future attacks. Previous work considered target selection by terrorists but where targets are implicitly within a single country so the costs of bringing the attackers and their weapons to their target are negligible (see e.g. [6]). In contrast, for transnational terrorism security measures and international logistics play a central role in attack planning [22, Chap. 3]. Other work considered the structure of the terrorist networks at the level of individual operatives or functions, rather than as the global network presented here (cf. [9, 14, 19, 29, 44].)

The model's findings imply that policymakers across the world should increase their coordination of counter-terrorism measures and be watchful for several emerging threats. The model also advances methodology in the analysis of terrorism. Based on this network model, future work could examine the risk from specific groups and to additional types of targets.

2   A Model of Transnational Terrorism

Transnational terrorist groups are characterized by their global aims, as opposed to regional conflicts; they recruit and attack in several countries. Transnational Islamic fundamentalist groups will serve as the central application of this model. These include al-Qaida, Hezbollah but also possibly extremist groups not currently thought to be violent, such as Hizb-ut-Tahrir. Those groups are at the focus because

they are probably the most potent present-day transnational terrorist threat. Instead of looking at any particular fundamentalist group, we will estimate the risk from a potential world-wide violent Islamist movement. Other violent transnational ideologies or even a specific existing group can be quantified through this model by re-estimating the model's parameters.

It is sometimes argued that estimation of terrorism risk is near impossible because terrorism is irrational behavior. Indeed, how can one explain the fanaticism of suicide bombers? However, the preponderance of evidence supports the alternative view – the rational choice theory (RCT) [40]. RCT claims that terrorist groups and leaders are rational agents capable of strategic decision-making. Their decisions are expressions of "instrumental rationality", that is, in line with their values and objectives [26]. The sophistication and technological adaptability of terrorists, such as in developing triggers for explosives, is strong evidence for their intelligence [21, 24]. More evidence for RCT comes from studies of target selection [37]. Those consistently find evidence for a substitution effect – as governments improve protection to certain targets, terrorists substitute them with less protected targets [2, 10, 21]. Indeed, the defining feature of terrorism – the use of violence against civilians rather than against military targets – is a strategic substitution effect because the latter are harder targets. Another line of evidence for rationality comes from analyzing the internal dynamics of terrorist groups. Rather like nonviolent organizations, they perform cost-benefit analyses and produce volumes of documents [38, 39]. Even the behavior of suicide bombers is sometimes viewed as expression of selfish utility maximization, where the benefits of such acts are either to the agent in paradise or to family members on earth (for a nuanced "rationally irrational" model see [7], for a critique of the terrorist strategy see [1]). The strength of the rational choice model is its predictive power. One can not only correctly anticipate which strategies terrorists would adopt (see e.g. [5]) but also how counter-terrorist policies would affect those strategies.

2.1   Operations Submodel

Suppose a transnational group controls a cell in a country, and must decide where to dispatch this cell (the cell might also be self-mobilizing, in which case it must solve its own targeting problem.) The three options are (1) do a domestic attack, (2) send the cell to attack another country (a transnational attack), and (3) do nothing. Option (1) entails certain risks and costs for collecting intelligence and preparing weapons. Dispatching the cell into another country, (2), incurs the additional cost and risk of interception due to security barriers, such a visas, intelligence collection in a foreign environment, and cultural difficulties. However, the other country might have more favorable security environment or offer better targets – more significant or less protected. Option (3) – abandon the attack and hide – has little or no risk or accounting cost and preserves the cell for future operations.
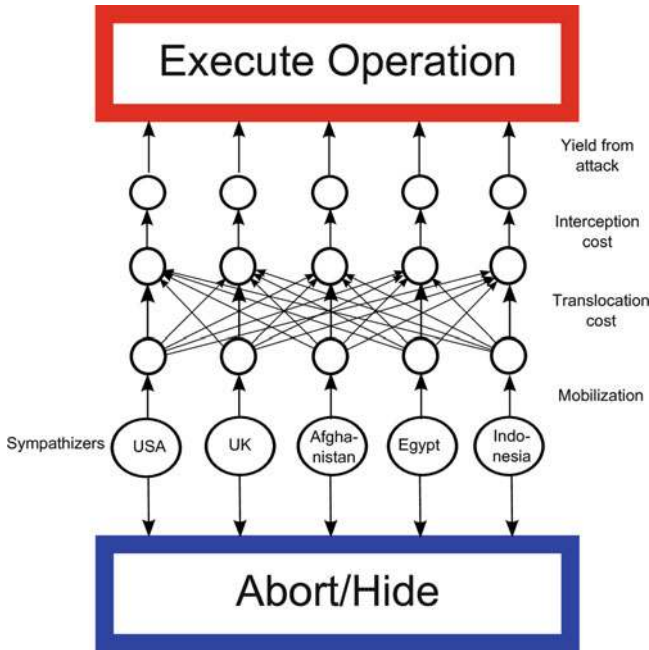
Fig. 2 Illustration of the model for five countries. The vertical direction represents countries while the horizontal represents different stages in execution of plots. Domestic attack plans correspond to motion upwards, while transnational attacks also make a diagonal transition. The full model includes many more countries

Any rational decision maker must weight the costs and benefits, and take the action offering the greatest net benefit. Surely then terrorists would also do such analysis, weighing at least the most obvious target choices and travel routes. A simple way of representing this is with an activity network, where nodes represent different stages of terrorist operations at different countries, and edges show the cost and risk involved in each stage (Fig. 2). In this figure, the vertical direction corresponds to the countries, including the country of origin, while the horizontal direction corresponds to the postures of the cells: the stages of the plots.

The network represents the options of the rational decision maker as directed paths – chains of nodes and directed edges that start in the source country node and lead to either the "attack" node or the "abandon/hide" node. If complete information is available about the costs and benefits of each option, then the rational decision is to select the path with the highest utility, that is the path with highest net benefit (benefit minus cost). For any path p, the cost $c(p)$ is found by adding the weights on the edges (tasks) constituting the path.

The edge weights of this network could represent resources like money or materiel that are consumed and produced by terrorist operations. The network could also be used to perform a probabilistic risk assessment (PRA): to evaluate

the gains from possible operations and the probabilities of successfully completing intermediate stages in the operations. Such a PRA is what we will do. In other words we will take the perspective of the terrorists: determine what they want and what they fear in order to anticipate how they will act.[1]

It is an open question whether terrorist groups can or will use such an algebraic method to analyze their operations. However, given their sophistication they may well come to the same decisions using other means or through operational experience. Of course, they might also intentionally avoid the most probable attacks to achieve surprise, but only at a cost (and models could also be constructed to anticipate that too).

Consider now the following specific model for transnational terrorist attacks, Fig. 2. A cell that was mobilized at country i experiences (1) the translocation cost/risk, T_{ij}, representing the barriers for moving from country i to country j; (2) the risk of interception at country j, I_j; and (3) the yield Y_j from attacks at country j. Yield reflects the gain to the terrorists from a successful attack, and so has the opposite sign from cost. A domestic attack at country i has cost $c(p) = T_{ii} + I_i + Y_i$ while a transnational attack has cost $c(p) = T_{ij} + I_j + Y_j$ ($i \neq j$). Because costs represent risks, the words "cost" and "risk" will be used interchangeably. Sometimes attackers reach country j through one or several intermediate countries (exploiting e.g. the Schengen treaty), a possibility we ignore for simplicity. From the counter-terrorism point of view, the likelihood of a particular plot depends also on the supply of operatives originating at each country. Therefore, we will estimate for each country i the number of cells that originate there, S_i. If the group decides to abandon, its path has cost $c(p) = A$. The parameter A may be a negative, representing the preservation of the cell, or positive, if the cell cannot be reactivated. It is possible to include in the model additional costs like cost of recruitment or training but this will be left for future studies because the data is hard to estimate.

The model's parameters can be estimated from open source information with a modest degree of confidence (see Sect. 3). Briefly, transit costs were estimated from data on global migration, the risk of interception from national expenditure on internal security and attack yields based on the political power of the targeted country, represented by its Gross Domestic Product (GDP). The supply of plots is estimated from public opinion surveys measuring support for terrorist attacks and from demographic data.

---

[1]Here is how PRA is represented by networks. Suppose in a multi-stage terrorist operation r_s is the probability of success at stage s (out of k stages in total) conditional on success at every previous stage. Suppose the gain from a successful operation is G (≥ 1). Then the *expected* gain from the operation is $E = r_1 r_2 \ldots r_k G$. Let us now relate r_s values to costs ($c_s \geq 0$) using exponentiation: $r_s = e^{-c_s}$, and let the gain be a function of yield Y: $G = e^{-Y}$. Thus, an attack has expected gain $E = \exp\left[-(c_1 + c_2 + \cdots + c_k + Y)\right]$. In the network representation of terrorist operations, we can compute the sums in the exponent by adding edge weights along network paths that trace through all the stages. Paths of lower weights translate to attacks of greater expected gain. By comparing such paths we could anticipate which attacks would have the highest expected gain.

2.2 Stochastic Decisions Submodel

If transnational terrorist groups could determine the values of the parameters precisely (the next section discusses this problem), then they should be able to plot the optimal attack from each country i by considering all possible options and finding the path that minimizes cost:

$$\min_{j} \left[\underbrace{T_{ij} + I_j + Y_j}_{c(p_{ij})} \right].$$

However, one of the general difficulties in decision making is uncertainties about costs and risks. Terrorists, like other decision makers should therefore occasionally identify the optimal attack incorrectly. Reliable risk assessment must therefore take into account the possibility that adversaries make mistakes (and ideally even the use of unpredictability to achieve surprise.) Fortunately, suitable stochastic prediction methods have already been developed for activity network models like in Fig. 2. With those methods probabilities can be assigned to different terrorist plans based on the costs of the corresponding paths on the network.

The methods to used here were introduced in [18, 20] and are based on Markov chains. In those chains, the path of least cost is typically assigned the highest probability but other paths have non-zero probabilities, and these probabilities can be quite high (for details see Appendix 2).[2] From this Markov chain model it is possible to compute the number of times any particular country would be targeted as well as to compute the changes in targeting due to various defensive actions, which are represented as increases in edge weights. It is also possible to determine whether defensive actions would materially increase the costs for the adversaries or merely lead them to change targets. Finally, the Markov chain model will incorporate our uncertainty about the parameters in the transnational network. Therefore, it automatically provides predictions that are robust under this uncertainty.

---

[2]One of the advantages of the stochastic model is that it can interpolate between the two extremes of complete ignorance and perfect information using a single parameter λ (≥ 0) that describes the amount of information available to the adversaries. For a given level of information, the probability that a path p would be selected is proportional to $\exp(-\lambda\, c(p))$. When λ is very large the path of least cost has a much higher probability than any of the alternatives, while λ close to 0 assigns all paths approximately the same probability. We set $\lambda = 0.1$ in the following but its value has a smooth effect on the predicted plots (i.e. the sensitivity is low). A value of 0.1 means that if the terrorist group learns of a increase in path cost by 10 units, its probability of taking the path will decrease by a multiplicative factor of ≈ 2.72. The exact amount of change depends on the original path probability: it is not as great a decrease when the original probability is high.

3 Estimation of Parameters

The model contains several sets of inputs: (1) the supply of plots at country i, S_i; (2) barriers for moving from country i to country j, T_{ij}; (3) risk of interception at country j, I_j; and (4) the yield from attacks at country j, Y_j. The yield of abandoning, A will be set to ∞ (no abandoned plots) and its effect will be analyzed separately. Because (1)–(4) contain security-related information that is also difficult to measure, the information is not public. Luckily, one can derive estimates from publicly-available demographic and economic data. Readers wishing to see the final results of estimates should open Tables 1–3 and skip the rest of this section.

Table 1 Information about countries: the interception cost (*left*) and the yields from attacks (*right*)

Country	Intercept. Cost I_j	Country	Yield Y_j
New Zealand	2.3	United States	−54.0
United Kingdom	2.1	Japan	−24.1
Czech Republic	1.6	Germany	−9.5
Hungary	1.6	United Kingdom	−7.8
United States	1.5	France	−6.8
Slovakia	1.5	Italy	−5.3
Estonia	1.5	Canada	−3.8
Portugal	1.4	Spain	−3.2
Italy	1.3	South Korea	−3.1
Spain	1.3	Australia	−2.2
Poland	1.2	Netherlands	−1.8
Netherlands	1.2	Sweden	−1.2
Israel	1.1	Belgium	−1.1
Belgium	1.1	Austria	−0.9
Slovenia	1.0	Poland	−0.9
Germany	1.0	Norway	−0.8
Canada	0.9	Denmark	−0.7
Austria	0.9	Greece	−0.6
Iceland	0.8	Finland	−0.6
Ireland	0.8	Ireland	−0.5
Japan	0.8	Portugal	−0.4
Sweden	0.7	Czech Republic	−0.2
Finland	0.6	New Zealand	−0.2
France	0.6	Hungary	−0.2
South Korea	0.6	Slovakia	−0.0
Greece	0.5	Luxembourg	−0.0
Denmark	0.3		
Luxembourg	0.2		
Norway	0.2		
Australia	0.0		

Table 2 Translocation costs T_{ij} (country i to country j) for select country pairs

Destination –>	Australia	Canada	France	Germany	Italy	Japan	South Korea	Spain	UK	US
Afghanistan	0.0	0.0	1.9	0.1	29.5	44.2	44.2	18.1	0.3	0.1
Algeria	0.2	0.1	0.1	9.9	15.1	32.4	32.4	9.2	6.2	1.9
Azerbaijan	0.6	0.3	10.3	4.3	50.2	71.6	71.6	9.3	5.1	0.2
Bangladesh	0.4	0.1	8.1	3.2	1.2	12.9	8.9	5.4	0.1	0.3
Burkina Faso	6.3	2.0	1.1	5.4	2.5	192.0	192.0	27.8	44.1	11.2
Chad	2.8	0.6	0.7	12.2	43.9	407.2	407.2	62.4	15.3	9.3
Cote d'Ivoire	2.1	0.5	0.1	2.8	0.8	17.7	17.7	8.1	1.5	1.2
Egypt	0.0	0.1	1.9	1.1	2.7	12.8	12.8	13.9	1.4	0.2
France	0.0	0.1	0.0	34.6	1.9	1.8	2.7	1.3	28.2	0.2
Guinea	2.2	0.4	0.3	0.4	3.4	4.8	4.8	0.6	8.3	1.2
India	0.2	0.1	4.9	3.5	5.9	64.4	153.1	9.4	0.3	0.2
Indonesia	0.3	0.2	2.5	0.9	7.6	5.0	3.2	7.0	1.3	0.3
Iran	0.0	0.0	1.2	0.5	4.4	3.1	3.1	4.0	0.5	0.1
Iraq	0.0	0.0	3.0	0.5	13.6	98.3	98.3	5.3	0.3	0.1
Jordan	0.0	0.0	3.5	1.4	2.9	9.2	9.2	1.2	0.8	0.0
Kazakhstan	0.5	0.1	7.4	23.1	5.9	81.6	81.6	8.8	2.9	0.4
Libya	0.0	0.1	6.3	4.3	0.8	55.4	55.4	15.1	0.7	0.4
Malaysia	0.0	0.0	0.8	0.1	4.1	1.8	14.8	3.7	0.0	0.1
Morocco	0.2	0.1	0.1	1.0	0.3	13.1	13.1	0.7	3.3	0.6
Mozambique	0.7	0.4	1.8	0.8	3.8	472.8	472.8	2.0	0.4	1.7
Nigeria	0.7	0.5	14.3	1.8	3.1	8.3	8.3	4.4	0.4	0.4
Pakistan	0.2	0.1	2.3	1.6	2.2	11.6	9.7	1.4	0.1	0.2
Palestine	0.0	0.0	2.7	1.5	15.5	1e+200	1e+200	3.9	0.7	0.0
Russia	0.1	0.1	6.8	38.3	6.8	10.3	10.3	3.1	7.5	0.2
Saudi Arabia	0.2	0.1	4.4	3.7	20.7	32.4	32.4	15.6	0.6	0.2
Senegal	0.4	0.4	0.0	1.3	0.1	6.3	6.3	0.4	4.7	0.8
Somalia	0.0	0.0	0.8	0.0	0.2	463.3	463.3	6.0	0.0	0.0
Sudan	0.1	0.1	7.9	2.8	23.6	34.5	34.5	34.5	0.8	0.5
Tajikistan	2.3	1.5	56.7	27.3	29.8	1967.0	1967.0	80.3	12.3	0.6
Tanzania	0.3	0.0	7.9	1.0	6.2	19.5	19.5	24.5	0.1	0.6
Tunisia	0.2	0.1	0.1	1.5	2.4	7.8	7.8	19.2	5.2	0.8
Turkey	0.0	0.2	0.4	0.1	24.4	10.8	10.8	38.2	1.1	0.3
UK	0.0	0.0	32.2	23.7	2.9	0.7	2.7	1.3	0.0	0.1
US	0.0	10.6	1.2	0.9	0.7	0.7	0.9	1.7	0.3	0.0

Rows are countries of departure, columns are the destinations. Notice that Japan has relatively large barriers, as estimated by its abnormally low population of immigrants. OECD data was key to those estimates, and when it was not available, it was sometimes possible to use neighboring countries to impute the missing information.

In building the estimates, it will be assumed for simplicity that each stage of terrorist operations carries about the same amount of risk. Namely, that the medians of T_{ij} ($i \neq j$) and of I_j both equal 1. Both T_{ij} and I_j show considerable variability around the median because some plots are much less risky than others. The yield from attacks is also normalized by its median.

Table 3 The supply of plots for the default weight, and change under two alternative weightings (high commitment and low commitment)

Country	Supply S_i	High commitment (%)	Low commitment(%)
Indonesia	52745.4	−46.2	24.6
Nigeria	52687.8	−33.3	17.8
India	49624.7	−43.1	23.0
Bangladesh	39960.8	−33.8	18.0
Iran	26723.7	−30.6	16.3
Egypt	24731.6	−41.0	21.8
Pakistan	16537.8	−22.1	11.8
Algeria	12387.6	−30.6	16.3
Iraq	11021.7	−30.6	16.3
Sudan	10910.5	−30.6	16.3
Afghanistan	10168.3	−30.6	16.3
Saudi Arabia	9037.1	−30.6	16.3
Yemen	8462.6	−30.6	16.3
Ethiopia	8278.6	−39.7	21.2
Mali	8247.4	−23.2	12.4
Uzbekistan	8161.3	−43.1	23.0
Syria	7315.4	−30.6	16.3
Morocco	6878.5	−26.5	14.1
China	6680.7	−43.1	23.0
Niger	6497.3	−32.2	17.2
Malaysia	6466.6	−47.7	25.4
Turkey	5521.4	−44.0	23.5
Russian Federation	5081.9	−43.1	23.0
Palestine	4632.0	−21.1	11.2
Burkina Faso	4004.9	−32.2	17.2
Tunisia	3700.5	−30.6	16.3
Senegal	3668.5	−40.3	21.5
Guinea	3664.4	−32.2	17.2
Cote d'Ivoire	3338.1	−32.2	17.2
Somalia	3258.2	−30.6	16.3

Certain countries are unusually dependent on the level of support the most radical segment provides, while others see relatively broad support for violence. Only the 30 largest sources are shown. For some countries in a particular region the sensitivity is identical because national survey data was not always available. In those countries, the radicalization values (σ^r, σ^s, σ^o) were imputed from regional averages.

Transformations of this kind on costs and yields are unavoidable if we wish to remove the effect of units, but they do reduce the reliability of the model. However, the core findings of the model regarding certain countries agree well with intuition, as will be seen. As well, the stochastic Markovian decision model is not sensitive to the exact values of the parameters. In computational simulations, we found the sensitivity to be low, with variation of ±50% in parameter values leading to only ±10% change in attack predictions (see Appendix 1).

3.1   *Estimating the Supply of Plots, S_i*

The task here is to determine the potential supply of recruits that a violent world-wide Islamic movement can command. Surprisingly, this task is easier than to determine the support for a particular group like al-Qaida: with a particular group, its past actions and current platform can significantly affect its support base. Focusing on a particular group, even a well-known group like al-Qaida could also lead to underestimation of the risk involved from its ideological pool. Another concern is self-mobilization of violent cells without any connection to an existing group [35].

A comprehensive picture on possible plots can be obtained from surveys. Over the last decade, most recently in 2009, the Pew charitable trusts run several global attitude surveys. Among other questions the survey asked Muslims about their support for suicide bombings [42]. In each of the surveyed countries, respondents were asked to state whether suicide bombings is "never justified" (σ^n), "rarely justified" (σ^r), "sometimes justified" (σ^s), and "often justified" (σ^o). These are given as fractions of the respondents. For some countries no data was available, so the quantities were extrapolated from countries in the same geographic region (e.g. Middle East, Americas, etc.) Pew also collected data on the Muslim population in 235 different countries and territories (J_i) [30] (of course, the overwhelming majority of Muslims everywhere are opposed to terrorism in the name of their religion.) The supply of violent plots can be estimated by taking the population and multiplying by the weighted fraction of respondents professing support for violence (the weights are s_r, s_s, s_o). One must also take into account that only a small fraction of those who profess radical ideology would actually be involved in a plot and that several people are involved in each plot (a factor Q):

$$S_i := Q \cdot J_i \cdot (s_r \sigma^r + s_s \sigma^s + s_o \sigma^o).$$

The support weights were set by default based on the assumption that every increase in professed support leads to an increase by a factor of 2 in the resources available to terrorist groups: $(s_r, s_s, s_o) = (0.25, 0.50, 1.00)$. We can explore the sensitivity of supply S_i to this assumption through two alternatives: most of the tangible support come from the narrow but committed minority, $(s_r, s_s, s_o) = (0.1, 0.2, 1.0)$, and a situation where even the least-committed supporters materially boost the terrorists, $(s_r, s_s, s_o) = (0.33, 0.66, 1.00)$. Note that the weights effect only the relative importance of countries as sources of terrorism, not the targets of plots originating in a given region.

In certain countries, such as Malaysia, Indonesia, and Turkey support for violence is relatively broad. This can be seen from the large decrease in supply under the high-commitment scenario, $(s_r, s_s, s_o) = (0.1, 0.2, 1.0)$, compared to the default weights $(s_r, s_s, s_o) = (0.25, 0.50, 1.00)$. In regions such as the Palestinian Territories, Pakistan, and Morocco the support is more dependent on the radical minority, as seen from the relatively small increase under the scenario $(s_r, s_s, s_o) = (0.33, 0.66, 1.00)$. Overall, the 10 largest sources of plots are not more sensitive to those parameters than other sources.

The factor Q enters as a multiplicative term at all source countries, and so its value has no bearing on the relative risk estimates. Nevertheless, it could be crudely estimated as follows. In 2006, the head of the British Security Service (MI5) reported that: "... my officers and the police are working to contend with some 200 groupings or networks, totaling over 1,600 identified individuals (and there will be many we don't know) who are actively engaged in plotting, or facilitating, terrorist acts here and overseas" [31]. Furthermore, "over 100,000 of our citizens consider that the July 2005 attacks in London were justified." This implies active participation at a rate of at least 1.6% and 8 people per plot ($Q = 0.002$).

3.2 Estimating the Barriers for Moving from Country i to Country j, T_{ij}

Barriers to transnational attacks include both deliberate barriers such as screening and intelligence and unofficial barriers such as differences in language and culture. Official barriers depend on factors such as the intelligence available on targets in the destination country, the cooperation the targeted country received from both the country of departure and the transport agent (e.g. airline). None of those figures are publicly available but a proxy measure can be found, as follows. Transnational terrorists often use tourism, education or immigration as cover to obtain travel documents and permits [17]. Indeed, travel in all of those categories became more difficult across the developed world as a result of the security measures introduced after the 9/11 attacks. Migration patterns thus provide an estimate of official barriers. Unofficial barriers to terrorism are likewise similar to the unofficial barriers to migration, including differences in language, culture, ethnicity, and others. Therefore, the foreign-born migrant population, suitably normalized, could be used as a proxy of transnational freedom of travel. Migration into most OECD countries is documented by the OECD [34].

It is to be expected that the number of migrants would be positively correlated with the population of the countries and negatively correlated with distance. This is known as a "gravity law" model. Many national and international relationships such as trade flows are well-approximated by gravity laws [13, 25], named for their similarity to Newton's law for the force of gravity. Therefore, an estimate for the number of migrants between country i and country j is the product of their populations (data: United Nations) divided by their distance squared (data: CEPII [32]): $\frac{p_i p_j}{d_{ij}^2}$. When the actual number of migrants, m_{ij}, falls below this estimate, that may indicate heightened official or unofficial barriers. Thus, we define the raw transnational terrorism barrier between countries i and j as:

$$\widehat{T_{ij}} := \frac{p_i p_j}{d_{ij}^2} \bigg/ m_{ij} .$$

The data must now be standardized, for several reasons: (1) the barrier data should be comparable with other costs considered by the terrorists (and in the model) by

removing the effects of units for population and distance and (2) since the barrier is a cost and risk, it must be represented by positive number in the activity network. Therefore, the quantity T_{ij} was computed from $\widehat{T_{ij}}$ by determining the minimum ($Min\widehat{T}$) and median ($Med\widehat{T}$). Because of (2) $MinT$ is subtracted from $\widehat{T_{ij}}$, and for (1) the quantity is divided by the median of the shifted values:

$$T_{ij} = \left(\widehat{T_{ij}} - Min\widehat{T}\right) / \left(Med\widehat{T} - Min\widehat{T}\right).$$

The resulting values are non-negative numbers with a median of 1.0. The more standard procedure of first removing the average and dividing by the standard deviation was rejected because the distribution is visibly non-Gaussian with large positive outliers (hence a skewed mean and large variance). Domestic operations will be assumed to have negligible barriers ($T_{ii} = 0$ for all countries i). Visa-free travel zones such as the European Schengen zone could be incorporated into T_{ij}, but were not because (1) they do not eliminate the considerable linguistic and cultural barriers, and (2) data on terrorism originating in Europe suggests a strong preference for domestic attacks (a catalog is found in [4]). Thus, plotters prefer targets in their homeland even though a neighboring European country might offer more valuable targets.

The OECD data lacks information about migration to non-OECD countries. Therefore set $T_{ij} = \infty$ in all such cases (effectively blocking such paths). While the OECD includes some of the most geopolitically-important nations of the planet, obtaining data on translocation costs to non-OECD countries would be valuable for two reasons. First, only with such data can we estimate the risk of terrorism to those nations (such as the July 11, 2010 bombings in Kampala), and second to estimate the effect of counter-terrorism policies in the OECD on terrorism in other countries. Indeed, in the ITERATE database of terrorist incidents [33], attacks on OECD countries that can be traced to Islamist groups account for only 22% of all Islamist attacks.

3.3 Estimating the Risk of Interception at Country j, I_j

The risk of interception can be estimated from OECD data on expenditure on public order and safety as a fraction of GDP. The relevant figure is the fraction of GDP rather than the raw figure because the number of valuable targets is related to the size of the economy, so the fractional figure indicates the level of security vulnerable sites can expect to receive. The GDP also correlates with the population size (in rich countries) and thus to the amount of police resources available to protect any human target. The extreme case of totalitarian police states is suggestive: in such countries internal security expenditures are disproportionately large relative to the GDP, and indeed terrorists have a lot of difficulties operating there [23]. This estimate of course neglects the efficiency of internal security force – a factor that is hard to estimate.

The internal security data is transformed almost exactly like the barrier data and for the same reasons: start with figures for internal security expenditure as a fraction of GDP for country j, SEC_j then compute the minimum ($MinSEC = \min_j SEC_j$) and median ($MedSEC$), and then normalize:

$$I_j = (SEC_j - MinSEC)/(MedSEC - MinSEC).$$

Generally speaking, we find that there is not much variation in the risk of interception in different countries (Table 1), as compared to variation in factors such as yield, discussed next. This suggests that interception risk plays a minor role in target choice.

3.4 Estimating the Yield from Attacks at Country j, Y_j

Transnational terrorist attacks attempt to influence policies. For example, one of al-Qaida's original objectives was to compel the withdrawal of US forces from Arabia, while Hezbollah forced France and the US to withdraw their peacekeepers from Lebanon in the 1980s. The precise value of the targets shifts with time and the political situation, but typically larger richer countries make for more powerful players in the international arena, and hence more important targets. Moreover, their homelands carry more targets of symbolic, political and economic significance. The economic damage from the loss of life and physical assets is also higher in richer countries because they tend to have higher productivity for labor and capital. Thus, it is expected that transnational terrorists would seek to attack larger richer countries. The weight of a country can be estimated from its dollar GDP figures at current exchange rates (data: UN).

Timing or political dynamics does play a role in transnational terrorism but its importance should not be overestimated. For example, the Madrid March 11, 2004 train bombings are often viewed as intending to pressure the Spanish government to withdraw its forces from Iraq, and they were timed with the Spanish elections. But surely an important factor was Spain's geopolitical weight (GDP is ranked 12th in the world) and its large contribution to the 2003 invasion. Otherwise, al-Qaida could have just as well pressured smaller countries such as El Salvador and Mongolia to withdraw their contributions to the invasion.

Here is how the yield Y_j was computed from the GDP figures. Recall that costs (barriers, internal security) are all positive, so yields must be negative. Let the minimum GDP be $MinGDP$, and the median $MedGDP$. The following formula produces negative values with a standardized median:

$$Y_j = \left(MinGDP - GDP_j\right)/\left(MedGDP - MinGDP\right).$$

The resulting values have a median of -1.0.
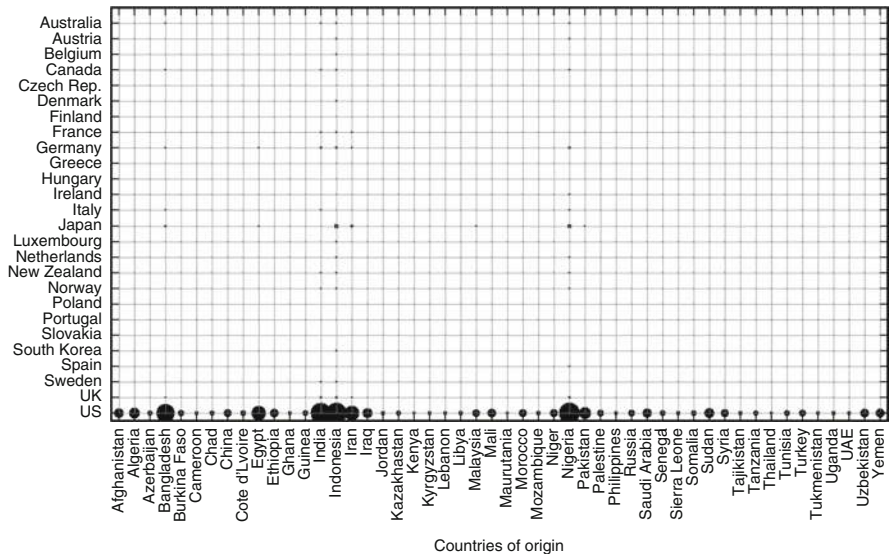
Fig. 3 Predicted attack risk matrix. The area of a circle at coordinates (i, j) is proportional to the number of plots from i to j. The *bottom row* indicates that the vast majority of plots target the US

4 Predictions

One way of representing the solution of the model is through an attack risk matrix, the counterpart of the historical data matrix in Fig. 1. The model predicts (Fig. 3) that the United States would attract the bulk of transnational terrorism – all other countries are almost free of terrorism (small circles). The reason the United States is such a magnet is because of its vast geopolitical weight and relatively open borders.

Examination of the sources for attacks exposes a number of risks. There is a considerable terrorist threat from Bangladesh, India, Indonesia, and Nigeria. They combine a large population and relatively high support for terrorism. It is notable that the 2009 Christmas bomber was Nigerian – one of the first attacks on OECD from that country. The risk from India stems from its large Muslim population (about 160 million) and the relatively large radicalization in the region (although the radicalization might be lower in India than in neighboring countries.)

The high burden of attacks borne by the US is directly related to the rational choice model: if there is a big prize to be won by attacking the US, no rational terrorist would attack other countries. The reasons why non-US attacks do occur (cf. Fig. 1) include: (1) some terrorist groups such as ethno-nationalist groups see as their enemy a particular country and lack a global agenda; (2) global Salafi groups have not yet expanded their recruitment channels in countries such as Nigeria and Bangladesh, so a large fraction of attacks is still carried out by groups with more narrow agendas; and (3) the US has deployed counter-terrorism measures commensurate with the threat it faces, making the US too costly to attack.

One implication of this finding concerns US policy. The matrix justifies in principle outlays by the US government towards countering international terrorism as a whole, without regard to its target. Investments in international counter-terrorism measures, such as nation building in unstable states, if effective in reducing the number of plots, are also efficient from the US perspective: because the US is the target of choice, it will retain most of the benefits from reducing the terror threat [28]. Unfortunately, many policies previously adopted were ineffective or had perverse effects on international terrorism (the so-called "blowback") (see e.g. [15, 16].)

4.1 National Fortresses

Consider now several alternative scenarios for the future, motivated by strengthening of counter-terrorism defenses, which may make transnational attacks less feasible. Suppose the US was successful in deterring attacks against itself by greatly increasing the barriers to entering US soil. If so, Fig. 4 shows the likely effect.

The protection of US frontiers will significantly increase the attack risk to most other OECD nations because transnational groups should then switch to more accessible targets. Perhaps surprisingly, Japan, now rarely mentioned as a target will see the largest absolute increase in terrorism. This prediction is due to its international profile, Japan being the second largest country in the OECD on
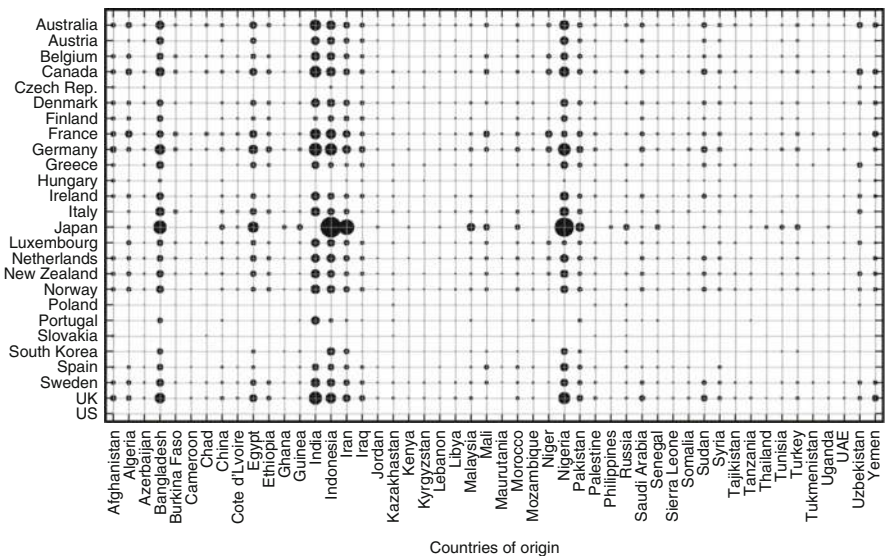
Fig. 4 Attack risk matrix in scenario where US becomes inaccessible to foreign plots. Terrorist plots increase in all other OECD countries

Fig. 5 Attack risk matrix in a scenario where OECD countries cannot be accessed by foreigners (both inside and outside the OECD). Countries with relatively large radicalized Muslim populations (e.g. France) rise in rank relative to their OECD peers. The total number of attacks on OECD countries does decrease significantly because foreign plots are blocked
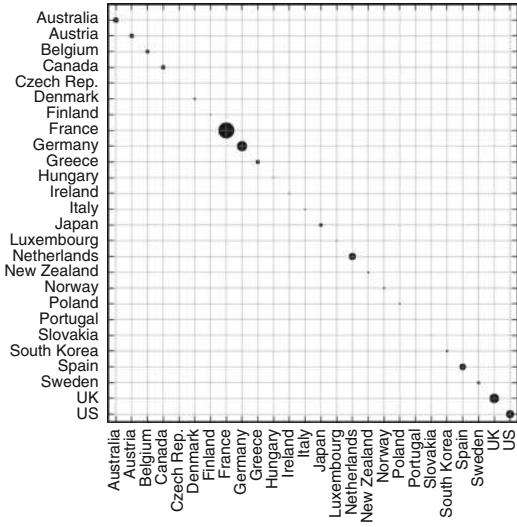
several measures. Japan's woes will be shared to some extent by most other major OECD countries, who will also see an increase in attacks (side-by-side numerical comparison with the baseline scenario is found in the tables of Appendix 1.)

Another possible scenario is where the security forces in each country are able to intercept the majority of external plots against their homelands. In other words, the translocation cost becomes very large ($T_{ij} = \infty$ for $i \neq j$). In this world, the dominant form of terrorism is home-grown. As Fig. 5 shows, this materially changes the risk matrix. Countries such as the France, with relatively large and relatively radicalized Muslim communities will see much more terrorism.

The two scenarios point to large conflicts of interest between OECD countries in tackling transnational terrorism. Helping the US intercept plots through advance warning will increase terrorism everywhere else. More broadly, country A will not always benefit from helping country B. Doing so might sometimes increase the chances that A's enemies, some of which even based in B, will shift to A. This factor may explain part of the difficulty achieving intelligence sharing and international police cooperation. Indeed, B could even come to an understanding with its home-grown terrorists in which they abstain from domestic attacks in return for non-intervention in their activities against A.

4.2 Deterrence

A number of defensive strategies are founded on deterrence. In terrorism, deterrence may involve convincing would-be groups or cells that operations are too risky or that the entire struggle they wage is hopeless. The model can express such conditions on
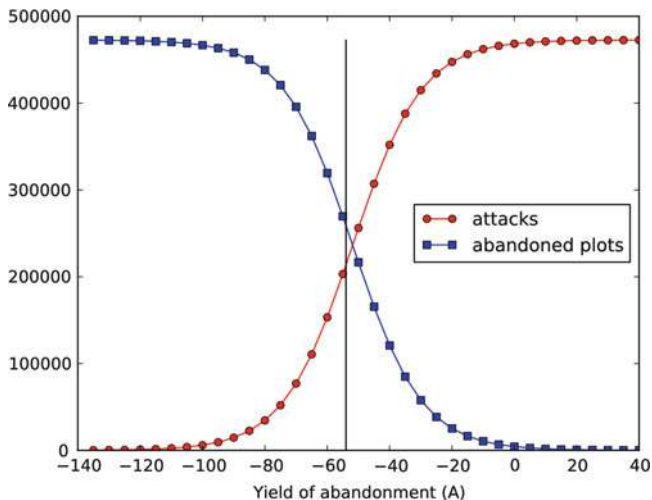
Fig. 6 The number of attacks as a function of the yield from abandoning. Negative values make abandoning more competitive and decrease attacks (*left side*) while positive values indicate that abandoning is costly and encourage attacks (*right side*). The *vertical black line* indicates the yield from attacking the US – the most valuable target. The sigmoid shapes suggest that the effect of deterrence is low until a threshold is reached, but the threshold must be close to the perceived value of attacking the US (*the vertical line*)

a global level by varying the parameter A, the perceived yield from abandoning. Raising this yield is equivalent to raising risks throughout the network. The results are in Fig. 6. The effect of A is non-linear with a threshold at around $A = -35$ beyond which attacks decline. Unfortunately, the threshold lies quite high, indeed nearer to the yield from attacking the US (-54) than countries such as France (-6.8), implying that it would be necessary to create a very large deterrence effect to reduce the number of plots.

If this level of deterrence is somehow achieved, the reduction in attacks will not occur at once in all countries because cells in some countries have lower translocation costs than cells in other countries. As a result, their perceived net benefit and probability of success are higher. Thus plots originating within the developed countries such as the G7 and especially home-grown plots will be the last to experience deterrence because they originate so close to high-value targets.

5   Discussion

The network model draws attention to differences between our past experience with terrorism and its possible future. Populous regions like Nigeria and Bangladesh are predicted to produce many plots although they have not participated significantly

in transnational terrorism yet. If those regions start producing terrorists at a level commensurate with their size and radicalization, the world will see many more attacks. Alternatively, it is possible that those regions have characteristics that hold back violent extremism. If so, future research should identify those characteristics and suggest policies that maintain and encourage them.

The model confirms the significant danger from substitution effects. Perceived successes in reducing the number of attacks against it may be due to a strategic redirection by terrorist groups that increases the risk to other countries. In the scenario where the US deters all attacks by foreign terrorists, many other countries would experience a large increase in threats. To an extent this has already been seen in Europe.

The current model considered the risk from a large-scale Islamist movement. Because such movement is yet to emerge, there is no record of attacks which could serve as validation data – a characteristic problem in risk analysis. However, the model could be re-estimated for extant groups and those variants could then be validated against the historical record. It would be particularly useful if databases such as ITERATE are augmented by the much larger record of attacks that were intercepted.

The model introduced above has limitations where perhaps the most significant is the assumption that a terrorist group's main resource are its human resources. In practice, attacks also require intelligence gathering, training and materials. Future work can develop network-theoretic methods to analyze how terrorist groups would bring those resources together while maintaining secrecy. It would also be possible to consider the risk of attacks by specific groups, including detailed model of their target preferences. Another particularly interesting extension are attacks on foreign representatives of a country (e.g. human representatives such as diplomats or journalists and physical installations such embassies or business offices). Such foreign representatives give terrorist groups high-value targets without the risk of transnational attacks. Similarly, one may consider attack on modes of international transit such as airplanes and ships.

6   Conclusions

The paper introduces a model of transnational terrorist groups that represents operations as a global activity network. It is possible to estimate the parameters of the model, and then predict the number of plots directed at OECD target countries from countries throughout the world. The model highlights the exceptionally high risk of attacks against the US. Yet if the US is successful in deterring attacks against itself without reducing the overall supply of terror then most OECD countries would see sharp increases in attacks because of a substitution effect.

The scale of the substitution effect should alert policy makers to the need to develop and execute multinational defensive strategies. Current strategies, focused on protecting national borders, are both inefficient at reducing the global supply of

plots and increase the threat to allied countries. Ideally, counter-terrorism strategies would reduce terrorism at its places of origin, creating benefits to the entire international community. Abstract solutions to this problem have actually been developed using the methods of graph theory [8, 41], where it is known as network interdiction. Those methods could identify where barriers should be erected in the transnational terrorist network to produce an increase in the costs to the terrorists in such a way that they cannot avoid it by shifting their plots to other countries. If those methods could be implemented as practical counter-terrorism strategies, then the threat of transnational terrorism would be greatly reduced.

Appendix

1 Global Sensitivity Analysis

Considerable uncertainty exists above the values of the model parameters: the supply of plots, translocation costs, interception costs and the yields. Indeed, is likely that the estimates of those parameters in Sect. 3 are different from the ones used by the terrorist groups. To explore the sensitivity of the predictions to this uncertainty we considered 100 realizations of the model, each with different parameter values. In each realization, each of the original values of the parameters was randomly changed: the value was multiplied by a random number sampled uniformly from [0.5, 1.5]. Thus, the values were varied through a range of ±50%.

Table 4 shows the sensitivity in the expected number of attacks against various countries. In the majority of cases, the expected number of attacks changes by ±10%, often less. The reason for this is that paths in the networks become sums of random quantities, and errors tend to partially cancel each other when summed (central limit theorem). The highest sensitivity was observed in the attacks against the US in the scenario when the US is protected against foreign plots. In such a case, the number of attacks in the US varies exactly as the number of plots that start in the US.

The values of λ (the determinism in the path selection, explained in Appendix 2) was kept at its default value (0.1) since λ directly expresses sensitivity. In the limit of $\lambda \to 0$, the terrorist groups is completely insensitive to risk or cost while in the limit $\lambda \to \infty$ the sensitivity is infinite and even small changes in costs can lead to arbitrarily large changes in path choice. The regime $\lambda \to \infty$ is the case where a decision maker can distinguish arbitrarily small differences in utility and never

Table 4 Expected number of plots and sensitivity to parameter estimates

Country	Plots	Range (%)	Country	Plots	Range (%)
Australia	2804.9	94.5–105.4	Australia	29229.1	96.4–101.5
Austria	1405.9	91.6–109.0	Austria	13691.0	94.5–104.5
Belgium	1817.1	93.4–108.1	Belgium	17954.6	94.6–103.5
Canada	3120.1	93.4–106.0	Canada	32647.7	94.4–102.8
Czech Republic	251.9	91.3–127.2	Czech Republic	2964.1	88.1–122.6
Denmark	1869.2	94.5–106.6	Denmark	18816.7	96.0–102.6
Finland	1183.4	92.4–110.8	Finland	11493.5	94.7–108.0
France	3049.7	94.1–110.2	France	31064.9	95.2–105.3
Germany	4272.6	93.3–107.4	Germany	42340.4	95.3–102.8
Greece	1344.2	93.8–109.7	Greece	14042.7	93.4–106.1
Hungary	281.5	92.4–130.2	Hungary	3035.2	91.3–122.5
Ireland	1694.0	93.9–107.6	Ireland	17422.9	94.6–104.1
Italy	2008.2	92.4–110.0	Italy	19648.6	93.2–109.7
Japan	6563.3	92.9–117.9	Japan	57477.1	92.3–115.1
South Korea	852.2	93.5–115.5	South Korea	7499.1	92.4–114.3
Luxembourg	1590.6	94.6–107.3	Luxembourg	16065.4	94.6–103.4
Netherlands	2015.1	93.2–107.8	Netherlands	20133.1	93.5–104.4
New Zealand	1832.9	91.1–110.1	New Zealand	19131.8	90.4–107.9
Norway	1966.9	95.3–106.3	Norway	20065.9	96.4–102.3
Poland	86.8	90.1–131.9	Poland	1003.9	89.3–131.5
Portugal	598.2	87.0–120.1	Portugal	6008.4	87.3–119.9
Slovakia	80.9	91.0–133.0	Slovakia	882.0	88.4–131.0
Spain	1379.5	93.9–113.3	Spain	13281.8	93.0–111.3
Sweden	2070.0	94.0–106.8	Sweden	21226.0	95.0–102.5
United Kingdom	3467.2	89.9–111.5	United Kingdom	35333.4	91.0–106.1
United States	425091.5	99.3–100.5	United States	238.9	58.8–144.0

When the supply of the plots and the costs on the edges are varied, the expected number of attacks against a country j, Z_j, changes. Shown are Z_j and the relative change coefficients: $[z_{j;min}, z_{j;max}]$. Namely, $z_{j;min} Z_j$ is the number of attacks at the bottom decile; $z_{j;max} Z_j$ is the number at the top decile. The baseline scenario (*left*). The scenario where the US is protected from foreign plots (*right*).

makes mistakes. Fortunately, clandestine and illicit decision makers like terrorist leaders are far from this omniscient regime.

2 Computation of Probabilities

In the framework of the theory of complex networks, attack plots could be represented as the motion of an adversary through a weighted network (the plot itself is the adversary we wish to stop). The adversary aims to find an attack path or to hide, whichever plan has the lowest cost. To map such a decision to the framework of activity networks, connect the "attack" and "abandon" nodes in Fig. 2 to a node termed "end" with edges of cost 0. Thus, an attack on a country j corresponds

to an adversary that starts at country i and goes through country j and then to the node "attack" and finally to "end". The decision to abandon corresponds to an adversary that starts at country i then goes to "abandon" and then to "end". The expected number of attacks on a particular target t can be computed by combining information about path costs with information about the supply of plots from a particular country S_i and the yield of abandonment A. Namely, it is the number of trips from all sources that arrive to the "attack" node from target country t.

The least-cost path corresponds to the optimal choice by the terrorists, but they can make mistakes. An attack plan under uncertainty could be described as a Markov chain [20]. The chain has initial distribution proportional to S_i, and a transition probability matrix $\mathbf{M}$ describing the likelihood of taking a particular edge on the network. The "end" node is the absorbing state of the chain. The $\mathbf{M}$ matrix can be computed using the least-cost guided evader model described in [20]. Briefly, for each edge (u, v) of the network, the transition probability through it is given by the formula

$$M_{uv} = \frac{\exp\left[-\lambda \left(c(p_{uv}) - c(p_{u*})\right)\right]}{\sum_w \exp\left[-\lambda \left(c(p_{uw}) - c(p_{u*})\right)\right]},$$

where $c(p_{u*})$ is the cost of the least-cost path from node u to the end node, and $c(p_{uv})$ is the cost of the path through edge (u, v): $p_{uv} = (u, v) \cup p_{v*}$. The sum in the denominator runs over the neighbors of node u. Thus, the model generalizes the least-cost path model[3]. The parameter λ was set to 0.1, in the reported data, but its value has a smooth effect on the predictions of the model because of the smoothness of the exponential function. The number of plots against a target country t is now found by taking the probability of a trip to that target multiplied by the total number of plots $(= \sum_i S_i)$.

References

1. Abrahms, M.: What terrorists really want: Terrorist motives and counterterrorism strategy. Int. Secur. **32**(4), 78–105 (2008). http://www.mitpressjournals.org/doi/abs/10.1162/isec.2008.32.4.78
2. Anderton, C.H., Carter, J.R.: On rational choice theory and the study of terrorism. Def. Peace Econ. **16**(4), 275–282 (2005)
3. Arquilla, J., Ronfeld, D.: Networks and Netwars: The Future of Terror, Crime, and Militancy. RAND Corporation, Santa Monica, CA (2001)
4. Bakker, E.: Jihadi terrorists in europe. Tech. rep., Netherlands Institute of International Relations, Dec 2006

---

[3]To compute the distances to the "end" node, $c(p_{uv})$, we use the Bellman–Ford algorithm because edge weights are negative for some edges (e.g. yield from attacks). Gutfraind et al. [20] uses the faster algorithm of Dijkstra because it treats only the case of positive weights.

5. Berman, E., Laitin, D.D.: Religion, terrorism and public goods: Testing the club model. J. Public Econ. **92**(10–11), 1942–1967 (2008). http://www.sciencedirect.com/science/article/B6V76-4S4TNTW-1/2/6a0e405748612f6c10b7db12ace22fc1

6. Bier, V.M., Oliveros, S., Samuelson, L.: Choosing what to protect: Strategic defensive allocation against an unknown attacker. J. Public Econ. Theory **9**(4), 563–587 (2007)

7. Caplan, B.: Terrorism: The relevance of the rational choice model. Public Choice **128**(1), 91–107 (2006). http://www.springerlink.com/content/041521274700t406

8. Corley, H.W., Sha, D.Y.: Most vital links and nodes in weighted networks. Oper. Res. Lett. **1**(4), 157–160 (1982)

9. Corman, S.R.: Using activity focus networks to pressure terrorist organizations. Comput. Math. Organ. Theory **12**, 35–49 (2006). http://www.springerlink.com/content/F36M527745172454

10. Dugan, L., Lafree, G., Piquero, A.R.: Testing a rational choice model of airline hijackings. Criminology **43**(4), 1031–1065 (2005). http://www.blackwell-synergy.com/doi/abs/10.1111/j.1745-9125.2005.00032.x

11. Enders, W., Sandler, T.: Distribution of transnational terrorism among countries by income classes and geography after 9/11. Report by the Center for Risk and Economic Analysis of Terrorism Events (CREATE) (2005)

12. Enders, W., Sandler, T.: What do we know about the substitution effect in transnational terrorism. In: Silke, A. (ed.) Researching Terrorism: Trends, Achievements, Failures, pp. 119–137. Frank Cass, Ilford (2004)

13. Erlander, S., Stewart, N.F.: The gravity model in transportation analysis: theory and extensions, Topics in transportation, vol. 3. VSP, Utrecht, The Netherlands (1990)

14. Finbow, A.S., Hartnell, B.L.: On designing a network to defend against random attacks of radius two. Networks **19**(7), 771–792 (1989). http://dx.doi.org/10.1002/net.3230190704

15. Ganor, B.: The Counter-Terrorism Puzzle: A Guide for Decision Makers. Transaction Publishers, Piscataway, NJ (2005)

16. Ganor, B.: Terrorist organization typologies and the probability of a boomerang effect. Stud. Confl. Terrorism **31**, 269–283 (2008)

17. Ginsburg, S.: Countering terrorist mobility: Shaping an operational strategy. Tech. rep., Migration Policy Institute, Feb 2006

18. Gutfraind, A.: Mathemaical Terrorism. Ph.D. thesis, Cornell University, Ithaca, New York, USA (2009)

19. Gutfraind, A.: Optimizing topological cascade resilience based on the structure of terrorist networks. PLoS ONE **5**(11), e13448 (2010)

20. Gutfraind, A., Hagberg, A.A., Izraelevitz, D., Pan, F.: Interdiction of a Markovian Evader. In: Dell, R., Wood, K. (eds.) Proceedings of the INFORMS Computing Society Conference. INFORMS (2011)

21. Hanson, M.A.: The economics of roadside bombs. SSRN eLibrary (2008)

22. Harmon, C.C.: Terrorism Today, 1 edn. Routledge, Oxford, UK, (2000)

23. Harrison, M.: Counter-terrorism in a police state : The KGB and codename blaster, 1977. The Warwick Economics Research Paper Series (TWERPS) 918, University of Warwick, Department of Economics (2009). http://ideas.repec.org/p/wrk/warwec/918.html

24. Hoffman, B.: Inside Terrorism. Columbia University Press, USA (2006)

25. Jung, W.S., Wang, F., Stanley, H.E.: Gravity model in the korean highway. EPL (Europhysics Letters) **81**(4), 48005 (2008). http://stacks.iop.org/0295-5075/81/i=4/a=48005

26. Lake, D.A.: Rational extremism: Understanding terrorism in the twenty-first century. Dialog-IO, pp. 15–29. Spring, Berlin (2002)

27. Landes, W.M.: An economic study of U. S. aircraft hijacking, 1961–1976. The J. Law Econ. **21**(1), 1 (1978). http://www.journals.uchicago.edu/doi/abs/10.1086/466909

28. Lee, D.R.: Free riding and paid riding in the fight against terrorism. Am. Econ. Rev. **78**(2), 22–26 (1988). http://www.jstor.org/stable/1818091

29. Lindelauf, R.H., Borm, P.E., Hamers, H.: The influence of secrecy on the communication structure of covert networks. Soc. Netw. (2009)

30. Lugo, L.: Mapping the Global Muslim Population. Pew Research Center, Oct 2009

31. Manningham-Buller, E.: The International Terrorist Threat to the UK. MI5 website (Nov 2006). http://www.mi5.gov.uk/output/the-international-terrorist-threat-to-the-uk-1.html
32. Mayer, T., Zignago, S.: Notes on CEPII's distances measures. http://www.cepii.fr/anglais-graph/bdd/distances.htm (May 2006), Centre d'Etudes Prospective et d'Informations Internationales
33. Mickolus, E.F.: International Terrorism: Attributes of Terrorist Events (ITERATE). http://www.icpsr.umich.edu/cocoon/ICPSR/STUDY/07947.xml (2009)
34. OECD: Immigrants and expatriates: Total population by nationality and country of birth. online (2006), Organisation for Economic Co-Operation and Development
35. Sageman, M.: Leaderless Jihad – Terror Networks in the Twenty-First Century. University of Pennsylvania Press, Philadelphia, PA (2008)
36. Sandler, T.: Collective action and transnational terrorism. World Econ. **26**(6), 779–802 (2003)
37. Sandler, T., Enders, W.: An economic perspective on transnational terrorism. Eur. J. Polit. Econ. **20**(2), 301–316 (2004). http://ideas.repec.org/a/eee/poleco/v20y2004i2p301-316.html
38. Shapiro, J.N., Siegel, D.A.: Underfunding in Terrorist Organizations. In: Memon, N., Farley, J.D., Hicks, D.L., Rosenorn, T. (eds.) Mathematical Methods in Counterterrorism, pp. 349–382. Springer, New York (2009)
39. Shapiro, J.N., Siegel, D.A.: The Greedy Terrorist: A Rational Choice Perspective on Terrorist Organizations. Terrorist Financing in Comparative Perspective. Inefficiencies and Vulnerabilities. Stanford University Press, California (2009)
40. Shughart, W.F. II: Terrorism in rational choice perspective (2009), working Paper
41. Washburn, A., Wood, K.: Two-Person Zero-Sum Games for Network Interdiction. Oper. Res. **43**(2), 243–251 (1995). http://or.journal.informs.org/cgi/content/abstract/43/2/243
42. Wike, R.: Pew global attitudes project. Pew Research Center (2006-2009)
43. Woo, G.: Understanding terrorism risk. online (2004), report for Risk Management Solutions
44. Woo, G.: Intelligence Constraints on Terrorist Network Plots. In: Memon, N., Farley, J.D., Hicks, D.L., Rosenorn, T. (eds.) Mathematical Methods in Counterterrorism, pp. 205–214. Springer, New York (2009)

A Framework for Analyst Focus
from Computed Significance

David Skillicorn and M.A.J. Bourassa

Abstract Attention is the critical resource for intelligence analysts, so tools that provide focus are useful. One way to determine focus is by computing significance. In the context of a known model, new data can be placed on a spectrum defined by: normal, anomalous, interesting, novel, or random; and significance is greatest towards the middle of this spectrum. However, significance also depends on the mental state of the analyst (and the organization). A framework for understanding significance is defined, and its impact on the knowledge discovery process is explored.

1 Motivation

Adversarial settings are those where the interests of those building models and some of those being modelled are not aligned. Such settings include counterterrorism, counterintelligence, law enforcement, fraud detection, financial tracking, and so on, but also more mainstream settings such as customer relationship management, email spam, and web ranking manipulation. Knowledge discovery in adversarial settings requires a process and model-building techniques that are explicitly aware that adversaries may be trying to conceal their traces and to manipulate or distort the results. These settings are examples of what Treverton [17, 18] calls *mysteries*, issues where framing the right questions is important and it is the quality of the analysis, rather than the amount of data, that is critical (in contrast to *puzzles* where questions can always be answered given enough data). Because adversary traces are often a small fraction of the available data, patterns associated with adversaries

D. Skillicorn (✉)
School of Computing, Queen's University, Kingston, ON, Canada
e-mail: skill@cs.queensu.ca

M.A.J. Bourassa
Royal Military College of Canada, Kingston, Ontario, Canada

are changing as adversaries attempt to evade detection, and the consequences are often severe, analysts face great challenges. Often, analyst/investigator attention is the critical resource.

Tools that can direct focus to the most significant aspects of data, and the models built from data, improve effectiveness, and reduce timelines. In high profile terrorism cases, e.g. the attempt by Umar Farouk Abdulmutallab to blow up a plane at the end of 2009, it often turns out that the data necessary to detect and prevent the attack had been collected, but insufficient analysis had taken place to enable detection beforehand. In other words, the problem is not so much "connecting the dots" as finding those dots that are important enough that their connections need consideration in a pool of dots that may be nine orders of magnitude larger. The biggest win, therefore, in tool design and implementation is improvement in the collaboration between analyst attention and tool-directed focus. This will obviously be easiest when tools exploit human strengths, e.g., using the high bandwidth of the human visual system by generating effective visualizations, or acknowledging the small size of human working memory by keeping focus small.

In adversarial settings, analysts and tools need to be aware of the implicit 'arms race' between adversaries' attempts to conceal themselves and manipulate the analysis process, and the quality, toughness, and timeliness of models. This implies that models will be updated regularly, and that update is a far more integral part of the process than typical business modelling. In particular, significance is a moving target, even independent of changes in the data or the analyst's mindset or knowledge.

A problem is that significance is not a property only of the data or models in themselves. It does not make sense to talk about a significant record, or a significant decision boundary, or a significant cluster without some context and usually some history. Significance is a relation between data or model on the one hand, and analyst state and history on the other. Indeed, what may be significant to one analyst may not be to another, depending on how much they already know and understand. Therefore, tools that use significance as a way to generate focus can only do well when they include some input that reflects the analyst/organization worldview.

There is, however, a danger of overshooting when analyst state is included – analysts can misinterpret new signals as something they already understand, when this is not the case. Another frequent refrain in terrorist attack post mortems is that there has been a "failure of imagination", essentially a claim that analyst mindset ruled out some possibilities. This can work in two ways: data or models are discounted as too unlikely, or discounted as something already understood; but, in both cases, they are discounted. It is important that tools using significance should be resistive to under- and over-reliance on analyst state; they should include it, but cautiously.

The contribution of this paper is to define a framework for significance, define how significance depends on the relationship between new data and existing models, and to explore how knowledge discovery techniques can integrate the idea of significance and analyst context. As usual, some techniques are significance-ready,

requiring only that some of their existing functionality be better used; others require substantial enhancement.

2   The Structure of Significance

Models that aim to find the "interesting" or "anomalous" or "novel" data or parts of a model often conceive of the problem, perhaps implicitly, as fundamentally two-class – objects are either normal or anomalous, known or new. We suggest that there is more structure to the problem, and that acknowledging and incorporating this structure provides a better way to design systems to recognize significance.

Every model is imperfect because it was built from less data than could have been collected and used, and because the model-building algorithm was restricted in form and so in power. The arrival of new data may reveal either or both of these limitations of the current model.

Figure 1 shows a very simple model in two dimensions, with records shown as points; implicitly, distance represents dissimilarity. The obvious structure of the model is three clusters, indicated by labelling the points with different symbols. Now consider new records whose attributes place them in the positions indicated by the lower-case letters. Each record is new and therefore potentially represents significant new knowledge to which an analyst might need to pay attention. However, these points are qualitatively different from each other, and it is this qualitative difference, we contend, that is the structure of significance. Their meanings are as follows:

- Point a is fully explained by the existing model – although it is new, it is not significant.
- Point b is *anomalous*; it lies just outside one of the existing clusters but is so close to it that it exerts no pressure for a new understanding. Its presence is explainable
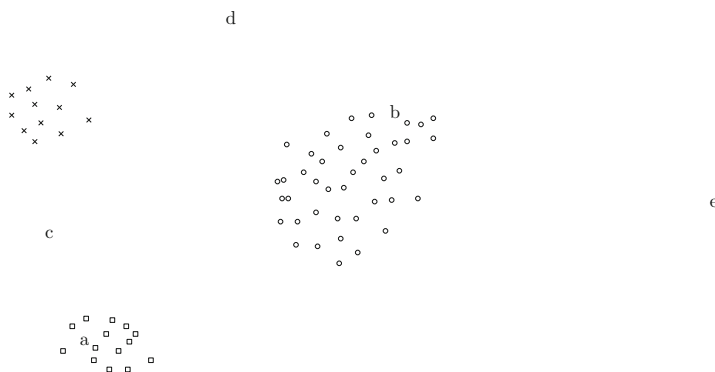
Fig. 1 Different possibilities for significant data in a simple dataset. Point a is ordinary; point b is anomalous; point c is interesting; point d is novel; and point e is random

by the finiteness of the particular sample that was used to build the model (the clusters). Each cluster might be considered to be wrapped by a boundary that falls just outside the points it contains, and an anomalous point lies within this boundary.

- Point c is *interesting* – its presence suggests a weakness in the current model, perhaps the presence of a yet-unrecognized fourth cluster, or the fact that the square cluster and the cross cluster are actually the same. The entire set of clusters might be considered to be wrapped by a boundary just outside the set of cluster boundaries. An interesting point lies within this boundary but not close to any existing cluster.
- Point d is *novel* – its presence does not indicate a weakness with the model as such because it is well outside the range of data from which the model was built. Its presence does, however, require it to be acknowledged and treated using a different process (which could be as simple as ignoring it, although this might be dangerous). A novel point lies outside the boundary wrapping the clusters.
- Point e is noise – its presence brings no new information because it is so far outside of the situation being modelled. Another, much more amorphous, boundary defines the difference between novel and noisy, perhaps constructed by considering multiples of the standard deviation of the entire dataset.

Of course, these categories and their boundaries are not hard and fast, and new points might be hard to unambiguously place in one or another category. Nevertheless, it seems helpful to consider them as qualitatively different when deciding what each *means* and what should be done as a result of the arrival of a new record of each kind.

These categories form a natural spectrum: *normal – anomalous – interesting – novel – random*, but what makes dealing with this technically difficult is that significance peaks in the middle of the spectrum. New data that is normal or random is not very significant, new data that is anomalous or novel is moderately significant, but data that is interesting is most significant of all. As the figure shows, it is relatively straightforward to detect data that is normal or random, but this becomes more difficult, the closer the data is towards the center of the spectrum. Not every technique for finding and displaying significance is congruent with this structure, but the more they are, the more effectively they work. There are connections here to, e.g., intrusion detection, but in that setting all data that is not normal is lumped together, and the issue of inadequacy of the model is not usually considered.

3 The Role of Context

Most knowledge discovery operates either on record-based data, or graph-structured data. Record-based data consists of a set of n records, each with m attribute values and so has a natural geometric representation in m-dimensional space, where each attribute is associated with a dimension, and each record with a point. Graph-based

data consists of n records, each regarded as a graph node, and (weighted) edges connecting some pairs of nodes. The advantage of graph representations is that they integrate the data globally (the structure depends on every edge) – this makes them resistant to adversarial manipulation, because many of their properties are emergent, but they are practically more difficult to work with.

Record-based data can be, and often is, converted to graph-structured data by constructing a local similarity between every pair of records. Local similarity may be derived from correlation between records, or Euclidean distance in the geometry. Local similarity is often thresholded, set to zero for records that are insufficiently correlated, too far apart, or not among each other's k nearest neighbors.

The significance of a record or a part of a model is usually not just a function of the model, but also of the context in which that model is being considered. In other words, significance is a stateful property and depends on what is already known and understood, and the (mental) weights assigned to each part of the model. This background state includes not just the mental state of each analyst but also aspects of the state of the entire organization.

There are connections here to work on e-learning, where individual learners interact with online material to 'teach themselves'. Building such systems requires not only presenting material, but creating different, and consistent, paths through the material to suit different learning styles (e.g., [19]). It also requires building a model of each learner to predict which path, and which next step, should be presented. Unfortunately, this work is not directly applicable to intelligence analysis because it relies on knowledge of the total content to be learned, often captured by an ontology, and on large-scale commonalities across learners. In other words, e-learning focuses on learning concepts that are understood by those building the system, and the learners are all intended to reach the same state; whereas in intelligence and law enforcement, the outcome is not yet known by anyone, and different analysts may reach different conclusions. E-learning is about solving a puzzle; most intelligence and law-enforcement analysis is about better framing a mystery.

There are three ways of addressing analyst context in significance computation:

1. Ignore context. Such systems implicitly answer the question "show me what is significant" and all users get the same answer.
2. A classical-physics view, in which the model underneath is not affected by the context, but the rendering, in its most general sense, is. Significance computation acts as a filter on the unchanging underlying model. Such systems implicitly answer the question "given what I know, show me what is significant (to me)".
3. A quantum-physics view, in which the model is reconstructed to reflect the analyst context. Such systems implicitly answer the question "using what I know to recalibrate the model, show me what is significant to me".

These methods are listed in increasing order of effectiveness, but also in increasing order of difficulty.

Of course, another way in which context plays a role in model building is by the inclusion of *domain knowledge*. Such knowledge is helpful to constrain the kinds of

models that are built, eliminate models or parts of models that are not helpful, focus attention on models or parts of models that are likely to be particular revealing, and guide the collection and use of high-value attributes. However, this aspect of context is usually taken into account at an earlier stage and by different people than analysts.

3.1 Non-contextual Modelling

Many existing knowledge discovery techniques provided some kind of significance indication, often as a side-effect. For example, some predictors (support vector machines, ensembles, random forests, some neural networks) provide not just a class label prediction but also some indication of how robust that prediction is. Predictions close to decision boundaries are *interesting*. On the other hand, predictors are poor at detecting *novel* records – because these are mostly far from decision boundaries they tend to be classified with high confidence, despite being unlike any of the data on which the predictor was trained.

Clustering algorithms also sometimes provide extra information. Distance-based clustering can identify records that are far from all clusters (*novel*) and equidistant from all/some clusters (*interesting*). Density-based clustering can identify records that are not close to any cluster, but has a harder time distinguishing *novel* from *interesting*.

Predictors and clustering algorithms can easily be improved by routinely adding an ancillary 1-class clustering algorithm whose role is to examine new records and decide whether or not they are *novel*. This could be done with differing levels of sophistication, ranging from wrapping the 'known' data region in a convex hull, to using a 1-class Support Vector Machine (SVM) [16]. This would prevent predictors silently making predictions for records unlike any they were trained on; and makes it easier to identify *interesting* records in clusterings [2] – they are far from existing clusters but not *novel*.

Determining significance without any context means that the process must be completely inductive. A useful property of significance is that, roughly speaking, frequent records cannot be significant for two reasons: first, adversaries tend to be rare, so records of their actions are not usually common; and, second, common records tend to be well accounted for by basic models. This underlies some of the mechanisms outlined above.

However, the uncommon records or parts of the model that represent them are not necessarily the most significant, and this is the technical challenge. Uncommon records are a mixture of *interesting*, *novel*, and *random* records, and perhaps records that reflect individual eccentricity. Separating these categories is not always straightforward.

A major issue is that, using only induction, the relative significance of different data depends heavily on how the values of the data are normalized. In general, attribute values are expressed in units that are not naturally comparable across attributes; and where the relationship between magnitude and importance is not

linear for each individual attribute. Without care, differences in significance can be artifacts of normalization choices – but there are seldom enough principled reasons for definitively choosing one normalization over the others.

A number of knowledge discovery tools address significance explicitly. The first, and most obvious, is Pagerank [4, 5], the algorithm used by Google to rank web pages. This is an example of a general approach based on ranking relative to the first eigenvector of a matrix.

Given a graph with positive weights on the edges, the first eigenvector of the adjacency matrix points from the origin towards nodes that are well-connected by high-weight edges. Projecting all nodes onto this vector generates a ranking where high significance is identified with projection far from the origin.

If significance actually is a single-factor property, then this works well. However, if it is not, it is difficult to extend to idea to include more factors. The first eigenvector passes from the origin through a hypercube in the positive hyperquadrant towards the centroid of the data. The second eigenvector is constrained to be orthogonal to the first, but this is not necessarily the direction of maximal remaining variation in the data.

The problem can be avoided, if necessary, by normalization. For record-based data, normalization that shifts the set of points so that the centroid is at the origin removes the misleading, and often useless, initial vector [15]; for graph-based data, replacing the adjacency matrix by one of the graph Laplacians does the same thing [20]. After these normalizations, an eigendecomposition or singular value decomposition can be truncated after some chosen k dimensions, projecting the data into a k-dimensional space where its structure is more easily seen. If a specific significance ranking is needed, points can be projected onto the vector passing through the point $(\sigma_1, \sigma_2, . \quad _k)$ (the first k eigenvalues or singular values of the decomposition). Extremal points in this ranking may be *random*, *novel*, or *interesting* depending on what other processing was done beforehand. However, the distributions of points along this line may itself inductively provide information on which category to assign each point to. For example, a small but modestly outlying cluster of points in the ranking is probably *interesting* rather than *novel*, because correlated unusual activity is a strong signal of adversaries.

For graph-based data, the analysis possibilities are richer. Suppose, for simplicity, that the graph has n nodes and is connected. The graph Laplacian is obtained from a weighted adjacency matrix by computing the row sums, replacing off-diagonal non-zero entries by the negation of their values divided by the row sum, and replacing the diagonal entries by 1's.[1] After a Singular Value Decomposition (SVD) of this matrix, the main structure is revealed in columns $n-1, n-2, .$ of the U matrix – e.g., plotting the points using these two columns provides a good drawing of the graph and can be used to partition it into subclusters. In particular, sorting the nodes by the magnitudes of the entries in column $n-1$ provides a global importance ranking roughly equivalent to that of Pagerank.

---

[1]This is the random walk Laplacian.

However, the SVD does not know that the initial matrix describes a graph, so information about variation is also captured in columns 1, 2, ... Ranking nodes by the magnitudes of the entries in column 1 provides a ranking by how unusual the local neighborhood of each node is, in particular how well it is connected to its neighbors.

Thus the columns at one end of the distribution provide information about the 'big' structure of the graph, and so which nodes are significant in the context of the large-scale structure of the graph (its clusters). The columns at the other end provide information about the local structure of the graph, and so which nodes are significant in the context of small-scale structure.

But wait, there's more. The columns in the 'middle' of the U matrix (where 'middle' depends on the precise structure of the graph, but can be found by looking at the absolute value of the sums of columns) reveal small, unusual regions of the graph [14]. Most of the values of these columns are zero; the non-zero entries provide information about the strong nodal domains which are made up of nodes from which the view of the rest of the graph is unusual [6].

If the graph represents relationship among people, one set of columns provides a view that focuses on power and who is leading the group; another focuses on those whose connection to the rest of the group is unusual; and the third on small subgroups that are unusual. Which of these views of significance is relevant depends on the data and the problem domain. However, all three reveal some latent structure in the data that focuses attention onto some nodes and edges.

These approaches rank the nodes (records) by significance. But what about the significance of the edges? In a graph, the edge weights already provide a measure of significance. But graph embeddings also provide an emergent indication of edge significance. In such an embedding, the distance between any pair of points reflects the (dis)similarity between them in a global sense – that is, by integrating the pairwise similarities across the whole graph. Thus, for nodes that are connected, differences between the pairwise edge weight and the distance apart they are in the embedding provides new second-order significance information. Edges are especially significant when they have high local similarity but low global similarity (they are placed far apart); and when they have low local similarity but high global similarity (they are placed close together). Thus, computing a suitably scaled difference between the distance implied by local similarity and the distance in the embedding allows edges to be ranked by a new second-order kind of significance.

A special case is when the local similarity is zero – the nodes are not directly connected. The two nodes concerned still end up at some distance apart in the embedding, so they always have a global similarity. Two unconnected nodes that end up sufficiently close in the global embedding are probably especially significant. For example, this may mean that there is an edge between them that failed to be captured when the data was collected; or that there are multiple *indirect* paths between them, suggesting that perhaps they are concealing their direct connection [11].

An alternative approach to discovering significance is to start from the 'other end' and try to remove information that is either spurious or trivially explainable. One recent approach to this, with special usefulness for data that represents

correlation, is random matrix theory. The distribution of eigenvalues of a random matrix, appropriately scaled, is known. Given a data matrix, its eigenvalues can be calculated and the eigenvectors associated with eigenvalues that match those of the known distribution can be discounted. The remaining structure is much more likely to be meaningful [10].

3.2   *"Classical" Contextual Modelling*

In the metaphorical classical-physics approach to modelling context, the analyst's state becomes a filter on the underlying model, so that the model as presented reflects what the analyst 'already knows'. Knowledge that is already understood may be discounted or de-emphasized, providing scope for heightening attention to the data or parts of the model that are more unexpected, given the current state. However, the underlying model does not change. This avoids several thorny problems: filtering is computationally much cheaper than recomputing the model, and the issue of rollback if the state (that is, the analyst's assumption) turns out to be erroneous is avoided.

Apart from changes in the data, there are three ways in which analyst context can change:

- The analyst may acquire new knowledge from some other source. For example, investigation of an apparently *interesting* record may show that its properties arise from some technical issue, perhaps in its collection. Knowing this, similar records may plausibly be discounted.
- The analyst may have a new hypothesis or opinion, and want to ask a "what if?" question. This new internally generated information, when taken into account, may alter the significance ranking, perhaps revealing some structure that was previously hidden.
- The analyst may have a change in perspective, as the meaning of something already visible in the data or model becomes apparent. Humans are poor at understanding the implications of well-known facts, so analyst state will change by introspection, without any external stimulus, and at unpredictable times as implications of the current state become apparent.

3.2.1   New Data Is Significant

The simplest analyst state is that she/he understood the data and model at some previous point in time ('yesterday'). Significance then becomes a surrogate for 'recent', and the system should emphasize data that has arrived since the previous point in time and the changes in the model that result. Although this is often relatively easy to implement, few systems seem to provide this functionality.

At its lowest level, all data newer than some checkpoint could be labelled as significant. However, when data is plentiful, this may already be overwhelming. A generic way to abstract from large data volume while still detecting when change is occurring is to compute statistics of batches of data, and compare them. Significance is derived from changes in statistics, rather than change in data. The particular statistics used can vary widely, from simple means, medians, and standard deviations of input data to statistics of the predictions across classes, and prediction margins [1, 3].

An important way in which new data becomes significant is if it relates to some previous issue or question with which the analyst was concerned. For example, suppose that an earlier search for a connection between two people did not find one. The arrival of new data that describes such a connection makes the data more than usually significant. If the search took place a long time ago, the analyst may not even remember it, and there may have been many such searches with a negative result, so system support compensates for human weakness well. Of course, this kind of significance can only be detected if the system records analyst activity and matches new data against it. As Jonas has cogently argued, all systems should do this [9].

3.2.2 Analysts Look at the Data from a Different Direction

Many analysis tools do provide ways for analysts to alter the view of the data or models to focus on some pieces or perspectives at the expense of others. Tools such as i2 Analyst Notebook or Netmap Analytics [7, 8] allow 'slice and dice' actions to segment data in different ways, map them to timelines, visualize records and/or their connections, including sophisticated rendering using color, multiple simultaneous projections, and fish-eye presentations that magnify detail in one region at the expense of others. So far, the limitations of these tools is that the actions are not inductively generated from the data, but rely on analysts to drive them. Thus they directly support significance ranking for hypothesis-driven investigation; and indirectly support it when analysts can work out and describe which parts of model have altered significance as the result of new knowledge or new perspectives – but this may be difficult, and puts a substantial cognitive burden on the analyst.

3.2.3 Analysts Have Existing High-Level Models of Significance

Analysts working in an area develop a sense of what kinds of data are likely to be more significant, and working with a particular dataset come to understand some parts that are especially significant or insignificant. These kinds of understandings can be captured by an analyst model. In contrast to an e-learner model, though, such models should be treated conservatively (analysts may think they understand more than they do) and temporary (because adversaries constantly try to exploit analyst blindspots).

If analysts label aspects of the data or models that they think they understand, then it becomes possible to build predictive models that will try to label new data as belonging to the same general class as the understood data (and so less significant) or not (and so more significant). All of the techniques described in the previous section can be reused for this, with "objects I understand" as the 1-class label, rather than "normal objects". In other words, analysts should be able to label some subset of records, or model pieces (e.g. clusters) as understood or especially interesting, and have this information incorporated into the rendering to discount or highlight them, and others like them.

There are several specialized techniques that do not fit well with determining significance (a multiclass problem) but can be useful for modelling analyst context (a two-class problem: 'things I understand' versus 'things I don't understand'). For example, autoassociative neural networks (AANNs) use standard neural network elements and learn by back-propagation, but have a small layer in the middle. They are trained to reproduce their inputs on their outputs. The presence of the small layer forces them to do this by learning a compact representation, rather than simply copying inputs to outputs. When an AANN has been trained, the difference between inputs and outputs will be small for any record that resembles those on which it was trained, but large for a record of any other kind.

Dictionary-based compressive techniques also provide a way to quickly determine anomaly of records. Known data is used to train a model that is actually a dictionary, mapping features of records to shorter representations [13]. Once trained, such a dictionary is able to compress records like those from which it was trained well, but other records will be compressed poorly. In contrast to AANNs, a dictionary is a model that can easily be refactored, so that updating to reflect changing analyst knowledge is cheaper.

Using techniques such as these, records that an analyst understands can be used to train a model of this understanding. For example, suppose the model is a clustering and the analyst understands all of the records in one cluster, and a subset of the records in a second cluster. A model trained on the understood records can be used to provide anomaly scores for all of the records, perhaps displayed as an overlaid color code (say, from green to red) on the current model. This should indicate that the understood records have low anomaly, and are colored green. The remaining records in the second cluster might either be also labelled green, in which case the second cluster is internally consistent, or labelled as orange, suggesting that there is important substructure within the cluster. Other clusters will also be labelled with colors ranging from orange to red indicating how anomalous they are.

The standard way in which to incorporate existing knowledge into a model is to use Bayesian techniques, in particular *priors* which encode the inherent probability of certain aspects of the data. Although Bayesian approaches are well-understood and powerful, they may not be of great use in adversarial settings. Prior knowledge is likely to be fragmentary, and its meaning and importance hard to assess. This kind of information is not easy to map to a representation as a probability distribution.

3.3 "Quantum" Contextual Modelling

In adversarial settings, models will often need to be rebuilt to reflect the changing actions of adversaries, and to prevent their discovering weaknesses in models, e.g. by probing. What we envisage here, though, is much smaller and more frequent recalibration of an existing model to reflect an analyst's (changing) external knowledge or hypotheses. Because such knowledge is not always correct, and because hypotheses are not necessarily correct by definition, it is critical to be able to roll back model changes.

3.3.1 Semisupervised Learning

The general field of semisupervised learning is concerned with learning models in settings where there is a large amount of available data, but class-labelled data or known associations among records are rare or expensive [21]. For prediction, a decision boundary is learned from the labelled data, but can exploit the presence of unlabelled data because, intuitively, boundaries should not pass through regions where records are dense. For clustering, providing knowledge that a pair of records *must* be in the same cluster, or *cannot* be in the same cluster can improve results. Partial information can also be applied to graph-structured data as a regularization that makes it likely that two nodes connected by an edges will have the same label (or be in the same cluster).

Such techniques can be adapted to alter or refine existing models to reflect extra information from an analyst, rather than to build totally new models. For example, the analyst may know that (or want to see what happens if) a particular record should have a different class label. An analyst may know of a connection between two nodes for reasons outside the data itself; or may decide that an apparent connection is spurious and should be removed.

3.3.2 Reweighting

Once data have been normalized, several algorithmic approaches allow extra information to be applied by changing the weights of some parts of the data. For example, the weight of a record, of an attribute, or of a graph edge may be changed – because this is done after normalization the change is a relative one, and plausible magnitudes can be estimated. Upweighting a value makes the record(s) seem more important. For models that consider correlation among data, this has the effect of altering the apparent importance of records that resemble the altered one(s). Thus, this process can be used, for records, to "show me more like this", or to increase the impact of an attribute, or to increase the local similarity of a set of nodes in a graph.

3.3.3 Parameter Setting

One way to alter a model is to change some of the parameters that were used to construct it. In adversarial settings, model-building techniques with few or no parameters are usually to be preferred, as it is often hard to know what choices of parameters are plausible. If adversaries can guess what parameter choices are likely, they can attempt to manipulate the resulting model. Nevertheless, there are sometimes ways to change parameters to explore ranges of models of the same general kind. For example, the threshold used when mapping pairwise similarities to adjacency matrices can be altered. Because the resulting embedding depends on the entire structure of the graph, changing this threshold is more than changing the rendering of the graph.

3.3.4 Additive Versus Subtractive Analyst Knowledge

There are some subtleties in the interaction between analyst and significance computation. If the analyst context adds information to the data, then the altered model will be more conservative than the original model. However, if the analyst context is inherently subtractive, e.g. by discounting some of the available data, then the model will become less conservative. As an analyst comes to understand, and so discount, more and more of the 'central' structure of the data, there is a tendency to focus increasingly on those parts of the data that are most random. These issues have already surfaced in research aimed at developing curious robots [12] which have to decide autonomously what to do next. Careful exploitation of the fact that significance peaks in the middle of the spectrum of possible forms of new data will help to avoid this pitfall.

Analysts would be helped if systems could point them not only towards areas where significance is high, but also to areas where there is the greatest payoff for understanding, that is areas where significance could most easily be decreased. As far as we know, no work in this direction has yet been attempted.

Overall, what is needed is ways to "quotient" one model by another, so that the most general difference between the two (and its derivatives) could be calculated.

4 Discussion and Conclusions

Current intelligence and law-enforcement investigation tools are analyst driven – they require analysts to generate hypotheses, and they then provide ways to test these hypotheses against the available data or low-level models. All of the inductive skill in constructing hypotheses about what the real world situation might be belongs to the analyst. This means that analysts must have a moderately rare skill set, and be carefully and thoroughly trained. It is no wonder that intelligence failures have

been (retroactively) associated with "failures of imagination"; avoiding this would require analysts to imagine every plausible scenario and evaluate the evidence for it.

The analyst-driven approach to modelling makes it straightforward to include the analyst's context – it is an integral part of the process – but it is relatively difficult to include the organisational context, especially in a timely way.

A better balance between what analysts do well and what algorithms do well is required. Analysts are good at assessing the plausibility of scenarios with which they are presented, bringing to bear their entire awareness of human behaviour, as well as knowledge of the wider context. Algorithms are good at inducing models from data, but are much weaker at assessing the quality of these models, except with the closed computational world where they are constructed.

A better process, therefore, is one where algorithmic systems generate models and 'push' them to analysts; analysts can then assess these models and provide feedback to the system; which can then revise the models accordingly. This symbiotic loop makes better use of the capabilites of each side than the current approach. Of course, there are substantial difficulties in building such system and making them reliable. As a partial step towards this level of sophistication, computing significance allows knowledge discovery tools to indicate which parts of their models (and also of the data) are of greatest importance to a particular analyst in a particular context. Analysts today are often overwhelmed with data, so this provides a way to guide their attention and so improve their productivity.

Significance has both an inductive and deductive component. Inductively, the data and models themselves provide signals about the meaning of new data. We have suggested that it is helpful to categorize these signals into a spectrum: *normal – anomalous – interesting – novel – random*, when deciding their meaning. Deductively, the context of analysts and organizations also signals what data and models mean, suggesting that some aspects are less significant because already understood.

The technical challenge is to incorporate these two components into algorithms that provide significance information as part of model building. Some techniques already include some calculation of significance, although not in a useful way; e.g., predictor confidence can indicate data that it *interesting* but not data that is *novel*. New algorithmic techniques are needed to include significance computation explicitly, and to create ways to feed context back into model rendering and recalibration.

References

1. Abdulsalam, H., Skillicorn, D., Martin, P.: Classification using streaming random forests. In: IEEE Transactions on Knowledge and Data Engineering, vol. 22. (2010)
2. Bourassa, M., Skillicorn, D.: Hardening adversarial prediction with anomaly tracking. In: IEEE Intelligence and Security Informatics 2009, pp. 43–48. (2009)
3. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)

4. Brin, S., Page, L., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Stanford Digital Libraries Working Paper, 1998
5. Bryan, K., Leise, T.: The $25,000,000,000 eigenvector: The linear algebra behind Google. SIAM Rev. **48**(3), 569–581 (2006)
6. Davies, E., Gladwell, G., Leydold, J., Stadler, P.: Discrete nodal domain theorems. Linear Algebr. Appl. **336**(51) (2001)
7. Galloway, J., Simoff, S.: Digging in the details: A case study in network data mining. In: Intelligence and Security Informatics. Lecture Notes in Computer Science 3495, pp. 14–26. Springer, Berlin (2005)
8. Galloway, J., Simoff, S.: Network data mining: Discovering patterns of interaction between attributes. In: Advances in Knowledge Discovery and Data Mining. Lecture Notes in Computer Science 3918, pp. 410–414. Springer, Berlin (2006)
9. Jonas, J., Sokol, L.: Data Finds Data, chapter 9. O'Reilly Media, New York (2009)
10. Laloux, L., Cizeau, P., Potters, M., Bouchaud, J.-P.: Random Matrix Theory and Financial Correlations. World Scientific, Singapore (1999)
11. Liben-Nowell, D., Kleinberg, J.: The link prediction problem for social networks. In: Proceedings of the Twelfth International Conference on Information and Knowledge Management, pp. 556–559 (2003)
12. Schmidhuber, J.: Developmental robotics, optimal artifical curiosity, creativity, music, and the fine arts. Conn. Sci. **18**(2), 173–187 (2006)
13. Schmidhuber, J.: Simple algorithmic principles of discovery, subjective beauty, selective attention, curiosity & creativity. In: Corruble, V., Takeda, M., Suzuki, E. (eds.) Proceedings of the 10th International Conference on Discovery Science (DS 2007), LNAI 4755, pp. 26–38. (2007)
14. Skillicorn, D.: Detecting anomalies in graphs. In: 2007 IEEE International Conference on Intelligence and Security Informatics, pp. 209–216. (2007)
15. Skillicorn, D.: Understanding Complex Datasets: Data Mining with Matrix Decompositions. CRC, Boca Raton (2007)
16. Tax, D.: One Class Classification. Ph.D. thesis, Technical University Delft (2000)
17. Treverton, G.: Reshaping National Intelligence for an Age of Information. Cambridge University Press, Cambridge (2001)
18. Treverton, G.: Risks and riddles. Smithsonian Magazine, June 2007
19. Türker, A., Görgün, I., Conlan, O.: The challenge of content creation to facilitate personalized elearning experiences. Int. J. E-Learning **5**(1), 11–17 (2006)
20. von Luxburg, U.: A tutorial on spectral clustering. Technical Report 149, Max Plank Institute for Biological Cybernetics, Aug 2006
21. Zhu, X., Goldberg, A.: Introduction to Semi-Supervised Learning. Synthesis Lectures on Artifical Intelligence and Machine Learning 6. Morgan & Claypool, San Rafael, CA (2009)

Interdiction of Plots with Multiple Operatives

Gordon Woo

Abstract On Christmas Day, 2009, a former president of the Islamic Society at University College London, Umar Farouk Abdulmutallab, came within a detonation of bringing down a transatlantic airliner flying into Detroit. He is the fourth president of a London student Islamic society to face terrorist charges in 3 years. A fortnight later, at London's Old Vic Theatre, a revival of John Guare's classic play 'Six Degrees of Separation' opened, reminding Londoners how small is the world of social networks, and encouraging the thought that terrorist networks might not be so hard to dismantle. The public discourse on counter-terrorism intelligence is usually presented in a qualitative way, which is natural given the administrative training of the leaders of intelligence and law enforcement agencies. A quantitative analysis is presented of the intelligence challenge of gaining entry into terrorist networks, and interdicting terrorist plots. Avoiding a further slide towards a surveillance society, counter-radicalization is shown to be the key to mitigating terrorism risk to citizens in Europe and North America.

1 Introduction

Since 9/11, the homelands of the western alliance (North America, Western Europe, and Australia), have been exposed to more than 60 macro-terror plots, aimed at causing significant loss. Yet only a few of these have not been interdicted by the collective efforts of the intelligence agencies, and of those where intelligence lapses allowed the target to be approached, only a few actually succeeded in causing significant loss: the Madrid rail bombings of March 11, 2004; and the London transport bombings of July 7, 2005. There have been some very close calls where attack weaponry failed, including the attempts in December 2002 and 2009 to blow

G. Woo (✉)
Risk Management Solutions, London, UK
e-mail: Gordon.Woo@rms.com

up transatlantic planes using explosives secreted in footwear and underwear; the London transport bomb plot of July 21, 2005; and the London nightclub car bomb plot of June 29, 2007.

Prior to the Al Qaeda Jihad, the basic empirical benchmark interdiction rates have been the 80% achieved by British counter-terrorism forces during the IRA campaign in England, and the 90% achieved by Israeli counter-terrorism forces during the second Palestinian Intifada. The public expectation in Israel of flawless security is termed the '90–10 paradox.' Even if 90% of plots are foiled, it is by the 10% which were not that the security services are remembered. Commenting after the London bombings of July 7, 2005, the director-general of MI5, Eliza Manningham-Buller, remarked that '*MI5 would have to be the size of the Stasi to have the chance of stopping every possible attack, and even then, it would be unlikely that it would succeed.*' This is a stark commentary on the superior capabilities and powers that police states have in combating dissidents.

Excellent communication is a paramount skill for the head of an intelligence agency. Clear writing is essential for clear analysis, and, as a former teacher of English, Manningham-Buller was punctilious over English grammar. Her female predecessor as head of MI5, Stella Rimington, has become a spy novelist. As the centennial historian of MI5, Christopher Andrew [1], has pointed out, it is not necessary for the head of an intelligence agency to have specific technical skills, such as in foreign languages or cryptography, since these can be delegated. But, in so far as the war on terror has become an exercise in risk management, as Christopher Coker [5] has argued and the Iraq War has spotlighted, higher priority might be accorded to quantitative terrorism risk analysis, in which scenarios are assigned likelihood values. Yet, it is qualitative risk analysis, based on evidence briefings [8], that remains the foundation of public policy decision-making. The fallibility of security against terrorism attack is expressed in the police chief dictum, 'It is not a matter of if, but when'. This is true of all extreme hazards. As for earthquakes, so also for terrorist attacks, the process of quantifying the likelihood of extreme events provides instructive insight for their cost-effective risk management.

Elaborate computer models for terrorism threat prediction, such as developed for US government contracts, have been criticized by Sageman [9] for preconceived false notions of how terrorist networks behave. The purpose of this paper is to present a basic quantitative model of plot interdiction which is grounded on terrorist network behavior, and captures the key elements of the interdiction process. The outcome can be used to support risk-informed public policy decisions, especially with regard to emphasis in progressing counter-radicalization campaigns alongside the tightening of security.

2 Interdiction

The efforts of the intelligence and law enforcement services to interdict plots is crucial for mitigating terrorism risk. As Magnus Ranstorp has pointed out [3]: 'Once a terrorist moves towards his target, half the battle is already lost'. All

conspiracies organize to maintain operational secrecy as far as possible. The internal organizational structure can vary from top-down to bottom-up, with some plots being directed from a centralized leadership, and others being inspired by the central leadership, but emanating from more decentralized planning.

Regardless of the internal structure of a terrorist cell, the key to its dismantling is the same as for storming a fortress, namely managing to find an entry point into the cell network. Once one member has been identified, then close surveillance of his movements and communications can unravel the network and ultimately expose most if not all of the other cell members. Thus, even though a Jihadi terrorist cell may be internally organized to minimize the potential for the leakage of plot information, a plot nevertheless can be compromised if a single entry point can be found into the cell network. Any cell member may inadvertently provide an entry point into the plot network through a link outside the network. A dangerous external link can occur in various ways, which are detailed below.

A cell member may come into contact with someone who happens to be an informant or member of the intelligence services, or who happens to be under surveillance by the intelligence services. This contact can arise via a personal meeting or phone or electronic communication within the community supportive of the Jihad, e.g. through the extended family, a mosque, an Islamist gathering such as organized by a college or Islamist society such as Hizb-ut-Tahrir, a Muslim training and fitness center, or even an Islamic bookshop as posited by ex-MI5 head, Stella Rimington.

Alternatively, the contact can arise through a virtual online meeting in a Jihadi forum. The Dark Web project addresses this sector of cyberspace (e.g. [4]). Almost all U.S. intelligence agencies, the Pentagon and the National Security Agency monitor Jihadi forums with a large number of Arabic-speaking agents. Likewise, Israel and neighboring Arab states have many Arabic-speaking agents tracking Jihadi forums. Intelligence agent postings may be outspoken in their advocacy of violence, and have been effective in luring potential terrorists.

2.1 Lucky Leads

An entry point into a plot network can also emerge through contacts outside the community supportive of the Jihad. There may be suspicions of ill-intent amongst the mainstream Muslim community. The Lakawanna group came under suspicion when an anonymous letter from an Arab–American arrived at the FBI building in Buffalo, New York. The letter named those who had received terrorist training and been recruited by Al Qaeda. In Britain, Isa Andrew Ibrahim, who was convicted in July 2009 of terrorist offences, was largely unknown to UK police until contacted by Muslims deeply concerned about his radical views and chemical experiments.

Public vigilance can also provide an entry point into a plot network. A Circuit City employee became alarmed while transferring the video of a Jihadi training session onto DVD for the Fort Dix plotters, convicted in 2008. The storage

of a large quantity of fertilizer for an inordinate length of time aroused public suspicion and assisted the interdiction of a major UK bomb plot. The police may also stumble across a plot through observing suspicious behavior, such as photographic reconnaissance or pseudo-military training, or through making a fortuitous discovery of terrorist attack plans. This happened in Torrance, California, during a routine police investigation of a gas station robbery in 2005. An apartment search uncovered documentation on modes of attack by 'Al Qaeda in California'.

Whilst head of the Metropolitan Police counter-terrorism unit, Peter Clarke noted that few prosecutions have yet originated from community intelligence. Based on US and UK experience since 9/11, between 10 and 15% of plots are interdicted through happenstance: a chance discovery or encounter, or timely tip off from the vigilant public. If plots were only foiled through such luck or in Jihadi parlance 'as God wills', terrorism would not be controllable. Indeed, in countries of the developing world where the security services are poorly funded and trained, ineffective, corrupt or compromised, the contribution of intelligence to supplement this basic chance level of interdiction can be very modest. By contrast, in the countries of the western alliance, where the intelligence services are professional, reliable and coordinated, systematic intelligence gathering is rewarded with a far superior interdiction rate.

But the level of interdiction is subject to a false negative error rate of intelligence leads which are not adequately followed up to prevent a terrorist attack. From observations since 9/11, this loss of efficiency more or less cancels out the gain in efficiency due to lucky leads arising. Thus good luck in receiving tip-off information is offset by the incorrect judgement call in not following through with initial contacts. Corresponding to a person such as Isa Andrew Ibrahim, arrested after a tip off, is a person of intelligence interest such as Nicky Reilly, who was monitored, but was discounted as a credible threat because of his autism. In May 2008, he went as far as setting off a nail bomb in a restaurant, but it malfunctioned and injured only himself. Circumstances can arise where a stroke of luck in plot discovery may not actually interdict a plot, because it is not followed up adequately. On May 2, 2004, an off-duty policeman stumbled across the July 21, 2005, London bombers training in the Lake District of Northwest England. A surveillance team was called in and photographed them, but the bombers were not kept monitored, and their attack took place.

3   Two Degrees of Separation

Valdis Krens [7] has been prominent in developing social network analysis as a mathematical method for connecting the intelligence dots. His analysis of the 9/11 terrorist network is particularly illuminating. Two suspected terrorists linked to Al Qaeda, Nawaf Al-Hazmi and Khalid Al-Midhar, were covertly photographed attending a meeting in Malaysia in January 2000. One of the chief suspects in the bombing of the USS Cole later in October 2000 happened also to be present at this

meeting. Krebs has pointed out that all 19 of the 9/11 hijackers were within two degrees of separation of these two suspects known to the CIA in early 2000. A key hub linking the hijackers was Mohammed Atta.

Current intelligence practise adopts the findings of this lesson on terrorist networks. According to a former staff member at the National Security Agency, '*Analysts start with a suspect and spider-web outward, looking at everyone he contacts, and everyone those people contact, until the list includes thousands of names. Before individuals are actually wiretapped, computers sort through flows of meta-data, information about who is contacting whom by phone or e-mail.*' By building up concentric circles of information, connecting each suspect to others, intelligence agencies can piece together a conspiracy, without needing the intrusive authority to mine information on entire civilian populations, which would jeopardize fundamental civil liberties.

Such is the extent of the social network of radicalized supporters of the Jihad who are within two degrees of separation of terrorist suspects, that any new conspiracy against the homelands of the western alliance has a significant chance of being discovered through a dangerous link. The likelihood of link detection can be estimated.

3.1   Likelihood of Link Detection

The overwhelming majority of Muslims deplore political violence. However, a small radicalized minority are supportive of the Jihad in that they would condone political violence, and turn a blind eye to any plot information of which they became aware. Omar Bakri Mohammed, the founder of Al Muhajiroun in UK flaunted his approval of martyrdom missions and his non-cooperation with British police terrorist investigations. Within the ranks of radicalized Muslims are the extremists who would be prepared to have some secondary or peripheral involvement with a plot, to assist the hard-core men of violence. Abdullah Azzam, the mentor of Osama bin Laden, defined this segmentation of the radicalized Muslim community as follows: '*A small group: they are the ones who carry convictions for this religion. An even smaller group are the ones who flee from this worldly life in order to act upon these convictions. And an even smaller group from this elite group are the ones who sacrifice their souls and their blood.*'

Recruitment of a Jihadi terrorist cell would draw upon the latter two categories. Apart from contacts within the cell, each terrorist cell operative will have some links with the broader radicalized community, quite confident they would not assist substantively with any terrorism probe. Consider a link between a terrorist cell operative and a member of the radicalized community. Of significance in terrorism threat assessment is the likelihood that such a link would be detected by the intelligence and law enforcement agencies.

First, a terrorist cell operative is likely to have links with some extremists, e.g. Jihadi leaders such as Omar Bakri Mohammed or radical clerics such as Anwar Al Awlaki, well known by the operative to be under intelligence watch. Communication

with such known extremists is only slightly hazardous, because they are major Jihadi network hubs, with a proliferation of links to a large body of supporters of the Jihad, making it hard to distinguish the few with actual violent intent.

Apart from linkage with well known extremists, a terrorist cell operative may have some links with other extremists who happen also to be under surveillance, although such intelligence activities remain secret and clandestine. Such links would provide opportune cell entry points, as would haphazard linkage with informants or covert intelligence officers.

In the UK, assessment of the likelihood of such link detection is greatly facilitated by the greater openness of MI5 after the July 7, 2005 bombings, for which they were criticized by the families of the victims. As of November 2007, MI5 admitted to keeping under constant surveillance 2000 people directly connected to Islamist terrorist plots. According to the director-general, Jonathan Evans, countless more are involved in fundraising, helping people to travel to Afghanistan, Pakistan and Somalia, and providing equipment, support and propaganda. Based on surveys, Al Qaeda is admired by about 60,000 UK Muslims, in the category most prone to become drawn into violent extremism.

MI5 currently employs around 3,500 people, the great majority of whom are engaged in counter-terrorism. Given that there are about 1,350 mosques in UK, there may be around a thousand agents or informants embedded within the radicalized Muslim community. So, apart from the 2000 individuals under surveillance, another 1,000 moles would also be dangerous for a terrorist cell operative to contact. In terrorist jargon code, they would be 'contagious'. The chance that a general member of the radicalized Muslim community might be either under clandestine surveillance or be an informant is estimated at about $(2,000 + 1,000)/60,000 = 1/20$. Thus, a link to this radicalized community from a terrorist cell operative has a 1/20 chance of leading to uncloaking. The larger the number of admirers of Al Qaeda, the smaller this chance is, and the better for conspirators. This situation is corroborated by the disclosure by MI5, in May 2009, that in the run up to the July 7, 2005 London transport bombings, because of the numerous leads from the radicalized community, they could properly track only 1/20 of the people it connected with terrorist cells.

This comparatively modest chance of spotting a terrorist link amidst the background noise of a crowd of Jihad supporters is reasonably consistent with the statistics of the Terrorist Screening Data Base (TSDB). The Christmas Day 2009 airliner plot has focused attention on the fact that fewer than 4,000 names from the 400,000 TSDB are on the no-fly list, while an additional 14,000 are on a selectee list that calls for mandatory secondary screening. The ratio of the numbers in the combined selectee and no-fly lists to the numbers in the Terrorist Screening Data Base is then slightly under 1/20.

3.2   Network Entry Likelihood

Consider a cell network with N operatives. Entry into the network may be gained if a cell operative has a link with an informant, or a radicalized Jihad supporter

unwittingly under surveillance. As a guide to the average number of links an operative might have with the radicalized Jihadi community, statistics on the ratio of terrorism convictions to arrests are relevant, because a sweep of a terrorist cell with the impounding of computers will tend initially to trawl in their radicalized contacts. As with Dr. Mohammad Asha, a Jordanian neurosurgeon working in an English hospital, who was acquitted of any involvement in the London and Glasgow terrorist attacks on June 29 and 30, 2008, an association with a suspect may be initial grounds for police questioning and arrest.

In the UK, which has suffered the largest number of Jihadi plots of any western country since 9/11, the most recent figures up to 2008 show there have been 200 terrorism convictions against a total number of terrorism-related arrests of over 1,450. This suggests that a cell of N operatives might, on average, have about 6N contacts within the radicalized community. Each operative is likely to have individual links with an extended family, Jihadi forums and websites relating to their designated function within the cell, Islamist societies, Muslim organizations, mosques, Islamic stores, etc. Accordingly, the minimum number of links of a cell to the radicalized community is estimated to be 3N. With 6N being the average number of cell links, prudence and vigilance suggest for any operative a practical upper bound of about 10 links to the radicalized community distinct from those of other cell members.

3.3   Cell Size Dependence

According to MI5, the 2000 people under surveillance form an estimated 230 networks suspected of links to terrorism. This corresponds to an average of about 8 operatives per network. This happens to be the number of terrorists in Dhiren Barot's network convicted in 2006/7 of terrorist offences. Included among the cell were specialists in bomb manufacture, architecture, building security systems, surveillance and driving. For major macro-terror attacks, involving meticulous preparation and professional terrorism trade-craft, a minimum of two operatives would be needed to generate a significant loss with a reasonably high chance of logistical and technical success.

For the failing state heartlands of Al Qaeda Central, namely Afghanistan/ Pakistan, Yemen and Somalia, the lack of state control of terrorist operations is accounted for by centralizing involvement in any western plot at the node of the mastermind. Al Qaeda freedom to develop and test weaponry and conduct training in these regions with little state impediment is of course the rationale for western involvement there.

For values of cell size N varying from 1 to 10, the interdiction probability for a cell plot is calculated by a computer simulation process which considers a large number of possible alternative external link configurations for the cell operatives. For each configuration, the total number of distinct external cell links to radicalized supporters of the Jihad and informants is enumerated, and the

Table 1 Dependence of plot interdiction probability on cell size

Cell size	1	2	3	4	5	6	7	8	9	10
Plot interdiction probability	0.26	0.46	0.60	0.70	0.78	0.84	0.88	0.91	0.93	0.95

interdiction probability is calculated as the complement of the likelihood that no cell link is dangerous. The chance of any dangerous 'contagious' link compromising the cell is taken to be 1/20. Thus if, for one configuration, there are K external links, then the interdiction probability for this configuration is $1 - (0.95)^K$ (Table 1).

4   Notable Non-interdicted Plots Since 2006

The high interdiction rates achieved by the western alliance since 9/11 reflect the fact that most of the major plots against the western alliance homelands have involved four or more operatives. Since the Madrid rail bombings of 2004 and the London transport attacks of 2005, the intelligence constraints on large plots have been considerably tightened. However, for plots involving small cells of two or three, with a compact profile and small external footprint, interdiction rates are lower.

4.1   The German Rail Attacks of July 2006

In July 2006, two bombs were found on trains in Dortmund and Koblenz in Germany. Because of a construction flaw, the propane gas tank bombs, which apparently used designs found on the internet, failed to detonate. Two Lebanese men, Hajj Dib and Jihad Hamad, who had come to Germany to study, were caught on CCTV placing the suitcase bombs on the trains while halted at Cologne station. Hajj Dib, who had been angered by the publication of Danish cartoons of the Prophet Muhammad, had links with extremist groups in Lebanon. His co-conspirator, Jihad Hamad, was arrested in his native Lebanon in August 2006 after fleeing Germany. Hajj Dib stood trial in Germany, was convicted and given a life sentence, reflecting the proximity of the safe German train system to suffering the terrorism mass casualty levels witnessed in London in 2005.

4.2   The London/Glasgow Attacks of June 29 and 30, 2007

Another small cell which succeeded in developing a plot which was not interdicted was that formed by Bilal Abdulla and Kafeel Ahmed. On June 29, 2008, they

attempted to bomb the Tiger Tiger nightclub in Central London, using a car loaded with propane gas tanks. When this failed, they returned to Scotland and rammed an SUV into Glasgow airport terminal the following day. Kafeel Ahmed subsequently died of his burns. A search of Kafeel Ahmed's computer retrieved from his Bangalore home turned up a large quantity of Jihadi material, which had been influential in shaping the plot strategy. For the material included downloads of the voluminous teachings of Abu Musab Al Suri, a noted theorist of the Jihad. His masterwork, 'Call for a Global Islamic resistance', espoused the operational value of network decentralization, epitomized by this plot. What few links the conspirators had with the radicalized Muslim community, including Hizb-ut-Tahrir, they were insufficient to draw the attention of MI5.

During the period 1992–1997, Al Suri lived mostly in Spain and then in London, where he worked with the media unit of the Algerian GIA. He was instrumental in setting up the notorious Al Qaeda cell led by Abu Dahdah in Spain. The Madrid rail bomb attack of March 11, 2004, is an example of a successful operation by a sizeable cell, which was actually quite close to interdiction. Abu Dahdah's phone number was known to one of the plot ringleaders, Jamal Zougam. Prior to the bombings, his apartment had been raided by Spanish police, following the arrest of Al Qaeda operative David Courtailler, who had Zougam's name in his address book.

4.3   The Airline Bomb Plot of December 25, 2009

The December 2001 airline shoe bomb plot involved two British operatives, Richard Reid and Saajit Badat. They were both in contact with Nizar Trabelsi, who was arrested in Brussels following information extracted from Djamel Beghal, a senior Al Qaeda officer, who had been arrested in Dubai earlier in the year. Analysis of phone card communication with Trabelsi was to lead to Badat's subsequent arrest and conviction, but might have helped earlier to interdict the plot, which proceeded despite Trabelsi's arrest.

Eight years later, another airline bomb plot might also have been interdicted, but was not. This plot was smaller, involving one suicide operative rather than two. The father of the alleged bomber, Umar Farouk Abdulmutallab, had tipped off the US embassy in Lagos with concern over his son's radicalization, and intelligence had been collected on the involvement of a Nigerian in a plot. Furthermore, whilst in London, he did have some contacts known as extremist targets of MI5 interest. But, as president of the Islamic Society at University College London, it would be unsurprising if the Nigerian banker's son shared the radical politics of speakers at an event he organised in January 2007, featuring talks on Guantánamo Bay, prison torture and the war on terror.

Six months before, at Islam Expo, a major Muslim congress in London, the assertion by British member of parliament, George Galloway, that 'Israel, UK, and USA are the axis of evil' was greeted with rapturous applause by a Muslim audience that largely sympathized with the Jihad, even if not sporting the Mujahid T-shirts

openly on sale. In the aftermath of the airline bomb plot, a public policy intelligence analyst, Amy Zegart, made the remark, echoed elsewhere that, 'It could only have been more obvious if the guy had been wearing a T-shirt, saying I'm a terrorist.' It is clear that errors in intelligence analysis were made, but all too many Muslims would condone such T-shirts. Popular sympathy for the Jihad would have made it very difficult for MI5 to have distinguished Abdulmutallab from numerous other Muslim college students in Britain who earnestly espoused Islamist ideals.

5   Conclusions

In the aftermath of 9/11, there was doubt as to whether as high a general interdiction rate in western countries as 80% might be achieved against Al Qaeda. According to the director-general of MI5, Jonathan Evans, in a speech delivered at his alma mater, Bristol University, in October 2009: '*After 9/11, the UK and other western countries were faced with the fact that the terrorist threat posed by Al Qaeda was indiscriminate, global and massive. We have a better understanding of the nature and scope of Al Qaeda's capabilities but we did not have that understanding in the period immediately after 9/11*'

Part of this better understanding of Al Qaeda's capabilities is an appreciation of the benefits of expanded intelligence operations in interdicting terrorist plots. Especially since the London transport bombings of 2005, the increased resources allocated to intelligence agencies have been instrumental in thwarting almost all the major plots against the western alliance homelands.

In an analysis of the patterns of links between Jihadi supporters, a tipping point was previously identified [10] in the number of operatives collectively involved in conspiracies: too many terrorists spoil the plot. The organization of large inter-linked plots, often favoured by Al Qaeda, makes interdiction much easier. This has since been demonstrated in the unravelling of a number of significant ambitious terrorist plots in Belgium, UK, and USA linked back to the Al Qaeda mastermind, Rashid Rauf, in Pakistan.

The analysis here affirms the high likelihood of interdiction of plots with sizeable cells. But it also highlights vulnerability to plots with only a few cell members. A plot by a lone terrorist would only have a comparatively modest chance of being interdicted, about 1 in 4. The loss that an individual can potentially cause is of tragic dimensions, as shown by Mohammed Bouyeri's brutal assassination of Theo van Gogh in Amsterdam in November 2004, and 5 years later by Major Nidal Hasan's shooting rampage at Fort Hood, Texas on November 5, 2009. However, the loss level is typically on the scale of potential criminal action. A fire attack on Milan's subway in 2002 was far less of a public safety concern than the audacious 2004 plan to set off C-4 explosives on multiple subway cars, as well as at the Duomo hub station in Milan.

A small cell of two or three could wreak havoc on a much larger scale, especially when advantage is taken of a force multiplier like civil aviation. Accordingly, for the

assurance of public safety, the interdiction rate for such plots needs to be increased from a little above evens to around 3 in 4, which is nearer to the level expected by the western public of counter-terrorism security.

In order to achieve a goal of an interdiction rate of 3 in 4 for plots emanating from small compact cells, simulation analysis indicates that the link detection likelihood needs to be doubled from 1/20 to 1/10. This can be realized through a combination of increasing the numbers under surveillance, enlisting more informants, lowering the error rate in intelligence analysis, and reducing the size of the radicalized Jihadi community. As policy options, these alternatives can be considered in turn.

Whilst sustained growth in western intelligence capability has been welcome over the past few years, a massive further expansion of surveillance activities, with the recruitment of more informants, would run counter to the civil libertarian principles of the western democracies, in the protection of which extremism is being fought, and might be counter-productive in alienating non-radicalized Muslim populations.

There is always room for some improvement in the accuracy and reliability of intelligence analysis, as is clear from retrospective assessment of past analysis failures. But the adaptive nature of terrorism, as exhibited particularly in the changing styles of aviation threat from hijacking to blowing up passenger and cargo planes, does bound the scope for improvement. The option which offers the greatest prospect for continuous sustained threat reduction is counter-radicalization. As embodied in the ICSR, the International Centre for the Study of Radicalization (www.icsr.info), inaugurated in January 2008, counter-radicalization has become a mainstream international academic and diplomatic focus. The rehabilitation and de-radicalization of militants and extremists is also a key topic of research across the world in Singapore [6]. Progressive diminution in the size of the radicalized Muslim population would sharpen the resolution of intelligence gathering, and raise the interdiction rate for small cell attacks of the kind assessed by Bergen [2] to be the predominant threat to the US homeland.

References

1. Andrew, C.: The Defence of the Realm: The Authorized History of MI5. Allen Lane, London (2009)
2. Bergen, P.: Reassessing the evolving Al Qaeda threat to the homeland. Testimony before the House of Representatives committee on Homeland Security (2009)
3. Bromfiel, G.: Homeland-security research: Mission impossible. Nature **419**, 10–11 (2002)
4. Chen, H.: IEDs in the dark web: Lexicon expansion and genre classification. In: ISI'09 Proceedings of the 2009 IEEE Conference On Intelligence and Security Informatics. (2009)
5. Coker, C.: War in an age of risk. Polity, Cambridge (2009)
6. ICPVTR: http://www.pvtr.org/pdf/Report/RSIS_PakistanReport_2010.pdf. (2010)
7. Krebs, V.: Connecting the dots. www.orgnet.com. (2002)
8. Omand, D.: Securing the State. Hurst & Co., London (2010)
9. Sageman, M.: Leaderless Jihad. University of Pennsylvania Press, Philadelphia (2008)
10. Woo, G.: Intelligence constraints on terrorist plots. RUSI/Jane's Sentinel (2007)

Understanding Terrorist Network Topologies and Their Resilience Against Disruption

Roy Lindelauf, Peter Borm, and Herbert Hamers

Abstract This chapter investigates the structural position of covert (terrorist or criminal) networks. Using the secrecy versus information tradeoff characterization of covert networks it is shown that their network structures are generally not small-worlds, in contradistinction to many overt social networks. This finding is backed by empirical evidence concerning Jemaah Islamiyah's Bali bombing and a heroin distribution network in New York. The importance of this finding lies in the strength such a topology provides. Disruption and attack by counterterrorist agencies often focuses on the isolation and capture of highly connected individuals. The remarkable result is that these covert networks are well suited against such targeted attacks as shown by the resilience properties of secrecy versus information balanced networks. This provides an explanation of the survival of global terrorist networks and food for thought on counterterrorism strategy policy.

1 Introduction

As researchers have begun to unravel the structure and dynamics of many different social, biological and other complex networks [9, 17, 20]. It is realized that the study of criminal and terrorist networks can also benefit from insights thus obtained [24]. Typically research on terrorist or criminal networks, i.e., covert networks, considers destabilization strategies [6, 8], organizational characterizations [7, 14, 15]

R. Lindelauf (✉)
Netherlands Defence Academy, Breda, The Netherlands
and
Tilburg University, Tilburg, The Netherlands
e-mail: rha.lindelauf.01@nlda.nl

P. Borm · H. Hamers
Tilburg University, Tilburg, The Netherlands
e-mail: p.e.m.borm@uvt.nl; h.j.m.hamers@uvt.nl

and methods for key player identification [5, 19]. Network oriented research in this domain is ordinarily done by either assuming a fixed network topology or by the use of empirical historical data [3, 13]. However, data on this topic is often inaccurate and anecdotal due to the widespread secrecy surrounding governmental data-sets. Mathematical models provide an alternative method for gaining insight into covert organizational structures. Once data becomes available these models can be evaluated and adjusted if necessary.

For many types of (overt) networks the position of their connection topology between the extremes of order and randomness has been established [21, 22]. However, little is known about the exact position of the connection topology of covert networks, and consequently about their resilience against disruption. The current study shows where covert networks are positioned by making use of their topological characterization as secrecy influenced communication structures [11, 12]. We find that the common characterization of social systems as small-world networks is generally not applicable to covert networks. This phenomenon can be explained by the fundamental dilemma such organizations have to solve: how to efficiently coordinate and exercise control while at the same time remaining secret. We corroborate our results with empirical findings of Jemaah Islamiyah's bombing of a Bali nightclub [10] and the active core of a heroin distribution network in New York City [16]. In addition, we show that a covert network topology is strongly resilient against disruption strategies focused on capturing and isolating highly connected individuals, partly explaining the difficulties in disrupting current terror networks.

Theoretical results and empirical investigations indicate that terrorist organizations have to make a trade-off between efficient coordination and control on the one hand and maintaining secrecy on the other. For instance, Enders and Su [7] model the process by which terrorists select 'between the competing ends of security versus the unbridled flow of information'. They argue that rational terrorists will attempt to counter increased efforts at infiltration and restructure themselves to be less penetrable, often by adopting certain network structures. That terrorist organizations take secrecy explicitly into account is well known: in a video lecture Mousab al Suri (an alleged Al Qaeda affiliate that was captured in November 2005) indicates that certain network structures should be avoided to ensure the secrecy of the organization [4]. Terrorists operating according to networked organizational forms are also observed in practice. For instance, the Nov. 26 Mumbai attack showed tactical commanders and individual team members using satellite and cell phones to connect to strategic commanders out of theatre. Multiple teams consisting of several individuals were able to communicate and direct each other as the attacks progressed. What sets apart such attacks is not the use of technology per se but the networked mode of operation that it enables. The organizational form of these attackers is not easily characterized as being 'hierarchical' or 'decentralized'. However, what is clear is that terrorist, insurgent and criminal organizations are increasingly able to cross borders, engage in fluent relationships and 'swarm' their objectives to achieve their goals. The underlying mechanism to all these operations

is the networked topology: information is exchanged on communication networks, weapons diffuse through trafficking networks and Shura councils meet in affiliation networks. It is, therefore, of paramount importance to understand these network structures.

Lindelauf et al. [11] introduce a multi-objective optimization framework to analyze the structure of terrorist networks taking the secrecy versus information tradeoff into account. That this tradeoff exists is intuitively clear: if everybody in the covert organization knows everybody else, then the security risk to the organization is very high because the exposure of an individual potentially exposes the entire organization. On the other hand, a very sparsely connected organizational network topology is difficult to coordinate and control, simply because efficient communication between individuals in such an organization is hard. We capture these critical considerations by use of an information measure I, a secrecy measure S, and a balanced trade-off measure μ. A detailed description of the methodology used can be found in the appendix. The information measure I reflects the fact that the ability to transfer information between individuals in a network is inversely proportional to the number of edges in the shortest path between those individuals. On the other hand, the secrecy measure S reflects the fraction of individuals in the network that is expected to remain unexposed upon capture of individuals according to a realistically chosen probability distribution. Finally, the total performance of a covert organization in dealing with the information versus secrecy trade-off dilemma is reflected by the multi-objective optimization based function $\mu = SI$. The higher the value of μ a network attains, the better it does in balancing secrecy and information.

Terrorist networks evolve, i.e., 'it would be naive to think that terrorists and their networks would remain invariant to measures designed to track and infiltrate the inner workings of their organizations' [7]. To what kind of structures do these terrorist networks evolve? Clearly, we argue that the proactive counterterrorism activities after 9/11 have resulted in terrorist networks adopting more decentralized, non-hierarchical networks, i.e., they have taken secrecy explicitly into account as design parameter. Thus these terrorist networks are a very special subset of general social networks about which a great deal is known. For instance, it is well known that many social networks can be characterized as small-worlds, i.e., most individuals in the network can be reached by a small number of steps. The evidence concerning terrorist network structures however is often anecdotal, providing an impetus for the development of theoretical models of covert networks. The aim of this article is therefore to analyze the structure of secrecy influenced terrorist networks, investigate their small-world properties and the resulting consequences on their survivability properties. The next section discusses the application of the secrecy versus information tradeoff characterization of terrorist networks to the analysis of their small-world structure. The important insight is that such terrorist networks do not appear to be small-worlds. A fact that can be motivated from a secrecy standpoint. In addition we will present empirical proof of this claim. Next, we investigate the resilience of these terrorist network structures against disruption. We find that such network structures perform well against disruption, actually they

outperform common social networks in case of targeted attacks. A fact that should have profound implications for counterterrorist strategies.

2   Small-World Network Analysis

Watts and Strogatz [21] quantified small-worlds as networks with low characteristic path length L and high clustering coefficient C relative to random networks with the same number of vertices. The characteristic path length L is a global property that measures the typical separation between two individuals in the network. Obviously, the characteristic path length L will be inversely related to the information measure I. This because a high separation between terrorists in the network will make it difficult for them to coordinate and control as reflected by a low information measure. The clustering coefficient C, a local property, measures the cliqueshness of a typical neighborhood. In many social networks an individual's friends are also friends among each other. Clearly, in covert networks this in general will not be the case because too many interconnections among individuals will degrade the secrecy of such an organization. The clustering coefficient C is based on the number of edges that exist between the neighbors of each vertex.

It is generally argued that covert organizations facing an exogenous threat transform into hybrid network structures that lie somewhere in between sparse networks (such as the star, ring, lattice or path) and the complete network in which everybody is connected to everybody else [2]. To simulate this transformation we interpolate between regular networks and the complete network and for each instance establish the optimality of the resulting network with regard to the secrecy versus information tradeoff characterization. To investigate whether the small-world characterization of various social networks also holds true for covert networks, we thus generate intermediate hybrid networks. Our procedure starts with an initial network (a star, ring, path or lattice) and with a probability p that each vacant edge is added. For fixed values of p several indicators relating to the small-world structure (L, C) and the secrecy versus information tradeoff (μ) of the network are computed and averaged over 20 realizations. In Fig. 1, we plot the normalized values of L, C and μ versus p for each of the four possible initial networks. It can be seen that the maximum value of μ, indicating approximate optimal covert network structures, is typically not attained at low characteristic path lengths and high clustering coefficients, features that characterize small-world networks. For instance, if the initial graph equals the path graph (Fig. 1 second from top), then it can be seen that μ attains its maximum around $p = 0.09$, where the value of L is small and C does not attain a high value. In particular, the tiny fraction of shortcuts that suffices to create small-worlds, although increasing the ability to communicate, increases the security risk to a covert network. Clearly, covert networks favor low clustering because this is in the interest of secrecy whereas low characteristic path lengths ensure the necessary communication and control abilities.

Fig. 1 Normalized characteristic path length L, clustering coefficient C and performance measure μ as a function of the probability p with which each vacant edge is added to an initial network which is a lattice (*top*), path (second from *top*), star (second from *bottom*) or ring (*bottom*). All networks have 100 vertices and L, C and μ are averaged over 20 realizations for each of the values for p
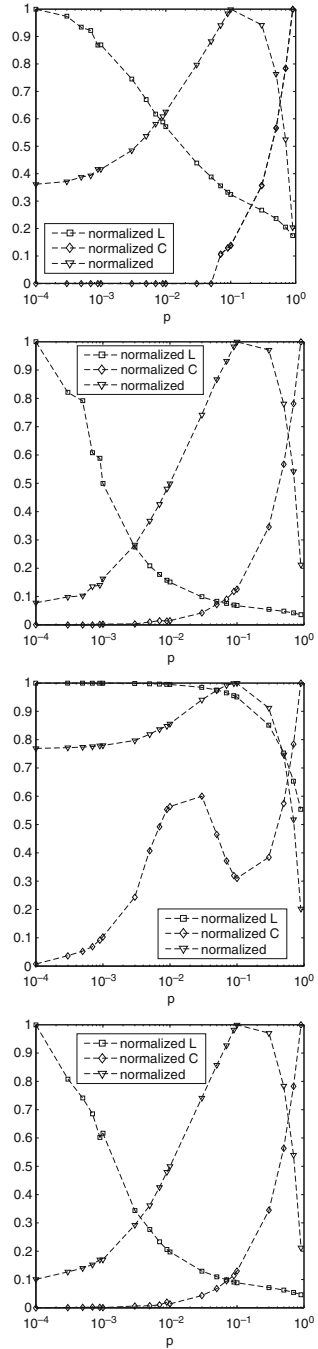
3   Empirical Examples

Our simulation shows that the small-world phenomenon is not characteristic of theoretically optimal covert networks. To obtain additional evidence we compute the characteristic path length and clustering coefficient for empirical covert networks: a heroin distribution network in New York city and the Jemaah Islamiyah cell responsible for the Bali bombings in 2002. We compare these values to the characteristic path length and clustering coefficient of a graph with the same number of vertices in which every possible edge occurs independently with probability $p = \frac{1}{2}$, i.e., a random graph. To compare these outcomes with networks that are small-worlds we present an empirical example of a film-actor network [21]. It can be seen that both empirical covert networks do not show the small-world phenomenon because their characteristic path lengths as well as clustering coefficients are comparable to those of a random network (Table 1). The film actor network, however, is a small-world: its characteristic path length is of similar order as the random graph on the same number of nodes whereas its clustering is much higher.

It is also interesting to investigate whether these empirical covert networks optimize their structure according to the theoretical framework on the secrecy versus information tradeoff dilemma. Therefore, we compute μ for both empirical covert networks (μ_{he} and μ_{ji} respectively) and we approximate the optimal value of μ on networks of the same order (μ_{he}^{opt} and μ_{ji}^{opt} respectively). We find that $\frac{\mu_{he}}{\mu_{he}^{\text{opt}}} = \frac{0.33}{0.39} = 0.85$ and that $\frac{\mu_{ji}}{\mu_{ji}^{\text{opt}}} = \frac{0.28}{0.38} = 0.74$. We may conclude that both networks attain empirical values for μ that are close to optimal and hence correspond to the region (Fig. 1) within which the existence of a possible small-world structure is contradicted. Thus we obtain further evidence for the fact that covert organizations are not small-worlds. In the next section we will explain the advantage of adopting structures differing from small-worlds.

4   Covert Network Resilience

To counter the terrorist threat it is essential to focus on the terrorists [18]. Generally speaking, in countering a covert network, the removal or isolation of individuals is a key strategy, the effect of which is in part determined by the network's robustness

Table 1 Comparison of characteristic path lengths and clustering coefficients of two empirical covert networks and an overt empirical network (film actors) and 100.000 randomly generated graphs with the same number of vertices

	L_{actual}	L_{random}	C_{actual}	C_{random}
Heroin network	4.74	4.93	0.44	0.13
Jemaah Islamiyah	3.18	3.11	0.89	0.46
Film actors	3.65	2.99	0.79	0.00027

properties. An example of this is the U.S. government's hope on decapitating Al Qaeda by pursuing high-value targets. In complex network theory it has been shown that networks with a few highly connected nodes (hubs) are resistant to random failures because these hubs dominate their topology [1]. However, this comes at the cost of vulnerability to deliberate attacks on such hubs. This appears one of the reasons why empirical covert organizations, instead of relying on a few hubs, have evolved into decentralized, non-hierarchical structures as theoretically quantified by our secrecy versus information trade-off performance measure. A case in point is Al Qaeda [18]: local groups self-organize by radicalization and interconnect, for instance through the internet. There is no top to bottom leadership or organization. What results is a sparsely connected network safeguarding secrecy, however with low separation (due to the internet's global reach). To understand the resilience of such an organizational form we investigate the effect of the removal of a fraction of vertices of an approximate optimal terrorist network on the basis of the secrecy versus information performance measure. More specifically, we compare two scenarios: a fraction f of vertices is either removed randomly from such an approximate optimal terrorist network or the same fraction f being removed consists of vertices with the highest degrees. Results are plotted in Fig. 2 (top) in case of random removal and in Fig. 2 (bottom) in case of targeted removal.

From Fig. 2 (top) it can be seen that the fraction of randomly removed vertices does not seem to affect the performance of the remaining network structure very much, i.e., μ is only slightly decreasing with increasing values of f. Only after a very large fraction of vertices has been removed ($f \approx 0.75$) does an effect take shape, which can be explained by the disintegration of the network. Figure 2 (bottom) on the targeted removal of high degree vertices shows a surprising result. Even though the information measure decreases rapidly with increasing f, the secrecy performance of the organization in fact increases. Balancing these two aspects, the total performance measure μ increases with respect to the fraction f of targeted removals. Thus, the more one focuses a destabilization strategy on the targeted removal of central individuals the more a covert organization's capacity to coordinate and control is reduced but the higher its performance in balancing the information versus secrecy trade-off will be. Only at very high values for the fraction $f \approx 0.8$ does the performance measure start to decrease. The implication for terrorist networks is obvious. Their evolution towards global, sparsely connected, leaderless networks has enabled them to survive the continuing targeted attack on their nodes. A similar finding can be found in Wiil et al. [23].

5   Conclusion

By studying terrorist networks from the small-world perspective we shed new light on their structures and the implications this holds for their survival. Since covert organizations are aware of their need to balance secrecy and information, our analysis, using the total performance measure as an evaluation criterion, shows that
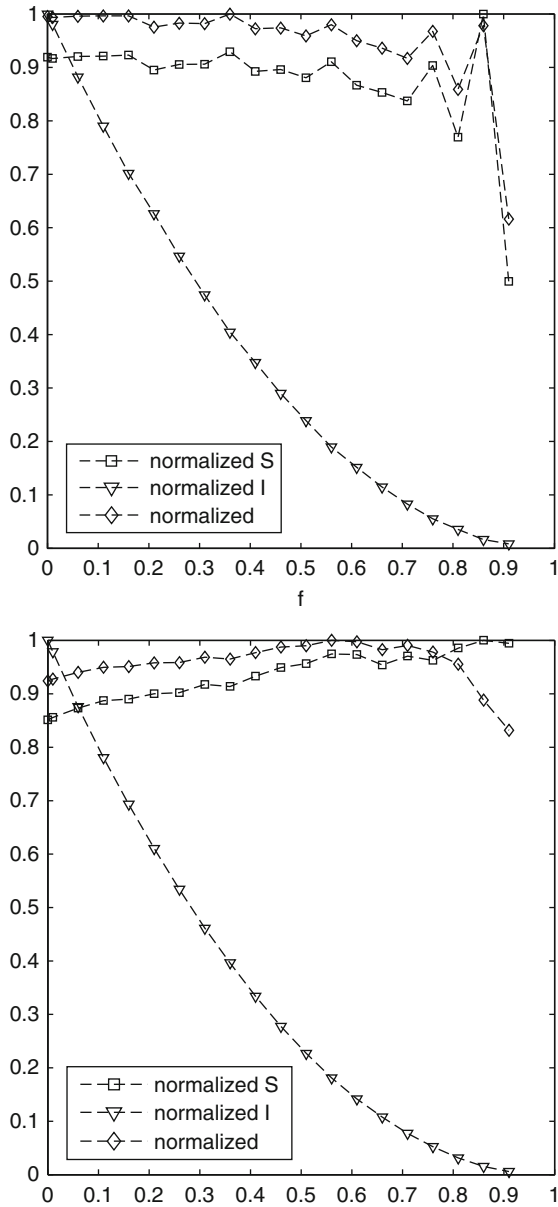
Fig. 2 Normalized values of S, I and μ as function of the fraction f of randomly (*top*) and targeted (*bottom*) removed vertices

their network topology will not satisfy a 'small-world' characterization as common in many social systems. In addition, we presented empirical evidence to support this claim. That terrorist networks avoid small-world structures can be explained by the low secrecy a highly clustered networked organizational form offers. Another reason for adopting a non small-world topology is found in the remarkable advantage the derived network topology offers against targeted removal. It is known that 'normal' social network topologies will show fast degradation in case of removal of hubs. However, we have shown that terrorist networks adopting secrecy and information balanced networks are perfectly capable to outlast targeted attacks. This may partly explain why current transnational terrorist networks appear to be so resilient: as long as disruption strategies do not completely disintegrate the network such efforts only strengthen their ability to attain a balance at remaining secret while being operationally effective – instead of disabling them to operate at all.

Appendix

A covert network is modeled by a graph $g = (N, E)$, where N represents the set of members (terrorists or terror cells) of the organization and E represents the links among these members. For instance, such links may represent the exchange of bomb making material or the communication over the internet. We set $|N| = n$ and $|E| = m$. The set of all such networks is indicated by $\mathbb{G}(n, m)$.

Information measure I

The information measure of a graph $g \in \mathbb{G}(n, m)$ is defined by the normalized reciprocal of the total distance in g, i.e.,

$$I(g) = \frac{n(n-1)}{T(g)}. \tag{A.1}$$

Here, $T(g)$ equals the total geodesic distance, i.e., $T(g) = \sum_{(i,j) \in N^2} l_{ij}(g)$ with $l_{ij}(g)$ the geodesic distance between vertex i and vertex j. It follows that $0 \leq I(g) \leq 1$. Thus, the information measure captures the ability of the terrorist organization to exchange information, i.e., to coordinate and control. The higher the value for I the better the organization can do so.

Secrecy measure S

The secrecy measure of a graph $g \in \mathbb{G}(n, m)$ is defined by

$$S(g) = \frac{2m(n-2) + n(n-1) - \sum_{i \in N} d_i^2(g)}{(2m+n)n}. \tag{A.2}$$

Here, $d_i(g)$ equals the degree of vertex i in g. It follows that $0 \leq S(g) \leq 1$. It can be seen that the secrecy measure equals the expected fraction of the organization that survives given that members of the organization are exposed according to a realistically chosen probability distribution.

Balanced trade-off performance measure μ

For $g \in \mathbb{G}(n,m)$ it holds that,

$$\mu(g) = S(g)I(g) = \frac{(n-1)(2m(n-2) + n(n-1) - \sum_{i \in N} d_i^2(g))}{(2m+n)T(g)}. \quad \text{(A.3)}$$

Following multi-objective optimization theory the terrorist organization, faced with trading off secrecy versus information, adopts those values of S and I that maximize their product. For a more thorough motivation of this measure, see Lindelauf et al. (2009a).

Small-world indicators

For $g \in \mathbb{G}(n,m)$ the characteristic path length is defined by

$$L(g) = \frac{1}{2}\frac{T(g)}{n(n-1)} = \frac{1}{2I(g)}, \quad \text{(A.4)}$$

and the clustering coefficient is defined by,

$$C(g) = \frac{1}{n}\sum_{i \in N} C_i, \quad \text{(A.5)}$$

where

$$C_i = \frac{|N_i(g)|}{|\Gamma_i(g)|(|\Gamma_i(g)| - 1)}. \quad \text{(A.6)}$$

Here, $\Gamma_i(g) = \{j \in N | l_{ij}(g) = 1\}$ is the set of neighbors of vertex i in network g, and $N_i(g) = \{\{k,l\} \in \Gamma_i(g) | l_{kl}(g) = 1\}$ is the set of neighbor pairs of vertex i that are connected in g. Small-world networks are characterized by low L and high C. When compared to random networks a small-world network satisfies $L \approx L_{\text{random}}$ and C is of a different order of magnitude than C_{random}.

Use of normalization

Since only relative comparison plays a role we normalized the indicators I, S, L, C and μ by dividing them by the maximum they attained at each relevant instance.

This avoids scaling differences in the corresponding figures but does not affect the resulting analysis.

Generating an approximate optimal covert network

A theoretically optimal covert network was approximated on $n = 100$ individuals as follows. We let $p \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$ and for each fixed p we generated 100.000 random graphs with each possible edge present independently and identically distributed with probability p. Among these 500.000 networks the one that attained the highest value for μ was selected.

References

1. Albert, R., Jeong, H., Barabási, A.-L.: Error and attack tolerance of complex networks. Nature **406**, 378–482 (2000)
2. Arquilla, J., Ronfeldt, D.: Networks and Netwars: The Future of Terror, Crime and Militancy. RAND, Santa Monica, CA (2001)
3. Asal, V., Nussbaum, B., Harrington, D.W.: Terrorism as Transnational Advocacy: An Organizational and Tactical Examination. Stud. Confl. Terrorism **30**, 15–39 (2007)
4. Bergen, P.: The Osama Bin Laden I Know: An Oral History of al Qaeda's Leader. Free Press, New York (2006)
5. Borgatti, S.P.: The Key Player Problem. Proceedings from National Academy of Sciences Workshop on Terrorism, Washington, D.C. (2002)
6. Carley, K.M.: Destabilization of covert networks. Comput. Math. Organ. Theory **12**(1), 51–66 (2006)
7. Enders, W., Su, X.: Rational terrorists and optimal network structure. J. Confl. Resolut. **51**(1), 33–57 (2007)
8. Farley, J.D.: Breaking Al Qaeda cells: A mathematical analysis of counterterrorism operations (A guide for risk management and decision making). Stud. Confl. Terrorism **26**, 399–411 (2003)
9. Jasny, B.R., Ray, B.: Life and the art of networks. Science **301**, 1863 (2003)
10. Koschade, S.: A social network analysis of Jemaah Islamiyah: The applications to counterterrorism and intelligence. Stud. Confl. Terrorism **29**, 559–575 (2006)
11. Lindelauf, R., Borm, P., Hamers, H.: The influence of secrecy on the communication structure of covert networks. Soc. Netw. **31**, 126–137 (2009a)
12. Lindelauf, R.H.A., Borm, P.E.M., Hamers, H.J.M.: In: Memon, N., Farley, J.D., Hicks, D.L., Rosenorn, T. (eds.) Mathematical Methods in Counter-terrorism. On Heterogeneous Covert Networks. Springer, New York (2009b)
13. Magouirk, J., et al.: Connecting terrorist networks. Stud. Confl. Terrorism **31**, 1–16 (2008)
14. McCormick, G.H., Owen, G.: Security and coordination in a clandestine organization. Math. Comput. Model. **31**, 175–192 (2000)
15. Morselli, K., Petit, K., Giguere, C.: The efficiency/security trade-off in criminal networks. Soc. Netw. **29**, 143–153 (2007)
16. Natarajan, M.: Understanding the structure of a large Heroin distribution network: A quantitative analysis of qualitative data. J. Quant. Criminol. **22**, 171–192 (2006)

17. Newman, M., Barabasi, A.L., Watts, D.J.: The Structure and Dynamics of Networks. Princeton University Press, Princeton, (2006)
18. Sageman, M.: Leaderless Jihad: Terror Networks in the Twenty-first Century. University of Pennsylvania Press, Philadelphia (2008)
19. Sparrow, M.: The application of network analysis to criminal intelligence: An assessment of the prospects. Soc. Networks **13**, 251–274 (1991)
20. Strogatz, S.H.: Exploring complex networks. Nature **410**, 268–276 (2001)
21. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature **393**, 440–442 (1998)
22. Watts, D.J.: The 'New' Science of Networks. Annu. Rev. Sociol. **30**, 243–270 (2004)
23. Wiil, U.K., Gniadek, J., Memon, N.: Measuring link importance in terrorist networks. Proceedings of the International Conference on Advances in Social Network Analysis and Mining (ASONAM 2010), pp. 356–359, IEEE. (2010)
24. Zacharias, G.L., et al.: Behavioural Modelling and Simulation: From Individuals to Societies. National Academy of Sciences, Washington, D.C. (2008)

Co-offending Network Mining

Patricia L. Brantingham, Martin Ester, Richard Frank, Uwe Glässer, and Mohammad A. Tayebi

Abstract We propose here a computational framework for *co-offending network mining* defined in terms of a process that combines formal data modeling with data mining of large crime and terrorism data sets as gathered and maintained by law enforcement and intelligence agencies. Our crime data analysis aims at exploring relevant properties of criminal networks in arrest-data and is based on 5 years of real-world crime data that was made available for research purposes. This data was retrieved from a large database system with several million data records keeping information for the regions of the Province of British Columbia. Beyond application of innovative data mining techniques for the analysis of the crime data set, we also provide a comprehensive data model applicable to any such data set and link the data model to the analysis techniques. We contend that central aspects considered in the work presented here carry over to a wide range of large data sets studied in intelligence and security informatics to better serve law enforcement and intelligence agencies.

1 Introduction

Mathematical methods and computational tools increasingly gain momentum in advanced studies of social phenomena not only in social sciences but also in emerging interdisciplinary research fields like Computational Criminology [6, 8]. Innovative research in criminology and counterterrorism indeed shows promising results [1, 24, 32] that underscore the enormous potential for serving practical needs

P.L. Brantingham · R. Frank
School of Criminology, Simon Fraser University, Vancouver, BC, Canada
e-mail: pbrantin@sfu.ca; rfrank@sfu.ca

M. Ester · U. Glässer (✉) · M.A. Tayebi
School of Computing Science, Simon Fraser University, Vancouver, BC, Canada
e-mail: ester@cs.sfu.ca; glaesser@cs.sfu.ca; tayebi@cs.sfu.ca
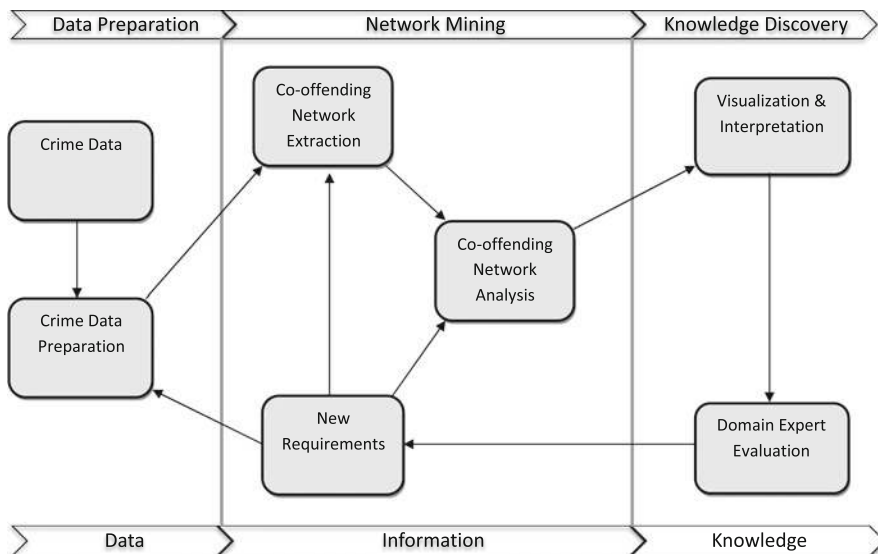
Fig. 1 Computational framework for the process of co-offending network mining

in crime analysis and prevention, namely as instruments in crime investigations, as an experimental platform for supporting evidence-based policy making and in experimental studies to analyze and validate theories of crime [21]. The work presented here has been inspired by practical experience with using mathematical modeling combined with computational analysis techniques in the study of crime events spanning a wide range of criminal activities, including opportunistic and violent serial crimes [7, 8, 31].

We propose here a comprehensive computational framework for *co-offending network mining* [2, 22, 38] defined in terms of a process that combines formal data modeling with data mining of large crime and terrorism data sets as gathered and maintained by law enforcement and intelligence agencies. Aiming at identifying common and potentially useful patterns in those data sets, the framework, as illustrated in Fig. 1, comprises three major phases: crime data preparation, co-offending network mining and knowledge–discovery.

The first phase of the framework, Data Preparation, is required to clean and transform the collected data into a format that is suitable for the co-offending network mining algorithms. Specifically, this phase includes data understanding and specification, protecting offenders' data privacy, data selection, detecting and handling missing values and semantic errors (where possible), and applying data preprocessing techniques that overall improve the efficiency of the mining algorithms. Not considered is unstructured text as part of the crime data records.

The output of the preparation phase feeds into the Network Mining phase. In the first step of this phase, the Co-offending Network Extraction component extracts the network from the 'cleaned' crime data set. Subsequently, different network analysis tasks are performed by the Co-offending Network Analysis Component, including network-level analysis, group-level analysis, node-level analysis and network evolution analysis.

Finally, in the Knowledge–Discovery phase, the information obtained in the Network Mining phase is first interpreted and then visualized as a basis for its evaluation by domain experts. Network visualization provides a synthetic and simple description of the important features by representing information about interrelationships of actors, groups and their connection patterns in the network. The Visualization & Interpretation component generates a visual representation of the co-offending network that facilitates comprehension of the obtained results and enables innovative methods for analytical reasoning using visual analytics.[1] The extracted information is then evaluated by criminologists. Knowledge–discovery naturally is a continuous and highly iterative process rather than a linear waterfall process. To this end, the Requirements Component seamlessly connects all three phases to enable a data-information–knowledge continuum. That is, new problems can be defined based on evaluations and feedback by experts, resulting in new tasks that are passed to one of the components in the first or second phase. The mining process then starts again from that component.

Crime data analysis as proposed here aims at exploring relevant properties of criminal networks in arrest-data. As a result of a research memorandum of understanding between ICURS[2] and "E" Division of Royal Canadian Mounted Police (RCMP) and the Ministry of Public Safety and the Solicitor General, 5 years of real-world crime data was made available for research purposes. This data was retrieved from the RCMP's Police Information Retrieval System (PIRS), a large database system keeping information for the regions of the Province of British Columbia which are policed by the RCMP. PIRS contains information about all reported crime events ($\approx$4.4 million) and all persons associated with a crime ($\approx$9 million), from complainant to charged. In addition, PIRS also contains information about vehicles used in crimes ($\approx$1.4 million), and businesses which were involved in crimes ($\approx$1.1 million). Of this dataset, only those offenders that were charged, chargeable, or had a charge recommended, were extracted and used for the following analysis. Being in one of these categories implies that the police were serious enough about the persons involvement in a crime as to warrant calling them 'offenders'. In total, there are over 50 groups of crime types. For the purposes of this study, however, only the four most important groups were considered:

---

[1]Visual analytics is an emerging field using computers to analyze and visually convey massive amounts of data in a form that human experts can more readily understand.

[2]The Institute for Canadian Urban Research Studies (ICURS) is a university research centre at Simon Fraser University.

- *Serious crimes:* crimes against a person, such as homicide and attempted homicide, assault, abduction
- *Property crimes:* crimes against property, such as burglary (break and enter onto a premises or real property, and theft
- *Moral crimes:* such as prostitution, arson, child pornography, gaming, breach
- *Drug crimes:* such as trafficking, possession, import/export

Beyond the application of innovative data mining techniques for the analysis of the crime data set, we also provide a comprehensive data model applicable to any such data set and link the data model to the analysis techniques. We contend that central aspects considered in the work discussed here carry over to a wide range of large data sets studied in intelligence and security informatics to better serve law enforcement and intelligence agencies.

The remainder of this manuscript is organized as follows. Section 2 provides some general background and discusses related work. Section 3 first presents the crime data model and then the co-offending network model, and also explains the extraction of the co-offending network from the crime data set. Section 4 focuses on the analysis of the resulting network. The visualization and interpretation of the results obtained in the network mining phase is illustrated and discussed in Sect. 5 in some detail. Section 8 concludes this manuscript.

2 Background and Related Work

Social networks represent relationships among social entities. Normally, such relationships can be represented as a network. Examples include interactions between members of a group (like family, friends or neighbors) or economic relationships between businesses. Social networks are important in many respects. Social influence may motivate someone to buy a product, to commit a crime, and any other decision can be interpreted and modeled under a social network structure. Spread of diseases such as AIDS infection and the diffusion of information and word of mouth also strongly depend on the topology of social contacts. In the following, we first provide some background on social network analysis (SNA) and mining and then discuss related work on mining co-offending networks.

2.1 Social Network Analysis

SNA focuses on structural aspects of networks to detect and interpret the patterns of social entities [35]. SNA essentially takes a network with nodes and edges and finds distinguished properties of the network through formal analysis. Data mining is the process of finding patterns and knowledge hidden in large databases [17]. Data mining methods are increasingly being applied to social networks, and there is substantial overlap and synergy with SNA.

New techniques for the analysis and mining of social networks are developed for a broad range of domains, including health [34] and criminology [38]. These methods can be categorized depending on the level of granularity at which the network is analyzed[5]: (1) methods that determine properties of the social network as a whole, (2) methods that discover important subnetworks, (3) methods that analyze individual network nodes, and (4) methods that characterize network evolution. In the following, we list the tasks that are most relevant for co-offending networks:

- *Centrality analysis* [35] aims at determining more important actors of a social network so as to understand their prestige, importance or influence in a network.
- *Community detection* [10] methods identify groups of actors that are more densely connected among each other than with the rest of the network.
- *Information diffusion* [26] studies the flow of information through networks and proposes abstract models of that diffusion such as the Independent Cascade model.
- *Link prediction* [20] aims at predicting for a given social network how its structure evolves over time, that is, what new links will likely form.
- *Generative models* [11] are probabilistic models which simulate the topology, temporal dynamics and patterns of large real-world networks.

SNA also greatly benefits from visual analysis techniques. Visualizing structural information in social networks enables SNA experts to intuitively make conclusions about social networks that might remain hidden even after getting SNA results. Different methods of visualizing the information in a social network providing examples of the ways in which spatial position, color, size, and shape can be used to represent information are mentioned in [35].

2.2 Mining Co-offending Networks

A co-offending network is a network of offenders who have committed crimes together [29]. With increasing attention to SNA, law enforcement and intelligence agencies have come to realize the importance of detailed knowledge about co-offending networks. Groups and organizations that engage in conspiracies, terroristic activities and crimes like drug trafficking typically do this in a concealed fashion, trying to hide their illegal activities. In analyzing such activities, investigations do not only focus on individual suspects but also examine criminal groups and illegal organization and their behavior. Thus, it is critical to identify and study criminal networks using information resources such as police arrest data and court data so as to apply SNA algorithms on these networks. SNA can provide very useful information about individuals as well. Investigators can determine those who play a more important role and make them their subjects of a closer inspection. In general, knowledge about co-offending network structures provides a basis for law enforcement agencies to make strategic or tactical decisions.

There have been some empirical studies that use SNA methods to analyze co-offending or terrorist networks. Reiss [29] concludes that the majority of co-offending groups are unstable, and their relationships are short-lived. But he also states that high frequency offenders are 'active recruiters to delinquent groups and can be important targets for law enforcement'. Reiss et al. [30] also found that co-offenders have many different partners, and are unlikely to commit crimes with the same individuals over time. McGloin et al. [23] showed that there is some stability in co-offending relationships over time for frequent offenders, but in general, delinquents do not tend to reuse co-offenders. However, the findings of these works may not be representative, since they were obtained on very small datasets: 205 individuals in [30], and 5,600 individuals in [23].

COPLINK [18] was one of the first large scale research projects in crime data mining that performed some excellent works on criminal network analysis. Xu et al. [37] employed the idea of a 'concept space' in order to establish links between individuals by comparing the activities of multiple offenders. The more two individuals were involved in the same criminal events, the more they are assumed to know each other. This method allows for the translation of event and narrative data into an undirected but weighted co-offending network. The goal was to identify central members and communities within the network, as well as interactions between communities. Their main contribution is the application of cluster analysis in order to detect subgroups within the network, and their ability to detect overall network structures which then can be used by the criminal investigators to further their investigations.

Xu et al. [38] presented CrimeNet Explorer, a framework for criminal network knowledge–discovery incorporating hierarchical clustering, SNA methods, and multidimensional scaling. The authors further expanded the research in [37] and designed a full-fledged system capable of incorporating outside data, such as phone records and report narratives, in order to establish stronger ties between individual offenders. Their results were compared to the domain knowledge offered by the Tucson Police Department, whose jurisdiction the data came from. Finally, Xu points out that the use of crime network analysis is highly impacted by laws, regulations and privacy issues over data collection, confidentiality and reporting.

Smith [33] presented a slight twist on crime network analysis, for the purposes of criminal intelligence analysis, where the network is enhanced by extra information. For example, vertices are not limited to offenders, but could be police officers, reports, or anything that can be represented as an entity. Links are associated with labels which denote the type of the relationship between the two entities, such as 'mentions' or 'reported by'. In this sense, their analysis is more representative of a database schema than a social network, which does have advantages as it is more expressive.

In [25], Kaza et al. explored the use of criminal activity networks to analyze information from law enforcement and other sources to provide value for transportation and border security. The criminal activity network is defined as a network of interconnected criminals, vehicles, and locations based on law enforcement records.

The authors concluded that including vehicular data in criminal activity network yields clear advantages, since vehicles provides new investigative leads that can be used to detect individuals and vehicles that might threaten the security of the border and transportation infrastructure.

3   Crime Data Model

This section introduces a unified formal model of crime data as a semantic framework for defining in an unambiguous way the meaning of co-offending networks and their constituent entities at an abstract level. Specifically, the formal model aims at bridging the conceptual gap between data level, mining level and interpretation level, and also facilitates separating the description of the data from the details of data mining and analysis. By reducing the unified model to more specific views, the co-offending network model is then obtained as one such view.

3.1   Unified Crime Data Model

Crime data can be modeled as a finite *attributed tripartite hypergraph* $\mathcal{H}$ with $\mathcal{V}, \mathcal{E}$ representing the vertices and the edges of $\mathcal{H}$. The vertex set $\mathcal{V}$ is partitioned into three pairwise disjoint sets, $A = \{a_1, a_2, \ldots, a_q\}$, $I = \{i_1, i_2, \ldots, i_r\}$ and $R = \{r_1, r_2, \ldots, r_s\}$, reflecting *actors* such as offenders, victims, witnesses, suspects and bystanders; *events* referring to crime incidents of a certain type; and *resources* used in a crime, like mobile phones, vehicles or weapons.

The set $\mathcal{E}$ consists of (hyper)edges such that each $e \in \mathcal{E}$ is a subset of vertices $\{v_1, v_2, \ldots v_p\} \subseteq \mathcal{V}$ with $|e \cap I| = 1$ and $|e \cap A| \geq 1$ and $|e \cap R| \geq 0$.[3]

Further, for any $e, e' \in \mathcal{E}$ with $e \cap I = e' \cap I$ it follows that $e = e'$. In other words, every edge e of $\mathcal{H}$ identifies a subset of actors $\{a_{i_1}, a_{i_2}, \ldots, a_{i_j}\} \subseteq A$ and a subset of resources $\{r_{i_1}, r_{i_2}, \ldots, r_{i_l}\} \subseteq R$ with any crime event $i_k \in I$, that is $e = \{i_k, a_{i_1}, a_{i_2}, \ldots, a_{i_j}, r_{i_1}, r_{i_2}, \ldots, r_{i_l}\}$. See Fig. 2 for an example.

Finally, attributes are defined on the vertices of $\mathcal{V}$ such that for each $v \in \mathcal{V}$ there is a finite list of pairs (α_v, β_v) where α_v is an attribute name and β_v is the value of α_v. Attributes of actors, for instance, include their name and address information, while attributes of events include the crime type, the location where, and the time when, this incident occurred.

---

[3]Every crime data record in the crime data set refers to a different crime incident.
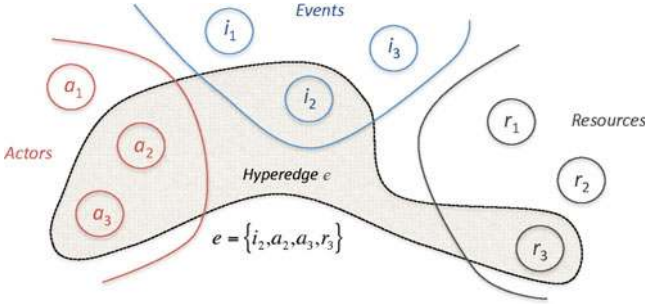
Fig. 2 Hyperedge in the crime data model

3.2 Co-offending Network Model

Co-offending networks are composed of individuals who commit crimes together. For analyzing and reasoning about co-offending networks, as well as other more specific aspects of crime data that can be described in terms of entities and their relations, the unified crime data model defined by the hypergraph $\mathcal{H}$ is decomposed into a number of simpler graph structures as follows.

Consider an attributed tripartite hypergraph $H = (V, E)$ where V is identical to $\mathcal{V}$ and $E = \{\{a, i, r\} | \exists e \in \mathcal{E}$ such that $\{a, i, r\} \subseteq e\}$ for $a \in A, i \in I, r \in R$. Note that H has the same attributes as $\mathcal{H}$. The hypergraph H is now further decomposed in a straightforward way into three *bipartite graphs* that model the relations between actors and events (graph *AI*), actors and resources (graph *AR*) and events and resources (graph *IR*).

3.2.1 Criminal Activity Graph

Starting from AI, a new graph $AI_O = (V_O, E_O)$, called *criminal activity graph*, is constructed as follows. V_O consists of vertices representing either *offenders* or *events*. That is, $V_O = A_O \cup I$, with $A_O \subset A$, where A_O identifies the offenders in the set of actors. Every edge in E_O either links an offender to an event or it links two offenders with one another. The latter type of edge means that two offenders have jointly committed one or more crimes in the past. To indicate multiple co-offenses, an attribute *strength* is associated with every edge $(a_i, a_j) \in E_O$, for $a_i, a_j \in A_O$, where $strength((a_i, a_j)) \geq 1$.

Figure 3 illustrates a criminal activity graph with three offenders a_1, a_2, a_3 for which it is known that a_1, a_2 and a_1, a_3 have jointly committed multiple crimes (some of the related incidents are not explicitly shown here). The resource in this example is not an integral part of the graph but derived information.
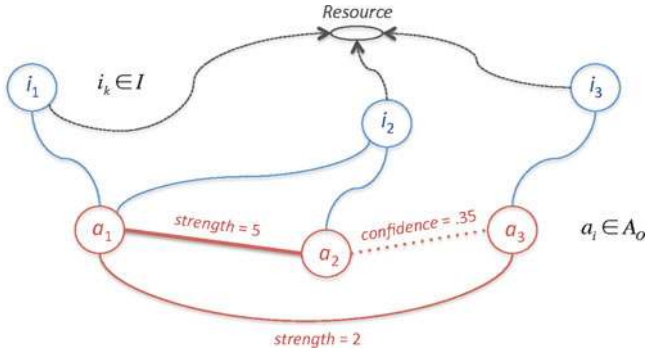
Fig. 3 Criminal activity graph with hidden links

3.2.2 Co-offending Network

For generating the co-offending network, we start from the criminal activity graph AI_O. Assuming k offenders and m events $(k, m > 1)$, we define a $k \times m$ matrix M such that $m_{uv} = 1$, if offender o_u is involved in event i_v, and "0" otherwise. Now, we define the co-offending network by means of the $k \times k$ matrix $N = MM^T$ and therefore have

$$n_{u,v} = \sum_{x=1}^{k} n_{ux} n_{xv}.$$

This matrix links offenders involved in the same crime events. For any two given offenders, the strength of a link is the number of co-offenses. The diagonal of this matrix shows for each offender the number of related crime events.

3.2.3 Hidden Links

Co-offenders often try to conceal their connections as much as they can. Also, the available data is police arrest data that contains only partial information of offender collaborations and their social interactions. Based on these two factors, one can expect that besides the links based on explicit facts in the crime data additional links can be derived by analyzing and mining the crime data using link prediction methods. Such links are called *hidden links*, which are probabilistic in nature as they are based on information that is considered uncertain. Thus, hidden links have an attribute *confidence*, the value of which is a positive real number from the interval [0..1]. A confidence value of "0" means that no link exists.

   The example illustrated in Fig. 3 assumes that in all three of the crime incidents i_1, i_2, i_3 a common resource, say a particular vehicle, was used by one of the offenders a_1, a_2, a_3. From this information, one may derive a hidden link (a_2, a_3) with some probability as stated by the value of the attribute *confidence*.
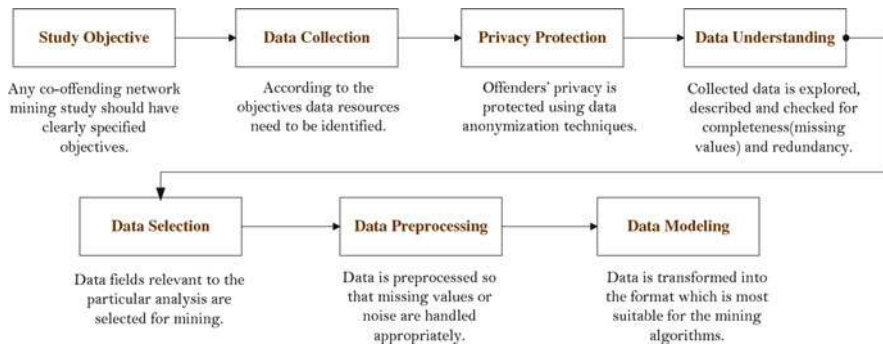
Fig. 4 Steps involved in the data preparation

Methods for link detection and link prediction are similar and essentially mean different interpretations. Link prediction determines links that have a high probability for creation in the next step of network evolution. Link detection assumes that detected links were hidden or removed from the data. A detected link is called missing link in link detection methods and forming link in link prediction approaches. In this work, the analysis focuses on explicit links.

3.3 Crime Data Preparation

This section addresses the data preparation phase covering all activities prior to analyzing the data for network mining. Figure 4 illustrates steps that are carried out in the preparation phase. A brief description of these steps follows below.

Study Objectives: Any co-offending network mining study should have clearly specified objectives. The objectives affect data selection and also the subsequent steps. For instance, the structure of the data and also the analysis process for static and for dynamic co-offending network mining are different.

Data Collection: Data for co-offending network mining can originate from different sources; for instance, police arrest data or court data. Additional resources may provide supplementary information like email contacts or phone calls that help filling gaps or resolving inconsistencies.

Privacy Protection: Preserving the privacy of individuals is a central issue when working with sensitive data sets. Data anonymization techniques normally have to be applied prior to sharing this kind of data with third parties.

Data Understanding: In this step, the data is explored, described and checked for completeness (missing values) and redundancy.

Data Selection: In this step, the target data that is relevant to the analysis is selected based on the study objectives, excluding redundant or useless data.

Data Preprocessing: Data is preprocessed so that missing values or noise are handled appropriately. For example, for the crime data studied here, about 40% of the offender home address fields were empty.

Data Modeling: Data is transformed into the format which is most suitable for the mining algorithms. See Sects. 3.1 and 3.2 for a detailed discussion of this important step.

4 Co-offending Network Analysis

In Sect. 2, we grouped the tasks of SNA into four main categories: (1) network-level analysis, (2) group-level analysis and (3) node-level analysis (4) network evolution analysis. In this section we represent the important analysis oriented concepts under these categories as well as the results of the analysis on the co-offending networks. We applied the analysis tasks on the co-offending networks extracted from different crime types and also on several snapshots of these networks[4] $G_u(t)$ denotes the co-offending network of a specific crime type u (a, s, p, d and m represents the all, serious, property, drugs and moral crimes types) from year 2001 to year t.

4.1 Network-Level Analysis

At the network-level the goal is to find properties of the network as a whole. Global properties such as degree distribution, clustering coefficient and average distance reflect network characteristics and can also be employed evaluate the similarity of different networks using these properties.

The degree of a node is the number of edges the node has. The degree distribution, $P(k)$, gives the probability that a randomly selected node has k links. Studies have shown that the most real world networks from divers fields ranging from sociology to biology to communication follow a power-law distribution: $P(k) = K^{-\lambda}$ [3], where λ is called the *exponent* of the distribution. Power-law distribution implies that nodes with few links are numerous, while very few nodes have very large number of links. Networks with this property are called scale free networks. Scale free property is one of the most documented networks property. There are some other network models like the Erdos–Renyi [14] and the Watts and Strogatz [36] models that are known as exponential networks and their degree distribution conforms to a Poisson distribution. In this type of networks there is a peak at the average degree of the network, therefor most of the nodes have the same

---

[4]In implementing the analysis tasks, we used SNAP library which is publicly available at http://snap.stanford.edu/.
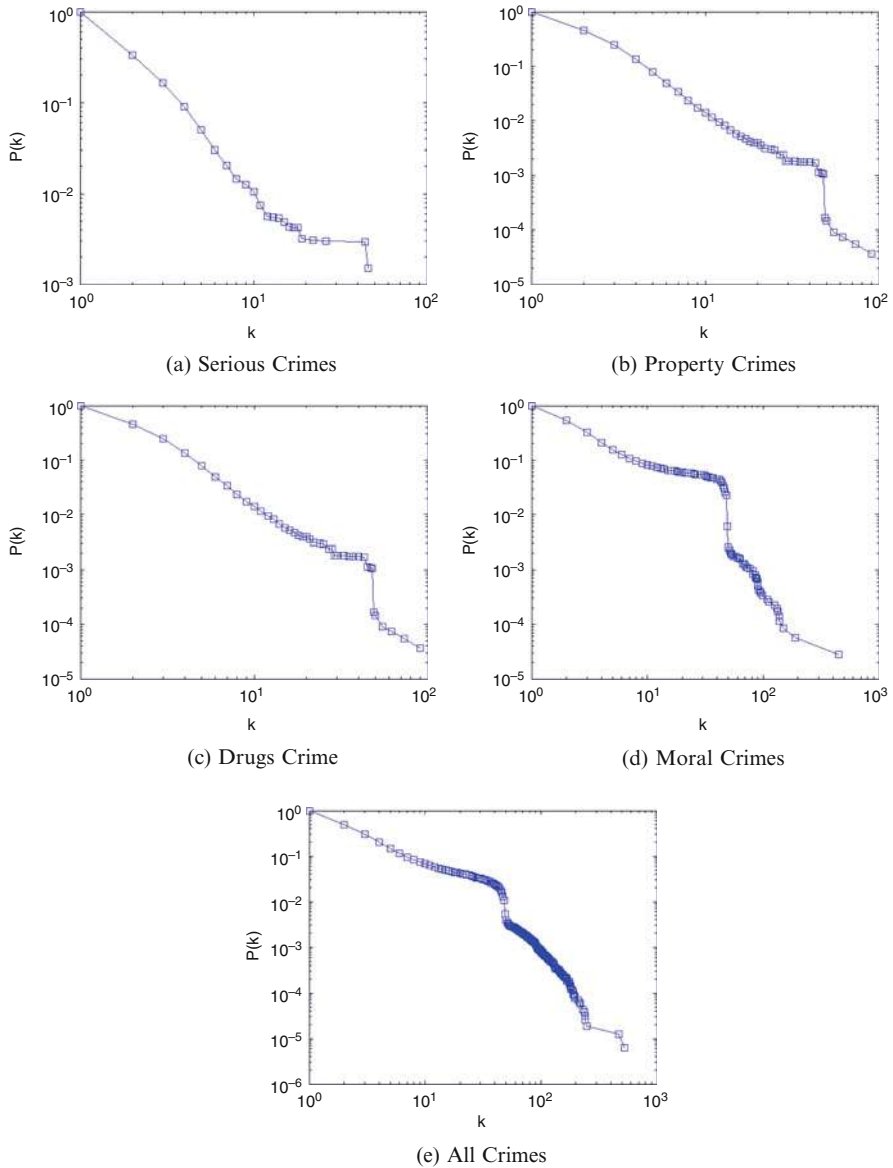
Fig. 5 Degree distribution of co-offending network for different crime types. (**a**) Serious crimes (**b**) Property crimes (**c**) Drugs crime (**d**) Moral crimes (**e**) All crimes

degree around average degree of the network and very few nodes have very small or very large node degrees.

The degree distributions of the co-offending network is also scale-free. Figure 5 demonstrates the cumulative degree distribution for different types of co-offending
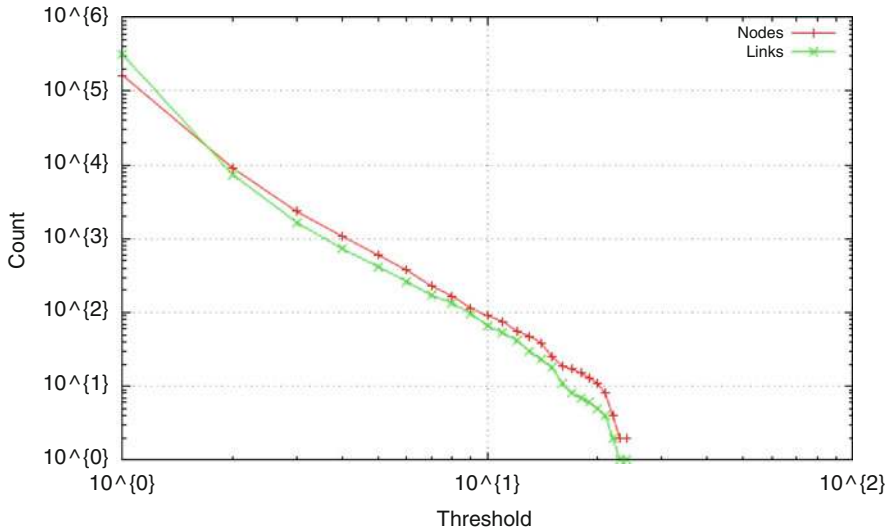
Fig. 6 Co-offending strength distribution

network. In all of these networks we can observe behavior consistent with a power-law network. It means the majority of the offenders have small degree, and a few offenders have significantly higher degree. To test how well the degree distributions are modeled by a power-law, we computed the best power-law fit using the maximum likelihood method [12]. The power-law parameter for all crimes, serious, property, drugs and moral co-offending network respectively are 2.29, 1.57, 1.42, 1.53 and 2.28.

Each link in the co-offending network is associated with a co-offending strength. The co-offending strength $S_{i,j}$ between two offenders i and j is equal to the number of crimes that two offenders have been involved. We then define a network $\bar{G}(V, E, \alpha)$ where E includes the links between the pairs of offenders i and j in V whose co-offending strengths $S_{i,j}$ exceed a specified threshold α. Then we will have a family of networks $\{\bar{G}(\alpha_1), \bar{G}(\alpha_2), \ldots, \bar{G}(\alpha_m)\}$. Figure 6 plots the distribution of number of nodes and links for the threshold networks. Again, a power law distribution of co-offending strength suggests that the vast majority of dyads only offended once or twice, but there are more than hundred dyads that offended with each other more than ten times over 5 years. When two offenders collaborate on multiple incidents, the likelihood of a strong relationship between the two offenders is high, therefore such offenders and their behaviors should be inspected more carefully by the crime investigators.

Law enforcement officers and intelligence analysts frequently need to determine if there is a possible association among a specific group of offenders in a co-offending network. So we need methods to determine if two offenders are connected and what is the shortest connecting path. Dijkstra's shortest path

algorithm [13] for weighted networks and Breadth First Search (BFS) algorithms for unweighted network, are applied to identify the shortest paths in the networks.

Average distance of the network $G(V, E)$ is defined as the average path distance of connected nodes pairs. Average path distance can show the speed of spreading a message in a co-offending network. Let l_{ij} denotes the number of links in the shortest path connecting nodes i and j in the case there is such a path and as infinity if there is not any path connecting nodes i and j. Therefore, the average distance in a network is defined as:

$$\mathrm{Avg}D(G) = \frac{\sum_{\{i,j\}:l_{ij} \neq \infty} l_{ij}}{|\{\{i, j\} : l_{ij} \neq \infty\}|}.$$

And the diameter is defined as $\mathrm{Diam}(G) = \mathrm{Max}(l_{ij} : \{i, j\} : l_{ij} \neq \infty)$. Diameter is an important property of the topology of a social network. The diameter represents the longest path within the network and describes the compactness and connectivity of the network. A network with a small diameter is very well-connected but a network with a large diameter will be very sparsely-connected. For removing the effect of outliers another measure called effective diameter is used. Effective diameter is the minimum number of hops for reaching at least 90% of all connected pairs of nodes in the network [28]. Table 1 shows the average path length, diameter and effective diameter for the five studied networks. The average path lengths and diameters for some of them are remarkably short. For the network $G_a(2006)$ average distance, diameter and effective diameter are 12.2, 36 and 16.87, respectively.

In many social networks, the friend of an actor is likely to be also her friend. In other words, actors tempt to create complete triangles of relationships. This property is called network clustering or transitivity. The clustering coefficient of a node in the co-offending network tells us how much a nodes collaborators are willing to collaborate with each other, and it represents the probability that two of its collaborators are involved in a crime together. Local Clustering Coefficient calculates the probability of neighbors of a node to be neighbors to each other is given by:

$$C_v = \frac{N_v}{k_v(k_v - 1)},$$

Table 1 Statistical properties of the studied networks

Metric	All crimes	Serious	Property	Drugs	Moral
# Co-offenders	157274	31132	44321	54286	35266
Avg. degree	4	1.85	1.95	2.15	4.8
Exponent (λ)	2.29	1.57	1.42	1.53	2.28
Avg. distance	12.2	1.69	8.45	22.17	3.41
Diameter	36	13	24	56	19
Effective diameter	16.87	4.1	14.36	36.14	5.68
Clustering coefficient	0.39	0.28	0.33	0.39	0.49
Largest Comp. percentage	25	10	32	23	21

where k_v is the number of neighbors of v, $k_v(k_v-1)$ is the maximum number of links that can exist between neighbors of v and N_v is the number of links that actually exist among neighbors of v. The average clustering coefficient per degree is shown in Fig. 10b. The clustering coefficient of the network is computed by averaging C_v for all nodes in the network [36]:

$$C = \frac{1}{|V|} \sum_{v \in V} C_v$$

The clustering coefficient of co-offending network of $G_s(2006)$, $G_p(2006)$, $G_d(2006)$, $G_m(2006)$ and $G_a(2006)$ are, respectively, 0.28, 0.33, 0.39, 0.49 and 0.39. Because the clustering coefficient in a network shows to what extent friends of a person are also friends with each other, we can conclude that in co-offending network of moral crimes with higher clustering coefficient, offenders have closer collaboration comparing to other types of co-offending networks.

4.2 Group-Level Analysis

Entities of a network are interested in forming groups and interact more closely to each other inside the group. The specific characteristic of a group is that there is a higher degree of connectivity inside the group than entities outside the group. Nowadays, in the field of criminology the idea that crimes are not always committed by offenders individually but that many crimes are planned and committed by several offenders working together, is becoming more important. Also in the last decade there have been more and more experimental studies into criminal activities that need specific forms of collaboration and organization [9]. For detecting these type of collaborations we need to mathematically formalize concepts such as group crime, gang, organized crime and corporate crime and then design efficient algorithms for this purpose. By inspecting relations between offenders to identify criminal groups, law enforcement organizations can track the origin and core of what may become an organized crime group or a gang. In this way, a criminal group can be identified prior to its formation and police can follow such offenders behavior.

As a first step, we studied the distribution of components of the co-offending network. A component is a connected subset of a graph in which there are paths between all pairs of nodes [35]. If two offenders were involved in a crime, there is a path between them. If a third offender was co-offended with any one of the first two offenders, a path can be built connecting the first offender with the third offender and so on. If a path between two offenders can be established, the two offenders are said to belong to the same component of the network. The notion of a component has particular significance for the study of the spread of epidemics on a network. In a co-offending network, we can find important situations for epidemic phenomena.

Having co-offending networks components structure can contribute on decreasing crime epidemics such as drug use epidemic in the society. Let $|c|$ represent the size of component c. Then we define three types of components: Large components $|c| \geq 1,000$, Medium components $100 \leq |c| < 1,000$ and small sized components $2 \leq |c| \leq 100$. In the network $G_a(2006)$, 25, 1 and 74% of the whole offenders are connected to each other respectively through large, medium, and small components.

In the second step, we studied the community structure in the co-offending network. We applied the Girvan–Newman algorithm [16] for detecting communities on the network $G_a(2006)$. The key idea behind this algorithm is that the edges that connect highly clustered communities have a higher edge betweenness. So, the communities are detected by progressively removing edges with highest betweenness from the network. After every removal, the betweenness of the edges is recalculated and the process is repeated until the social network is divided into a specified number of subnetworks, the communities. Figure 7 shows the size distribution of detected communities and also components. The largest extracted community size has about 4 thousand members, which is relatively small compared to the largest component with more than 39 thousand nodes. This group size is too large to be meaningful from a criminological point of view. There is a need for novel community extraction methods that particularly address the special requirements of co-offending networks.
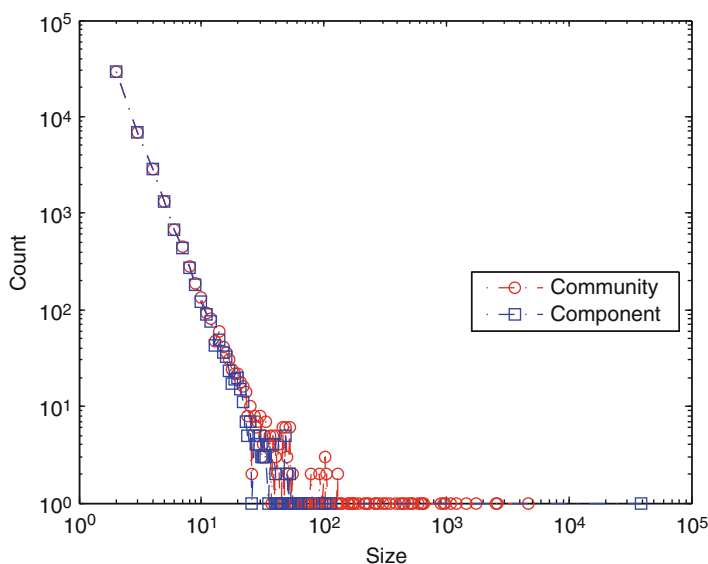
Fig. 7 Size distribution of components and communities

4.3   Node-Level Analysis

In addition to partition social networks into groups, we can categorize them based on the set of relations they have in the network. Such actors take similar *positions* within an organization, community the whole social network [35]. Establishing and breaking relationships happens very often in co-offending networks. Therefore positions, roles, and power of offenders in a co-offending network change consequently. At the node level of analysis, centrality is the most comprehensively studied concept. Centrality reveals how important, influential or powerful a node is, which may reflect the roles of actors in a network. The centrality of offenders can be determined using different measures such as:

- *Degree Centrality:* In this definition centrality of a node refers to the number of links incident upon the node.
- *Betweenness Centrality:* is based on the number of shortest path connections between any two nodes in the network that the node in question lies along.
- *Closeness Centrality:* The idea is that nodes are more central if they can reach other nodes easily which is measured by averaging the length of the shortest path from a node to all other nodes.
- *Eigenvector Centrality:* The main idea behind eigenvector centrality is that nodes receiving many links from other nodes with high centrality measures, are more central nodes.

As an example, we present the centrality analysis on the second largest component. Table 2 lists the ranks of the key nodes A to L (see Fig. 8) according to the chosen centrality measure. The numbers within the table indicate the ordering of the Top 5 offenders identified by each measure. For example, offender E was identified as the second most important offender by eigenvector centrality, but only 4th with betweenness centrality.

Although all the different centrality measures tended to identify different individuals in varying order, all but one measure agreed that offender A was most important. This strong result was somewhat surprising given that in total 11 offenders were identified to be in the Top 5 by different measures. Offender A does seem to be an important offender in the network, as this offender has the largest number of edges, is involved in 3 cliques of at least size 5, and would fragment the network into 4 pieces if removed. Without this information, a police force could

Table 2   Centrality measures on the second largest component

Centrality measure	Offenders										
	A	B	C	D	E	F	G	H	J	K	L
Degree	1			3					4	5	2
Betweenness	1	3			5	4	2				
Closeness	2	4					1	3	5		
Eigenvector	1		5	3	2				4		
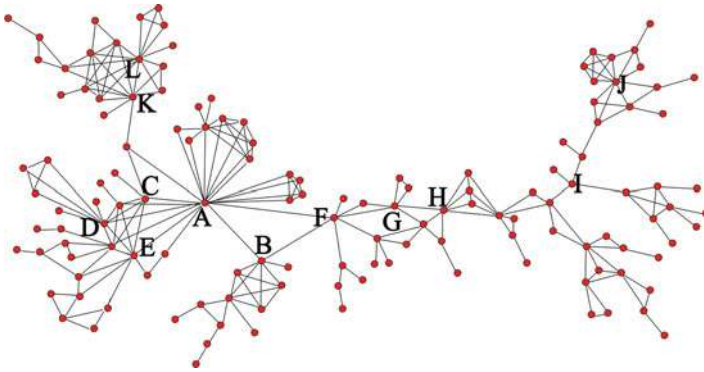
Fig. 8 Visualization of the second largest component

capture multiple offenders, not realizing that specifically targeting only offender A in the network would have a huge impact on the network and remove by far the most important offender. It is also interesting to see that the above 4 measures identified quite a few offenders along the shortest path between the two furthest nodes. This path travels through offenders K, C, A, F, G, H and J.

4.4 Network Evolution Analysis

Like other social networks, a co-offending network is not a static network and keeps changing over time: offenders may leave or join the network and their positions may change by obtaining or losing power. Links between offenders may form or disappear. Offenders groups may appear, split, merge or disappear. Network structure may change from decentralized to centralized, flat to hierarchical or vise versa. Analysis of all these evolutions is important and also complicated. But detecting the evolution patterns of a co-offending network, can represent important information to law enforcement organizations.

We study how the co-offending network evolves over time based on multiple snapshots of the network. For this purpose we generated five snapshots of the co-offending network for the years 2001–2006. Each snapshot contains the extracted co-offending network from events that happened from 2001 up to that time. For example $G_a(2004)$ is the co-offending network of all crimes from 2001 to 2004. Below, we examine the evolution of co-offending network based on these five snapshots for various network structural properties.

Figure 9a demonstrates the evolution of size and number of components over time. The most interesting observation is that, after 1 year, in the network $G_a(2002)$ there is no large component but it grows in a nearly linear trend. On the other hand, in all networks not many offenders are connected to the medium sized components. The reason is that the medium sized components are merged with
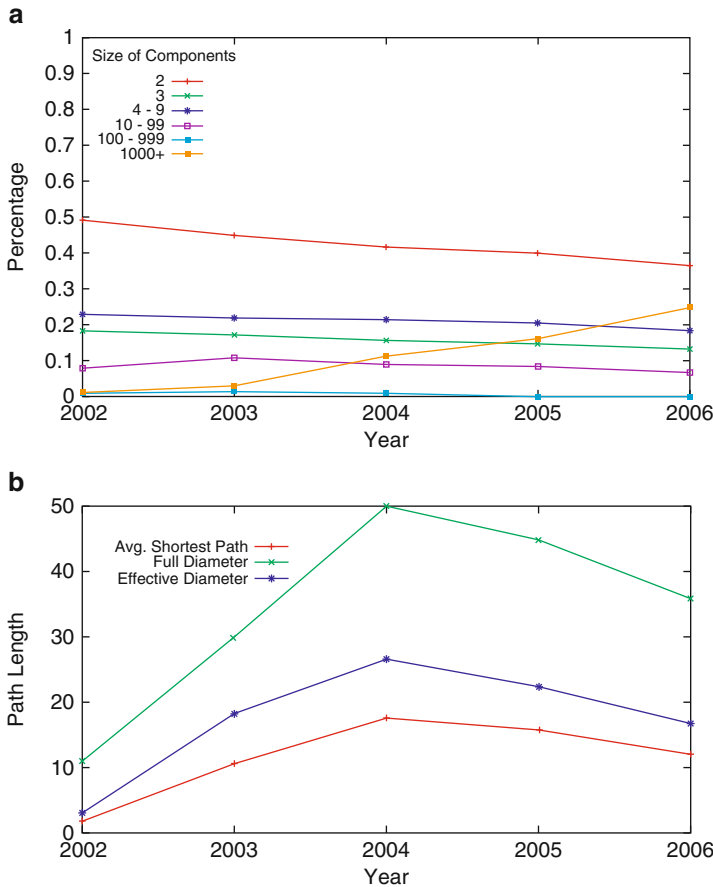
Fig. 9 (**a**) Clustering coefficient vs time; (**b**) Average distance vs time

the large components through some of their nodes and we do not have them as independent components. In other words, medium sized components blend in the large components very soon and make them richer; therefore, we do not observe their existence in the network for a long time period. There exists a similar phenomenon in other real-world social networks, a large component tends to form with the remaining being singletons and smaller components [27]. The number of nodes that belong to the small components are almost constant in all 5 years. The reason is that always some of the small components are connected to the medium or large components and simultaneously some new small components appear in the network.

In Fig. 9b, we plot the evolution of the average distance, diameter and effective diameter of the co-offending network between 2001 and 2006. This finding may be surprising because of the increasing size of the co-offending network, as network

(a) Clustering coefficient vs time.

(b) Average clustering coefficient per degree.
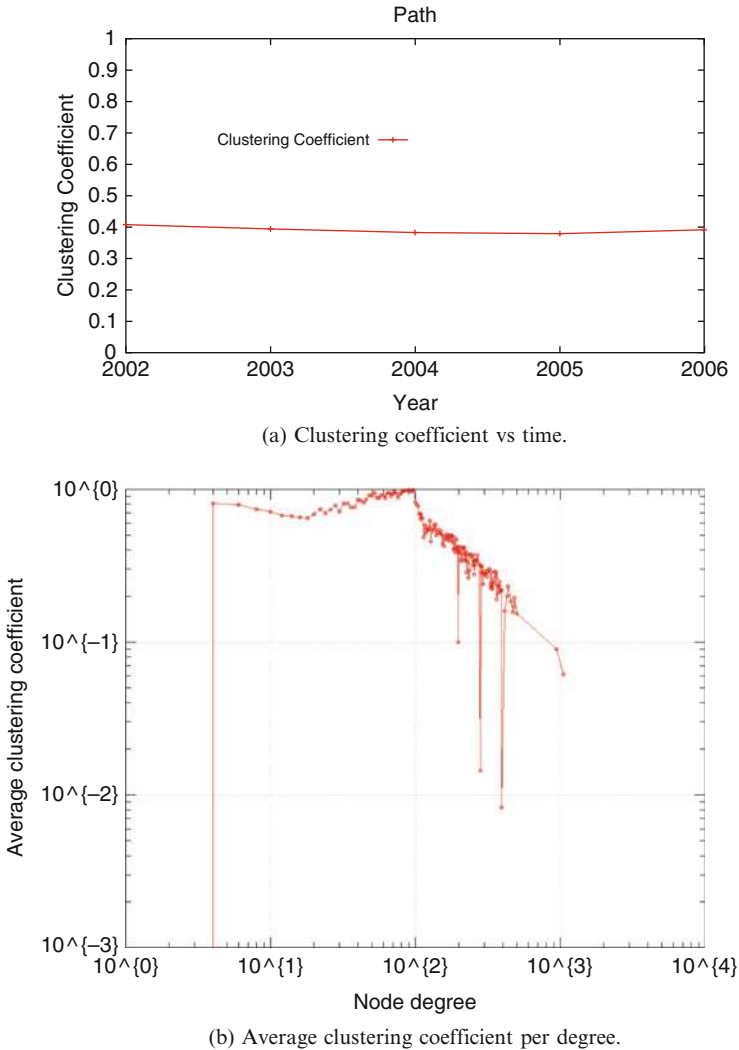
Fig. 10 (**a**) Clustering coefficient vs time; (**b**) Average clustering coefficient per degree

models generally suggest that average distance and diameter should increase with network size [4]. In our case, all these three measures are increasing in the first 3 years and then they start decreasing in last 2 years. There are studies which report similar results [19].

Figure 10a shows the evolution of the clustering coefficient. There are three observations. Firstly, clustering is stationary in all 5 years. As expected, clustering is higher than the expected clustering of a random network with same number of

vertices and edges. Finally, our results is opposite to the empirical studies of some of social networks [4], where clustering was found to decrease over time.

5 Network Visualization and Interpretation

Visualization can facilitate SNA. Visualization allows the investigators to discover patterns of interactions among the offenders, including detecting criminal subgroups, central offenders and their roles, and discovering patterns of interactions among offenders. Visualization also can provide new insights into network structures for investigators while helping them communicate with others [35]. Definitely, besides just creating images, the process of visualizing a social network can generate learning situations. But effective visualization mostly should be accompanied with a comprehensive and detailed interpretation. This is needed more in multidisciplinary projects like co-offending network mining. Well-done visualization and interpretation can fill the gap that exists between SNA experts and law enforcement officers. This is the reason of having a component called "visualization and interpretation" in the knowledge–discovery phase of the proposed framework. For applying our visualization tasks and interpreting them, we selected the second largest component of the co-offending network, see Fig. 8. Before looking at the properties of the offenders within the network, a high-level description of the entire network is in order. The selected network, Fig. 8, contained 138 vertices, 266 edges and was created from 189 distinct criminal events and all offenders associated to those events. The network does not contain any isolates; they would have been included in another 'network'. Thus, reachability is 1.

5.1 Crime Type

The network was limited to four different crime-types: serious crimes, property crimes, moral crimes and drug crimes. The below sections analyze the network by focusing on only a specific crime-type during each analysis. Note that for easy analysis, all the vertices are fixed, the links however do change depending on the crime-type being shown. Isolates in any of the networks simply imply that the offender did not commit the type of crime being analyzed.

Serious crimes: Of the original 138 vertices and 266 edges, limiting the analysis to co-arresters involved in a serious crime fragmented the original network into many very small pieces, see Fig. 11a. Instead of a single cohesive network, now there are 11 sub-networks, varying in size from 2 to 8 offenders. In total 39 vertices are active, with 36 active edges, although only 34 edges are visible since 2 of the edges are repeat co-arresters in serious crimes. For an edge to be active, both vertices involved with the edge must be charged with a serious crime during the same event.

(a) Serious Crimes.   (b) Property Crimes.

(c) Drugs Crime.   (d) Moral Crimes.
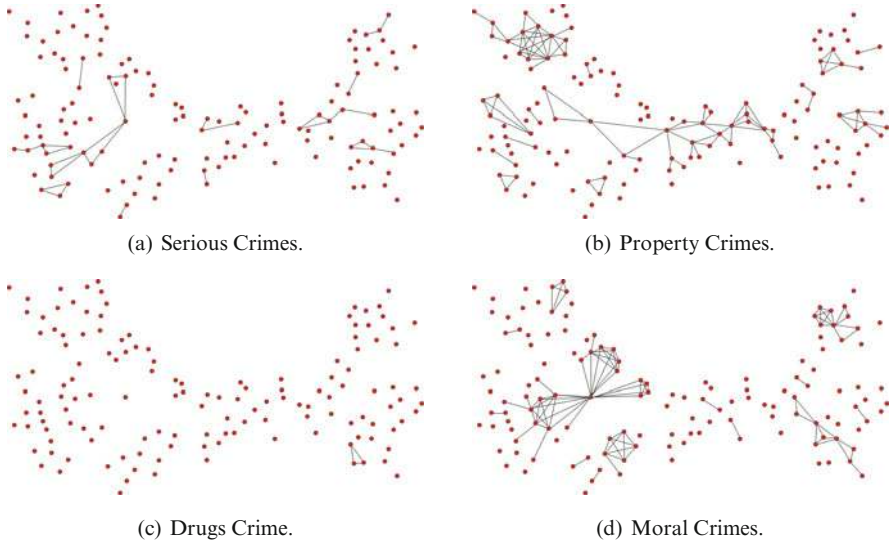
Fig. 11 Visualization of the second largest component for the crime types (**a**) Serious crimes (**b**) Property crimes (**c**) Drugs crime (**d**) Moral crimes

The average number of edges per vertex is 0.872, meaning, overall, offenders tend to commit crimes once or twice, each time with a different offender.

As expected though, the majority of the resulting sub-networks were small. This could be because people do not commit these types of crimes as frequently as other offenses, or people who do commit these types of crimes are sent to jail for longer periods of time and hence their inability to commit further offenses with others is taken away from them.

Property Crimes: Since property crimes are easier to commit than crimes against other people, intuition would tell us that this type of crime would be more common, leading to more extensive and probably more connected sub-networks than that for serious crimes. This was indeed so. See Fig. 11b for an illustration.

This restriction also created 11 smaller sub-networks, similar to that for serious crimes. However, the largest property-crime co-arrester network had 27 offenders. 74, or roughly half of the 138 original vertices remained active, and 123, or roughly half of the edges of the original 266 remained active. Of the 123 edges, only 106 were visible, implying that there were quite a few repeat co-arresters in the sub-network. There are also several sections of the graph which involve cliques of 6 (M) or 4 (N, O and P) vertices. Compared to the network produced by serious crimes this network is much more connected.

The difference in networks for serious and property crimes is interesting, in that the network for property crimes is much more extensive, implying that property offenders co-offend much more frequently and serious crimes tend to involve much

smaller networks. Information on people offending individually was not available; hence, if a person committed a crime alone, they would not be part of this analysis.

Morals: Restricting the network to moral crimes, such as prostitution, arson, child porn, gaming, breach, resulted in a sub-network that had only 104 edges and 65 vertices active (Fig. 11d). There are 2 cliques of 6 (points Q and R) and 3 cliques of 5 (points S, T and U). Offender A is obviously very highly connected, with 16 different edges, and is involved in 3 sizable cliques. Just the fact that this person alone is within 3 totally distinct cliques illustrates how well this person is connected within the network. If the police wanted to disrupt the network, this person would be a good target: prolific and pretty well embedded in the network.

It is interesting to see that there is an edge (V) between two individuals which exists in both the Moral and Property sub-networks implying that Offender A is not only prolific, but does not restrict themselves to a single type of crime.

Figure 11c illustrates the sub-network created when the crime-type is restricted to just drug crimes (such as trafficking, possession, import/export). The resulting 'network' is not large, just a single event involving 3 offenders. Note that this proportion of crimes which are drug-related is not reflective of the other networks in general that were created (there is a single network of 100 offenders that shows just drug offenses).

5.2   Spatial Mapping of Co-offenders

For detecting the patterns between co-arresters home location of offenders are visualized (Fig. 12a). As a result, only 3 very tiny clusters were seen in this visualization, one in Prince George, one on Vancouver Island and finally on in the Greater Vancouver Regional District (GVRD). The clusters were separated by huge distances (since there is nothing there but either mountains or water), but differences within the clusters were lost due to the small scale. Thus, the most important cluster was chosen, the one located in the GVRD, and analysis was restricted to that. Further, offenders with no reported location, or a location which could not be geo-coded, had to be eliminated from this analysis. The resulting set of people can be seen in Fig. 12a, which shows the co-arresting relationships for offenders in the GVRD. Multiple offenders with the same reported home location, in this image, would overlap and hence would not be visible. In total, 81 vertices were left of the network, creating a total of 162 edges.

This analysis brought up the question whether offenders tend to co-arrest with others outside of their own city when committing an offense. For each event, the home location of each co-arrester involved in the event was compared to all the other co-arresters in the event. Of all the 155 events, 38 of them had all the offenders involved living in the same city, and 14 events had all offenders living within 2 cities. This implies that only about 25% of the population tends to get arrested with someone else from the same city.
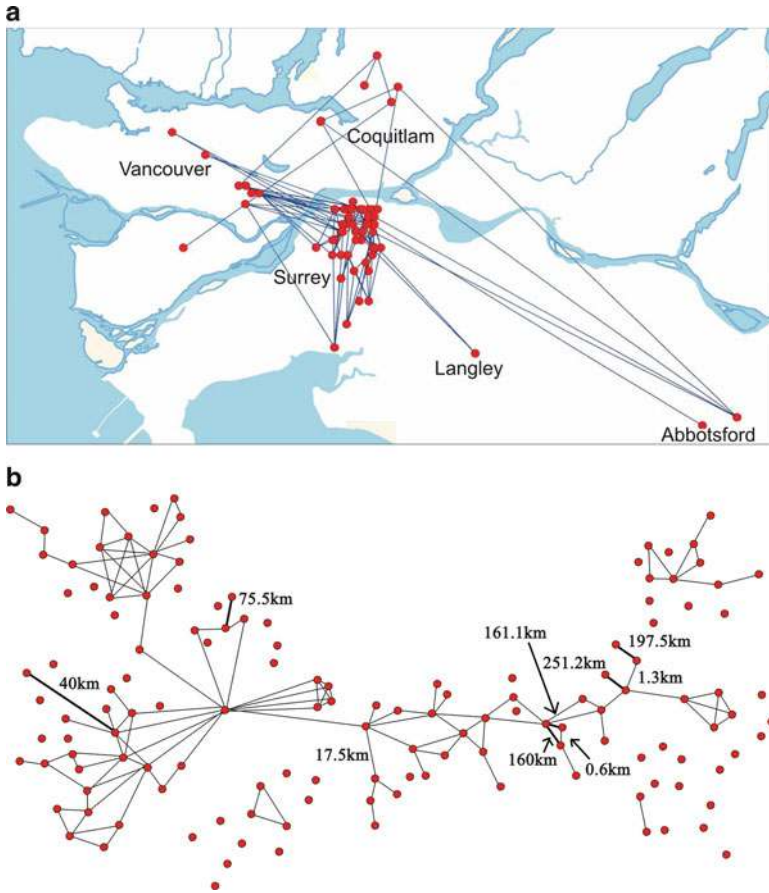
Fig. 12 (**a**) Relationships for offenders in the GVRD; (**b**) Network showing distance between home locations of offenders (*thickness* indicates magnitude of distance. *Missing links* indicate missing information)

5.3 Distances Between Home Locations

For co-arresting relationships, distances between the home location (at the time of crime) of the two offenders was calculated and used to construct the network, see Fig. 12b. Missing edges indicate that at least one of the two vertices involved in that edge did not have a properly geocodable location. The thickness of the edge indicates the distance between the two offenders.

Upon visual inspection it can be concluded that co-arresters do not live that far away from each other. A detailed analysis of the raw distance numbers, Table 3, confirms this visual result. Taking all the distance measures, the average distance an offender travels to get to their co-arrester's residence is 10.7 km. That is quite

Table 3 Distance and age difference between co-offenders

	Distance difference		Age difference	
	All data (km)	Top/bottom 10% removed (km)	All data (y)	Top/bottom 10% removed (y)
Min	0	0.3	0	0
Max	251	16.8	35	9
Average	10.7	4.3	3.25	1.7
Median	2.8	2.8	1	1
StdDev	31.2	4	6	1.6

a distance, especially given that the standard deviation is very large compared to this, at 31.2 km. This indicates that the extreme measures dominate the calculation. Thus the median, at 2.8 km, is a more reasonable number to look at. 2.8 kms is not a distance which will faze most people from driving it, and would most likely put the co-arresters within the same city.

In order to remove outliers from the data, the top and bottom 10% of the data were removed. In this instance, the maximum distance fell to 16.8 kms, indicating that 10% of the people actually co-arrest with someone living very far away. Similarly, the minimum rose from 0 to 0.3 km, indicating that 10% of the people live very close, as in the same building. The average value fell significantly. All the results indicate that co-arresters tend to live within a few minutes driving of each other.

Why this is so is not clear. It is possible that convenience dictates how far apart offenders live and people do not wish to offend with others who live far away since they do not wish to drive that far. It is also possible that offenders establish co-offending relationships with others more readily if they live in the same neighborhood, simply due to the chance of meeting someone randomly is much larger if both people live in the same neighborhood, than if they live far apart. Either way, spatial distance does seem to be a barrier to co-offending.

5.4   Offenders Age Differences

The setup of this sub-network is similar to the previous sub-network. The difference in ages between co-arrester was calculated for all people who co-arrested together. Figure 13a shows the results. Thickness of the edges shows size of age difference (thickest line is 35 y). All people had birth-dates recorded, hence there are no missing edges.

The network, after a visual analysis, has a lot of very thin edges. In fact, the median difference in ages was 1 y, Table 3. Thus, at least 50% of the people commit crimes with people that are (for all intents and purposes) the same age as them. Even when the top and bottom 10% of the data was removed, the average age difference between co-arresters was 1.7 y, and the maximum difference was only 9 y.
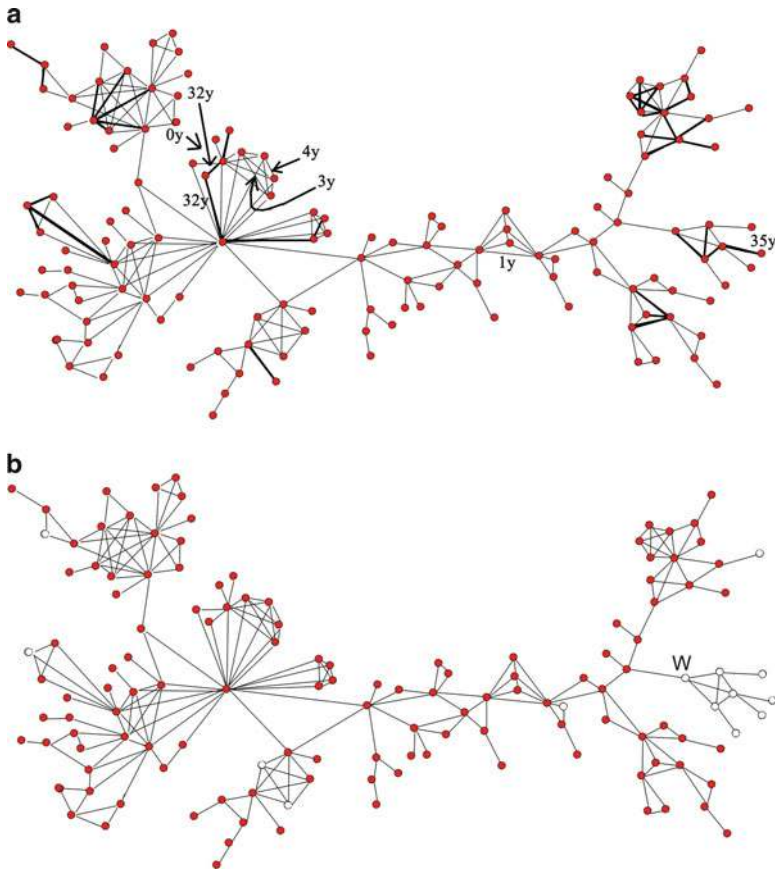
Fig. 13 (**a**) Network showing distance between home locations of offenders (*thickness* indicates magnitude of distance. *Missing links* indicate missing information); (**b**)Network showing gender of offenders (*solid nodes*=male)

How do these people meet? Given that their ages are very close together, it would imply that they meet in school, where they are enclosed with other people of similar ages. This is not true for work-places, where the co-workers could be of any age. If co-offending relationships are established in school, then perhaps more focus could be placed on school activities, such that these would-be offenders are occupied with other things and do not turn their attention to mischief. Perhaps if these co-offending relationships are not established so early, they will not be established later.

The biggest difference in ages was 35 y. This prompted the question, are co-arresters with a large age difference family members? The names of the co-arresters in these large age-difference relationships was compared and the answer, surprisingly, is no. After this discovery, all co-arresters were compared for family-ties (based on encrypted last names) and 4 pairs of co-arresters were found to be

probable family members and had very close ages. one pair of co-arresters was found to have the same family name, but were 17 years apart in age. Thus, it does not seem that family co-arresting is a big problem, but again, due to the limited comparisons based on encrypted family names, this approach has its problems. A wife, who takes her husbands name, would not be tied back to her parents, for example. Two different offenders with last name 'Smith' would be connected although it is a very popular name.

5.5 *Offenders' Gender*

Finally, the network was analyzed by focusing on the gender of the offender. Analysis of this property of the network, Fig. 13b, reveals that most of the offenders (89.8%) are male, which is consistent with literature. However, the distribution of the females in the network does not seem to be random due to the small cluster of female offenders connected to offender W. The linkages between these offenders are visible in the 'Property Crime' sub-network, thus the cluster (not clique) of 8 offenders were involved in property crimes. The link between the female and male offender I however is shown on the 'Serious Crime' sub-network, implying that W is a relatively important offender, involved in at least 2 offenses, and of larger severity than the offenders connected to W.

6 Concluding Remarks

Research in co-offending network mining often lacks access to large real-world crime data sets. One reason for this limitation is the highly sensitive nature of such data and the related privacy issues demanding strict security protocols as well as data storage and processing facilities that meet exceptionally high security standards. An interesting open question is to what extend advanced anonymization techniques can help solving this problem by making secure data more widely available without compromising privacy.

In our study, we have extracted co-offending networks for a number of most important types of crime, including serious, property, drugs and moral crimes. The analysis of the co-offending network revealed several interesting insights. In particular, the ranking of offenders with respect to various centrality measures agreed well, which allowed a robust discovery of the most important offenders to be targeted by the police. Surprisingly, the average distance and the diameter of the co-offending network have shrunken in the last few years, indicating a densification of the social network. These results are in line with known studies that show similar phenomena in other types of social networks [19].

The proposed formal model of crime data and co-offending networks provides a well defined semantic framework for describing in an unambiguous way the

meaning of co-offending networks and their constituent entities at an abstract level. Specifically, the formal model aims at bridging the conceptual gap between data level, mining level and interpretation level, and also facilitates separating the description of the data from the details of data mining and analysis.

Our analysis also pointed out directions that require future research. While a state-of-the-art method for community detection produced more meaningful results than a simple baseline method, the communities detected were too large to be meaningful from a criminological point of view. There is a need for novel community detection methods addressing the special requirements of co-offending networks. Finally, the role of visualization as an enabling factor for analytical reasoning as part of the knowledge–discovery process is crucial in any practical use of the proposed framework by law enforcement and intelligence agencies. We intend to closely collaborate on innovative visualization techniques with the recently founded Vancouver Institute for Visual Analytics (VIVA),[5] a joint initiative by the University of British Columbia and Simon Fraser University.

References

1. Abbasi, A., Chen, H.: Applying authorship analysis to extremist-group Web forum messages. IEEE Intell. Syst. **20**(5), 67–75 (2005)
2. Adderley, R., Musgrove, P.: Modus operandi modelling of group offending: A data-mining case study. Int. J. Police Sci. Manage. **5**(4), 265–276 (2003)
3. Barabasi, A.L., Albert, R.: Emergence of scaling in random networks. Science 286 (1999)
4. Barabasi, A.L., Jeonga, H., Neda, Z., Ravasz, E., Schubert, A., Vicsek, T.: Evolution of the social network of scientific collaborations. Physica **311**, 590614 (2002)
5. Brandes, U., Erlebach, T.: Fundamentals. In: Network Analysis: Methodological Foundations. Springer, Berlin (2005)
6. Brantingham, P.L.: Crime pattern theory. In: Fisher, B., Lab, S. (eds.) Encyclopedia of Victimology and Crime Prevention. Sage Publishing, Beverly Hills (2010)
7. Brantingham, P.L., Glässer, U., Jackson, P., Kinney, B., Vajihollahi, M.: Mastermind: Computational modeling and simulation of spatiotemporal aspects of crime in Urban environments. In: Liu, L., Eck, J. (eds.) Artificial Crime Analysis Systems: Using Computer Simulations and Geographic Information Systems. IGI Global, Hershey, PA (2008)
8. Brantingham, P.L., Glässer, U., Jackson, P., Vajihollahi, M.: Modeling criminal activity in Urban landscapes. In: Memon, N., et al. (eds.) Mathematical Methods in Counterterrorism. Springer, Berlin (2009)
9. Bruinsma, G., Bernasco, W.: Criminal groups and transnational illegal markets. Crime Law Soc. Change **41**(1) (2004)

---

[5]See also www.sfu.ca/~viva/.

10. Chen, J., Zaine, O.R., Goebel, R.: Detecting communities in social networks using max-min modularity. In: Proceedings of the SIAM International Conference on Data Mining, SDM 2009, Sparks, Nevada, USA (2009)
11. Chakrabarti, D., Faloutsos, C.: Graph mining: Laws, generators, and algorithms. ACM Comput. Surv. **38**(1) (2006)
12. Clauset, A., Shalizi, C.R., Newman, M.E.J.: Power-law distributions in empirical data. http://arxiv.org/abs/0706.1062v1. (2007)
13. Dijkstra, E.: A note on two problems in connection with graphs. Numer. Math. **1** 269271 (1959)
14. Erdos, P., Renyi, A.: On random graphs. Publ. Math. **6** (1959)
35. Freeman, L.C.: Visualizing social networks. J. Soc. Struct. **1**(1) (2000)
16. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA **99** 7821–7826 (2002)
17. Han, J., Kamber, M.: Data Mining: Concepts and Techniques, 2nd edn. Morgan Kaufmann, San Francisco, CA (2006)
18. Hauck, R.V., Atabakhsh, H., Ongvasith, P., Gupta, H., Chen, H.: Using Coplink to analyze criminal-justice data. IEEE Comput. **35**(3), 3037 (2002)
19. Leskovec, J., Kleinberg, J.M., Faloutsos, C.: Graph evolution: Densification and shrinking diameters. ACM TKDD **1**(1), 2 (2007)
20. Liben-Nowell, D., Kleinberg, J.M.: The link prediction problem for social networks. In: Proceedings of the Twelfth Annual ACM International Conference on Information and Knowledge Management, CIKM 2003, Nov 2003
21. Liu, L., Eck, J. (eds.): Artificial Crime Analysis Systems: Using Computer Simulations and Geographic Information Systems. IGI Global, Hershey, PA (2008)
22. Malm, A., Bichler, G., Van de Walle, S.: Comparing the ties that bind criminal networks: Is blood thicker than water?. Secur. J. **23**, 5274 (2010)
23. McGloin, J.M., Sullivan, C.J., Piquero, A.R., Bacon, S.: Investigating the stability of co-offending and co-offenders among a sample of youthful offenders. Criminology **46**(1) (2008)
24. Memon, N., Farley, J.D., Hicks, D.L., Rosenorn, T. (eds.): Mathematical Methods in Countert-errorism. Springer, New York (2009)
25. Kaza, S., Xu, J., Marshall, B., Chen, H.: Topological analysis of criminal activity networks: Enhancing transportation security. IEEE Trans. Intell. Transform. Syst. **10**(1) (2009)
26. Kempe, D., Kleinberg, J.M., Tardos, E.: Influential nodes in a diffusion model for social networks. In: Proceedings of Automata, Languages and Programming, 32nd International Colloquium, ICALP 2005, Lisbon, Portugal (2005)
27. Kumar, R., Novak, J., Tomkins, A.: Structure and evolution of online social networks. In: KDD 06: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 2006
28. Palmer, C., Gibbons, P., Faloutsos, C.: ANF: A fast and scalable tool for data mining in massive graphs. In: SIGKDD (2002)
29. Reiss, A.J.: Co-offending and criminal careers. Crime Justice: A Review of Research. University of Chicago Press, Chicago (1988)
30. Reiss, A.J., Farrington, D.P.: Advancing knowledge about co-offending: Results from a prospective longitudinal survey of London males. J. Crim. Law Criminol. **82**(2) (1991)
31. Rossmo, D.K.: Geographic Profiling. CRC, New York (2000)
32. Short, M.B., Brantingham, P.J., Bertozzi, A.L., Tita, G.E.: Dissipation and displacement of hotspots in reaction-diffusion models of crime. PNAS **107**, 3961–3965 (2010)
33. Smith, M.N., King, P.J.H.: Incrementally visualising criminal networks. In: Sixth International Conference on Information Visualisation (IV'02), iv, pp. 76. (2002)
34. Valente, T.W.: Social Networks and Health: Models, Methods and Applications. Oxford University Press, New York (2010)

35. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Cambridge University Press, Cambridge (1994)
36. Watts, D.J., Strogatz, S.H.: Collective dynamics of small-world networks. Nature **393** (1998)
37. Xu, J.J., Chen, H.: Untangling Criminal Networks: A Case Study. In: ISI 2003, pp. 232–248. (2003)
38. Xu, J.J., Chen, H.: CrimeNet explorer: A framework for criminal network knowledge discovery. ACM Trans. Inform. Syst. **23**(2), 201–226 (2005)

Part II
Tools and Techniques

Region-Based Geospatial Abduction
with Counter-IED Applications

Paulo Shakarian and V.S. Subrahmanian

Abstract Geospatial abduction problems (GAPsfor short) were introduced in
Shakarian et al. (Gaps: Geospatial abduction problems. ACM Transactions on
Intelligent Systems and Technology. (2011)). Given a set $\mathscr{O}$ of observations,
GAPstry to find a set of "partner" locations (points) that best explain those obser-
vations. For instance, the observations may refer to improvised-explosive device
(IED) attacks (or burglaries) and the partner locations may refer to caches supporting
those attacks (or the burglar's house/office). A *region-based* GAP(or RGAP) tries
to find a set of regions that best explain the observations. We study the complexity
and mathematical properties of region-based GAPswhere we vary the shape of the
region(s) we are seeking. We develop several exact and approximate algorithms for
RGAPs, often with guarantees – we also explore practical implementation issues.
We performed experiments where we attempted to use RGAPsto locate weapons
caches in Baghdad based on IED attack locations. Our implementation was able to
find regions that contained multiple weapons caches (on average 1.7 cache sites) as
well as a significantly higher density such caches (8 caches per square kilometer
vs. the city-wide average of 0.4). Further, the algorithm ran quickly, performing
computation in just over 2 s on commodity desktop hardware.

1 Introduction

Suppose $\mathscr{O}$ is a set of locations where we observed some event and we want to
infer regions on the ground where some related events or entities (called partners)
are located. For instance, Criminologists use *geographic profiling* [24] and *crime
pattern theory* [1] to determine locations of criminals (partners) based on where the
criminal activity occurred (observations). In the same vein, military analysts try to

P. Shakarian · V.S. Subrahmanian (✉)

Department of Computer Science, University of Maryland, College Park, MD, USA
e-mail: pshak@cs.umd.edu; vs@cs.umd.edu

find terrorist safe houses and weapons locations (partners) by studying where the attacks occurred (observations). In both cases, detectives and military analysts use their knowledge of the geography of the region involved, as well as knowledge of the socio-cultural-economic aspects of different subregions of the region. All these problems are instances of *geospatial abduction problems* and this class of problems was introduced in [25, 26] by the authors.

In this paper, we introduce a variant of these problems called *region-based geospatial abduction problems* (RGAPs). In RGAPs, we are given a map, a set $\mathcal{O}$ of observations, and a set of subregions of the map (this could include all subregions of the map in the worst case or can be defined by some logical condition). We want to find a set of regions that best "explain" the observations and includes, for each observation, at least one partner.

In this paper, we make several contributions. In Sect. 2, we introduce multiple possible formal definitions of RGAPs – including cases where the regions are determined by a given radius from each observation, regions are non-convex, and when regions are of irregular shape due to terrain restrictions. We then perform a detailed complexity analysis in Sect. 3, proving that most of these problems are NP-complete. This leads us to use approximation techniques in Sect. 4. We also describe some practical implementation issues. Section 5 describes our implementation and includes an experimental evaluation on a real-world data-set consisting of IED attacks in Baghdad, Iraq and related weapons cache sites. In our evaluation, regions outputted by the algorithm contained, on average, 1.7 cache sites, with an average cache density of over 8 caches per square kilometer (significantly higher than the city-wide average of 0.4). Further, the algorithm ran quickly, performing computation in just over 2 s on commodity desktop hardware. Finally, we survey related work in Sect. 6.

2 Technical Preliminaries

We assume the existence of a real-valued $M \times N$ space $\mathcal{S}$ whose elements are pairs of real numbers from the set $[0, M] \times [0, N]$. An observation is any member of $\mathcal{S}$. We use $\mathcal{O}$ to denote an arbitrary, but fixed, finite set of *observations*. We assume there are real numbers $\alpha \leq \beta$ such that for each observation o, there exists a partner p_o (to be found) whose distance from o is in the interval $[\alpha, \beta]$.[1] Without loss of generality, we also assume that all elements of $\mathcal{O}$ are over β distance away from the edge of $\mathcal{S}$. Example 1 presents a neighborhood as a space and locations of illegal drug sales as observations.

Example 1 (Illegal Drug Sales). A criminal gang is selling illegal drugs. Consider the space $\mathcal{S}$ depicted in Fig. 1. Drug dealers were arrested by police at points

---

[1]Shakarian et al. [25, 26] describes methods to learn α, β automatically from historical data.
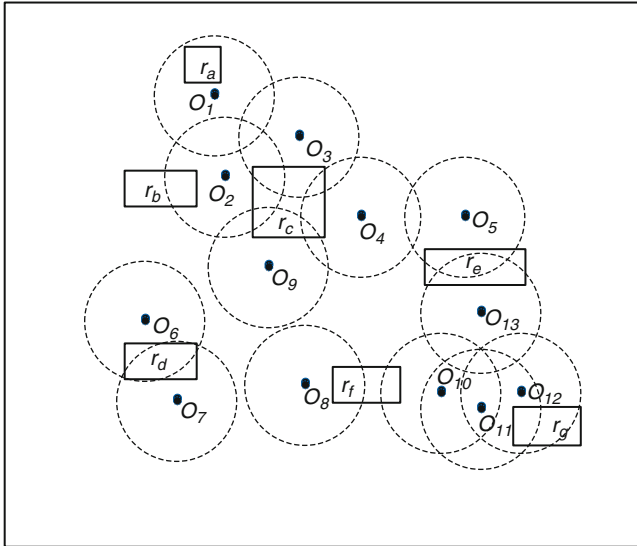
Fig. 1 Locations of illegal drug sales and suspected support zones $\{r_a, r_b, r_c, r_d, r_e, r_f, r_g\}$. The β distance for each observation is shown with a *dashed circle*

$\mathcal{O} \equiv \{o_1, \dots, o_{13}\}$. Historical data suggests that safe houses are located within 5 km of such transactions (i.e. $\alpha = 0$ and $\beta = 5$ km). Note that in Fig. 1, circles of radius 5 km are drawn around the observation points. Police are interested in locating such safe-houses.

Throughout this paper, we assume the notion of a *distance function d* on $\mathcal{S}$ satisfying the usual properties of such distance functions.[2] We now define a region and how they relate to the set of observations. Our intuition is simple – a region *explains* an observation if that region contains a partner point for that observation.

Definition 1 (Region / Super-Explanation / Sub-Explanation). A *region r* is a subset of $\mathcal{S}$ such that for any two points $(x, y), (x', y') \in r$, there is a sequence of line segments from (x, y) to (x', y') s.t. no line segment lies outside r.

1. A region r *super-explains* point o in $\mathcal{S}$ iff there exists a point $p \in r$ such that $d(o, p) \in [\alpha, \beta]$.
2. A region r *sub-explains* some point o in $\mathcal{S}$ iff $(\forall p \in r) \, d(o, p) \in [\alpha, \beta]$.

First, note that regions can have any shape and may overlap. Throughout this paper, we assume that checking if some point o is sub-(super-) explained by region r can be performed in constant (i.e. $O(1)$) time. This is a reasonable assumption for most regular shaped regions like circles, ellipses and polygons. The following result follows immediately from Definition 1.

---

[2] $d(x, x) = 0; d(x, y) = d(y, x); d(x, y) + d(y, z) \geq d(x, z)$.

Observation 2.1 *If region* $r \neq \emptyset$ *sub-explains point o, then r super-explains point o.*

We would like to explain observations by finding regions containing a partner. In some applications, the user may be able to easily search the entire region – hence, a super-explaining region would suffice. In other applications, we may want to be sure that any point within the region can be a partner as not to waste resources – so only a sub-explanation would make sense in such a case. Often, these situations may depend on the size of the regions. We shall discuss the issue of restricting region size later in this section. For now, we shall consider regions any shape or size. Example 2 shows regions that super- or sub-explain various observations.

Example 2. Consider the scenario from Example 1 and the regions $R = \{r_a, r_b, r_c, r_d, r_e, r_f, r_g\}$ shown in Fig. 1. Suppose these regions correspond with "support zones" for the drug sales – i.e. places that may contain a safe-house. Consider region r_a. As it totally lies within the α, β distance of o_1, it sub- and super- explains this observation. Conversely, region r_d super-explains both o_6 and o_7 but sub-explains neither.

This paper studies following decision problems.

Sub-(Super-)Region Explanation Problem (Sub/Sup-REP)
INPUT: Given a space $\mathscr{S}$, distance interval $[\alpha, \beta]$, set $\mathscr{O}$ of observations, set R of regions, and natural number $k \in [1, |\mathscr{O}|]$.
OUTPUT: Set $R' \subseteq R$, where $|R'| \leq k$ and for each $o \in \mathscr{O}$, there is an $r \in R$ s.t. r sub-(super-) explains o.

The fact that a set R of regions is part of the input is not an assumption, but a feature. A user might set R to be all the regions associated with $\mathscr{S}$; alternatively, he might use a logical condition to define regions, taking into account, the terrain and/or known aspects of the population living in the area of interest. For instance, when trying to identify regions containing IED caches in Baghdad used for attacks by Shi'ite groups, regions were defined to be places that were not predominantly Sunni and that did not contain US bases or bodies of water. Other kinds of logical conditions may be used when dealing with burglaries or drug trafficking. Thus, the set R of regions allows an analyst to specify any knowledge he has, and allows the system to benefit from that knowledge. If no such knowledge is available, R can be taken to be the set of all regions associated with $\mathscr{S}$. R can also be used to restrict the size of the region (e.g. only considering regions whose area is less than 5 sq. km.).

There are two different associated optimization problems associated with both Sub-REP and Sup-REP. The first deals with finding a subset of regions of minimal cardinality that explains all observations.

Sub-(Super-)Region Explanation Problem-Minimum Cardinality (Sub/Sup-REP-MC)
INPUT: Given a space, $\mathscr{S}$, distance interval $[\alpha, \beta]$, set of observations $\mathscr{O}$, and set of regions R.

OUTPUT: Set $R' \subseteq R$ of minimum cardinality, where for each $o \in \mathcal{O}$, there is an $r \in R$ s.t. r sub-(super-) explains o.

Our second optimization problem fixes the number of regions returned in the solution, but maximizes the number of observations that are explained.

Sub-(Super-)Region Explanation Problem-Maximum Explaining (Sub/Sup-REP-ME)

INPUT: Given a space $\mathcal{S}$, distance interval $[\alpha, \beta]$, set $\mathcal{O}$ of observations, set R of regions, and natural number $k \in [1, |\mathcal{O}|]$.

OUTPUT: Set $R' \subseteq R$, where $|R'| \leq k$ s.t. the number of $o \in \mathcal{O}$ where there is an $r \in R$ s.t. r sub-(super-) explains o is maximized.

Consider the following Example.

Example 3. Consider the scenario from Example 2. Consider an instance of Sup-REP with $k = 7$. The set $\{r_a, r_b, r_c, r_d, r_e, r_f, r_g\}$ is a solution to this problem. Now consider Sup-REP-MC with $k = 6$, the set $\{r_a, r_c, r_d, r_e, r_f, r_g\}$ is a solution to this problem. Finally, consider Sup-REP-ME with $k = 2$. The set $\{r_c, r_d\}$ is a solution to this problem.

We now consider a special case of these problems that arises when the set R of regions is created by a partition of the space based on the set of observations ($\mathcal{O}$) and concentric circles of radii α and β drawn around each $o \in \mathcal{O}$. We can associate regions in such a case with subsets of $\mathcal{O}$. For a given subset $\mathcal{O}'$, we say that there is an associated set of *induced regions* (denoted $R_{\mathcal{O}'}$), defined as follows:

$$R_{\mathcal{O}'} = \{\{x| \ \forall o \in \mathcal{O}', d(x, o) \in [\alpha, \beta] \wedge$$
$$\forall o' \notin \mathcal{O}', d(x, o') \notin [\alpha, \beta]\} \}$$

We note that for a given subset of observations, it is possible to have a set of induce regions, $R_{\mathcal{O}'}$ that has more than one element. For example, consider set $R_\emptyset = \{r_1, r_{12}\}$ in Fig. 2. For a given set of observations $\mathcal{O}$, we will use the notation $R_\mathcal{O}$ do denote the set of all induce regions. Formally,

$$R_\mathcal{O} = \bigcup_{\substack{\mathcal{O}' \in 2^\mathcal{O} \\ R_{\mathcal{O}'} \neq \emptyset}} R_{\mathcal{O}'}$$

We illustrate the idea of induce regions in the following example.

Example 4. In order to identify locations of drug safe-houses, police create 33 *induced* regions in $\mathcal{S}$ by drawing 5 km radius circles around all observations (see Fig. 2), the set of which is denoted $R_\mathcal{O} = \{r_1, \ldots, r_{33}\}$.

For the special case where $R_\mathcal{O}$ is the set of regions, we have the following result.
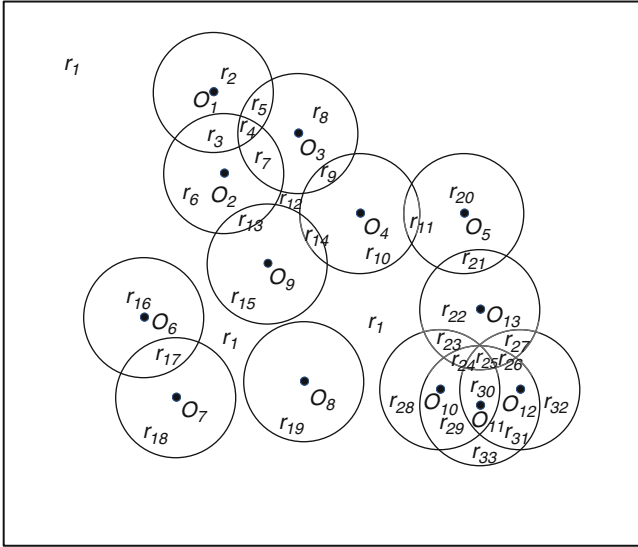
Fig. 2 Space $\mathscr{S}$ and the regions in set $R_{\mathscr{O}}$

Lemma 1. *Suppose $\mathscr{O}$ is a set of observation and $R_{\mathscr{O}}$ is the induced region. A region $r \in R_{\mathscr{O}}$ sub-explains an observation $o \in \mathscr{O}$ iff it super-explains o.*

Note that the proof of this lemma, as well as all lemmas and theorems in this paper, can be found in the appendix.

By this result, for the special case of induced regions, we only need one decision problem.

Induced Region Explanation Problem (I-REP)
INPUT: Given a space, $\mathscr{S}$, distance interval $[\alpha, \beta]$, set $\mathscr{O}$ of observations, and natural number $k \in [1, |\mathscr{O}|]$.
OUTPUT: Set $R' \subseteq R_{\mathscr{O}}$, where $|R'| \leq k$ and for each $o \in \mathscr{O}$, there is an $r \in R$ s.t. r sub-explains o.

As mentioned earlier, the sizes of regions can be regulated by our choice of R. However, we may also explicitly require that all regions must be less than a certain area. Consider the following variant of Sup-REP.

Area-Constrained Super-Region Explanation Problem (AC-Sup-REP)
INPUT: Given a space, $\mathscr{S}$, distance interval $[\alpha, \beta]$, set $\mathscr{O}$ of observations, set R of regions, area A, and natural number $k \in [1, |\mathscr{O}|]$.
OUTPUT: Set $R' \subseteq R$, where $|R'| \leq k$ and each $r \in R'$ has an area $\leq A$ and for each $o \in \mathscr{O}$, there is an $r \in R$ s.t. r super-explains o.

The following proposition tells us that AC-Sup-REP is at least as hard as I-REP, yet no harder than Sup-REP (an analogous result can easily be shown for an area-constrained version of Sub-REP). We note that essentially, we eliminate the regions whose area is above area A, which gives us an instance of Sup-REP. To go the other direction, we directly encode I-REP into an instance of AC-Sup-REP and have A be larger than the area of any region.

Theorem 1. *I-REP is polynomially reducible to AC-Sup-REP. AC-Sup-REP is polynomially reducible to Sup-REP.*

In our final observation of this section, we note that the set $R_\mathcal{O}$ can be used as a "starting point" in determining regions. For instance, supplemental information on area that may be restricted from being partnered with an observation may also be considered and reduce the area of (or eliminate altogether) some regions in the set. Consider the following example.

Example 5. Consider the scenario from Example 4. The police may eliminate a river running through the area and certain other ares from their search. These "restricted areas" are depicted in Fig. 3. Note that several regions from Fig. 2 are either eliminated or have decreased in size. However, by eliminating these areas, the police have also pruned some possibilities from their search. For example, regions r_9, r_{13} were totally eliminated from consideration.
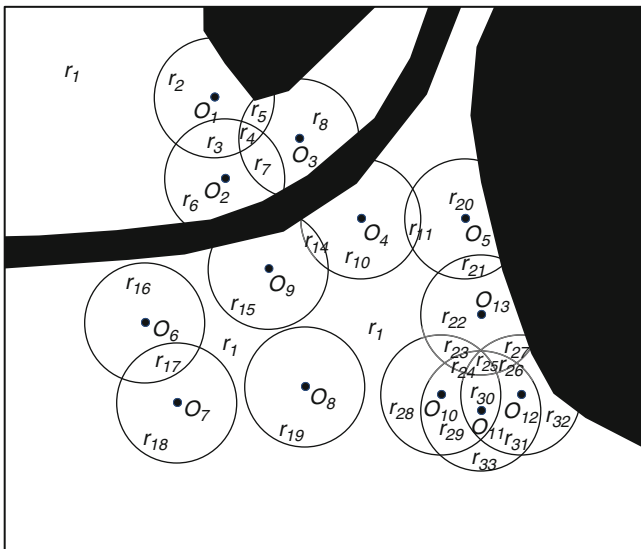
Fig. 3 A set of regions in $\mathscr{S}$ created based on the distance $\beta = 5$ km as well as restricted areas (shown in *black*)

3   Complexity

In this section, we show that Sub-REP, Sup-REP, and I-REP are NP-Complete and that the associated optimization problems are NP-Hard. We also show that the optimization problems Sub-REP-MC, Sup-REP-MC, and I-REP-MC cannot be approximated by a fully polynomial-time approximation scheme (FPTAS) unless $P = NP$. We also note that the complexity of the area-constrained versions of these problems follows directly from the results of this section by the reduction of Theorem 1 (page 111).

   We first prove that I-REP is NP-complete, which then allows us to correctly identify the complexity classes of the other problems by leveraging Lemma 1. First, we introduce the problem "circle covering" (CC) that was proven to be NP-complete in [18].

Circle Covering (CC)
INPUT: A space $\mathscr{S}'$, set P of points, real number β', natural number k'.
OUTPUT: "Yes" if there is a set of points, Q in $\mathscr{S}'$ such that all points in P are covered by discs centered on points in Q of radius β' where $|Q| \leq k'$ – "no" otherwise.

   The theorem below establishes that I-REP is NP-complete.

Theorem 2. *I-REP is NP-Complete.*

Proof Sketch. *Clearly, a solution to I-REP can be verified in PTIME. To show NP-hardness, we show that $CC \leq_p$ I-REP by the following transformation: $\mathscr{S} = \mathscr{S}'$, $\mathscr{O} = P$, $\beta = \beta'$, $\alpha = 0$, and $k = k'$. ($\Leftarrow$) Given a solution to the instance of I-REP, we can simply pick a point in each returned region, and center a circle on it of radius β' – which will be a solution to CC. Likewise, ($\Rightarrow$) given a solution to CC, we can be assured that each point in the solution is enclosed by exactly one region from the set $R_{\mathscr{O}}$, which would ensure a solution to I-REP.* □

   Further, as the optimization version of circle covering is known to have no FPTAS unless $P = NP$ [13], by the nature of the construction in Theorem 2, we can be assured of the same result for I-REP-MC.

Corollary 1. *I-REP-MC cannot be approximated by a fully polynomial-time approximation scheme (FPTAS) unless $P = NP$.*

   So, from the above Theorem and Corollary and Lemma 1, we get the following results:

Corollary 2. *1. Sub-REP and Sup-REP are NP-Complete.*
2. Sub-REP-MC, Sup-REP-MC, I-REP-MC, Sub-REP-ME, Sup-REP-ME, and I-REP-ME are NP-Hard.
3. Sub-REP-MC, Sup-REP-MC cannot be approximated by a FPTAS unless $P = NP$.

4 Algorithms

In this section we devise algorithms to address the optimization problems associated with Sup-REP, Sub-REP, and I-REP. First, we show that these optimization problems reduce to either instances of set-cover (for Sub/Sup-REP-MC) or max-k-cover (for Sub/Sup-REP-ME). These problems are well-studied and there are algorithms that provide exact and approximate solutions. We then provide a new greedy-algorithm for Sub/Sup-REP-MC that also provides an approximation guarantee. This is followed by a discussion of approximation for I-REP-ME for the case where $\alpha = 0$. Finally, we discuss some practical issues dealing with implementation.

4.1 Exact and Approximate Solutions by Reduction

In this section, we show that the -MC problems can reduce to set-cover and that the -ME problem can reduce to max-k-cover. First, we introduce the two problems in question. First, we present set-cover [20].

Set-Cover
INPUT: Set of elements S, family of subsets of S, $\mathscr{H} = H_1, \ldots, H_m$.
OUTPUT: Subset $\mathscr{H}' \subseteq \mathscr{H}$ of minimum cardinality s.t. $\bigcup_{H_i \in \mathscr{H}'} H_i \supseteq S$.

Next, we present max-k-cover [6], which is often regarded as the dual of set-cover:
Max-k-Cover
INPUT: Set of elements S, family of subsets of S, $\mathscr{H} = H_1, \ldots, H_m$, natural number $k \leq |S|$.
OUTPUT: Subset $\mathscr{H}' \subseteq \mathscr{H}$ s.t. $|\mathscr{H}'| \leq k$ where $|\bigcup_{H_i \in \mathscr{H}'} H_i \cap S|$ is maximized.

The key to showing that Sub/Sup-REP optimization problems can reduce to one of these problems is to determine the family of subsets. We accomplish this as follows: for each region $r \in R$, we find the subset of $\mathscr{O}$ that can be partnered with r. We shall refer to this set as $\mathscr{O}_r$. This gives us the following algorithm for the optimization problems (we simply omit the k parameter for the -MC problems that reduce to Set-Cover):

REDUCE-TO-COVERING($\mathscr{O}$ *set of observations, R set of regions, k natural number*) returns instance of covering problem $\langle S, \mathscr{H}, k \rangle$

1. For each $r \in R$, find $\mathscr{O}_r$ (i.e. o is in $\mathscr{O}_r$ iff r sub/super-explains o)
2. Return $\langle \mathscr{O}, \bigcup_{r \in R} \{\mathscr{O}_r\}, k \rangle$

Proposition 1. *REDUCE-TO-COVERING requires* $O(|\mathcal{O}| \cdot |R|)$ *time.*

The following theorem shows that REDUCE-TO-COVERING correctly reduces a Sub/Sup-REP optimization problem to set-cover or max-k-cover as appropriate.

Theorem 3. *Sub/Sup-REP-MC polynomially reduces to Set-Cover and Sub/Sup-REP-ME polynomially reduces to Max-k-Cover.*

This result allows us to leverage any exact approach to the above optimization problems to obtain a solution to an optimization problem associated with Sub/Sup-REP. A straightforward algorithm for any of the optimization problems would run in time exponential in $|\mathcal{O}|$ or k and consider every $|\mathcal{O}|$ or k sized subset of $\bigcup_{r \in R}\{\mathcal{O}_r\}$. Clearly this is not practical for real-world applications. Fortunately, there are several well-known approximation techniques for both these problems. First, we address the Sub/Sup-REP-ME problems, which reduce to Max-k-Cover. As the Max-k-Cover problem reduces to the maximization of a submodular function over a uniform matroid, we can leverage the greedy approximation algorithm of [19] to our problem. We do so below.

GREEDY-REP-ME($\mathcal{O}$ *set of observations, R set of regions, k natural number*) returns $R' \subseteq R$

1. Let $\mathbf{O} = \bigcup_{r \in R}\{\mathcal{O}_r\}$ (obtained by REDUCE-TO-COVERING)
2. Let $\mathcal{O}' = \mathcal{O}$, set $R' = \emptyset$
3. While $k \neq 0$ loop

    a. Let the element $\mathcal{O}_r$ be the member of $\mathbf{O}$ s.t. $|\mathcal{O}_r \cap \mathcal{O}'|$ is maximized.
       $R' = R' \cup r$
       $\mathcal{O}' = \mathcal{O}' - (\mathcal{O}_r \cap \mathcal{O}')$
       $k - -$

4. Return R'

Suppose 'f' denotes the maximum number of observations that can be partnered with a given region.

Proposition 2. *GREEDY-REP-ME runs in* $O(k \cdot |R| \cdot f)$ *time and returns a solution such that the number of observations in* $\mathcal{O}$ *that have a partner region in* R' *is within a factor* $\left(\frac{e}{e-1}\right)$ *of optimal.*

Example 6. Consider Example 2 (page 108), where the set of regions is $R = \{r_a, r_b, r_c, r_d, r_e, r_f, r_g\}$. Suppose the police want to run GREEDY-REP-ME to solve an instance of Sup-REP-ME associated with this situation with $k = 3$. Initially set $\mathcal{O}' = \{o_1, \ldots, o_{13}\}$. On the first iteration of the outer loop, it identifies set $\mathcal{O}_{r_c} = \{o_2, o_3, o_4, o_9\}$ where the cardinality of $\mathcal{O}_{r_c} \cap \mathcal{O}'$ is maximum. Hence, it picks region r_c. The set $\mathcal{O}' = \{o_1, o_5, \ldots, o_8, o_{10}, \ldots o_{13}\}$. On the second iteration, it identifies $\mathcal{O}_{r_e} = \{o_5, o_{13}\}$, which intersected with $\mathcal{O}'$ provides a maximum

cardinality, causing r_e to be picked. Set $\mathscr{O}'$ is now $\{o_1, o_6, \ldots, o_8, o_{10}, \ldots, o_{12}\}$. On the last iteration, it identifies $\mathscr{O}_{r_g} = \{o_{11}, o_{12}\}$, again the maximum cardinality when intersected with $\mathscr{O}'$. The element is picked and the solution is r_c, r_e, r_g, and the observations super-explained are $\{o_2, o_3, o_4, o_5, o_9, o_{11}, o_{12}, o_{13}\}$.

Likewise, we can leverage the greedy algorithm for set-cover [20] applied to Sub/Sup-REP-MC.

GREEDY-REP-MC($\mathscr{O}$ *set of observations, R set of regions,*) returns $R' \subseteq R$

1. Let $\mathbf{O} = \bigcup_{r \in R} \{\mathscr{O}_r\}$ (obtained by REDUCE-TO-COVERING)
2. Let $\mathscr{O}' = \mathscr{O}$, set $R' = \emptyset$
3. While not $\mathscr{O}' \equiv \emptyset$ loop

    a. Let the element $\mathscr{O}_r$ be the member of $\mathbf{O}$ s.t. $|\mathscr{O}_r \cap \mathscr{O}'|$ is maximized.
      $R' = R' \cup r$
      $\mathscr{O}' = \mathscr{O}' - (\mathscr{O}_r \cap \mathscr{O}')$

4. Return R'

Proposition 3. *GREEDY-REP-MC runs in $O(|\mathscr{O}| \cdot |R| \cdot f)$ time and returns a solution whose cardinality is within a factor of $1 + \ln(f)$ of optimal.*

Example 7. Consider the scenario from Example 6. To explain all points, the police can create an instance of Sup-REP-MC and use GREEDY-REP-MC. The algorithm proceeds just as GREEDY-REP-ME in the first three steps (as in Example 6, but will continue on until all observations are super-explained. So, GREEDY-REP-MC proceeds for three more iterations, selecting r_f ($\mathscr{O}_{r_f} = \{o_8, o_{10}\}$), r_d ($\mathscr{O}_{r_d} = \{o_6, o_7\}$), and finally r_a ($\mathscr{O}_{r_a} = \{o_1\}$). The solution returned is $\{r_c, r_e, r_g, r_f, r_d, r_a\}$.

We now focus on speeding up the set-cover reduction via the GREEDY-REP-MC2 algorithm below.

In the rest of this section, we use 'Δ' to denote the maximum number of different regions that can be partnered with a given observation.

Proposition 4. *GREEDY-REP-MC2 runs in $O(\Delta \cdot f^2 \cdot |\mathscr{O}| + |\mathscr{O}| \cdot \ln(|\mathscr{O}|))$ time and returns a solution whose cardinality is within a factor of $1 + \ln(f)$ of optimal.*

Although GREEDY-REP-MC2 considers regions in a different order than GREEDY-REP-MC, it maintains the same approximation ratio. This is because the region (in set GRP_o) that is partnered with the greatest number of uncovered observations is selected at each iteration, allowing us to maintain the approximation guarantee. There are two selections at each step: the selection of the observation (in which we use a minimal key value based on related observations) and a greedy selection in the inner loop. Any selection of observations can be used at each step and the approximation guarantee is still maintained. This allows for a variety of different heuristics. Further, the use of a data structure such as a Fibonacci Heap allows us to actually obtain a better time complexity than GREEDY-REP-MC.

GREEDY-REP-MC2($\mathcal{O}$ set of observations, R set of regions,) returns $R' \subseteq R$

1. Let $\mathbf{O} = \bigcup_{r \in R}\{\mathcal{O}_r\}$ (obtained by REDUCE-TO-COVERING)
2. For each observation $o \in \mathcal{O}$, let $GRP_o = \{\mathcal{O}_r \in \mathbf{O} | o \in \mathcal{O}_r\}$
3. For each observation $o \in \mathcal{O}$, let $REL_o = \{o' \in \mathcal{O} | o' \in \bigcup_{\mathcal{O}_r \in GRP_o} \mathcal{O}_r\}$ and let $key_o = |REL_o|$
4. Let $\mathcal{O}' = \mathcal{O}$, set $R' = \emptyset$
5. While not $\mathcal{O}' \equiv \emptyset$ loop

    a. Let o be the element of $\mathcal{O}$ where key_o is minimal.
    b. Let the element $\mathcal{O}_r$ be the member of GRP_o s.t. $|\mathcal{O}_r \cap \mathcal{O}'|$ is maximized.
    c. If there are more than one set $\mathcal{O}_r$ that meet the criteria of line 5b, pick the set w. the greatest cardinality.
    d. $R' = R' \cup r$
    e. For each $o' \in \mathcal{O}_r \cap \mathcal{O}'$, do the following:

        i. $\mathcal{O}' = \mathcal{O}' - o'$
        ii. For each $o'' \in \mathcal{O}' \cap REL_{o'}$, $key_{o''} --$

6. Return R'

Example 8. Consider the situation in Example 4 where the police are considering regions $R_{\mathcal{O}} = \{r_1, \ldots, r_{33}\}$ that are induced by the set of observations and wish to solve I-REP-MC using GREEDY-REP-MC. On the first iteration of the loop at line 5, the algorithm picks o_8, as $key_{o_8} = 1$. The only possible region to pick is r_{19}, which can only be partnered with o_8. There are no observations related to o_8 other than itself, so it proceeds to the next iteration. It then selects o_6 as $key_{o_6} = 2$ because $REL_{o_6} = \{o_6, o_7\}$. It then greedily picks r_{17} which sub-explains both o_6, o_7. As all observations related to o_6 are now sub-explained, the algorithm proceeds with the next iteration. The observation with the lowest key value is o_5 as $key_{o_5} = 3$ and $REL_{o_5} = \{o_4, o_5, o_{13}\}$. It then greedily picks region r_{21} which sub-explains o_5, o_{13}. The algorithm then reduces the key value associated with o_4 from 4 to 3 and decrements the keys associated with o_{10}, o_{11}, o_{12} (the un-explained observations related to o_{13}) also from 4 to 3. In the next iteration, the algorithm picks o_9 as $key_{o_9} = 3$. It greedily picks r_{12} which sub-explains o_9, o_2. It then decreases key_{o_4} to 2 and also decreases the keys associated with o_1 and o_3. At the next iteration, it picks o_1 as $key_{o_1} = 2$. It greedily selects r_4, which sub-explains o_1, o_3 and decreases the key_{o_4} to 1 which causes o_4 to be selected next, followed by a greedy selection of r_{11} – no keys are updated at this iteration. In the final iteration, it selects o_{10} as $key_{o_{10}} = 3$. It greedily selects r_{25}, which sub-explains all un-explained observations. The algorithm terminates and returns $\{r_{11}, r_{12}, r_{17}, r_{19}, r_{21}, r_{25}\}$.

4.2 Approximation for a Special Case

In Sect. 3, we showed that circle covering is polynomially reducible to I-REP-MC. Let us consider a special (but natural) case of I-REP-MC where $\alpha = 0$, i.e. there is no minimum distance between an observation and a parter point that caused it.

We shall call this special case I-REP-MCZ. There is a great similarity between this problem and circle-covering. It is trivial to modify our earlier complexity proof to obtain the following result.

Corollary 3. *I-REP-MCZ is polynomially reducible to CC.*

Further, we can adopt any algorithm that provides a constructive result for CC to provide a result for I-REP-MCZ in polynomial time with the following algorithm. Given some point p, it identifies the set $\mathcal{O}_r$ associated with the region that encloses that point.

FIND-REGION($\mathscr{S}$ *space*, $\mathcal{O}$ *observation set*, β *real*, p *point*) returns set $\mathcal{O}_r$

1. Set $\mathcal{O}_r = \emptyset$
2. For each $o \in \mathcal{O}$, if $d(p, o) \leq \beta$ then $\mathcal{O}_r = \mathcal{O}_r \cup \{o\}$
3. Return $\mathcal{O}_r$.

Proposition 5. *The algorithm, FIND-REGION runs $O(|\mathcal{O}|)$ time, and region r (associated with the returned set $\mathcal{O}_r$) contains p.*

By pre-processing the regions, we can compute $\mathcal{O}_r$ a-priori and simply pick a region r associated with the output for FIND-REGION. While there may be more than one such region, any one can be selected as, by definition, they would support the same observations.

Example 9. Paleontologists working in a 30×26 km area represented by space $\mathscr{S}$ have located scattered fossils of prehistoric vegetation at $\mathcal{O} = \{o_1, o_2, o_3, o_4\}$. Previous experience has led the paleontologists to believe that a fossil site will be within 3 km of the scattered fossils. In Fig. 4, the observations are labeled and circles with radius of 3 km are drawn (shown with solid lines). Induced regions $r_1, \ldots, r_6$ are also labeled. As the paleontologists have no additional information, and $\alpha = 0$, they can model their problem as an instance of I-REP-MCZ with $k = 3$. They can solve this problem by reducing it to an instance of circle-covering. The circle-covering algorithm returns three points – p_1, p_2, p_3 (marked with an 'x' in Fig. 4). Note that each point in the solution to circle-covering falls in exactly one region (when using induced regions). The algorithm FIND-REGION returns the set $\{o_1, o_2\}$ for point p_1, which corresponds with region r_2. It returns set $\{o_3\}$ for p_2, corresponding with r_6 and returns set $\{o_4\}$ for p_3, corresponding with r_5. Hence, the algorithm returns regions r_2, r_6, r_5, which explains all observations.

Any algorithm that provides a constructive result for CC can provide a constructive result for I-REP-MCZ. Because of this one-to-one mapping between the problems, we can also be assured that we maintain an approximation ratio of any approximation technique.

Corollary 4. *An a-approximation algorithm for CC is an a-approximation algorithm for I-REP-MCZ.*

Fig. 4 Given the instance of
I-REP-MCZ for Example 9 as
input for circle-covering, a
circle-covering algorithm
returns points p_1, p_2, p_3
(points are denoted with an
"x", *dashed circles* are the
area of 3 km from the point)
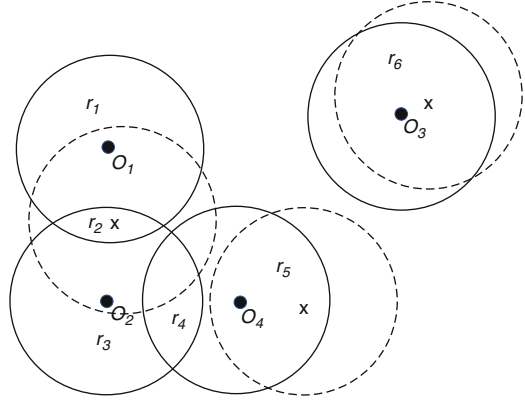
This is useful as we can now use approximation algorithms for CC on I-REP-MCZ. Perhaps the most popular approximation algorithms for CC are based on the "shifting strategy" [13]. To leverage this strategy, we would divide the space, $\mathscr{S}$, into strips of width $2 \cdot \beta$. The algorithm considers groups of ℓ consecutive strips – ℓ is called the "shifting parameter." A local algorithm **A** is applied to each group of strips. The union of all solutions is a feasible solution to the problem. The algorithm then shifts all strips by $2 \cdot \beta$ and repeats the process, saving the feasible solution. This can be done a total of $\ell - 1$ times, and the algorithm simply picks the feasible solution with minimal cardinality. In [13], the following lemma is proved (we state it in terms of I-REP-MCZ – which is done by an application of Corollary 4):

Lemma 2 (Shifting Lemma [13]). *Let $a_{S(A)}$ be the approximation factor of the shifting strategy applied with local algorithm A and a_A be the approximation factor for the local algorithm. Then:*

$$a_{S(A)} = a_A \cdot \left(1 + \frac{1}{\ell}\right).$$

Further, the shifting strategy can actually be applied twice, solving the local algorithm in squares of size $2 \cdot \beta \cdot \ell \times 2 \cdot \beta \cdot \ell$. This gives the following result:

$$a_{S(S(A))} = a_A \cdot \left(1 + \frac{1}{\ell}\right)^2.$$

A good survey of results based on the shifting strategy can be found in [7], which also provides a linear-time algorithm (this result is later generalized by [10] for multiple dimensions). The following result leverages this for I-REP-MCZ by Corollary 4 (and is proved in [10]).

Proposition 6. *I-REP-MCZ can be solved with an approximation ratio of $x \cdot \left(1 + \frac{1}{\ell}\right)^2$ in $O(K_{\ell,\rho} \cdot |\mathscr{O}|)$ time. Where p is the maximum number of points in a*

finite lattice over a square of side length $2 \cdot \beta \cdot \ell$ s.t. each observation in such a square lies directly on a point in the lattice and $x \in \{3, 4, 5, 6\}$ (and is determined by β, see [7] for details) and $K_{\ell,\rho}$ is defined as follows.

$$K_{\ell,\rho} = \ell^2 \cdot \sum_{i=1}^{\lceil \ell \cdot \sqrt{2} \rceil^2 - 1} \binom{p}{i} \cdot i$$

An alternative to the shifting strategy leverages techniques used for the related problem of geometric dominating set. In [16], the authors present a $1 + \epsilon$ approximation that runs in $O(|\mathcal{O}|^{O(\frac{1}{\epsilon^2} \cdot \lg^2(\frac{1}{\epsilon}))})$ time.

4.3 Practical Considerations for Implementation

We now describe some practical implementation issues. Our primary aim is to find a set of regions that resembles the set of induced regions, $R_{\mathcal{O}}$. There are several reasons for doing this. One reason is to implement a fast heuristic to deal with I-REP optimization problems, specifically when $\alpha \neq 0$. Another, is that such a set of induced regions in the space may be a starting point for creating a set of regions that may include other data, such as that shown in Example 5.

As most GIS systems view space as a set of discrete points, we discretized the space using the REGION-GEN algorithm below. The parameter g is the spacing of a square grid that overlays the space.

REGION-GEN($\mathcal{S}$ *space*, $\mathcal{O}$ *observation set*, α, β, g *reals* returns set R

1. Overlay a grid of spacing g on space $\mathcal{S}$. With each grid point, p, associate set $\mathcal{O}_p = \emptyset$. This can easily be represented with an array.
2. Initialize list L of pointers to grid-points.
3. For each $o \in \mathcal{O}$, identity all grid points within distance $[\alpha, \beta]$. For each point p meeting this criteria, if $\mathcal{O}_p = \emptyset$, add p to L. Also, set $\mathcal{O}_p = \mathcal{O}_p \cup \{o\}$
4. For some subset $\mathcal{O}' \subset \mathcal{O}$, let $str(\mathcal{O}')$ be a bit string of length $|\mathcal{O}|$ where every position corresponding to an element of $\mathcal{O}'$ is 1 and all other positions are 0.
5. Let T be a hash table of size $\lceil |\mathcal{O}| \cdot \frac{\pi \cdot \beta^2}{g^2} \rceil$ regions indexed by bit-strings of length $|\mathcal{O}|$
6. For each $p \in L$, do the following:

   a. If $T[str(\mathcal{O}_p)] = $ null then initialize this entry to be a rectangle that encloses point p.
   b. Else, expand the region at location $T[str(\mathcal{O}_p)]$ to be the minimum-enclosing rectangle that encloses p and region $T[str(\mathcal{O}_p)]$.

7. Return all entries in T that are not null.

Proposition 7. *REGION-GEN has a time complexity* $\Theta(|\mathcal{O}| \cdot \frac{\pi \cdot \beta^2}{g^2})$.

Example 10. Consider the scenario from Example 9. Suppose the paleontologists now want to generate regions using REGION-GEN instead of using induced regions. The algorithm REGION-GEN overlays a grid on the space in consideration. Using an array representing the space, it records the observations that can be explained by each grid point (Fig. 5, top). As it does this, any grid point that can explain an observation is stored in list L. The algorithm then iterates through list L, creating entries in a hash table for each subset of observations, enclosing all points that explain the same observation with a minimally-enclosing rectangle. Fig. 5 (bottom) shows the resulting regions $r_1, \ldots, r_6$.

One advantage to using REGION-GEN is that we already have the observations that a region super-explains stored – simply consider the bit-string used to index the region in the hash table. Another thing that can be done, for use in an algorithm such as GREEDY-MC2, where the regions are organized by what observation they support, can also be easily done during the running of this algorithm at an additional cost of f (the number of observations that can be partnered with a given region) – by updating an auxiliary data structure at line 6a.

5 Experimental Results

We implemented REGION-GEN and GREEDY-MC2 in approximately 3,000 lines of Java code and conducted several experiments on a Windows-based computer with an Intel $x86$ processor. Our goal was to show that solving the optimization problem Sup-REP-MC would provide useful results in a real-world scenario. We looked at counter-insurgency data from [14] that included data on improvised-explosive device attacks in Baghdad and cache sites where insurgents stored weapons. Under the assumption that the attacks required support of a cache site a certain distance away, could we use attack data to locate cache sites using an instance of Sup-REP-MC solved with GREEDY-MC2 using regions created with REGION-GEN? In our framework, the observations were attacks associated with a cache (which was a partner). The goal was to find relatively small regions that enclosed partners (caches). We evaluated our approach based on the following criteria:

1. Do the algorithms run in a reasonable amount of time?
2. Does GREEDY-MC2 return regions of a relatively small size?
3. Do the regions returned by GREEDY-MC2 usually contain a partner (cache)?
4. Is the partner (cache) density within regions returned by GREEDY-MC2 significantly greater than the partner density of the space?
5. How does the spacing between grid points affect the runtime and accuracy of the algorithms?

Overall, the experiments indicate that REGION-GEN and GREEDY-MC2 satisfactorily meet the requirements above. For example, for our trials considering locating regions with weapons cache sites (partners) in Baghdad given recent IED
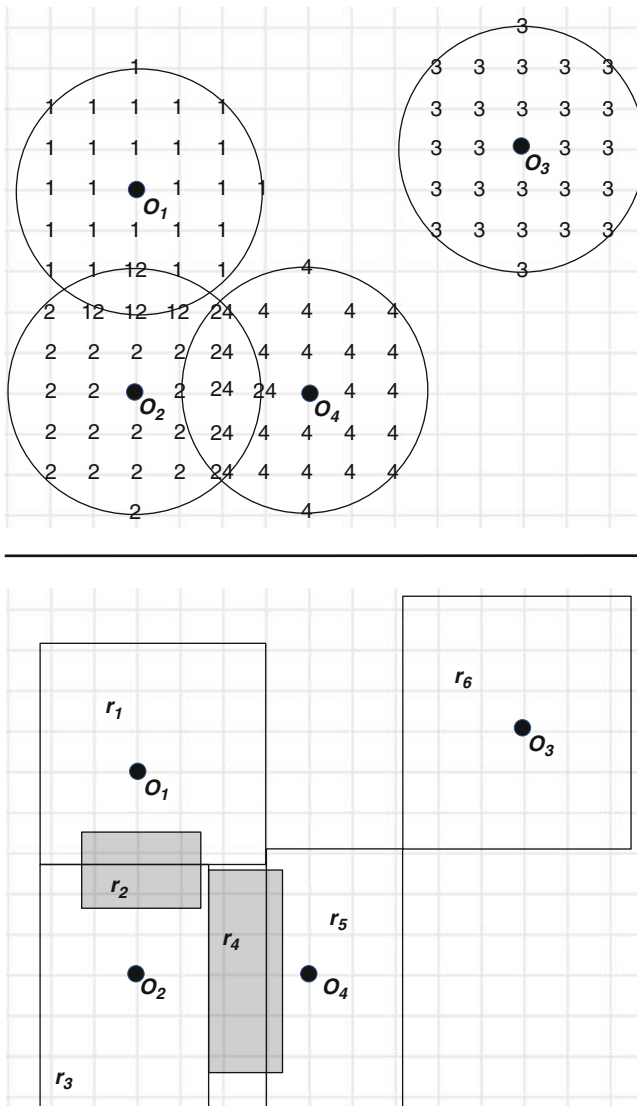
Fig. 5 REGION-GEN applied to the paleontology example (Example 9). First, it identifies observations associated with grid points (*top*). It then creates minimally-enclosing rectangles around points that support the same observations (*bottom*)

attacks (observations), with a grid spacing $g = 100$ m, the combined (mean) runtime on a Windows-based laptop was just over 2 s. The algorithm produced (mean) 15.54 regions with an average area of 1.838 km². Each region, on average, enclosed 1.739 cache sites. If it did not contain a cache site, it was (on average) 275 m away

from one. The density of caches within returned regions was 8.09 caches/km^2 – significantly higher than the overall density for Baghdad of 0.488 caches/km^2.

The rest of this section is organized as follows. Sect. 5.1 describes the data set we used for our tests and experimental set-up. Issue 1 is addressed in Sect. 5.2. We shall discuss the area (issue 2) of the regions returned in Sect. 5.3 and follow this with a discussion of issue 3 in Sect. 5.4. We shall discuss issue 4 in Sect. 5.5. Throughout all the sections, we shall describe results for a variety of different grid-spacings, hence addressing issue 5.

5.1 Experimental Set-Up

We used the *Map of Special Groups Activity in Iraq* available from the Institute for the Study of War [14]. The map plots over 1, 000 insurgent activities attributed to what are termed as "Special Groups" – groups with access to certain advanced weaponry. This data set contains events for 21 months between February 2007 and November 2008. The activity types include the following categories:

1. Attacks with probable links to Special Groups
2. Discoveries of caches containing weapons associated with Special Groups
3. Detainments of suspected Special Groups criminals
4. Precision strikes against Special Groups personnel

We use this data for two geographic areas: the Baghdad urban area and the Sadr City district. In our experiment, we will view the attacks by the special groups (item 1) as observations and attempt to determine the minimum set of cache sites (item 2), which we shall view as partners. Hence, a region returned by GREEDY-MC2 encloses a partner iff a cache falls within the region.

For distance constraints, we used a simple algorithm to learn the parameter β (α was set to zero). This was done using the first 7 months of attack data ($\frac{1}{3}$ of the available months) and 14 months of cache data. We used the following simple algorithm, FIND-BETA, to determine these values. Note we set β_{max} to 2.5 km.

We ran the experiments on a Lenovo T400 ThinkPad laptop with a 2.53 GHz Intel Core 2 Duo T9400 processor and 4GB of RAM. The computer was running Windows Vista 64-bit Business edition with Service Pack 1 installed.

As the relationship between attacks and cache sites may differ varied on terrain, we ran tests with two different geographic areas. First, we considered the entire Baghdad urban area. Then, we considered just the Sadr City district. We ran FIND-BETA with a β_{max} of 2.5 km on both areas prior to testing the algorithms. There were 73 observations (attacks) for Baghdad and 40 for Sadr City. Table 1 shows the exact locations and dimensions of the areas considered.

We conducted two types of tests: tests focusing on GREEDY-MC2 and tests focusing on REGION-GEN.

Algorithm 1 Determines β value from historical data

FIND-BETA($\mathcal{O}_h$ *historical, time-stamped observations*,
$\mathcal{E}_h$ *historical, time-stamped partners*, β_{max} *real*)

1. Set $\beta = \beta_{max}$
2. Set Boolean variable *flag* to TRUE
3. For each $o \in \mathcal{O}_h$, do the following:

    a. For each $p \in \mathcal{E}_h$ that occurs after o, do the following.

        i. Let d be the Euclidean distance function.
        ii. If *flag*, and $d(o, p) \le \beta_{max}$ then set $\beta = d(o, p)$
        iii. If not *flag*, then do the following:

            A. If $d(o, p) > \beta$ and $d(o, p) \le \beta_{max}$ then set $\beta = d(o, p)$

4. Return real β

Table 1 Locations and dimensions of areas considered

Area	Lower-left latitude	Lower-left longitude	E–W (km) distance	N–S (km) distance
Baghdad	33.200° N	44.250° E	27	25
Sadr City	33.345° N	44.423° E	7	7

For the tests of **GREEDY-MC2**, we used multiple setting for the grid spacing. We tested grid spacings at every 10 meter interval in the range of [70, 1000] meters – giving a total of 93 different values for g. Due to the fact that **REGION-GEN** produces a deterministic result, we ran that algorithm only once per grid setting. However, we ran 100 trials of **GREEDY-MC2** per each parameter g. This was done for both Baghdad and Sadr City – giving a total of 18, 600 experiments.

To study the effects of grid-spacing on the run-time of **REGION-GEN**, we also ran 25 trials for each grid spacing setting for both geographic areas – giving a total of 4, 650 experiments. To compare the algorithms running with different settings for g in a statistically valid manner, we used ANOVA [9] to determine if the differences among grid spacings are statistically significant. For some test results, we conducted linear regression analysis.

5.2 Running Time

Overall, the run-times provided by the algorithms were quite reasonable. For example, for the Baghdad trials, 73 attacks were considered for an area of 675 m². With a grid spacing $g = 100$ m, **REGION-GEN** ran in 2,340 ms and **GREEDY-MC2** took less than 30 ms.

For **GREEDY-MC2**, we found that run-time generally decreased as g increased. For Baghdad, the average run times ranged over [1.39, 34.47] ms. For Sadr City, these times ranged over [0.15, 4.97] ms. ANOVAs for both Baghdad and Sadr
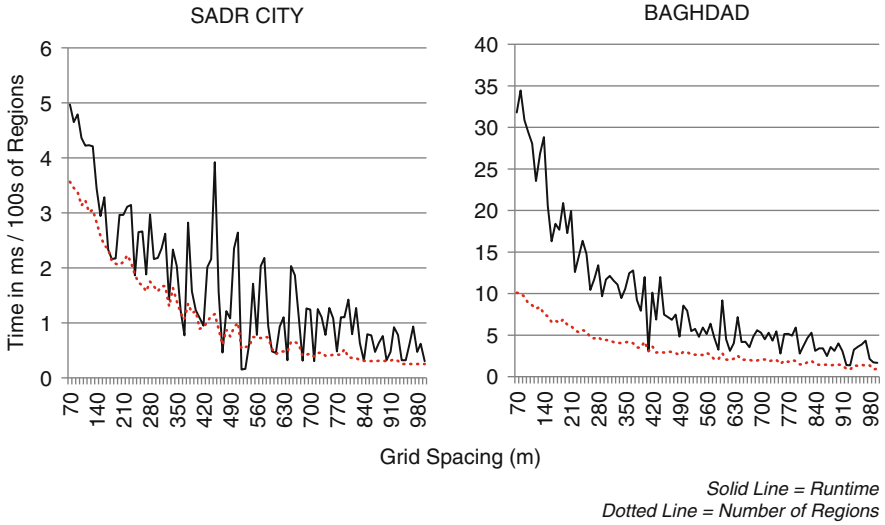
Fig. 6 The run-time of GREEDY-MC2 in ms compared with the number of regions considered

City run-times gave p-values of 2.2×10^{-16}, which suggests with well over 99% probability that the algorithm run with different grid settings will result in different run-times. We also recorded the number of regions considered in each experiment (resulting from the output of REGION-GEN). Like run-times, we found that the number of regions considered also decreased as the grid spacing increased. For Baghdad, the number of considered regions ranged over [88, 1011]. For Sadr City, this number ranged over [25, 356]. ANOVAs for both Baghdad and Sadr City number of considered regions gave p-values of 2.2×10^{-16}, which suggests with well over 99% probability that the algorithm run with different grid settings will result in different numbers of considered regions. Note that this is unsurprising as REGION-GEN run deterministically. We noticed that, generally, only grid spacings that were near the same value would lead to the same number of considered regions.

The most striking aspect of the run-time/number of regions considered results for GREEDY-MC2 is that these two quantities seem closely related (see Fig. 6). This most likely results from the fact that the number of regions that can be associated with a given observation (Δ) increases as the number of regions increases. This coincides with our analysis of GREEDY-MC2 (see Proposition 4).

We also studied the average run-times for REGION-GEN for the various different settings for g. For Baghdad, the average run times ranged over [16.84, 9184.72] ms. For Sadr City, these times ranged over [0.64, 308.92] ms. ANOVAs for both Baghdad and Sadr City run-times gave p-values of 2.2×10^{-16}, which suggests with well over 99% probability that the algorithm run with different grid settings will result in different run-times. Our analysis of REGION-GEN (See Proposition 7) states that the algorithm runs in time $O(\frac{1}{g^2})$. We found striking similarities with this analysis and the experimental results (see Fig. 7).
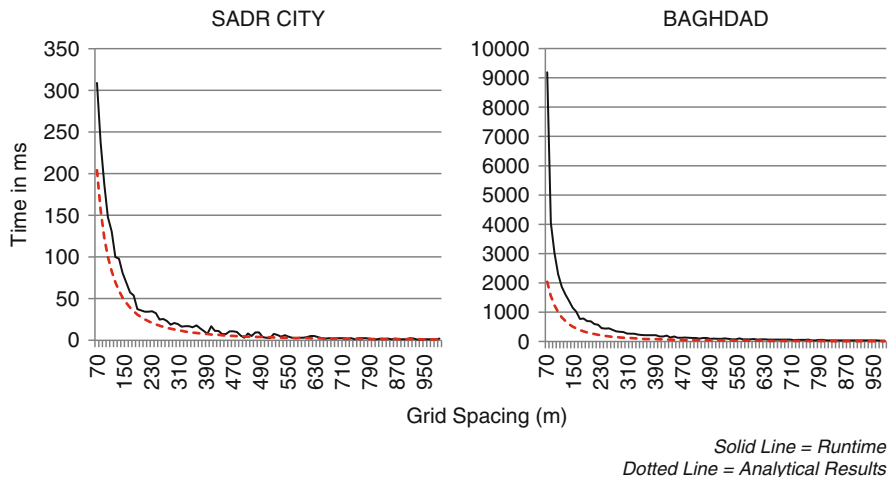
Fig. 7 A comparison between analytical ($O(\frac{1}{g^2})$) and experimental results for the run-time of REGION-GEN compared with grid spacing (g)

5.3 Area of Returned Regions

In this section, we examine how well the REGION-GEN/GREEDY-MC2 suite of algorithms address the issue of returning regions that are generally small. Although not inherently part of the algorithm, our intuition is that the Sup-REP-MC optimization problem will generally return small regions based on the set R produced by REGION-GEN. The reason for this is that we would expect that smaller regions generally support more observations (note that this is not always true, even for induced regions, but our conjecture is that it is often the case for induced regions or the output of REGION-GEN).

To define "small" we look at the area of a circle of radius β as a basis for comparison. As different grid settings led to different values for β, we looked at the smallest areas. For a given trial, we looked at the average area of the returned regions.

For Baghdad, the average areas ranged over [0.611, 2.985] km^2. For Sadr City, these times ranged over (0.01, 0.576] km^2. ANOVAs for both Baghdad and Sadr City run-times gave p-values of 2.2×10^{-16}, which suggests with a 99% probability that the algorithm run with different grid settings will result in different average areas. Plotting the areas compared with the established "minimum area" described earlier in this section clearly shows that REGION-GEN/GREEDY-MC2 produce solutions with an average area that is about half of this value – refer to Fig. 8.

Overall, there seemed to be little relation between grid spacing and average area of the returned set of regions – based on grid spacings in [70, 1000] m. As an example, we provided screen-shots of GREEDY-MC2 for $g = 100$ and $g = 1000$ (Fig. 9). Anecdotally, we noticed that larger grid spacing led to more "pinpoint"
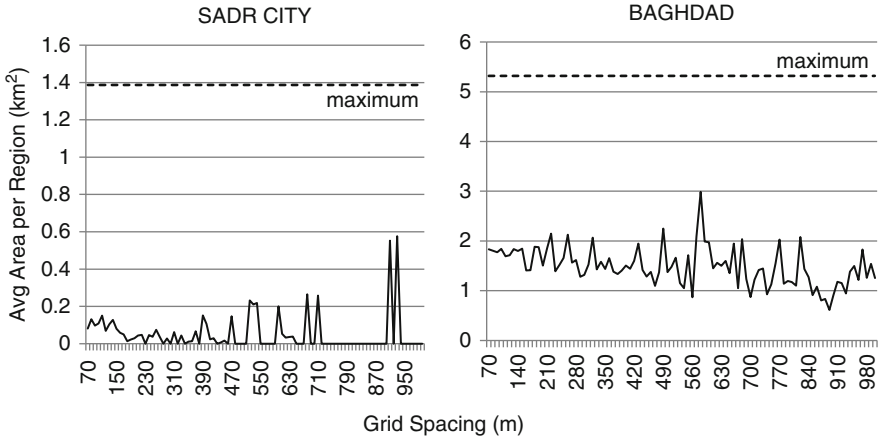
Fig. 8 Average areas for solutions provided by REGION-GEN/GREEDY-MC2 for Baghdad and Sadr City

Fig. 9 Results from two runs of GREEDY-MC2 –
$g = 100$ m (*top*),
$g = 1,000$ m (*bottom*).
Pinpoint-regions are denoted with plus-signs. Notice that the average areas of the results are comparable
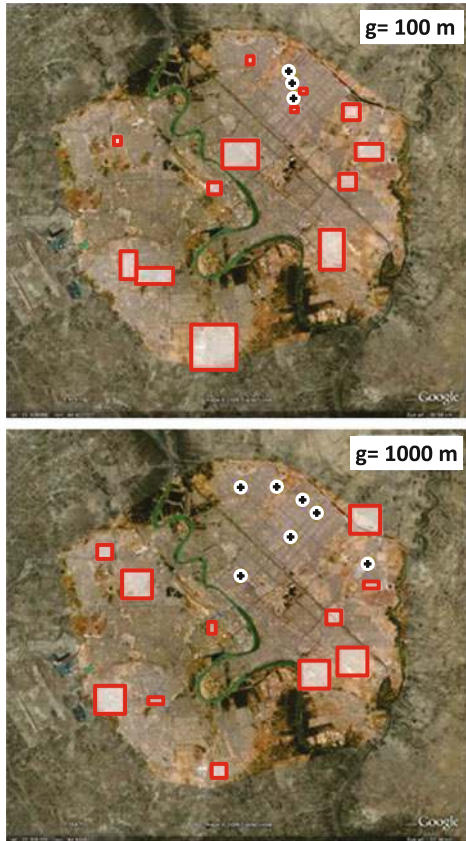
regions – regions encompassing only one point in the grid (and viewed as having an area of 0). This is most likely due to the fact that overlaps in the circles around observations points would overlap on fewer grid points for larger values of g. Another factor is that different settings for g led to some variation of the value β – which also affects accuracy (note for our analysis we considered only the smallest values of β as an upper bound for the area – see Fig. 8.

5.4   Regions That Contain Caches

In this section, we discuss the issue of ensuring that most of the returned regions enclose at least one partner (cache in the case of our experiments). One measure of this aspect is to look at the average number of caches enclosed per region in a given result. We found, that for Baghdad, we generally enclosed more than 1 cache per region in a given result – this number was in the range $[0.764, 3.25]$. The results for Sadr City were considerably lower – in the range $[0, 0.322]$. ANOVAs for both Baghdad and Sadr City gave p-values of 2.2×10^{-16}, which suggests with a 99% probability that the algorithm run with different grid settings will result in different average number of enclosed caches. However, we did not observe an obvious trend in the data (see Fig. 10).

As an alternative metric – we look at the number of regions in provided by GREEDY-MC2 that contain at least one region. Figure 12 shows the number of regions returned in the output. For Baghdad, generally less than half the regions in the output will enclose a cache – the number of enclosing regions was in $[1, 8]$, while the total number of regions was in $[10.49, 22]$. This result, along with the average number of caches enclosed by a region – may indicate that while sometimes
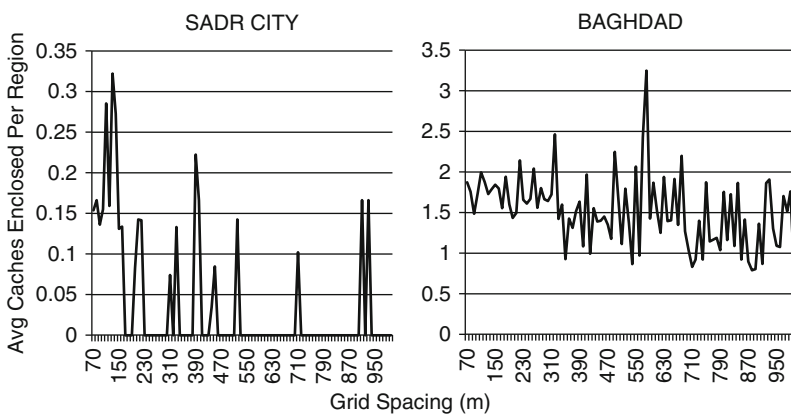
Fig. 10 Average caches enclosed per region for Baghdad and Sadr City for various grid-spacing settings
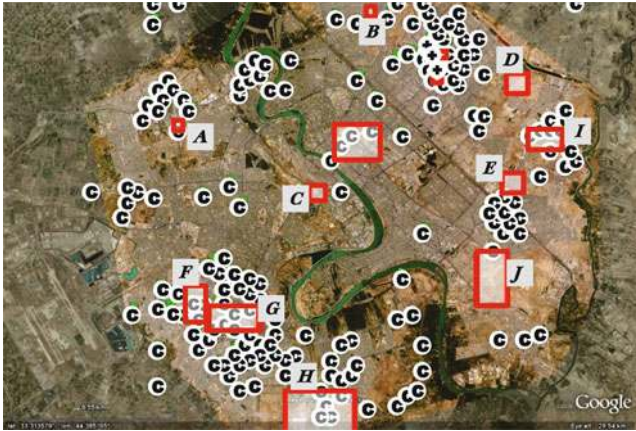
Fig. 11 The output of GREEDY-MC2 for Baghdad with $g = 100$ m compared with the locations of actual cache sites (denoted with a "C"). Notice that regions A–E do not contain any cache sites while regions G–I all contain numerous cache sites

GREEDY-MC2 may find regions that enclose many caches, there are often regions that enclose no caches as well. This may indicate that for Baghdad, some attacks–cache relationships conform to our model and others do not – perhaps there is another discriminating attribute about the attacks not present in the data that may account for this phenomenon. For example, perhaps some attacks were preformed by some group that had a capability to store weapons in a cache located further outside the city, or perhaps some groups had the capability to conduct attacks using cache sites that were never found. We illustrate this phenomenon with an example output in Fig. 11. Note that in the figure, regions A–E do not contain any cache sites while regions G–I all contain numerous cache sites.

For Sadr City, the number of caches that contain one region was significantly lower – in the range $[0, 2]$, while the total number of returned regions was in $[3, 9.8]$. ANOVAs for both Baghdad and Sadr City gave p-values of 2.2×10^{-16}, which suggests with well over 99% probability that the algorithm run with different grid settings will result in different number of caches that enclose a region.

We believe that the low numbers for caches enclosed by regions for Sadr City were directly related to the smaller areas of regions. However, the mean of the average area of a returned set of regions was 0 for 49 of the 94 different grid settings (for Sadr City). This means that for the majority of grid settings, the solution consisted only of "pinpoint" regions (see Sect. 5.3 for a description of "pinpoint" regions).

Obviously, it is unlikely for a pinpoint region to contain a cache site merely due to its infinitesimally small area. To better account for this issue – we develop another metric – distance to nearest cache. If a region contains a cache, the value for this metric is 0. Otherwise, it is the distance to the closest cache outside of the region. For Baghdad, we obtained distances in $[0.246, 0.712]$ km, for Sadr City,
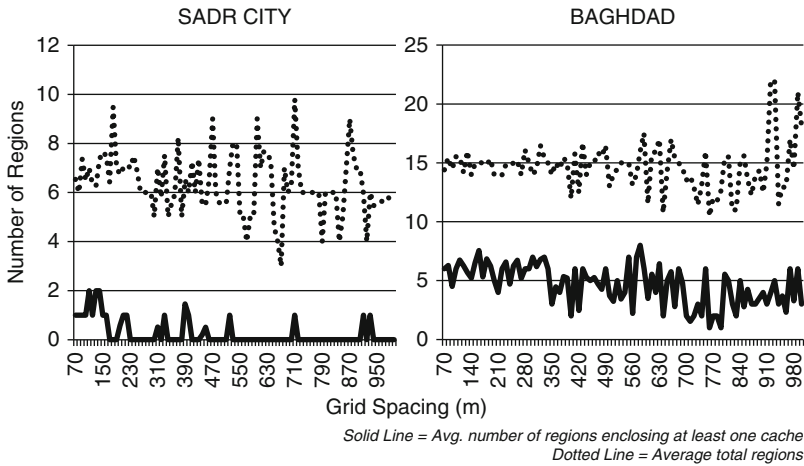
Fig. 12 Regions in the output that enclose at least one partner (cache) and total number of regions returned for Baghdad and Sadr City
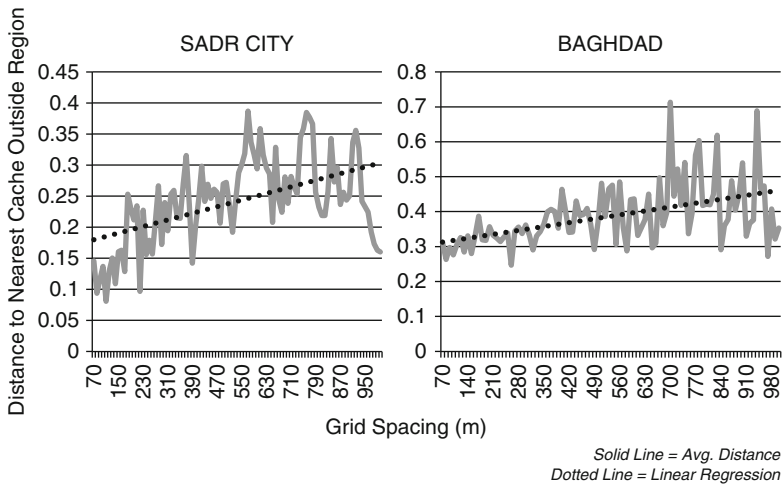
Fig. 13 Distance to nearest cache vs. grid spacing

[0.080, 0.712] km. ANOVAs for both Baghdad and Sadr City gave p-values of 2.2×10^{-16}, which suggests with well over 99% probability that the algorithm run with different grid settings will result in different distances to the nearest cache. Using linear regression, we observed that this distance increases as grid spacing increases. For Baghdad, we obtained $R^2 = 0.2396$ and $R^2 = 0.2688$ for Sadr City. See Fig. 13 for experimental results and the results of the linear regression analysis.

5.5 Partner Density

T consider the density of partners in the regions, we compare the number of enclosed partners to the overall partner density of the area in question. For Baghdad, there were 303 caches in an area 27×24 km – giving a density of 0.488 caches/km^2. For Sadr City, there were 64 caches in an area 7×7 km – giving a density of 1.306 caches/km^2. In our experiments, we looked at the cache density for each output. For Baghdad, the density was significantly higher – in [0.831,34.9] cache/km^2. If we consider $g \in [70, 200]$, the density is in [7.19,32.9] cache/km^2. For $g = 100$, the density was 8.09 caches/km^2. Most likely due to the issue of "pinpoint" regions described in Sect. 5.3, the results for Sadr City, were often lower than the overall density (in [0,31.3] cache/km^2). For $g = 100$, the density was 2.08 caches/km^2. We illustrate these results compared with overall cache density in Fig. 14.

ANOVAs for both Baghdad and Sadr City gave p-values of 2.2×10^{-16}, which suggests with well over 99% probability that the algorithm run with different grid settings will result in different cache densities. Using linear regression, we observed that this cache density decreases as grid spacing increases. For Baghdad, we obtained $R^2 = 0.1614$ and $R^2 = 0.1395$ for Sadr City. See Fig. 14 for experimental results and the results of the linear regression analysis.

Although partner density is a useful metric, it does not tell us anything about partners that lie close to a region – although still outside. For example, consider Fig. 11. Although region A does not enclose any caches, there is a cache just outside – region B is similar. Also consider the cluster of caches south of region E and north of region J – in this situation it appears as though GREEDY-MC2
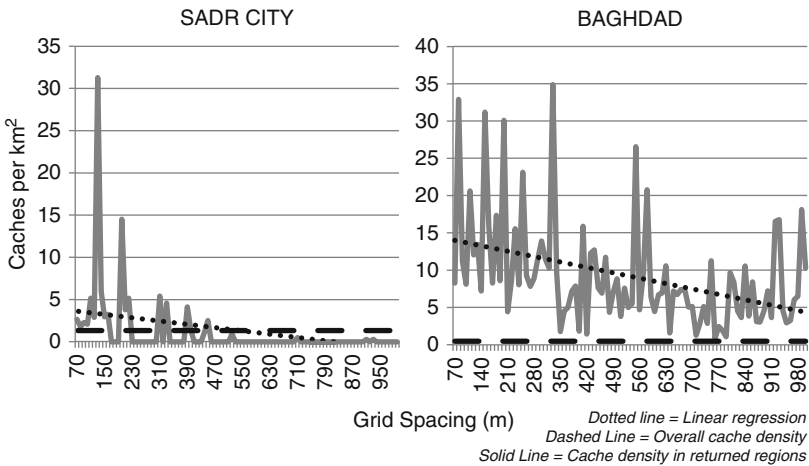
Fig. 14 Cache density of outputs produced by GREEDY-MC2 for Baghdad and Sadr City compared with overall cache density and linear regression analysis

Fig. 15 Close-up of region F
from Fig. 11. While region F
contains 1 cache, there are 4
other caches < 250 m from
the boundary of that region.
The area-quadrupling metric
helps us account for such
scenarios
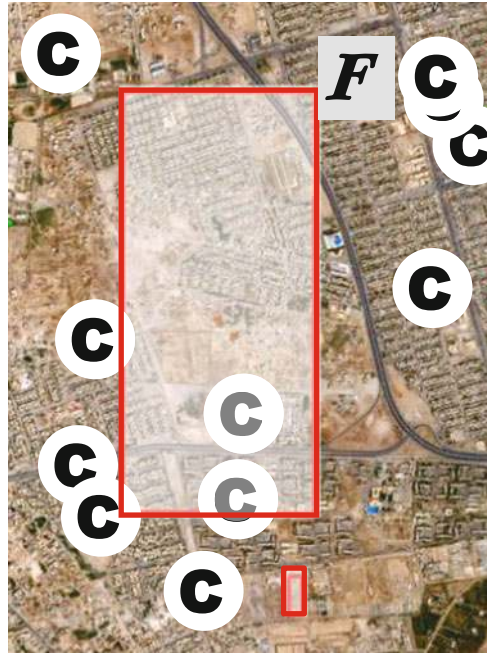
mis-positioned a region. We include a close-up of region F in Fig. 15, which encloses
a cache, but there are also 4 other caches at a distance of 250 m or less.

In order to account for such phenomena, we created an area-quadrupling
metric – that is we uniformly double the sides of each region in the output. Then,
we calculated the density of the output with area-quadrupled regions. For Baghdad,
this density was in $[0.842, 30.3]$ caches/km^2. For Sadr City, this density was in
$[0, 12.3]$ caches/km^2. These results are depicted in Fig. 16.

As the regions for Sadr City were often smaller than those in Baghdad, we found
that the cache density for area-quadrupled regions was often higher for Sadr City
(i.e. a region in Sadr City would have nearby cache sites). An example is shown in
Fig. 15.

ANOVAs for both Baghdad and Sadr City gave p-values of 2.2×10^{-16}, which
suggests with well over 99% probability that the algorithm run with different grid
settings will result in different cache densities for area-quadrupled regions. We also
conducted linear regression analysis, and like the normal partner density, we found
that cache density decreases as grid spacing increases. However, this linear analysis
was more closely correlated with the data than the analysis for non-area quadrupled
density. For Baghdad, we obtained $R^2 = 0.3171$ (for non-area quadrupled, we
obtained $R^2 = 0.1614$) and $R^2 = 0.3983$ (for non-area quadrupled, we obtained
$R^2 = 0.1395$) for Sadr City. See Fig. 16 for experimental results and the results of
the linear regression analysis.

Solid Line = Cache density in quadruple-size regions
Dotted Line = Linear regression
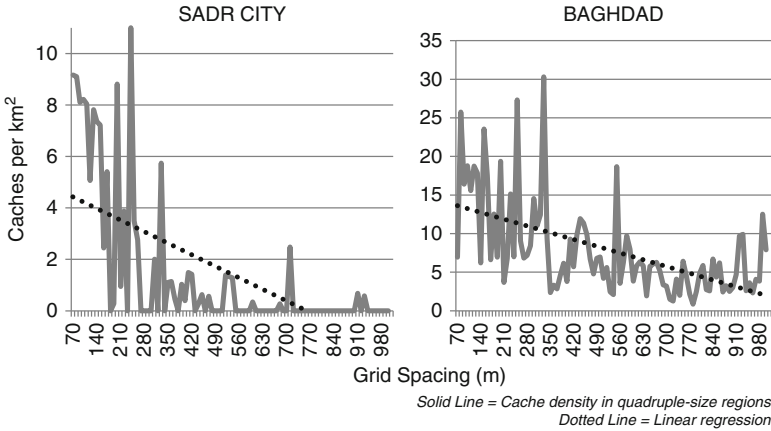
Fig. 16 Area quadrupled cache density of output produced by GREEDY-MC2 with linear regression analysis

6 Related Work

This work builds upon the concept of *geospatial abduction* introduced in [25, 26]. In that paper, the authors consider a space represented by a set of integer points and seek to find a set of *points* that explains a set of observations w.r.t. an objective function. This work differs primarily in that we consider regions and real-valued coordinates, which has several advantages over the approach in [26]. First, for a given optimization problem in [26], there are often multiple solutions. Often, points that differ among solutions are in the same induced region. This has the effect of reducing non-determinism in the algorithms as well as suggesting a single area to consider that intuitively represents the variance among different solutions. Another advantage to regions is that many potential users of a spatial abduction system already use regions – i.e. military analysts use "named areas of interest" and paleontologists use "fossil sites" – both represented as regions and not points. Regions can also have possible computational advantages – providing potentially better approximation guarantees theoretically for certain special cases (i.e. induced regions when $\alpha = 0$). Finally, we also introduce the maximal-explaining optimization problems, in which we do not have to sub/super-explain all observations, just most. This is not studied in [26] and is useful if an application permits outliers or if the user has constrained resources.

Abduction [21] is a form of reasoning used to generate possible explanations for observed phenomenon. It has been used in many applications, including medical diagnosis [22] and planning [15]. Typically, abduction problems are either logic-based [5] or set-covering [3]. Shakarian et al. [26] provides a brief survey of the various forms of abduction. Abducing regions, as in this paper, clearly falls in the category of set-covering abduction and has some similarities with the reasoning used in [22] for medical diagnosis. However, this type of work often seeks to find all

possible explanations, (the counting of which is $\#P$-complete in the general case, as shown in [3]), in this work, we focus on finding a single set of regions that explains all or most of the observations, based on an approximation algorithm that leverages intuitions specific to the problem domain. However, other than [26], we are unaware of any form of abduction that considers space in the way done here. We have already noted the significant difference between this work and [26].

Facility location [27] may also appear similar to this work. However, facility location problems normally seek to locate a facility at an infinitesimal point with respect to some minimality criteria – not identify a region. Further, in a facility location problem, distance is often sought to be minimized – so a "closer" facility may be more optimal. In our formulation, we restrict distance with α, β, but a more optimal region is not necessarily closer to its associated observation. Rather, a region is often more optimal provided if it supports multiple partners. This may, in fact, make regions further from their observations. Another problem, which influences some facility location work, is the *k-means* problem [17]. This type of "clustering" technique looks to group points together and possibly locate a "center." While there is an implicit grouping of observations by the algorithms of this paper, we are attempting to find regions that explain them rather than simply group them. Moreover, [26] shows experimentally that the methods for solving GAPs significantly outperform simply applying k-means algorithms. This fact illustrates that the problem of this paper (and other work in geospatial abduction) is fundamentally different from work in clustering. Perhaps some of the closest work to our problem is in the study of the circle-covering problem [2, 12, 13, 18]. The problem of this paper is more general than circle-covering although special case of the region-explanation problem does reduce to circle-covering, as described in Sect. 4.2 (page 116).

7   Conclusions

In this paper, we explored a variant of "geospatial abduction" (which was introduced in [26]) called *region-based geospatial abduction problems* where the user wishes to identify a set of regions that best explain a given set of observations. This has several important applications including criminology [24], marketing [11], natural science [23], and the military [28]. We explored properties and the complexity of several variants of this problem, including variants where the space is induced by a distance from the observations, as well as when the regions are irregular shapes (including non-convex). As most of the problems were NP-hard, we illustrated a variety of approximation techniques, often with guarantees, to address these problems. We also implemented some of our algorithms and evaluated with a real-world counterinsurgency [14] data-set to find weapons cache sites based on attack data in Baghdad, Iraq and produced regions that had an average density of over 8 caches per square kilometer, significantly higher than the city wide density of 0.4.

There are many interesting open questions relating to this type of abduction problem. Future work may include studies of the counting version of the problem,

where we may consider all possible solutions to a given region explanation problem according to a probability distribution and determine the "most probable" regions. Another aspect to consider would be time – perhaps in some applications the locations of the partners are in a certain region only at a certain time.

From a more practical standpoint, there are other avenues for future work – particularly when dealing with the IED cache location problem. Some of the more obvious extensions in this direction would be to study the use of different distance functions (as opposed to Euclidean distance used here) and/or methods to generate regions (as opposed to REGION-GEN). For example, perhaps considering elevation and/or terrain in the distance function may provide a more realistic picture of the α, β bounds. Alternatively, perhaps we could consider terrain and socio-cultural variables that restrict partner placement (as depicted in Fig. 3) when we generate our regions – which would necessitate an extension or alternative to REGION-GEN. Both of these practical considerations may greatly improve our described counter-IED applications and may allow for RGAPsto significantly aide military operations.

Another direction in future work – that would both extend the study of geospatial abduction and improve counter-IED applications – would be to consider the case where the observations were caused by more than one agent. For example, in our experiments, we considered attacks and caches from Iranian-sponsored militants in Iraq. We implicitly assumed that these groups conducting attacks would operate in a similar manner and would share areas used for caches. As the results of our experiments were generally encouraging, this was most likely a valid assumption. However, suppose we have a set of attacks that could come from a variety of groups, which may not all operate in a similar manner and may not cooperate with each other. In such a case, it may not be appropriate to apply the algorithms of this paper as-is. There are two approaches to this variant: (1) cluster the attacks beforehand and solve an RGAPfor each cluster or (2) extend RGAPsto a the probabilistic and/or multi-agent case. Both raise interesting technical and practical issues. In short, we believe that region-based geospatial abduction offers a wide space of research topics to explore that would be interesting to study as well as relevant to security operations.

Appendix

Proofs

Proof of Lemma 1

Given observations $\mathcal{O}$ and the set of regions $R_{\mathcal{O}}$, then a region $r \in R_{\mathcal{O}}$ sub-explains an observation $o \in \mathcal{O}$ iff is super-explain o.

Proof. CLAIM 1: Any point in a region $r \in R_\mathcal{O}$ is either within distance $[\alpha, \beta]$ or outside the distance $[\alpha, \beta]$ from each $o \in \mathcal{O}$.
As $R_\mathcal{O}$ is created by drawing circles of radii α, β around each observation, the statement follows by the definition of $R_\mathcal{O}$.

CLAIM 2: ($\Leftarrow$) There is no $r \in R_\mathcal{O}$ that super-explains some $o \in \mathcal{O}$ but does not sub-explain the observation.
Suppose, BWOC, there is some $r \in R_\mathcal{O}$ that super-explains some $o \in \mathcal{O}$ but does not sub-explain it. Then, there must be at least one point in r that can be partnered with $\mathcal{O}$ and at least one point in r that cannot be partnered with o. However, by claim 1, this is not possible, hence a contradiction.
CLAIM 3: ($\Rightarrow$) There is no $r \in R_\mathcal{O}$ that sub-explains some $o \in \mathcal{O}$ but does not super-explain the observation.
Follows directly from Observation 2.1.

Proof of Theorem 1

I-REP $\leq_p$ AC-Sup-REP.
AC-Sup-REP $\leq_p$ Sup-REP.

Proof. CLAIM 1: I-REP $\leq_p$ AC-Sup-REP.
Set up an instance of AC-Sup-REP with the input for I-REP plus the parameter $A = \pi \cdot (\beta^2 - \alpha^2)$. For direction $\Leftarrow$, note that a solution to this instance of I-REP is also a solution to AC-Sup-REP, as any region that sub-explain an observation also super-explains it for the set of region $R_\mathcal{O}$ (Lemma 1) and the fact that, by definition, all regions in the set $R_\mathcal{O}$ must have an area less than A. For direction $\Rightarrow$, we know that only regions that can be partnered with observations are considered by the area restriction, and by Lemma 1, the all regions in the solution are also super-explanations for their corresponding observation.

CLAIM 2: AC-Sup-REP $\leq_p$ Sup-REP.
Consider the set R from AC-Sup-REP and let set
$R' = \{r \in R|$ *the area of r is less than or equal to A*$\}$.
Set up an instance of Sup-REP where the set of regions is R' and the rest is the input from AC-Sup-REP. or direction $\Leftarrow$, it is obvious that any solution to AC-Sup-REP is also a solution to Sup-REP, as $R - R'$ are all regions that cannot possibly be in the solution to the instance of AC-Sup-REP. Going the other direction ($\Rightarrow$), we observe that by the definition of R', all regions in the result of the instance of Sup-REP meet all the requirements of the AC-Sup-REP problem.

Proof of Theorem 2

I-REP is NP-Complete.

Proof. CLAIM 1: I-REP is in-NP.

Given a set of regions, $R' \subseteq R_{\mathcal{O}}$ we can easily check in polynomial time that for each $o \in \mathcal{O}$ there is an $r \in R$ that is a partner for o. Simply check if each r falls within the distance $[\alpha, \beta]$ for a given $o \in \mathcal{O}$. The operation will take time $O(|\mathcal{O}| \cdot |R'|)$ - which is polynomial.

CLAIM 2: I-REP is strongly NP-hard.
We show that for an instance of the known strongly NP-complete problem, circle covering (CC), $CC \leq_p I - REP$ by the following transformation.

- Set $\mathcal{S} = \mathcal{S}'$
- Set $\mathcal{O} = P$
- Set $\beta = \beta'$
- Set $\alpha = 0$
- Set $k = k'$

This transformation obviously takes polynomial time. We prove correctness with the following two sub-claims.

CLAIM 2.1: If there is a k-sized solution R' for I-REP, then there is a corresponding k'-sized solution for CC.
Consider some $r \in R'$. Let $\mathcal{O}'$ be the subset of $\mathcal{O}$ (also of P) such that all points in $\mathcal{O}'$ are partnered with r. By definition, all points enclosed by r are of distance β or less away from each point in $\mathcal{O}'$. Hence, we can pick some point enclosed by r and we have the center of a circle that covers all elements in $\mathcal{O}'$. The statement follows.

CLAIM 2.2: If there is a k'-sized solution Q for CC, then there is a corresponding k-sized set solution for I-REP.
Consider some point $q \in Q$. Let P' be the subset of P (also of $\mathcal{O}$) such that all points in P' are of distance β' from q. As p is within β of an element of $\mathcal{O}$, it is in some region of the set $R_{\mathcal{O}}$. Hence, the region that contains p is a partner region for all elements of P'. The statement follows.

Proof of Corollary 1

I-REP-MC cannot be approximated by a fully polynomial-time approximation scheme (FPTAS) unless $P == NP$.

Proof. Follows directly from [18] and Theorem 2.

Proof of Corollary 2

1. Sub-REP and Sup-REP are NP-Complete.
2. Sub-REP-MC, Sup-REP-MC, I-REP-MC, Sub-REP-ME, Sup-REP-ME, and I-REP-ME are NP-Hard.

3. Sub-REP-MC, Sup-REP-MC cannot be approximated by a FPTAS unless $P ==$ NP.

Proof. All follow directly from Lemma 1, Theorem 2, and Corollary 1.

Proof of Theorem 3

Sub/Sup-REP-MC $\leq_p$ Set-Cover Sub/Sup-REP-ME $\leq_p$ Max-k-Cover

Proof. CLAIM 1: Sub/Sup-REP-MC $\leq_p$ Set-Cover
Consider the instance of set-cover $\langle \mathcal{O}, \bigcup_{r \in R}\{\mathcal{O}_r\}\rangle$ obtained from **REDUCE-TO-COVERING**$(\mathcal{O}, R)$. Let $\mathcal{H}'$ be a solution to this instance of set-cover. ($\Leftarrow$) If R' is a solution to the instance of Sub/Sup-REP-MC, then the set $\bigcup_{r \in R'}\{\mathcal{O}_r\}$ is a solution to set-cover. Obviously, it must cover all elements of $\mathcal{O}$ and a smaller solution to set-cover would indicate a smaller R' – a contradiction. ($\Rightarrow$) Given set $\mathcal{H}'$, let $R'' \equiv \{r \in R | \mathcal{O}_r \in \mathcal{H}'\}$. Obviously, R'' provides a partner for all observations in $\mathcal{O}$. Further, a smaller solution to Sub/Sup-REP-MC would indicate a smaller $\mathcal{H}'$ is possible – also a contradiction.

CLAIM 2: Sub/Sup-REP-ME $\leq_p$ Max-k-Cover
Consider the instance of max-k-cover $\langle \mathcal{O}, \bigcup_{r \in R}\{\mathcal{O}_r\}, k\rangle$ obtained from **REDUCE-TO-COVERING**$(\mathcal{O}, R, k)$. Let $\mathcal{H}'$ be a solution to this instance of max-k-cover. ($\Leftarrow$) If R' is a solution to the instance of Sub/Sup-REP-ME, then the set $\bigcup_{r \in R'}\{\mathcal{O}_r\}$ is a solution to max-k-cover. Obviously, both have the same cardinality requirement. Further, if there is a solution to max-k-cover that covers more elements in $\mathcal{O}$, this would imply a set of regions that can be partnered with more observations in $\mathcal{O}$ - which would be a contradiction. ($\Rightarrow$) Given set $\mathcal{H}'$, let $R'' \equiv \{r \in R | \mathcal{O}_r \in \mathcal{H}'\}$. Obviously, R'' meets the cardinality requirement of k. Further, a solution to Sub/Sup-REP-ME that allows more observations in $\mathcal{O}$ to be partnered with a region would indicate a more optimal solution to max-k-cover – a contradiction.

Proof of Proposition 1

REDUCE-TO-COVERING requires $O(|\mathcal{O}| \cdot |R|)$ time.

Proof. Follows directly from Line 1.

Proof of Proposition 2

GREEDY-REP-ME requires $O(k \cdot |R| \cdot f)$ time and returns a solution whose where the number of observations in $\mathcal{O}$ that have a partner region in R' is within a factor $\left(\frac{e}{e-1}\right)$ of optimal.

Proof. The complexity proof mirrors that of Proposition 3 while the approximation guarantee follows directly from the results of [19].

Proof of Proposition 3

GREEDY-REP-ME requires $O(|\mathscr{O}| \cdot |R| \cdot f)$ time and returns a solution whose cardinality is within a factor of $1 + \ln(f)$ of optimal.

Proof. The outer loop of the algorithm iterates no more than $|\mathscr{O}|$ times, while the inner loop iterates no more than $|R|$ times. The time to compare the number of elements in a set $\mathscr{O}_r$ is $O(f)$.

The approximation factor of $1 + \ln(f)$ follows directly from [20].

Proof of Proposition 4

GREEDY-REP-MC2 runs in $O(\Delta \cdot f^2 \cdot |\mathscr{O}| + |\mathscr{O}| \cdot \ln(|\mathscr{O}|))$ time and returns a solution whose cardinality is within a factor of $1 + \ln(f)$ of optimal.

Proof. CLAIM 1: GREEDY-REP-MC2 runs in $O(\Delta \cdot f^2 \cdot |\mathscr{O}| + |\mathscr{O}| \cdot \ln(|\mathscr{O}|))$ time. The pre-processing in lines 1-4 can be accomplished in $O(\Delta + \Delta \cdot f)$ as the size of each GRP_o is bound by Δ and the size of each REL_o is bound by $\Delta \cdot f$.

The outer loop of the algorithm iterates $\mathscr{O}$ times. In each loop, the selection of the minimal element (line 5a) can be accomplished in constant time by use of a Fibonacci heap [8] (i.e. storing observations in $\mathscr{O}'$ organized by the value key_o). The next lines of the inner loop (lines 5b-5c) can be accomplished in $O(\Delta)$ time. The next line, line 5d requires $O(\ln(|\mathscr{O}|))$ time per observation using a Fibonacci heap, as observations partnered with . However, we can be assured that, during the entire run of the algorithm, this operation is only performed $|\mathscr{O}|$ times (hence, we add an $|\mathscr{O}| \cdot \ln(|\mathscr{O}|)$). The final loop at line 5e occurs after the inner loop and iterates, at most f times. At each iteration, it considers, at most $f \cdot \Delta$ elements. Hence, the overall complexity is:

$$O(|\mathscr{O}| \cdot (\Delta + f^2 \cdot \Delta) + |\mathscr{O}| \cdot \ln(|\mathscr{O}|))$$

The statement of the claim follows.

CLAIM 2: GREEDY-REP-MC2 returns a solution whose cardinality is within a factor of $1 + \ln(f)$ of optimal.
The proof of this claim resembles the approximation proof of the standard greedy algorithm for set-cover (i.e. see [4] page 1036).

Let $r_1, \ldots, r_i, \ldots, r_n$ be the elements of R', the solution to **GREEDY-REP-MC2**, numbered by the order in which they were selected. For each iteration (of the outer loop), let set COV_i be the subset of observations that are partnered for the first time with region r_i. Note that each element of $\mathscr{O}$ is in exactly one COV_i. For each $o_j \in \mathscr{O}$, we define $cost_j$ to be $\frac{1}{|COV_i|}$ where $o_j \in COV_i$. Let R^* be an optimal solution to the instance of Sub/Sup-REP-MC.

CLAIM 2.1: $\sum_{r_i \in R^*} \sum_{o_j \in \mathscr{O}_{r_i}} cost_j \geq |R|$

By the definition of $cost_j$, exactly one unit of cost is assigned every time a region is picked for the solution R. Hence,

$$COST(R) = |R| = \sum_{o_j \in \mathscr{O}} cost_j$$

The statement of the claim follows.

CLAIM 2.2: For some region $r \in R$, $\sum_{o_j \in \mathscr{O}_r} cost_j \leq 1 + \ln(f)$.

Let P be the subset of $\mathscr{O}$ that can be partners with p. At each iteration i of the algorithm, let $uncov_i$ be the number elements in P that do not have a partner. Let *last* be the smallest number such that $uncov_{last} = 0$. Let $R_P = \{r_i \in R | (i \leq last) \wedge (COV_i \cap P \not\equiv \emptyset)\}$. From here on, we shall renumber each element in R_P as $r_1, \ldots, r_{|R_P|}$ by the order they are picked in the algorithm (i.e. if an element is picked that cannot partner with anything in P, we ignore it and continue numbering with the next available number, we will COV_i and the iterations of the algorithm as well, but do not re-define the set based on the new numbering).

We note that for each iteration i, the number of items in P that are partnered is equal to $uncov_{i-1} - uncov_i$. Hence,

$$\sum_{o_j \in \mathscr{O}_r} cost_j = \sum_{i=1}^{last} \frac{uncov_{i-1} - uncov_i}{|COV_i|}$$

At each iteration of the algorithm, let $PCOV_i$ be the subset of observations that are covered for the first time if region p is picked instead of region r_i. We note, that for all iterations in $1, \ldots, last$, the region p is considered by the algorithm as one of its options for greedy selection. Therefore, as p is not chosen, we know that $|COV_i| \leq |PCOV_i|$. Also, by the definition of $ucov_i$, we know that $|PCOV_i| = ucov_{i-1}$. This gives us:

$$\sum_{o_j \in \mathscr{O}_r} cost_j \leq \sum_{i=1}^{last} \frac{uncov_{i-1} - uncov_i}{ucov_{i-1}}$$

Using the algebraic manipulations of [4] (page 1037), we get the following:

$$\sum_{o_j \in \mathscr{O}_r} cost_j \leq H_{|P|}$$

Where H_j is the jth harmonic number. By definition of the symbol f (maximum number of observations supported by a single partner), we obtain the statement of the claim.

(Proof of Claim 2): Combining claims 1-2, we get $|R| \leq \sum_{r_i \in R^*} (1 + \ln(f))$, which gives us the statement.

Proof of Proposition 3

I-REP-MC-Z $\leq_p$ CC

Proof. Follows directly from Theorem 2.

Proof of Proposition 5

The algorithm, FIND-REGION runs $O(|\mathscr{O}|)$ time, and region r (associated with the returned set $\mathscr{O}_r$) encloses p.

Proof. PART 1: FIND-REGION consists of a single loop that iterates $|\mathscr{O}|$ times.

PART 2: Suppose, the region enclosing point p has a different label. Then, there is either a bit in *label* that is incorrectly set to 1 or 0. As only observations which are $\leq$ from β have the associated bit position set to 1, then all 1 bits are correct. As we exhaustively consider all observations, the 0 bits are correct. Hence, we have a contradiction.

Proof of Proposition 4

An α-approximation algorithm for CC is an α-approximation for KREP.

Proof. Follows directly from Theorem 2.

Proof of Proposition 7

REGION-GEN has a time complexity $\Theta(|\mathscr{O}| \cdot f \cdot \frac{\pi \cdot \beta^2}{g^2})$.

Proof. For any given observation, the number of points in the grid that can be in a partnered region is $\frac{\pi \cdot \beta^2 - \alpha^2}{g^2}$. Hence, the first loop of the algorithm and the size of L

are both bounded by $|\mathcal{O}| \cdot \frac{\pi \cdot \beta^2}{g^2}$. We note that the lookup and insert operations for the hash table T do not affect the average-case complexity - we assume these operations take constant time based on [4], hence the statement follows.

References

1. Brantingham, P., Brantingham, P.: Crime Pattern Theory. In: Wortley, R., Mazerolle, L. (eds.) Enviromental Criminology and Crime Analysis, pp. 78–93. Willan Publishing, UK (2008)
2. Brönnimann, H., Goodrich, M.T.: Almost optimal set covers in finite vc-dimension. Discrete Comput. Geom. **14**, 293–302 (1995)
3. Bylander, T., Allemang, D., Tanner, M.C., Josephson, J.R.: The Computational Complexity of Abduction. Artif. Intell. **49**(1-3), 25–60 (1991)
4. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to Algorithms, 2nd edn. MIT, Cambridge (2001). http://mitpress.mit.edu/catalog/item/default.asp?tid=8570&ttype=2
5. Eiter, T., Gottlob, G.: The complexity of logic-based abduction. J. ACM **42**(1), 3–42 (1995)
6. Feige, U.: A threshold of ln n for approximating set cover. J. ACM **45**(4), 634–652 (1998)
7. Franceschetti, M., Cook, M., Bruck, J.: A geometric theorem for network design. IEEE Trans. Comput. **53**(4), 483–489 (2004)
8. Fredman, M.L., Tarjan, R.E.: Fibonacci heaps and their uses in improved network optimization algorithms. J. ACM **34**(3), 596–615 (1987). http://dx.doi.org/10.1145/28869.28874
9. Freedman, D., Purves, R., Pisani, R.: Statistics, 4 edn. W.W. Norton and Co., New York (2007)
10. Fu, B., Chen, Z., Abdelguerfi, M.: An almost linear time 2.8334-approximation algorithm for the disc covering problem. In: AAIM '07: Proceedings of the 3rd international conference on Algorithmic Aspects in Information and Management, pp. 317–326. Springer, Berlin (2007)
11. Gijsbrechts, E., Campo, K., Goossens, T.: The impact of store flyers on store traffic and store sales: a geo-marketing approach. J. Retailing **79**(1), 1–16 (2003). http://www.sciencedirect.com/science/article/B6W5D-4893MB0-1/2/db6c17461f5e9d38b3856a6b742e5a7a
12. Gonzalez, T.F.: Covering a set of points in multidimensional space. Inf. Process. Lett. **40**(4), 181–188 (1991)
13. Hochbaum, D.S., Maass, W.: Approximation schemes for covering and packing problems in image processing and vlsi. J. ACM **32**, 130–136 (1985)
14. ISW: Map of Special Groups Activity in Iraq, Institute for the Study of War (2008)
15. doLagoPereira, S., deBarros, L.N.: Planning with abduction: A logical framework to explore extensions to classical planning. In: Lecture Notes in Computer Science Advances in Artificial Intelligence – SBIA (2004)
16. Liao, C., Hu, S.: Polynomial time approximation schemes for minimum disk cover problems. J. Comb. Optim. http://dx.doi.org/10.1007/s10878-009-9216-y. **20**(4), 399–412 (2009)
17. MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: Cam, L.M.L., Neyman, J. (eds.) Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability. vol. 1, pp. 281–297. University of California Press, Berkeley, CA (1967)
18. Megiddo, N., Supowit, K.J.: On the complexity of some common geometric location problems. SIAM J. Comput. **13**(1), 182–196 (1984)
19. Nemhauser, G.L., Wolsey, L.A., Fisher, M: An analysis of approximations for maximizing submodular set functions. Mathematical Programming **14**(1), 265–294 (1978)
20. Paschos, V.T.: A survey of approximately optimal solutions to some covering and packing problems. ACM Comput. Surv. **29**(2), 171–209 (1997)
21. Peirce, C.S.: In: Buchler, J. (ed.) Philosophical Writings of Peirce. Dover Publications, New York (1955)

22. Peng, Y., Reggia, J.A.: Abductive Inference Models for Diagnostic Problem-Solving. Springer, New York (1990)
23. Rich, T.H., Fenton, M.A., Fenton, C.L.: The fossil book: A record of prehistoric life, 2nd edn. Dover Publications, New York (1996)
24. Rossmo, D.K., Rombouts, S.: Geographic Profiling. In: Wortley, R., Mazerolle, L. (eds.) Enviromental Criminology and Crime Analysis, pp. 136–149. Willan, Portland, OR (2008)
25. Shakarian, P., Subrahmanian, V., Spaino, M.L.: SCARE: A Case Study with Baghdad. In: Proceedings of the Third International Conference on Computational Cultural Dynamics. AAAI (2009)
26. Shakarian, P., Subrahmanian, V., Spaino, M.L.: Gaps: Geospatial abduction problems. ACM Transactions on Intelligent Systems and Technology. (2011)
27. Stollsteimer, J.F.: A working model for plant numbers and locations. J. Farm. Econom. **45**(3), 631–645 (1963)
28. US Army: Intelligence Preparation of the Battlefield (US Army Field Manual), FM 34-130 edn. (1994)

Finding Hidden Links in Terrorist Networks by Checking Indirect Links of Different Sub-Networks

Alan Chen, Shang Gao, Panagiotis Karampelas, Reda Alhajj, and Jon Rokne

Abstract Modeling and analyzing criminal and terrorists networks is a challenging problem that has attracted considerable attention in the academia, industry and government institutions, especially intelligence services. Criminals try to keep their communications and interactions uncovered as much as possible in order not to be discovered and resolved. Their success is our society failure and vice versa. Hence, it is essential to thoroughly study such networks to investigate their details. However, incompleteness of criminal networks is one of the main problems facing investigators, due to missing links in the network; and social network methods could be effectively used to analyze and hopefully prevent, avoid or stop criminal activities. Social network analysis can be applied to criminal networks in order to elaborate on good strategies to prosecute or prevent criminal activities. Having all this in mind, our research provides a method to identify hidden links between nodes in a network using the current information available to investigators. The method presented generates networks that represent all the possible hidden links, and the links of these generated networks represent the number of times the two entities are indirectly connected in each relationship type. The method was tested on multiple

A. Chen · S. Gao · J. Rokne
Computer Science Department, University of Calgary, Calgary, AB, Canada

P. Karampelas
Department of Information Technology, Hellenic American University, Manchester, NH, USA
and
Hellenic Air Force Academy, Athens, Greece
e-mail: pkarampelas@gmail.com

R. Alhajj (✉)
Computer Science Department, University of Calgary, Calgary, AB, Canada
and
Department of Computer Science, Global University, Beirut, Lebanon
and
Department of Information Technology, Hellenic American University, Manchester, NH, USA
e-mail: alhajj@ucalgary.ca

small terrorism data sets and the results demonstrate that the proposed method is capable of finding interesting hidden links. This is a valuable technique in criminal network analysis, because it can assist investigators in finding hidden links in the network and reduce the amount of missing data.

1 Introduction

Social network analysis is an emerging field originated from sociology and anthropology; it is heavily based on graph theory, statistics, mathematics, and more recently computer science techniques which much influenced the scalability of modeling and analysis; the main theme is to model and investigate connections between social entities [14]. For instance, in order to find out who plays the most central role in an organization, we can construct a social network with individuals being social entities and the link between two entities representing certain social connections, like working on the same project. Once the social network has been constructed, it could be analyzed for knowledge discovery, such as finding the most influential individuals, the most homogeneous group, etc. Social networks are used in different problem domains and for various circumstances as the underlying abstract model. For example, the work described in [2, 17] constructed social networks to analyze the organizational structure of terrorists, and the work described in [7] discussed a popular problem of predicting links in social networks. The power of social network modeling is the interpretation of relationships or connections between entities, and the ease of visualization with graphical models.

Social network analysis can be applied to criminal[1] networks in order to elaborate on good strategies to prosecute or prevent criminal activities [1, 5, 6, 10–12]. For social network analysis to be more effective for criminal networks, where there are likely missing links, the amount of missing links needs to be reduced. The goal of this research is to provide a method to help identify hidden links between nodes in a network with the current information available to investigators.

1.1 Problem

Sparrow [15] identified incompleteness as one of the three main problems for using social network analysis to analyze criminal activities. Incompleteness is having missing nodes and links in the network, because the investigators are not able to discover all the nodes and links. This is due to criminals attempting to hide the ties

---

[1]We will interchangeably use criminal and terrorist networks, though the two are different and have different targets, but the same analysis applies to both networks.

to each other, in order to minimize any compromises to the network. An example of this is the network created by Krebs of the nineteen dead hijackers of the events during September 11th, 2001 [5]. The network was very sparse, and members on the same team were not directly linked to all other members of the same team. The works described in [2, 9] deal with communication networks, and do not take into consideration other types of data to be used for identifying hidden groups. Another piece of work described in [13] provides another solution to infer missing links, which uses a sampling technique on the network along with Bayes' theorem. Although similar to our solution, the difference is that our solution breaks down the network into many different sub-networks based on relationships, rather than take samples of the network; also, their approach predicts links based on Bayes' theorem, whereas our approach predicts links based on the number of times are indirectly connected in the different sub-networks. In terrorist networks, hidden groups or structures can be mined [8, 16] and social network measures can be used to analyze the structure [3, 4]. In other words, criminals and terrorists try to keep as much minimal information as possible in order to avoid being traced and dissolved.

1.2 Solution

Although criminals will try to minimize information about the criminal organization to avoid detection, our solution uses the information recorded about the individual, such as where they work, where they have studied, who their friends are, who their family members are, what events they participated in, etc. The method will find the number of times two individuals are indirectly connected in the different relationships. For example, if there is no connection between person A, and person B, but they have a chain of friends in common, attended the same schools, been to the same cities, are members of the same club, etc. This cannot be considered coincidental that entities A and B are not linked together if the number of indirect links is very high. This would mean that it is likely there's a hidden link between A and B, because each indirect connection implies there's a possibility of a hidden link, and a large number will indicate a higher likelihood. It is like each indirect connection is flipping a coin to determine whether there is a hidden link. The more indirect connections, the more coin flips to determine if there is a hidden link, thus this is why there's a higher likelihood of a hidden link if there's more indirect connections.

In this paper, the goal is to provide a method to help identify hidden links between nodes in a network with the current information available to investigators. The method presented generates networks that represent all the possible hidden links, and the links of these generated networks represent the number of times the two entities are indirectly connected in each relationship type. The test results reported in this paper are encouraging; they demonstrate the effectiveness and applicability of the proposed approach.

The rest of the paper is structured as follows. In Sect. 2, we introduce the novel graph-based algorithm for decomposing the social networks by constructing conflict vertex sets with a running example; following the main algorithm, we discuss the re-construction of terrorist networks and the issue of finding hidden links by different relationship types in Sect. 3; we present the experimental results on different terrorist networks using the data from [4] and demonstrate the mined results in Sect. 4. Our paper is concluded in Sect. 5 with the summary of the proposed approach and some future research directions.

2 Finding Non-contradictory Vertex Set

Given a network, represented with a graph $G = (V, E)$, where V is the vertex set and E is the set of edges connecting pairs of vertices. In social network, vertices are referred as *nodes*, *actors*, or *individuals* and edges are termed *links*, *ties*, or *relationships* between nodes. The method presented in this paper makes use of a network partitioning algorithm we researched and developed to partition the vertex set of a graph into non-contradictory sets. The network partitioning algorithm will return results similar to graph coloring techniques, we adapt the algorithm in mining the terrorist networks whose network structure is typically compact. The specialty of the terrorist network prompts the need for a graph traversal method to effectively find non-conflict node sets.

2.1 Graph Partitioning

We demonstrate the algorithm with a running example in this paper. Suppose the graphical structure of a network is as shown in Fig. 1; for partitioning the network of entities, start at the node E, because this node has the lowest degree, a degree of one. It is also possible to start at node D to end up with the same hidden sets, because D also has a degree of one. The algorithm needs to start at the node with the lowest degree, because it is a recursive algorithm that calls itself on the subgraphs. Starting at the node with the lowest degree, allows for omission of the nodes with the least links in the subgraphs, and makes sure that all possible sets are generated.

The node E does not have a set number assigned to it, so assign it the first available set number, which is one. Then assign all of its neighbors the same set number is not available, which in this case is the assignment of the set number one is not available to its neighbor C. The assignment of the set number being unavailable to its neighbors will make sure that no connected nodes will appear in the same set. After, assign all of its non-neighbors the same set number one. Assigning the same set number to all of the non-neighbors will make sure that every possible set
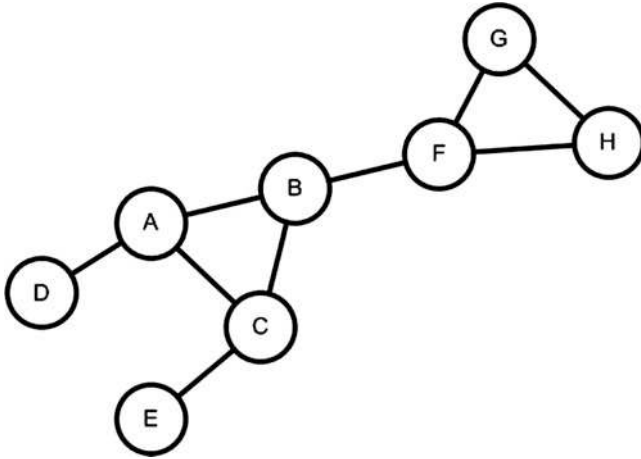
Fig. 1 Partitioning network example

Algorithm 2 Graph partitioning to find non-contradictory vertex set
___

INPUT: Graph $G = (V, E)$
OUTPUT: Subsets of V that are non-contradictory
for each component of G **do**
   $N=\{$any node with the lowest degree$\}$
   for each node in N **do**
      if node is not assigned a set number (denoted as ℵ) **then**
         assign node to the next available ℵ
         mark neighbors the same ℵ is n/a
         assign non-neighbors the same ℵ
      else
         for each ℵ assigned to the nodes **do**
            mark all neighbors the same ℵ is n/a
         end for
      end if
      add all the nodes's neighbors to N
   end for
   partition into vertex sets based on the max ℵ
   for each node in N **do**
      for every ℵ assigned **do**
         if ℵ is not marked n/a **then**
            add the node to the set indicated by ℵ
         end if
      end for
   end for
   for each ℵ **do**
      create a subset
      apply the algorithm to subgraphs of all nodes with the same ℵ and nodes marked n/a
   end for
end for

is generated, because this will add the non-neighbor nodes to the set, or make sure the non-neighbor nodes are considered in the subgraphs for other iterations.

Next, move onto a neighboring node C, and since it also does not have a set number assigned to it, assign it the first available set number, which is two, as the set number one was assigned as not available. Then assign its neighbors E, A, and B, the set number two is not available. After, assign all of its non-neighbors the same set number two.

Next, move onto a neighboring node A, but it is also possible to move onto the node B to end up with the same hidden sets. Because it already has a set number assigned to it, assign its neighbors D, B, and C, the set number one is not available, because this node is assigned the set number one. This assignment of the assigned set number is not available to the node's neighbors in this case will make sure that all the possible subsets where there are no connected nodes are consider.

Next, move onto a neighboring node D, and since it already has a set number assigned to it, assign its neighbor A, the set number one is not available, because this node is assigned the set number one.

Now that there are no neighbors for D to move onto, we will move onto a previous neighboring node that was skipped, which was B. B already has a set number assigned to it, assign its neighbors A, C, and F, the set number one is not available, because this node is assigned the set number one.

Next, move onto a neighboring node F, and since it already has a set number assigned to it, assign its neighbors G, B, and H, the set number one and two is not available, because this node is assigned the set number one and two.

Next, move onto a neighboring node G, and since it already has a set number assigned to it, assign its neighbors F, and H, the set numbers one and two are not available, because this node is assigned the set number one and two.

Next, move onto a neighboring node H, and since it already has a set number assigned to it, assign its neighbors F, and G, the set numbers one and two are not available, because this node is assigned the set number one and two.

Now that all the nodes have been traversed, the first iteration is finished, and we will start building the sets (see Fig. 2). In the figure, the negative sign is used to denote that a set number is assigned as not available. Build sets for each set number, and assign the entities with nodes that are only assigned to a set number to that set number. The entity E is assigned to set one, and the entities C and D are assigned to set two (see Fig. 2). The other entities also have nodes that are assigned set numbers, but in addition, they are also assigned the same set number is not available. They will not be assigned to a set at this moment, because they will be assigned to a subset in the following iterations of the algorithm on subgraphs.

Next, consider the subsets for set one by using the same procedure on the subgraph of nodes that were assigned set number one, and assigned set number one is not available (see Fig. 3). In the case of the second iteration, all the nodes except for E and C are omitted in the subgraph, because we know from the set assignment, whether or not E and C will appear in all the subsets of set one. The sets generated from the subgraph will be subsets of the original set one (see Fig. 3). This recursion
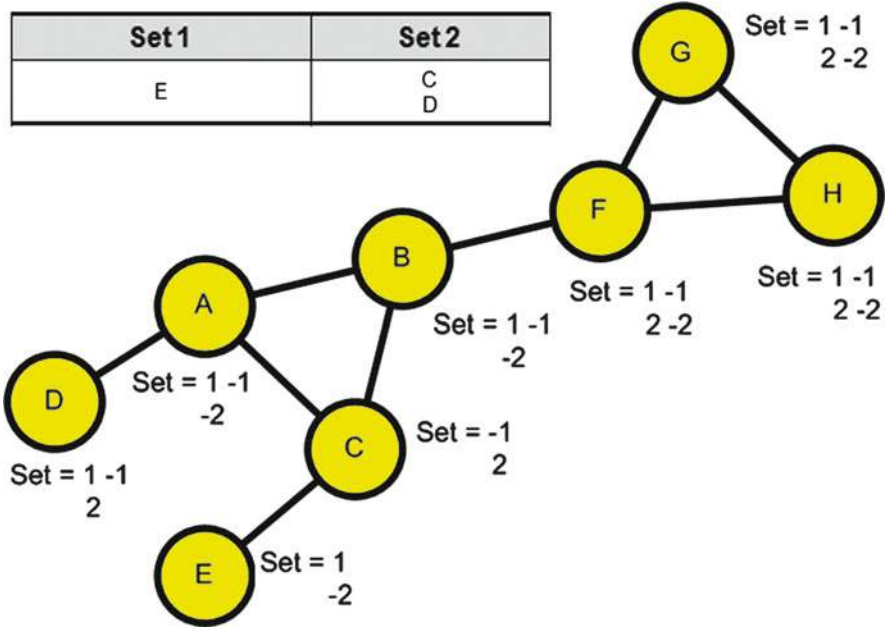
Fig. 2 First iteration of the partitioning network example
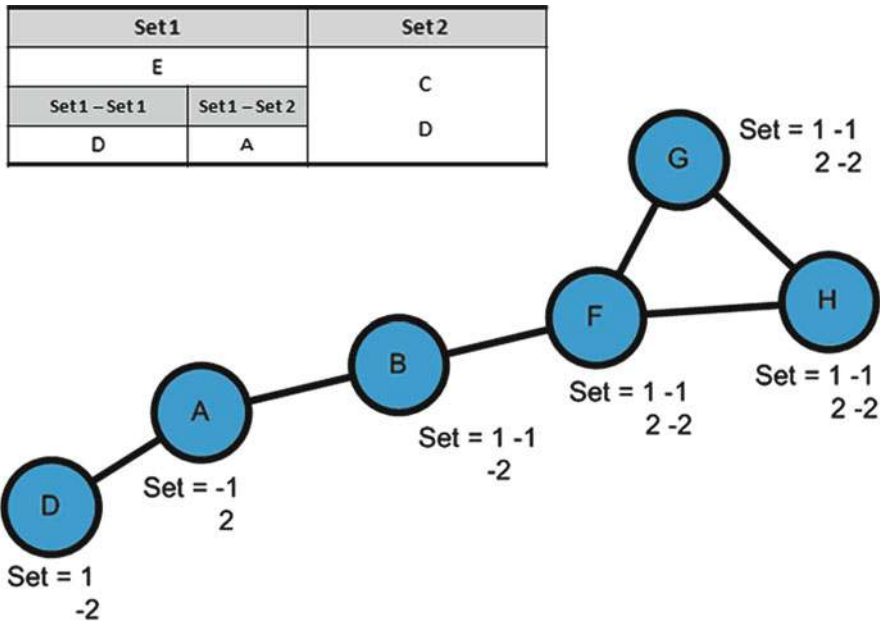
Fig. 3 Second iteration of partitioning network example

Set1						Set2		
E						C D		
Set1 – Set1			Set1 – Set1			Set2 – Set1	Set2 – Set2	Set2 – Set3
D			A					
Set1 – Set1 – Set1		Set1 – Set1 – Set2	Set1 – Set2 – Set1	Set1 – Set2 – Set2	Set1 – Set2 – Set3	G	F	H
B		F						
Set1 – Set1 – Set1 – Set1	Set1 – Set1 – Set1 – Set2							
G	H		G	F	H			

Fig. 4 Hidden set hierarchy

will continue until the subgraph is empty, and then the process moves to the other sets to begin the same procedure.

At the end of the process, a set hierarchy is generated, as shown in Fig. 4. From this, we can see that there are nine possible hidden sets. They can be built up from the bottom up from Fig. 4:

1. G, B, D, E
2. H, B, D, E
3. F, D, E
4. G, A, E
5. F, A, E
6. H, A, E
7. G, D, C
8. F, D, C
9. H, D, C

The sets have no links between any entities in the set in each individual set, and any other possible entity sets with no links will be a subset of one of the nine sets. This makes the hidden sets the maximal. The maximal hidden sets are important, because they hold the most entities with no links. This means there is no possible set of entities without links to each other that is not covered by one of the hidden sets, so all the indirect connections are explored.

3 Reconstruction of Terrorist Network

3.1 General Idea

In general, the main idea is to generate networks that represent the possible hidden links, and have the links between the nodes represent the number of times the two entities are indirectly connected in each relationship type. The relationship type of

networks can be anything that links two entities together, such as friend of, family of, coworker of, been to, studied at, eats at, participated in, plays in, and member of. The higher the weights, the more times the nodes are indirectly connected; thus the more likely there is a hidden link between them. The idea is to start off by building a main network of all the entities, connected based on whether there is any type of relationship between the entities. The entities are the nodes, and they can be anything, such as people, places, companies, accounts, etc. The relationships are the links, and they can be any type of association, such as friend of, been to, works for, deposited in, etc. In addition, subnetworks will be build of all the entities, connected based on the same relationship types. Afterwards, we can use the network partitioning algorithm on the main network to find all the possible hidden groups in the network. These groups can be represented as weighed networks, and the weights can be updated based on the number of indirect links for each relationship type. At the end, the investigators can create the hidden links based on different criteria, such as having weights above a certain threshold, or being in the top portion of the highest weighted links. Sociological studies would be needed and if well integrated into the outcome may lead to determine what values and criteria would be best used here.

3.2 Main Network Creation

We may create the main social network by making each entity in E a node and linking two nodes demarcated by the relationship type in the data. For example, in a terrorist network, the types of relationship can be sibling, friends, correspondence, etc.

The process is illustrated in Fig. 5. Assume we have the data presented in Fig. 5. Make each entity a node. Afterwards, start making the links by processing each relationship. There are relationships between A and B, A and C, as well as A and D, so there are links with A and all the other nodes (see Fig. 5). The relationships
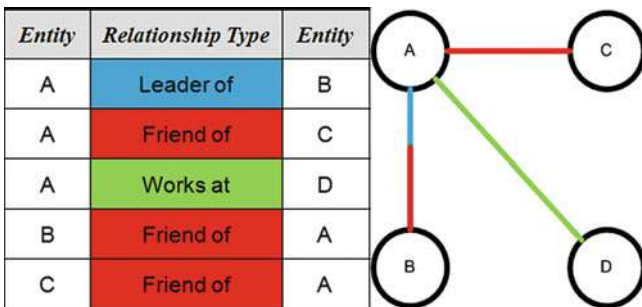
Fig. 5 Main network creation example

between B and A, as well as C and A, also indicate there are links, but the links are already formed by the previous relationships processed.

3.3 Sub-network Creation

To create sub-networks based on the main network from Sect. 3.2, we divide the network into separate components, as described in Algorithm 3.

Algorithm 3 Sub-network Creation

INPUT: Main network from Sect. 3.2
OUTPUT: Subnetwork by relationship types
for each relationship type in data **do**
    make each entity a node
    for each relationship **do**
        if the relationship is the same as the relationship type **then**
            link the two entities together
        end if
    end for
end for

To exemplify the process, assume we have the data presented in Fig. 5. For relationship 1, there is a relation of type 1 between A and B, so there is a link between A and B (see Fig. 6). The other relationships are not of type 1, so there are no more links to be formed for this sub-network. For relationship 2, there is a relation of type 2 between B and A, as well as A and C, so there is a link between A and B, as well as A and C. The relationship between C and A, also indicates that there is a link, but that link is already formed by a previous relationship already processed. The other relationships are not of type 2, so there are no more links to be formed for this sub-network. For relationship 3, there is a relation of type 3 between
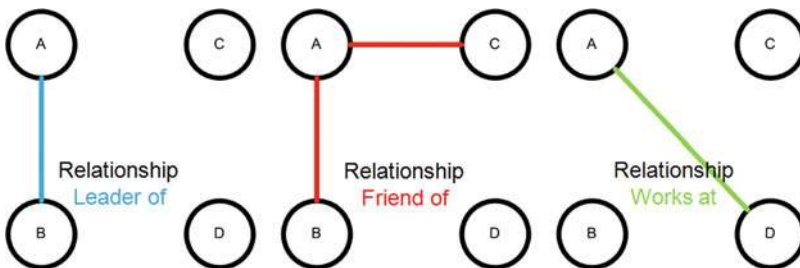
Fig. 6 Sub-networks creation example

Fig. 7 Hidden networks
creation example
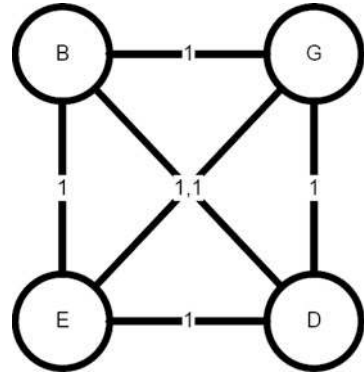
A and *D*, so there is a link between *A* and *D*. The other relationships are not of
type 3, so there are no more links to be formed for this sub-network.

3.4 Hidden Networks Creation

To unravel hidden networks, we employ Algorithm 4.

Algorithm 4 Find Hidden Networks

INPUT: Non-contradictory vertex set
OUTPUT: Hidden networks
for each non-contradictory vertex set **do**
    apply Algorithm 2
    make each entity a node
    create a link between every node of weight 1
end for

Assume we obtain the non-contradictory vertex set $\{B, G, E, D\}$ from the
application of Algorithm 2. Make each entity a node (see Fig. 7). Afterwards, create
every possible link between each node of weight 1. The conversion of the set to a
network representation needs to be applied to each hidden set.

3.5 Compute Hidden Networks Weights

In order to differentiate different relationship types in terrorist networks, we use
Algorithm 5 to weight the hidden links.

The variable δ can be any number. It can be 1 to count the number of times two
nodes are indirectly connected in each network of different relationship types. The

---

Algorithm 5 Compute Hidden Weights

INPUT: Non-contradictory vertex set
OUTPUT: Hidden networks with weights
for each non-contradictory vertex set **do**
   for each hidden network from Algorithm 4 **do**
      for all links in the hidden network **do**
         if the two nodes linked together are in the hidden set, and the link weight has not been
         increased during this run through each sub network **then**
            Increase the weight of the link by δ
         end if
      end for
   end for
end for

Fig. 8 Compute hidden
networks weights example
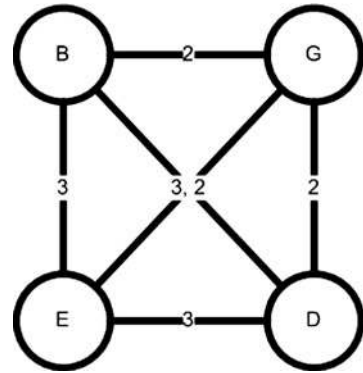
variable x can take positive values to check whether different relationship types have more significance. For example, subnetworks for the relationship type "met" will have $\delta = 1$, whereas subnetworks of the relationship type "friend of" will have $\delta = 2$. This is to signify that indirect connections through being a "friend of" are more likely to have a hidden link, than indirect connections through having "met" the different entities. In addition, the variable δ can also have negative values to take into account if certain relationship types represent that links are less likely between indirectly connected entities. Different values of δ will need sociological studies and may be expert involvement to determine the values for the different relationship types.

Assume we have the hidden network presented in Fig. 7; also we obtained the hidden sets $\{B, D, E\}$, and $\{B, D\}$ for relationship type 1, along with the hidden set $\{B, G, E, D\}$ for relationship type 2. We will use $\delta = 1$ to count the number of times unconnected nodes are indirectly linked. Start by increasing by 1 the weights of the links between B and D, B and E, as well as D and E, because the entities B, D, and E are part of one of the hidden sets for relationship type 1. Relationship type 1 has another hidden set with B and D; but we do not increase the weight for B and D again, because it has already been increased for this subnetwork for

Table 1 Experimental datasets

	2002	2005	London	Madrid	WTCB
Number of entities	166	14	31	67	19
Number of relationship types	36	13	11	12	14
Number of links	260	21	44	88	23

Fig. 9 Hidden group network example
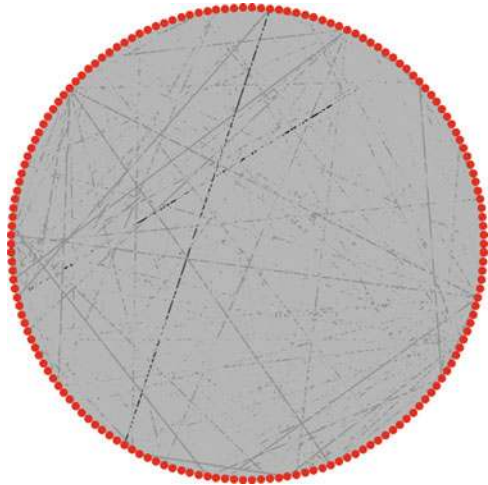
relationship type 1. Next, increase all the weights of all the links by 1, because the entities B, G, E, and D are part of the hidden set for relationship type 2. The new weights for the hidden network will be the ones presented in Fig. 8.

4 Experimental Results

The proposed approach has been applied to small terrorist data sets based on 2002, 2005, London, Madrid, and WTCB information (see Table 1) [4]. The value for δ, used in all the relationship types was set to 1. The conducted experiments used the proposed approach to measure the number of indirect links between possible hidden links. The highest weights were then noted to be the most interesting.

All components of the hidden group networks generated will always be cliques, but with different weights that give information on the likelihood of the hidden link based on the data. Figure 9 shows a generated hidden network in a circle layout with the links colored to be lighter if they are closer to 1, and darker if they are closer to the highest value based on the 2002 data set. The majority of the cells are light gray, but there are several that stand out. There is a black link of weight 7 and a dark gray link of weight 6, and they indicate interesting links, as the other link weights are

Fig. 10 Hidden group
sub-network example
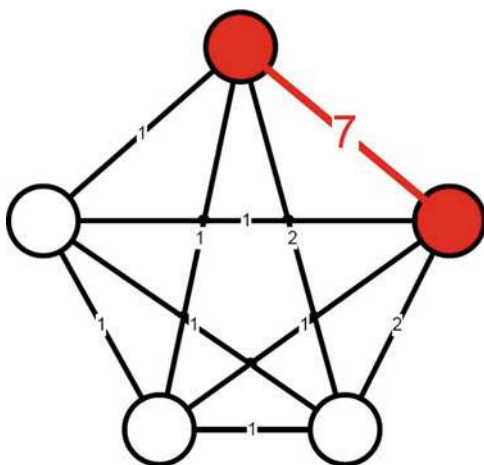
quite low in comparison. Figure 10 shows a smaller graph visualization of several
nodes in the network in Fig. 9, and this again highlights the interesting link, because
there is a large weight of 7, and the other weights are 1 and 2.

For the 2002 terrorist data set, applying the proposed approach generated hidden
networks with the majority of the link weights between 1 and 3. There was one
hidden link with a weight of 7 between Azahari Husin and Wan Min Wan Mat; thus,
this suggests that there is a high likelihood that these two entities are linked, than
other entities based on this data, because the weight of 7 was a very high weight that
occurred only once.

For the 2005 terrorist data set, applying the proposed approach generated hidden
networks with the majority of the link weights as 1 or 2. There was no hidden link
weight that was interesting, thus this suggests that there is likely no hidden links in
the data set based on this data, because all the weights were low and remained in the
same low range.

For the London terrorist data set, applying the proposed approach generated
hidden networks with the majority of the link weights between 1 and 3. There was
no hidden link weight that was interesting, thus this again suggests that there is
likely no hidden links in the data set based on this data, because all the weights
were low and remained in the same low range.

For the Madrid terrorist data set, applying the proposed approach generated
hidden networks with the majority of the link weights between 1 and 3. There was
no hidden link weight that was interesting; thus this again suggests that there is
likely no hidden links in the data set based on this data, because all the weights
were low and remained in the same low range.

For the WTCB terrorist data set, applying the proposed approach generated
hidden networks with the majority of the link weights to be 1. There were three
hidden links with a weight of 2 between Abd al-Karim Yousef and Abd al-Mun'im

Yousef, Konsonjaya and Mohamed Jamal Khalifa, as well as Mohamed Salameh and Sheikh Omar Abdul Rahman, thus this suggests that there is a higher likelihood that these three hidden links exist compared to the other links based on this data, because the weight of 2 was a higher weight that rarely occurred.

5   Summary and Conclusions

The proposed method creates networks that provide valuable information on the likelihood of hidden links between two nodes based on the current information available. The proposed method considers all the possible hidden connections, and allows investigators to compare the likelihood of hidden links against other hidden links. In addition, the proposed method allows for adjustments to the measure used based on the relationship type of the data presented, by modifying the δ value in the algorithm. The problem with the proposed method is that no information on the relationship type of the hidden link is generated. Depending on the situation, the relationship type may be a critical component. There are also situations when the relationship type can be solved easily, if there is only one logical relationship type between the entity types. Another problem is that the runtime for the proposed method can be much longer for larger sets of data with many different entities and relationship types. This can make it difficult to use large sets of data if it requires too much memory to store the network information. In addition, the proposed method addresses only part of the problem of incompleteness by identifying hidden links. The problem of missing nodes still remains, and thus there is likely also some hidden links associated with the hidden nodes, which will not be uncovered. The experiments form a simple application of the proposed method on small data sets which demonstrate that the proposed method is capable of finding interesting hidden links. This is a valuable technique in criminal network analysis, because it can help investigators find hidden links in the network, and reduce the amount of missing data. The δ value in the algorithm will allow for this method to be used with different types of data, and the weights generated give investigators a measure of the likelihood of the hidden links. This weight can be used for a variety of analysis methods, such as certainty of the existence and comparison against other hidden links.

We are currently investigating the applicability of pattern mining methods to identify what the relationship types of the hidden links are. We are also conducting research to improve the runtime of this method on real world data, and if runtime turns out to be an issue, then different methods of partitioning the entities into hidden sets by giving up accuracy or completeness for speed can be researched to reduce the memory load. In addition, we want to study the effectiveness of different method to identify hidden nodes to completely solve the incompleteness problem of using social network analysis for criminal activities.

References

1. Baker, W., Faulkner, R.: The social organization of conspiracy: Illegal networks in the heavy electrical equipment industry. Am. Sociol. Rev. **58**(6), 837–860 (1993)
2. Baumes, J., Goldberg, M., Magdon-Ismail, M., Al Wallace, W.: Discovering hidden groups in communication networks. In: Chen, H., Moore, R., Zeng, D.D., Leavitt, J. (eds.) Intelligence and Security Informatics. Lecture Notes in Computer Science, vol. 3073, pp. 378–389. Springer, Berlin/Heidelberg (2004)
3. Farely, D.: Breaking al qaeda cells: A mathematical analysis of counterterrorism operations. Stud. Confl. Terrorism **26**(6), 399–411 (2003)
4. Dawoud, K., Alhajj, R., Rokne, J.: A global measure for estimating the degree of organization of terrorist networks. In: International Conference on Advances in Social Networks Analysis and Mining, pp. 421–427. 9–11 Aug 2010
5. Krebs, V.: Mapping networks of terrorist cells. Connections **24**, 43–52 (2002)
6. Latora, V., Marchiori, M.: How the science of complex networks can help developing strategies against terrorism. Chaos Solitons Fractals **20**(1), 69–75 (2004)
7. Liben-Nowell, D., Kleinberg, J.: The link-prediction problem for social networks. J. Am. Soc. Inform. Sci. Technol. **58**(7), 1019–1031 (2007)
8. Shaikh, M., Wang, J.: Discovering hierarchical structure in terrorist networks. In: Proceedings of the International Conference on Emerging Technologies, pp. 238–244. (2006)
9. Magdon-Ismail, M., Goldberg, M., Wallace, W., Siebecker, D.: Locating hidden groups in communication networks using hidden markov models. In: Chen, H., Miranda, R., Zeng, D., Demchak, C., Schroeder, J., Madhusudan, T. (eds.) Intelligence and Security Informatics. Lecture Notes in Computer Science, vol.2665, pp.958. Springer, Berlin/Heidelberg (2010)
10. Memon, N., Wiil, U., Qureshi, A.: Design and development of an early warning system to prevent terrorist attacks. In: Proceedings of the International Conference on Artificial Intelligence and Neural Networks, pp. 222–226. (2009)
11. Klerks, P.: The network paradigm applied to criminal organizations. Connections **24**(3), 53–65 (2001)
12. Qin, J., Xu, J., Hu, D., Sageman, M., Chen, H.: Analyzing terrorist networks: A case study of the global salafi jihad network. pp. 287–304 (2005)
13. Rhodes, C.J., Jones, P.: Inferring missing links in partially observed social networks. J. Oper. Res. Soc. **60**(10), 1373–1383 (2009)
14. Strogatz, S.: Exploring complex networks. Nature **6825**(410), 268–276 (2002)
15. Sparrow, M.: The application of network analysis to criminal intelligence: An assessment of the prospects. Soc. Networks **13**(3), 251–274 (1991)
16. Tsvetovat, M., Carley, K.: Structural knowledge and success of anti-terrorist activity: The downside of structural equivalence. J. Soc. Struct. **6**(2) (2005)
17. Xu, J., Chen, H.: Crimenet explorer: A framework for criminal network knowledge discovery. CM Trans. Inform. Syst. **23**(2), 201–226 (2005)

The Use of Open Source Intelligence
in the Construction of Covert Social Networks

Christopher J. Rhodes

Abstract Open source intelligence is playing an increasing role in helping agencies responsible for national security to determine the characteristics, motivations and intentions of adversary groups that threaten the stability of civil society. Analytic methods that are able to assimilate and process the emergent data from such rich sources in a timely fashion are required in order that predictive insights can be made by intelligence specialists. The methods of social network analysis (SNA) have proved particularly useful in organising and representing covert network organisations however, it is a particularly data-hungry technique. Here, we present recent work on a statistical inference method that seeks to maximise the insight that can be gained into the structure of covert social networks from the limited and fragmentary data gathered from intelligence operations or open sources. A protocol for predicting the existence of hidden "key-players" covert in social networks is given.

1 Introduction

In recent years, social network analysis (SNA) has begun to play an increasingly prominent role as means of organising and representing intelligence data relating to terrorist, insurgency and organised-crime groups. Network representations are a highly efficient and concise presentation of large complex data sets and admit an immediate visual comprehension of the extent and internal organisation of many of these groups [1–3]. This is important, because network structures are often used to assist in the development of assessments of the capabilities and resilience of such groups by government agencies responsible for ensuring the safety of

C.J. Rhodes (✉)

Networks and Complexity Programme, Institute of Mathematical Sciences and Imperial College Institute for Security Science and Technology, Imperial College London, London, UK
e-mail: c.rhodes@imperial.ac.uk

civil society. Furthermore, the structural nature of these network topologies invites further, more detailed, quantitative analysis. The application of network measures, such as centrality or betweenness, or the application of algorithms to detect clusters or sub-groups within networks, for example, promises the possibility of further insight on the relative importance of different individuals or different parts of the network, and also for comparisons to be made between different networks.

In most cases, it is straightforward to construct social network topologies and perform analysis or apply algorithms from available data. However, SNA is a data-hungry science. In practice, much of the difficulty lies in the data collection process, long before any quantitative SNA is attempted. Significant time and resources have to be expended in order to gather social network data. It is necessary to collect detailed information about the characteristics of individuals within the network and the presence or absence of connections between them and also, if possible, the characteristics of the links between those individuals. This activity is manageable if the network is not too large and has a fixed structure, that is, it is not subject to frequent internal topological reconfigurations. In conventional SNA, much of the research that has been undertaken on empirical data has used data-sets assembled by direct questioning or surveying of a well-defined number of participants in recognisable stable social contexts. This methodology often leads to rich sources of data that can be analysed using diverse SNA methods.

By contrast, SNA as applied to counter-terrorism or organised-crime faces multiple difficulties. The most significant difficulty is the covert nature of many of the groups involved in these activities. Participants wish to evade arrest or detention and therefore take active steps to conceal their membership of or involvement in such groups and their operations. This means that information gathered during the intelligence collection phase may well be patchy or of limited usefulness, and sometimes contradictory. An additional difficulty is that of knowing who to include in a representative list of members of the network. Knowing where to draw the "boundary" of the network is not a well-defined process. Also, these groups are very flexible and task-focussed and will re-organise and re-configure in response to external threats or to exploit new opportunities. Taken together, this makes it very challenging to identify network participants, to establish the existence and nature of the ties between individuals and to construct candidate network topologies on a rapid time-scale. This last point is particularly pertinent because, unlike conventional SNA, if network analysis is to be of any use in combating terrorism or organised-crime it must be able to deliver reliable network structures on a time-scale driven by operational (either military or law enforcement) tempos.

Therefore, it is desirable to develop SNA methods that can exploit the sort of limited and variable quantities of data that are typically generated by intelligence collecting campaigns against covert groups on a rapid timescale. It is at this point that the possibility of utilising open-source intelligence (OSINT) becomes desirable and advantageous. Publicly accessible data-bases from a diverse range of sources are becoming increasingly available, particularly on-line, permitting the acquisition of the sort of freely available information relating to individuals that could be of use in constructing social networks. As we shall see below, methods have recently been

developed that facilitate the aggregation of data that in or itself would be considered of limited usefulness yet, when it is combined with other data, yields results that have predictive value. Much of this information could, in principle, be derived from open-sources, so it can be argued legitimately that these unrestricted sources of data will play an increasing role in intelligence-led counter-terrorism operations in the future.

In this chapter, we will outline some recent work on inferring candidate topologies for covert social networks. Three different calculations will be discussed: (a) inferring where individuals fit into an existing network, (b) predicting missing links in sampled networks, and (c) predicting the possibility of hidden "key players" in covert social networks. All the examples use data that was obtained from open-sources (in this case media reports) following public judicial proceedings against leading members of a terrorist organisation. However, in this case, much of the data we use is unlikely to have been in the public domain beforehand. That said, the methods presented here are completely general and can be applied to OSINT alone, or applied to OSINT combined with restricted or confidential data sources to yield predictive value.

2 Inferring Network Topologies

The objective is to construct a candidate network structure for a covert social network using the sort of data that is generally obtained during an intelligence gathering campaign. The process of inferring links depends on assessing evidence for interaction between two given individuals against a known sample of positive (and negative) links in the network. Our approach follows that of Jansen et al. [4] who used Bayes' Theorem [5] to investigate the topologies of protein interactomes. A positive link is a link that connects any two individuals in the population, whereas a negative link is simply the absence of a link.

Each individual in the network has a number of possible independent attributes assigned to them. This attribute data is obtained from the intelligence-gathering phase of the SNA process. Attribute data for individuals might be information such as age, location of residence, school attended, role in group, etc. Due to the difficulties of collecting data on covert groups it is unlikely in practice that a fully populated attribute list would ever be available for every individual of interest, but the method presented here is robust against that limitation. Rhodes and Keefe [6] showed how several attributes, such as the role played in the organization, which faction they belonged to and the resources they controlled could be exploited to infer the network structure of a covert group

In the Bayesian approach (following the standard definition of statistical odds), for a given probability of a link $P(\text{pos})$, the "prior" odds of finding a positive link is given by

$$O_{\text{prior}} = \frac{P(\text{pos})}{1 - P(\text{pos})}. \tag{1}$$

which can be written as

$$O_{\text{prior}} = \frac{P(\text{pos})}{P(\text{neg})}. \tag{2}$$

By contrast, the "posterior" odds is the odds of finding a positive link after we have considered N pieces of evidence (in our case the attributes) with values $A_1 \ldots \ldots A_N$, and is given by

$$O_{\text{post}} = \frac{P(\text{pos}|A_1 \ldots \ldots A_N)}{P(\text{neg}|A_1 \ldots \ldots A_N)}. \tag{3}$$

According to Bayes' rule the prior and posterior odds are related by

$$O_{\text{post}} = L(A_1 \ldots \ldots A_N) O_{\text{prior}}. \tag{4}$$

where the likelihood ratio L is given by

$$L(A_1 \ldots \ldots A_N) = \frac{P(A_1 \ldots \ldots A_N|\text{pos})}{P(A_1 \ldots \ldots A_N|\text{neg})}. \tag{5}$$

When the pieces of evidence (attribute data) under consideration are conditionally independent the likelihood ratio factorises into a product of individual likelihoods, i.e.

$$L(A_1 \ldots \ldots A_N) = \prod_{i=1}^{N} \frac{P(A_i|\text{pos})}{P(A_i|\text{neg})}. \tag{6}$$

The probabilities of positive $P(A_i|\text{pos})$ links given an attribute A and negative $P(A_i|\text{neg})$ link given an attribute A are calculated from the sampled network data and the likelihood is simply the product of the ratios for each attribute.

Estimating the prior odds requires an assumption to be made about the number of positive links one would expect to see in the network. Using this estimate, and by evaluating the likelihood ratio (6), it is possible to calculate the posterior odds of there being a link – the probability that individuals are linked following consideration of the available evidence.

We now show how this inference approach can be used to handle attribute data from an intelligence collection process and make predictions about covert network structures.

3   Three Applications to Social Network Data

Extant data on covert networks is, understandably, minimal. To illustrate the analysis process we use data relating to the (believed defunct) Greek terrorist organisation "Revolutionary Organisation November 17" (N17). Following public trial of many of the members of this group much data emerged into the public domain and could

be used for the basis of analysis [7, 8]. The following table shows the attribute data that became available following the public trial for the members of the network.

Each individual who was brought to trial had been engaged in a number of different activities in relation to the operations of N17. Some handled money or raised money; others were responsible for weapons or security, or the acts of terror and murder. Moreover, the evidence available suggested that each individual could be assigned a more generic role within the organisation, as either leadership figures, L, (concerned with ideology and public communications) or purely operational functionaries, O, (concerned only with organising and executing acts or terror). Finally, the organisation appeared to be broadly split into three distinct sub-groups (or factions) with each individual (but not all) associated with one or the other factions. Each faction was focussed on a particular key individual in the organisation – Giotopoulos, Koufontinas or Sardanopoulos, though there was some overlap between two of the factions. Each of the entries in Table 1 is a distillation

Table 1 Attribute data for N17

	Money	Weapons	Safe houses	Attacks	Drugs	Human trafficking	Weapons	Cash acquisition	Weapon acquisition	Role	Faction
Alexandros Giotopoulos	1	0	0	1	0	0	0	1	1	L	G
Anna	0	0	0	1	0	0	0	0	0	L	G
Christodoulos Xiros	0	1	1	1	0	0	0	1	1	L	K
Constantinos Karatsolis	0	1	1	1	1	1	1	0	0	O	S
Constantinos Telios	0	1	1	1	0	0	0	1	1	O	K
Dimitris Koufontinas	1	1	1	1	0	0	0	1	1	L	K
Dionysis Georgiadis	0	1	1	1	0	0	0	1	1	O	K
Elias Gaglias	0	1	1	1	0	0	0	1	1	O	K
Iraklis Kostaris	0	1	1	1	1	1	1	0	0	O	S
Nikitas	0	0	0	1	0	0	0	0	0	L	G
Ojurk Hanuz	0	1	1	1	0	0	0	1	1	O	–
Patroclos Tselentis	0	1	1	1	1	1	1	0	0	O	S
Pavlos Serifis	0	0	0	1	1	1	1	0	0	L	S
Sardanopoulos	1	0	0	1	1	1	1	0	0	L	S
Savas Xiros	0	1	1	1	0	0	0	1	1	O	K
Sotirios Kondylis	0	1	1	1	0	0	0	1	1	O	–
Fotis	0	0	0	1	0	0	0	0	0	L	G
Thomas Serifis	0	1	1	1	1	1	1	0	0	O	S
Vassilis Tzortzatos	0	1	1	1	0	0	0	1	1	O	K
Vassilis Xiros	0	1	1	1	0	0	0	1	1	O	K
Yiannis	0	1	1	1	0	0	0	1	1	O	–
Yiannis Skandalis	0	1	1	1	0	0	0	1	1	O	–

of the trial information and is, therefore, essentially the attribute data A_i introduced in the previous section. A "1" indicates that there was evidence that an individual participated in a particular activity, and a "0" indicates there is no evidence for participation. In this example, there is a more-or-less complete set of attribute data. In practice, this is an atypical situation but one of the attractive features of the Bayesian inference method presented here is that it degrades gracefully under the condition of missing attribute data. Using the trial reporting summaries it was also possible to construct what was believed to be the network topology for N17. Whilst it will never be known whether this structure is wholly accurate we used it as the basis for comparison and calculation as it was the best available representation of a network for a reasonably-sized covert organisation. It must be borne in mind that all the results presented below may well be limited by this assumption. Also, we focus only on the absence or presence of links and do not make any attempt to characterise the nature of the link between individuals, though this itself may yield useful insight.

Next, we show three different uses of the Bayesian inference method as applied to empirical data from a covert social network. The three scenarios are typical of those encountered when undertaking the rapid analysis of limited amounts of network data on operational timescales.

3.1 Predicting the Structure of Covert Networks

In the first example, we assume that preliminary detailed data gathering has identified a limited number of individuals for whom we have good intelligence data. For N17 we assume that the network shown in Fig. 1 has been obtained from such an action.

The objective is to predict, on the basis of attribute data, where the other individuals might fit into this core network. The procedure for the calculation is presented in detail in Rhodes and Keefe [6] – here we simply illustrate the result.

Figure 2 shows the resulting network on the basis of an inference calculation. By insisting upon a 50% chance of a predicted link being a real link the calculation draws in an additional four individuals in to the network. The black lines indicate the correct prediction of links that are believed to be present in the whole network. The dotted lines indicate predicted links for which there was no evidence in the actual N17 network. It is apparent that the resulting network topology is a probabilistic one. By varying the accepted confidence level in the existence of predicted links more or fewer individuals and links are included in the predicted network structure. Note that the confidence level in the link should not be equated with the link strength or weight. The empirical data did not permit independent assessment of link weights between individuals, so it is not possible to explore any such correlation with this information. Experience with other network configurations (for N17 data) and other covert networks (unpublished work) indicate that the result presented here is representative of the performance of the Bayesian inference technique.
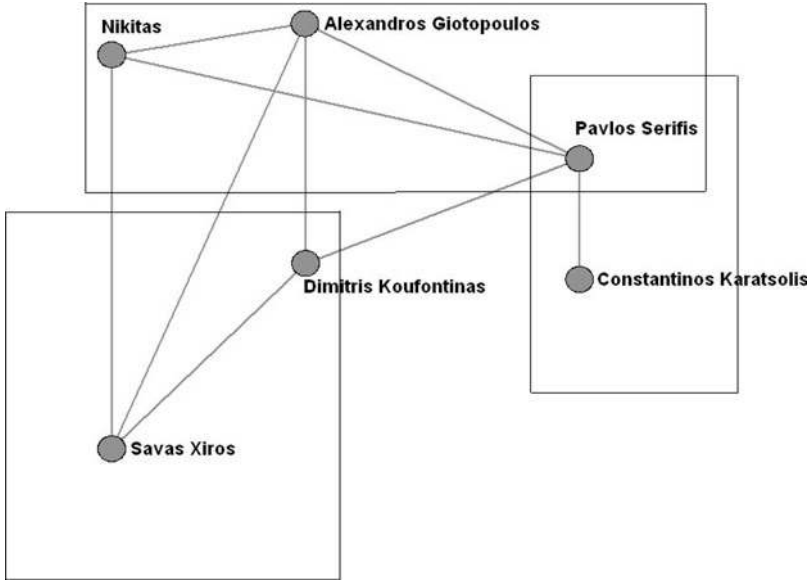
Fig. 1 Preliminary data gathered on a small number of individuals in the N17 network. Three factions in N17, described in the main text, are shown Giotopoulos (*top*), Koufontinas (*left*) and Sardanopoulos (*right*)
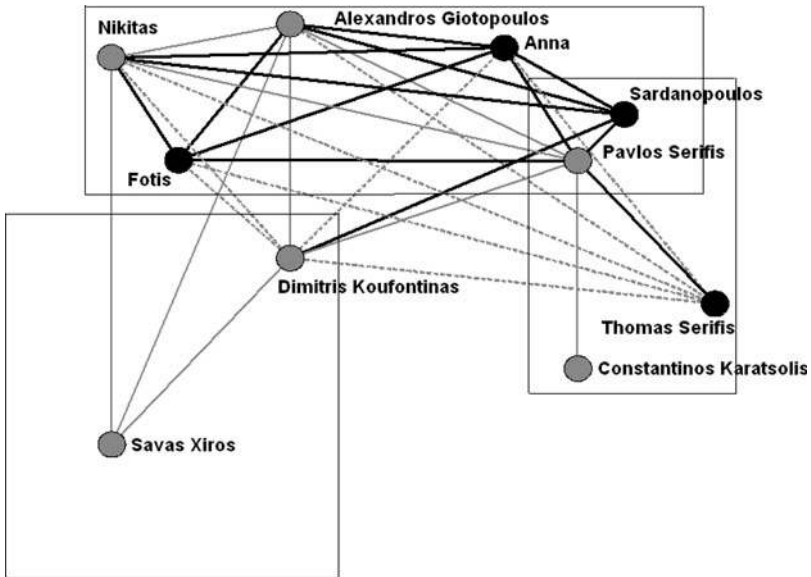
Fig. 2 Inferred network composition and structure for N17 at the 50% confidence level building on the preliminary data shown in Fig. 1. *Black lines* indicate the prediction of links that exist in the network (true positives); *dotted lines* indicate predicted links for which there is no evidence (false postives)

In this example, we are "growing" the network as attribute information on new individuals of interest becomes available. This allows network structures to be proposed without the need for further intelligence gathering activity aimed at establishing the presence or absence of links. It is also possible, but not shown here, to calculate where a given individual with a given attribute set is most likely to fit into an existing network. This could be useful if it was desirable to maximise the likelihood that a given individual succeed in deliberately joining or infiltrating the network.

3.2   Predicting Missing Links in Network Structures

Often it is readily apparent who the constituent members of a network are but it may be less apparent how they are connected. In this situation, a finite amount of intelligence-gathering resource deployed over a fixed and limited period of time will yield a sample of the network. This sample will, inevitably, be incomplete as not all the links will be observed during this finite interval. However, it is often desirable to know, or have to have an indication, whether two given individuals are connected or not as this may influence future intelligence gathering activity and resource allocation.

It is straightforward to combine attribute data on individuals with the observed pattern of links derived from an initial observation of the network. Figure 3 shows a sample of the N17 network. This structure is generated by randomly selecting around 50% of the links for inclusion.
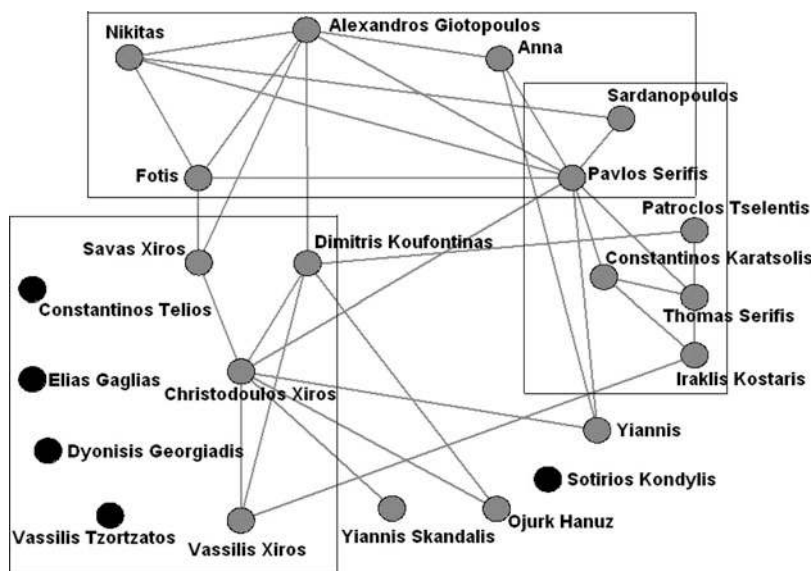
Fig. 3 Preliminary data gathered by a random sampling of the N17 network. The objective is to predict those extant links that are missed in this sampling process
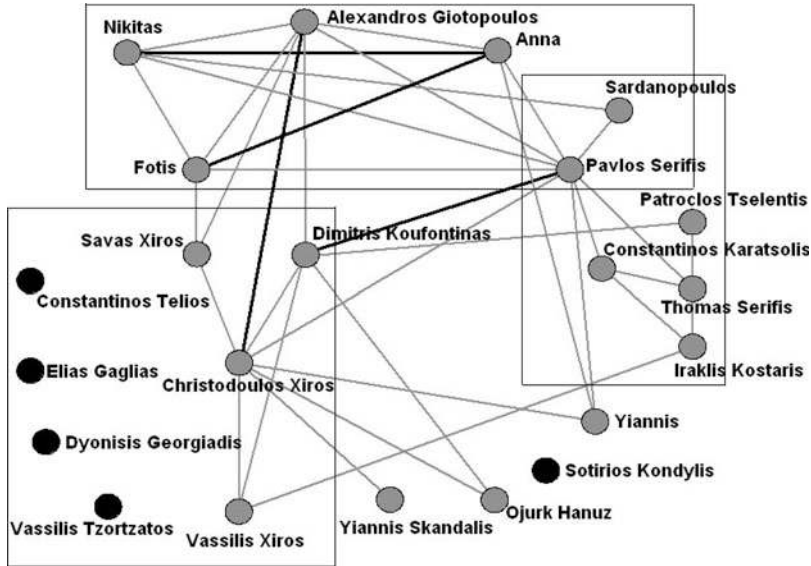
Fig. 4 *Black lines* indicate predictions of links in the sampled network of Fig. 3 at the 80% confidence level

The network resulting from this sampling process connects the individuals indicated in grey. Those individuals coloured black are of interest and are known (and possess attribute data) but there is no evidence of links to the network. The details of the calculation used to aggregate the individuals' attribute data and to predict the presence of missing links is presented in Rhodes and Jones [9]. If we insist on there being greater than 0.8 probability of there being a predicted link, the inference method generates the structure shown in Fig. 4.

The heavy black lines indicate where the method predicts links that were absent in the sampled network. In this instance, no missing links were incorrectly predicted. At this high level of confidence it is noteworthy that peripheral figures were not linked back into the network, but examples shown in [9, 10] show how lowering the confidence level fills out the network and makes it more inclusive.

The usefulness of this calculation is that it can extract further insight from existing samples of network data and crucially give some indication of which individuals are likely to be connected (in the absence of any evidence) and where to deploy intelligence collection resources to confirm or discount these insights.

3.3 Predicting the Presence of Missing "Key Players" in Covert Social Networks

One theme of recurring concern when analysing covert social networks is whether intelligence gathering has revealed the presence of all the essential figures in a

network or whether there is a missing "key player" who is not revealed by detailed monitoring of the network. The issue is whether investigation of the structural and attribute properties of the network can reveal the presence of these hidden, but nevertheless important, individuals.

Using the Bayesian inference approach it is possible to make some progress with this issue and here we describe an outline of a method that is under development[1] which appears to show some promise in addressing this difficult problem. In order to make some progress with predicting missing nodes, it is necessary to introduce the concept of a "two-path". The inference approach implicitly assumes that links form between individuals because of their attribute properties. We now extend this idea to individuals that are connected via an intermediate node. A two-path is defined to be a set of two links that join two nodes via an intermediate node.

If the adjacency matrix $\overline{A}$ of the social network is squared then the elements of the matrix $\overline{A^2_{ij}}$ give the number of two paths between nodes i and j. By dichotomising this matrix the resultant matrix has an entry of 1 where a two-path exists between two nodes and a 0 where there is no two-path. However, this matrix still contains the "one-paths" as well, so if we subtract these out we are left with the matrix of two-paths, i.e. $\overline{M} = \mathrm{dichot}\left(\overline{A^2}\right) - \overline{A}$. Using the matrix M as the basis for Bayesian inference we can now calculate the likelihood of any two nodes (given their attributes) being connected via a two-path. If we denote the likelihood matrix of all nodes by $\overline{L2}$ and we have a cut-off likelihood above which we accept the presence of a link, then if $\overline{L2_{ij}} > L_{\mathrm{cut}}$ and $M_{ij} = 0$ then we should be alert to the possibility of a link between nodes i and j.

There are three ways a two-path can link two nodes:

1. Via a known intermediate node to which one of the pair (i, j) is already linked (denoted ML1)
2. Via a known intermediate node to which neither of the nodes already connected (denoted ML2)
3. Via an unknown intermediate node (denoted MN)

These possibilities are shown below:

If we find a high likelihood two-path predicted between two nodes but no such path exists on the network then an intermediary node must exist yet be missing from the intelligence picture (Fig. 5).

An algorithmic approach to locating missing nodes in networks would proceed as follows:

1. Find pairs of nodes *(i, j)* for which $\overline{L2_{ij}} > L_{\mathrm{cut}}$ and $M_{ij} = 0$.
2. ML1: Look for nodes to which *(i, j)* are connected and find the likelihood using Bayesian inference on the matrix of one-paths that these intermediary nodes could be connected to the other node by a link (denoted $L1_{ik}$ – where k is a node connected to j).
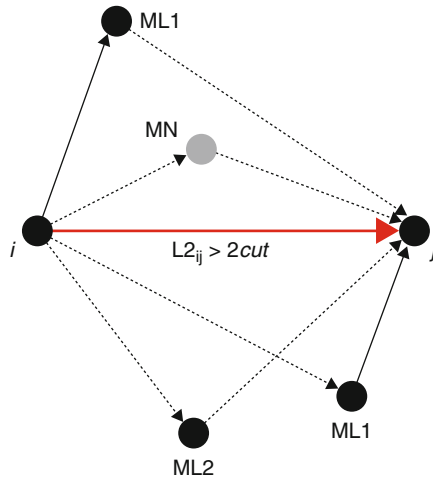
Fig. 5 The three different ways (ML1, ML2, MN) in which a two-path can connect two nodes in a network

3. ML2: Look for nodes to which *(i, j)* are not connected and find the likelihood using Bayesian inference on the matrix of one-paths that these intermediary nodes could be connected to the other node by a link (denoted $L1_{ik}L1_{kj}$ – where k is a node connected to j).
4. MN: Construct a set of individuals with all possible attribute combinations and calculate their likelihoods of linking *(i, j)*.
5. Rank the options from steps 2 to 4 by likelihood to find the best explanation for the missing two-path.

A preliminary analysis of the N17 data set described above, and some other social network data sets obtained from open sources, indicates that this approach has promise as a means of detecting missing nodes. It should be noted that in the case of option 4 (MN), not only is the presence of a missing node flagged-up, but its attribute set is also predicted. This should assist in mounting a focussed intelligence gathering campaign to locate the missing individual, as the attribute set will give indicators about their likely characteristics.

4   Conclusions

SNA is now being widely used as a key method in analysing and understanding the capabilities of covert groups such as terrorist organisations, insurgency factions and organised crime groups. Despite its acknowledged usefulness SNA demands plentiful, timely and detailed data in order for its full benefits to be realised, particularly when used against rapidly evolving or frequently re-configuring networks.

It is imperative that methods are developed that can integrate and exploit all available data in a flexible and robust analytic framework. The methods presented here are motivated by that requirement. Three different analysis scenarios were presented that are representative of the sort of analysis that can be undertaken during operationally-driven SNA. Of particular note is the potential for the inference method to be able to predict the presence of (and attribute set of) missing "key-players" in covert social networks. In this regard, OSINT represents a vast and largely under-utilized source of data for SNA purposes and the inference approach outlined here is capable of integrating categorical and numerical data from a diverse range of sources. OSINT should be increasingly exploited to contribute data for downstream analysis using network methods. Steps should be taken to develop and refine techniques such as those outlined in this chapter to increase confidence in their applicability to the analysis of covert social networks.

References

1. Krebs, V.: Uncloaking Terrorist Networks. First Monday **7**(4) (2002)
2. Sparrow, M.K.: The application of network analysis to criminal intelligence: an assessment of the prospects. Soc. Netw. **13**(3), 251–274 (1991)
3. Carley, K.M., Lee, J., Krackhardt, D.: Destabilising Networks. Connections **24**(3), 31–34 (2001)
4. Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J.F., Gerstein, M.: A Bayesian networks approach to predicting protein–protein interactions from genomic data. Science **302**, 449–451 (2003)
5. Sivia, D.S.: Data Analysis: A Bayesian Tutorial. Oxford University Press, Oxford (2004)
6. Rhodes, C.J., Keefe, E.M.J.: Social network topology: a Bayesian approach. J. Op. Res. Soc. 58, 1605–1611 (2007)
7. Irwin, C., Roberts, C., Mee, N.: Counter Terrorism Overseas. Dstl Report, Dstl/CD053271/1.1 (2002)
8. Abram, P.J., Smith, J.D.: Modelling and analysis of terrorist network disruption. MSc thesis, Cranfield University, Shrivenham, United Kingdom (2004)
9. Rhodes, C.J., Jones, P.: Inferring missing links in partially observed social networks. J. Oper. Res. Soc. 60, 1373–1383 (2009)
10. Rhodes, C.J.: Inference approaches to constructing covert social network topologies. In: Memon, N., Farley, J.D., Hicks, D.L., Rosenoorn, T. (eds.) Mathematical Methods in Counter-Terrorism. Springer, Berlin (2009)

A Novel Method to Analyze the Importance of Links in Terrorist Networks

Uffe Kock Wiil, Jolanta Gniadek, and Nasrullah Memon

Abstract A terrorist network is a special kind of social network with emphasis on both secrecy and efficiency. Such networks are intentionally structured to ensure efficient communication between members without being detected. A terrorist network can be modeled as a generalized network (graph) consisting of nodes and links. Techniques from social network analysis (SNA) and graph theory can be used to identify key entities in the network, which is helpful for network destabilization purposes. Research on terrorist network analysis (TNA) has mainly focuses on analysis of nodes, which is in contrast to the fact that the links between the nodes provide at least as much relevant information about the network as the nodes themselves. This paper presents a novel method to analyze the importance of links in terrorist networks inspired by research on transportation networks. The link importance measure is implemented in CrimeFighter Assistant and evaluated on known terrorist networks harvested from open sources.

1 Introduction

A terrorist network is a special kind of social network with emphasis on both secrecy and efficiency. Such networks are intentionally structured to ensure efficient communication between members without being detected [1–3].

U.K. Wiil (✉) · J. Gniadek
Counterterrorism Research Lab, The Maersk Mc-Kinney Moller Institute, University of Southern Denmark, Odense, Denmark
e-mail: ukwiil@mmmi.sdu.dk

N. Memon
Counterterrorism Research Lab, The Maersk Mc-Kinney Moller Institute, University of Southern Denmark, Odense, Denmark
and
Hellenic American University, Manchester, NH, USA
e-mail: memon@mmmi.sdu.dk

Knowledge about the structure and organization of terrorist networks is important for both terrorism investigation and the development of effective strategies to prevent terrorist attacks. Theory from the knowledge management field plays an important role in dealing with terrorist information. Knowledge management processes, tools, and techniques can help intelligence analysts in various ways when trying to make sense of the vast amount of data being collected in relation to terrorism [4]. The collected data needs to be analyzed and visualized in order to gain a deeper understanding of the terrorist network.

A terrorist network can be modeled as a generalized network (graph) consisting of nodes and links. Nodes are entities (people, places, events, etc.) with attributes allowing relevant information to be stored about the entities. Links are relationships between the entities. Links have attributes that describe properties about the relationships.

Techniques from social network analysis (SNA) and graph theory [5] can be used to identify key nodes in the network, which is helpful for network destabilization purposes [6]. Taking out key nodes will decrease the ability of the terrorist network to function normally.

However, research on terrorist network analysis (TNA) has mainly focuses on analysis of nodes. Links are seldom first class objects with the same properties as nodes. This is in contrast to the fact that the links between the nodes provide at least as much relevant information about the network as the nodes themselves [7].

A terrorism domain model with both nodes and links as first class objects will allow additional features to be built into the TNA and visualization tools. Hence, a possible response to the above mentioned issue is to develop new measures for TNA focusing on links. Aiming for a better balance between analysis of nodes and analysis of links, results in additional and more precise knowledge about the terrorist network.

This paper presents a novel method to analyze the importance of links in terrorist networks inspired by research on transportation networks. The link importance measure is implemented in CrimeFighter Assistant and evaluated on known terrorist networks harvested from open sources: 9/11 attacks (2001), Bali night club bombing (2002), Madrid bombings (2004), and 7/7 London bombings (2005).

The remainder of this paper is structured as follows. Section 2 describes various techniques that can be used to analyze terrorist networks. We start by looking at general techniques related to analysis of social networks and continue by looking at specific techniques that are related to analysis of terrorist networks. In Sect. 3, we present and evaluate our new method to analyze the importance of links in terrorist networks. Section 4 concludes the paper and discusses future work.

2   Terrorist Network Analysis

This section presents techniques that are useful to analyze terrorist networks. The starting point for TNA is the existence of a network structure. Hence, much knowledge management work needs to take place prior to network analysis. Data

needs to be gathered, data needs to be processed (filtered, mined, etc.) to create useful information in the form of a network structure. These prerequisite knowledge management processes are not the focus of this paper.

The network analysis phase should result in new insights (knowledge) about the network, the entities, and the relations. Also, the achieved knowledge must be visualized in a human comprehensible way to enable the intelligence analysts to make informed decisions (recommendations) about possible actions to destabilize the network.

2.1 General Social Network Analysis Techniques

Since a terrorist network is a special kind of social network, many techniques useful to analyze social networks are also applicable to TNA. SNA relies to a large extent on a mathematical model in the form of a graph and a set of algorithms that traverses the graph in various ways to analyze the network.

A graph G consists of two sets of information: a set of nodes, $N = \{n_1, n_2, \ldots, n_n\}$, and a set of links $L = \{l_1, l_2, \ldots, l_l\}$ between pairs of nodes. There are n nodes and l links. In a graph, each link is an unordered pair of distinct nodes, $l_k = \{n_i, n_j\}$.

Small graphs can provide visual information about the network, but for larger graphs it is difficult to perform analysis visually. Graph theory provides several ways to measure social networks:

- *Size* is defined as the number of nodes (n) in the network.
- *Density* is the number of links (l) in proportion to the number of links that are possible in G (if all nodes where connected to each other).
- *Nodal degree* is defined as the number of links that are incident with the node.
- A *cluster* is a part of the graph with high density of nodes and links between them.
- The *average shortest path* is the average length of the geodesic between two nodes.
- *Node degree centrality*. A node is central when it has many ties (links) to other nodes in the network. This kind of centrality is measured by the degree of the node. The higher the degree, the more central the node is.
- *Node closeness centrality* indicates that a node is central when it has easy access to other nodes in the network. This means that the average distance (calculated as the shortest path) to other nodes in the network is small.
- *Node betweenness centrality*. Usually, not all nodes are connected to each other in a network. Therefore, a path from one node to another may go through one or more intermediate nodes. Betweenness centrality is measured as the frequency of occurrence of a node on the geodesic connecting other pairs of nodes. A high frequency indicates a central node.

- *Eigenvector centrality* is like a recursive version of node degree centrality. A node is central to the extent that the node is connected to other nodes that are central. A node that is high on eigenvector centrality is connected to many nodes that are themselves connected to many nodes.

Wasserman and Faust [5] provide additional details about SNA.

2.2 Specific Terrorist Network Analysis Techniques

Terrorist networks are covert networks. Covertness is the major difference between terrorist and regular social networks. In terrorist network, ties between participants are usually strong, but not transparent and visible in every day routine. Relations are long-term; participation in a terrorist plot requires a high level of trust in the network. Terrorist networks are often "sleeping"; they are prepared, but remain inactive. This way they are more difficult to uncover.

As Krebs [8] noticed, a covert network must be active at times. It is during these periods of activity that they may be most vulnerable to discovery. Social networks in covert organizations tend to structure themselves towards better efficiency or robustness [9]. According to the definition of efficiency in [10], the most efficient network is a clique of the size of the network (a complete graph with density equal to 1). Yet, this structure makes the terrorist networks vulnerable to detection. If one suspect is uncovered, observation of this suspect allows investigators to determine all suspects that are connected to this suspect. Hence, the whole network would be easily disrupted. This example shows that terrorist networks cannot operate in the same way as regular social networks. Terrorists want to keep their actions (attacks are an exception) and relations hidden from the public. According to Baker and Faulkner [11], the need for secrecy is crucial to covert networks. Thus, terrorist networks have to find a balance between efficiency and secrecy.

What then determines the secrecy of a network? Lindelauf et al. [3] proposes a measure of secrecy which is defined by two parameters: the exposure probability and the link detection probability. The exposure probability applies to individual nodes and depends on the location in the structure. It is defined as the probability of a member of the network to be detected as a terrorist. Link detection probability represents the chance of exposure of a part of the network if a member is detected.

Considering the above measure of secrecy, the safest structure of a terrorist network would be a path graph, where all the nodes know only two neighboring nodes. Looking at this structure from an information exchange perspective, the weakness is obvious. Information has to travel a long distance from one part of the network to another and that decreases the efficiency of the network. The lower the efficiency, the worse communication and coordination in the network – to the point when launching successful operation becomes impossible.

According to the definition in [10], *efficiency* is a measure to quantify how efficiently the nodes of a network can exchange information. To calculate the

efficiency of network, all the shortest path lengths between any pair of nodes in the graph must be calculated. The assumption is made that every link can be used to transfer information in the network. The efficiency is calculated in two parts: (1) the inverse of the sum of the shortest paths between any pair of nodes are calculated; (2) the result from (1) is divided by the possible number of pairs of nodes to find the average efficiency of the network.

Various measures used in SNA and TNA will be exemplified below by describing them in relation to the 9/11 hijackers and associates network presented in [8] (Fig. 1).

Fig. 1 The 9/11 hijackers and associates network [8]

Table 1 Node centralities of 9/11 network

Person	Degree	Closeness	Betweenness	Eigenvector
Atta	0.361	0.587	0.588	0.412
Al-Shehhi	0.295	0.466	0.088	0.399
Hanjour	0.213	0.445	0.126	0.249
Jarrah	0.164	0.424	0.017	0.258
Moussaouri	0.131	0.436	0.232	0.084
Khemais	0.180	0.433	0.252	0.059
Al-Omari	0.148	0.424	0.023	0.237
Al-Shibh	0.164	0.436	0.048	0.233
Ahmed	0.131	0.407	0.026	0.201
Bahaji	0.115	0.399	0.002	0.198

SNA measures of the 9/11 hijackers and associates network:

- *Size.* The full 9/11 network consisted of 62 persons. The core of the network (only the hijackers) consisted of 19 persons divided into 4 groups of strongly connected members.
- *Density.* The density of the full network is very low (0.08). It shows that the network overall focused on few connections between members to ensure a high level of secrecy.
- *Nodal degree.* The average nodal degree of the full network is 4.9 which mean that detection of one member would potentially comprise 5 other members.
- *Clusters.* The core of the 9/11 network has a high density (0.585). It shows that inside the trusted core there was a focus on a high level of efficiency.
- *Average shortest path.* The diameter of the network (the longest shortest path) is 5. The average shortest path is 2.92. Thus, on average information needs to travel through 3 links to reach the target.
- *Node centralities.* When calculating the node centralities (degree, closeness, betweenness, and eigenvector), a small group of people was pointed out as central in the network. Table 1 ranks members according to the combined score of the centrality measures – only the top 10 ranked members out of 62 are listed. The scores marked in bold face are the top three scores for that measure (again out of 62).
  A closer look at the numbers in Table 1 shows that the first 4 on the list (Atta, Al-Shehhi, Hanjour, and Jarrah) each belonged to a different hijacked plane. The next 2 on the list (Moussaoui and Khemais) are central in the part of the network that connects the hijackers to the associates.

The above example based on the full 9/11 network shows how traditional measures of SNA can play a role in TNA. These kinds of measures are sensitive to changes in the network. We must assume that there are missing nodes and links in the network (members and ties that were not discovered in the investigation). Thus, uncovering new members and ties will change centrality measures. However, these methods still give a good measure of the importance of members in the part of the network that was detected.

According to the efficiency measure in [10], the efficiency of 9/11 network is 0.395. According to the secrecy measure in [3], the secrecy of the 9/11 network is 0.86, while the secrecy of the hijackers part of the network is lower (0.77) as this part has a higher density.

In addition to the above SNA measures, various measures related to TNA have been proposed in the literature.

Memon [12] proposed several new analysis measures and destabilization methods including:

- *Position Role Index* (PRI) is a measure aimed at making a distinction between the gatekeeper and follower roles. PRI evolved from testing efficiency of a network based on the assumption that a network without followers has a higher efficiency as followers are less connected within the structure. PRI is measured as the change of network efficiency after removal of a node. A high PRI value indicates a large loss in efficiency, if a particular node is removed.
- *Detecting hidden hierarchy.* This method aims to identify hidden hierarchical structures in horizontal networks. The method uses SNA measures and graph theory to indicate parent–child relationships of nodes in the network.
- *Subgroup detection.* A terrorist network can often be partitioned into cells (subgroups) consisting of individuals who interact closely with each other. This method uses SNA measures and graph theory to indicate clusters (subgroups) in relation to a particular node and the diameter from that node.

These methods can be used to provide a richer and deeper understanding and insight into terrorist networks to enable better approaches to destabilize them.

Rhodes [13] proposed the use of Bayesian inference techniques to predict missing links in a covert network demonstrated through a case study of the Greek terrorist group November 17. The assumption is that during the analysis of terrorist networks it is unlikely that the intelligence analysts have an overview of the full terrorist network. Prediction of missing links can be a useful method to gain deeper understanding and conduct detailed analysis of the terrorist network.

One of a very few metrics that includes the property of links is link (edge) betweenness centrality. It measures the frequency of link occurrence on the geodesic connecting pairs of nodes [14]. Link betweenness indicates how much information flows via a particular link. The assumption is that communication flows along the shortest path. A high frequency indicates a central link. Newman [15] has proposed a variant of link betweenness centrality based on random walks instead of shortest paths.

2.3 Summary

This section has reviewed various techniques for analyzing terrorist networks. The usefulness of a number of these techniques was demonstrated through an example using the full 9/11 network as presented by Krebs [8]. The techniques primarily

focus on estimating the importance of nodes of the networks. Few techniques focus on estimating the importance of links in the network. Thus, the importance of links in terrorist networks remains to a large degree an unexplored issue.

3 Link Importance

As mentioned, current methods of TNA assign high importance to nodes. Role analysis, centrality, and clustering measures focus on positions of nodes in the network. The goal from an intelligence analysis perspective is to come up with informed decisions regarding possible actions to destabilize the network. Carley et al. [16] proposed the following criteria to evaluate if a network has been destabilized:

- The rate of information flow through the network has been seriously reduced, possibly to zero.
- The network, as a decision-making body, can no longer reach consensus, or takes much longer to do so.

Information flow in terrorist networks takes place through links and nodes. Removing an important node will destabilize the network. Removing a node will also remove all the links connecting that node. To be able to make informed decisions regarding which nodes to remove to destabilize a network, both the importance of nodes and links should be considered. By focusing on link importance, a node that is connected to other nodes through important links becomes important itself. The goal of this paper is to determine which links are crucial to the network and how removing a particular link would influence the network in terms of secrecy and efficiency.

3.1 Secrecy and Efficiency

According to Lindelauf et al. [3], secrecy depends on the number of links, the number of nodes, and their degree. The higher the degree of nodes, the lower the secrecy is in the network. Therefore, in order to keep a high level of secrecy in the network, connections between nodes should be sparse and there should be a low level of redundancy of connections.

Lindelauf et al. [3] has also proposed a definition of information performance which in many ways is similar to the definition of efficiency proposed in [10]. Comparing these two methods using the 9/11 network (Fig. 1), results in an efficiency of 0.395 (as mentioned in the previous section) and an information performance of 0.342.

Lindelauf et al. [3] also proposed a measure of an overall performance of a network as the product between secrecy and information performance. This measure

is used to assess the performance of the network in the light of the goals of terrorist network to reach a balance between secrecy and efficiency.

Looking from a link importance perspective, link removal will in most cases lead to increased secrecy (as a side effect) and to decreased efficiency (which is the goal of destabilization). However, in some cases link removal will not cause changes in network efficiency (if redundant paths of the same length exist). In this case, link removal will only result in increased secrecy (which is not the goal of destabilization).

3.2   Link Importance in Transportation Networks

TNA has its origin in SNA and uses related metrics. A social network can be referred to as a "pure network" because only its topology and connectivity is considered. If a network is characterized by its topology and flow characteristics (such as capacity constraints, path choice, and link cost functions) it is referred to as a flow network [17]. A flow network is a directed graph where each link has some capacity and can receive a flow – lower than its capacity. A transportation network is a flow network representing the movement of people, vehicles, or goods [18].

In terrorist networks, links represent relations between entities; people communicate with each other and the outcome of those relations is information flow. Communication intensity can be the analogy of movement in transportation networks. Some links and nodes will be used more than others in the information flow depending on their position in the structure of a network and also on the source and destination of the information. However, unlike transportation network, the same information can travel at the same time in different parts of the network. If we consider a covert network as a special case of a transportation network, where all links have infinite capacities, then measures of transportation networks can be used for TNA.

In transportation networks links are first class objects. A flow between two nodes is dependent on links and their capacity. An illustrative example can be a traffic network. Roads are pictured as links between cities; some roads are heavily used while traffic on others is light. Figure 2 shows an example involving traffic between different parts of Sweden. The thickness of links represents the density of the traffic – the thicker the link, the higher the density. Nodes presented as big circles symbolize major origins/destinations of movement. Relating this example to terrorist networks, "heavy traffic" can be understood as a high level of information exchange between nodes, and nodes can be described according to centrality measures.

Based on transportation networks, Jenelius et al. [19] proposed the measure of link importance. It is based on the concept of vulnerability, which in a road traffic context is defined as the susceptibility to incidents that can result in considerable reductions in network capacity during a given period. Bienenstock and Bonacich define vulnerability in the context of SNA as the loss of efficiency resulting from
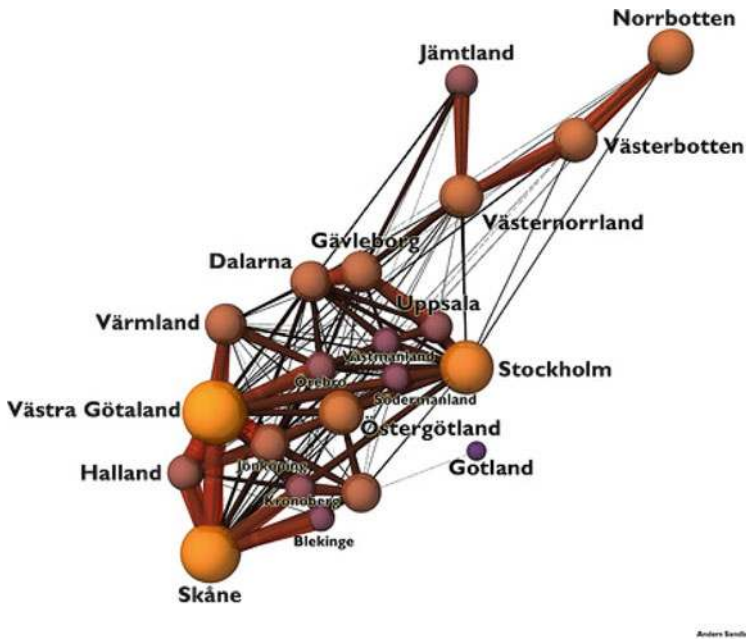
Fig. 2 Example transportation network

the elimination of nodes [20]. In transportation networks, the focus is placed on what effect the removal of a link will have on functionality. In social networks, the effect of removing a node is considered.

In both cases, removal of a node or a link leads to decreased network functionality. Looking at Fig. 2, it is clear that closing the road between Västra Götaland and Skåne would lead to bigger problems than closing the road between Stockholm and Västernorrland. Therefore, the link between the first two regions is more important than the link between the latter two. Based on this assumption, the conclusion can be made that the difference between current performance and performance after link removal can be an indicator of link importance in the network.

In transportation networks, performance is considered as the sum of cost of travel from node i to node k. Based on this, a definition of link importance for transportation networks has been proposed [19] which takes three things into consideration (1) travel demand from node i to node j; (2) the generalized cost of travel from node i to node j in the initial, undamaged network; and (3) the generalized cost of travel from node i to node j when link k is closed.

3.3 Link Importance in Terrorist Networks

The measure of link importance in terrorist networks is inspired by transportation networks. According to Carley et al.'s [16] proposed criteria to measure network

destabilization, we use overall performance (as defined by Lindelauf et al. [3]) to measure how well a terrorist network is functioning. P^0 denotes the initial performance of the network and P_k denotes the performance after removal of link k. Therefore, the performance change when removing link k can be described as:

$$\triangle P_k = P^0 - P^k \tag{1}$$

The goal of link removal in terrorist networks is to lower the performance of the network. Thus, the higher the value of performance change when removing a link, the more important is the link for the network.

The definition of link importance for transportation networks accounted travel demand between nodes as a weight for the performance change. Demand for information flow between two nodes in terrorist network can be expressed by link betweenness. The higher the value of this metrics, the more information travels via the link.

The change in link betweenness when removing link k will be used as a weight for the performance change. The weight is calculated as follows:

$$w(k) = \sum C_b(G) \Big/ \left(\sum C_b(G) - C_b(k) \right) \tag{2}$$

Where $\sum C_b(G)$ denotes the sum of link betweenness for all links in the graph G and $C_b(k)$ denotes link betweenness for link k.

This weight will give higher importance to links with higher link betweenness, yet it will still keep performance change as the major factor of link importance. Based on the presented factors, the importance of link k is measured as follows:

$$LI(k) = \triangle P_k * w(k) \tag{3}$$

The algorithm for calculating link importance consists of the following steps:

1. *Efficiency.* The sum of the shortest paths connecting each pair of nodes is computed. Based on this the efficiency (information performance) of the network is calculated according to the definition by Lindelauf et al. [3].
2. *Secrecy.* The sum of the square degrees of nodes is computed. Based on this and the number of links and nodes, the secrecy of the network is calculated according to the definition by Lindelauf et al. [3].
3. *Performance.* The overall performance of the network is computed as the product of secrecy and efficiency according to the definition by Lindelauf et al. [3].
4. *Weight.* The weight measure for link importance is computed for each link according to (2) above.
5. *Link importance.* The link importance for each link is computed according to (3) above.

Calculation of link importance results in positive and negative values. A positive value for a link means that after its removal, the performance of the network will

decrease (the increase in secrecy will be lower than the decrease in efficiency of the network). A negative value for a link means that the increase of secrecy is higher than the decrease of efficiency – hence, the performance of the network is increased.

The link importance measure helps intelligence analysts to understand which links are important for communication in covert networks and how their removal will influence the rest of the network. Naturally, links with high importance should be taken under consideration in order to destabilize the network.

3.4   Scenario 1: Link Importance in a Small Network

We use a part of the 9/11 terrorist network (Fig. 3) as an example to illustrate how link importance works.

The importance of the individual links in the network is shown in Table 2.

The results of link importance are depicted in Fig. 4 together with efficiency, secrecy, initial efficiency (before link removal) and initial secrecy (before link removal). Removal of link number 13 results in the highest decrease of efficiency, while the increase of secrecy is insignificant. Hence, link 13 is the most importance link from a network destabilization point of view.

3.5   Scenario 2: Link Importance in the Full 9/11 Network

The network of 9/11 hijackers and their associates (Fig. 1) is a medium sized network consisting of 62 nodes and 153 links. Figure 5 shows the results of link
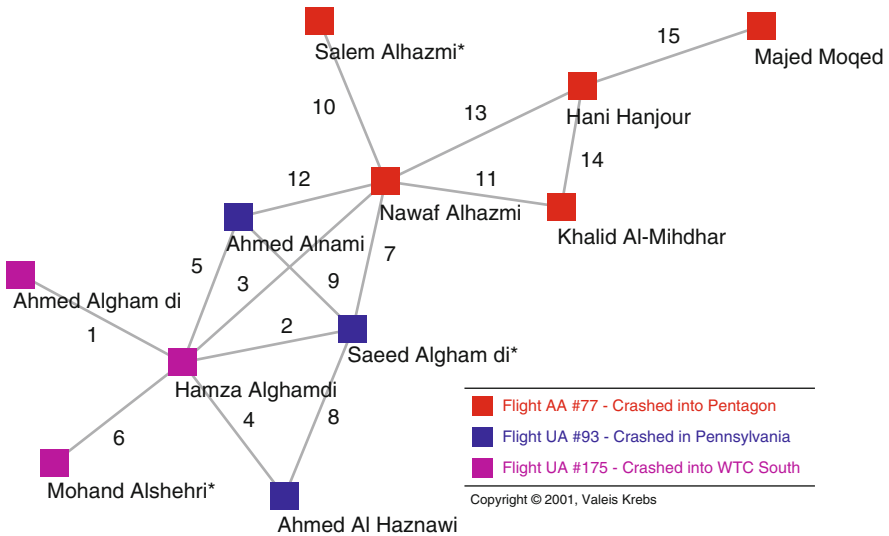
Fig. 3 Part of 9/11 network

Table 2 Link importance in part of 9/11 network

Link	Link importance	Secrecy	Efficiency
13	0.027759679	0.6060606	0.42307693
3	0.018798648	0.6200466	0.42635658
11	0.009842231	0.6013986	0.45081967
14	0.001790676	0.5874126	0.47413793
12	0.000169219	0.6060606	0.46218488
7	−0.002133225	0.6107226	0.46218488
4	−0.002570151	0.6013986	0.47008547
8	−0.002997429	0.5920746	0.47826087
5	−0.004772992	0.6060606	0.47008547
9	−0.005190209	0.5967366	0.47826087
2	−0.007002236	0.6107226	0.47008547
10	−0.088121056	0.5967366	0.6043956
1	−0.092513749	0.5967366	0.6111111
6	−0.092513749	0.5967366	0.6111111
15	−0.106107757	0.58275056	0.64705884
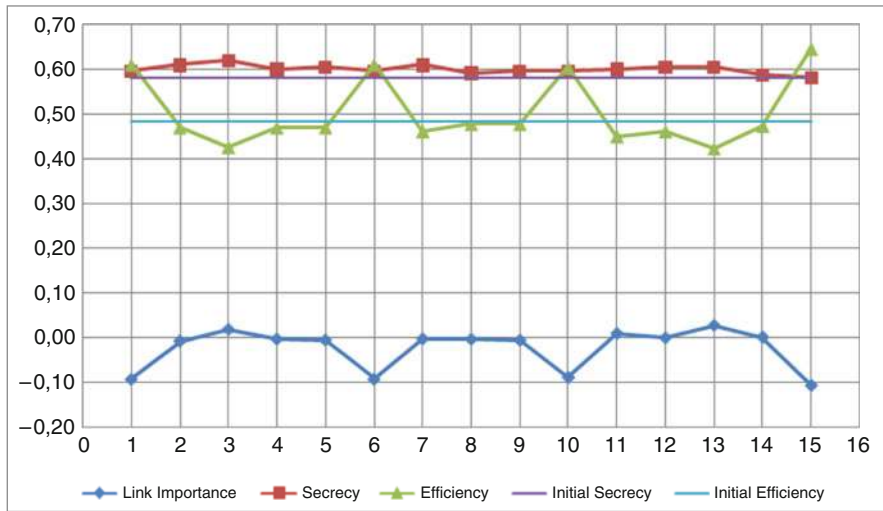
Fig. 4 Link importance measures

importance by coloring the 10 links with the highest (red) and the 10 links with the lowest importance (blue).

It is visible that the 10 most important links connect the network in a way that no node is further than two steps away from these links. These 10 links connect different segments of the network and maps out the "information backbone" of the network. The 10 links are focused around Atta, Hanjour, and Moussaoui, who were among the five most important nodes in the network based on calculation of node centralities (as pointed out in Sect. 2). Therefore, it is not a surprise that the 10 most

Fig. 5 Link importance in the 9/11 network

important links are primarily connecting those nodes indicating the significance of the communication between them. The most important link connects Atta to Khemais, who was the head of the Italian cell of Al-Qaeda. It is also visible that the 10 least important links are the ones that connect peripheral nodes to the network – except for the link connecting Hanjour and Al-Shehhi. The reason for the latter is the redundancy of links in the core of the network. Hence, removal of the particular link does not decrease the efficiency of the network.
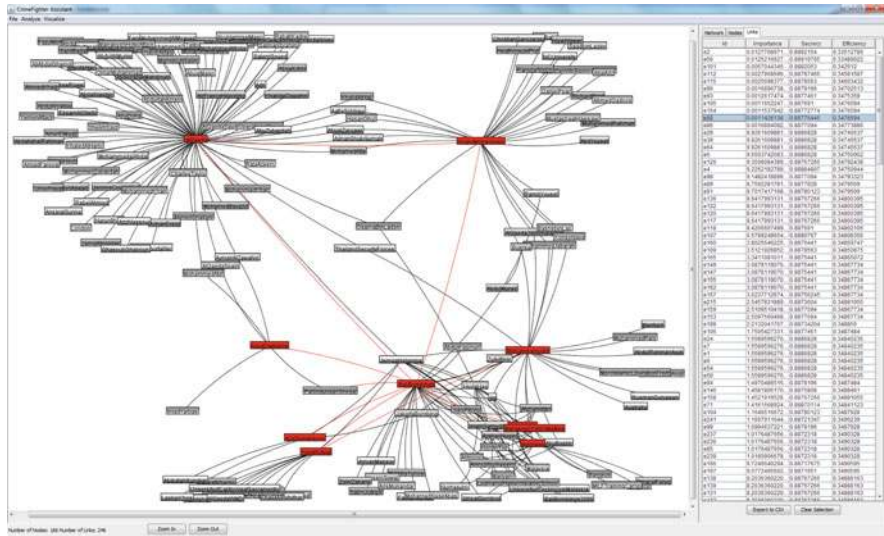
Fig. 6 Screenshot of CrimeFighter assistant

3.6 Evaluation

The algorithm for link importance has been evaluated based on the 9/11 network (as described in this paper), the Bali night club bombing network (as described in [21]) as well as on the Madrid bombings network (67 nodes and 88 links) and the 7/7 London bombings network (50 nodes and 61 links). The algorithm is implemented in the CrimeFighter Assistant together with various other network, node, and link measures. Figure 6 shows a screenshot of the CrimeFighter Assistant when analyzing the Bali night club bombing network (166 nodes and 246 links).

The network visualization in the left part of CrimeFighter Assistant highlights the 10 most important nodes (according to the PRI measure) and the 10 most important links (according to the link importance measure). In the four cases, the link importance and link betweenness measures identify some of the same links as their top ranked links. The two measures have the same 5 links (Madrid and London cases) respectively 6 links (9/11 and Bali cases) as part of their top 10 links, but not in the exact same order.

A similar observation can be made regarding the different node centrality measures used for SNA (see Table 1). They identify some of the same nodes as their top ranked nodes. In other words, the node centrality measures complement each other and together they provide a balanced view on how important the individual nodes are. In the same manner, the measures of link importance and link betweenness complement each other and together they provide a balanced view on how important the individual links are.

Thus, the measure of link importance offers new insights into terrorist networks by pointing out links that are important to the performance of the network. The measure of link importance maps out the information backbone of a terrorist network and can (together with node and other link measures) point to new ways to destabilize a network.

4 Conclusion

This paper has proposed link importance as a new measure for destabilizing terrorist networks. The usefulness of the method was demonstrated based on an analysis of four terrorist networks harvested from open sources. The presented work provides the following contributions:

- Description and evaluation of a novel method for measuring link importance in terrorist networks, which is inspired by research on transportation networks. It uses the measures of secrecy and efficiency proposed by Lindelauf et al. [3] together with the measure of link betweenness.
- An implementation of the proposed link importance measure in CrimeFighter Assistant, which to our knowledge also provides the first implementation of the secrecy and efficiency (information performance) measures as proposed by Lindelauf et al. [3].

Future work will further investigate, evaluate, and improve the measure of link importance. We are currently looking into how link weights can be incorporated, since not all links are equally important. We are also looking into how link direction can be incorporated. We think that incorporation of link weights and direction will result in a more precise measure of link importance.

References

1. Baccara, M., Bar-Isaac, H.: Interrogation methods and terror networks. Mathematical Methods in Counterterrorism, pp. 271–290. Springer, Berlin (2009)
2. Lindelauf, R., Borm, P., Hamers, H.: On heterogeneous covert networks. Mathematical Methods in Counterterrorism, pp. 215–228. Springer, Berlin (2009)
3. Lindelauf, R., Borm, P., Hamers, H.: The influence of secrecy on the communication structure of covert networks. Soc. Netw. **31**, 126–137 (2009)
4. Wiil, U.K., Memon, N., Gniadek, J.: Knowledge management processes, tools and techniques for counterterrorism. In: Proceedings of the International Conference on Knowledge Management and Information Sharing, pp. 29–36. INSTICC Press, Funchal, Portugal, Oct 2009
5. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Cambridge University Press, Cambridge (1994)
6. Memon, N., Wiil, U.K., Alhajj, R., Atzenbeck, C., Harkiolakis, N.: Harvesting covert networks: The case study of the iMiner database. Int. J. Netw. Virtual Organ. **8**(1/2), 52–74 (2011)

7. Gloor, P.A., Zhao, Y.: Analyzing actors and their discussion topics by semantic social network analysis. Inform. Vis. 130–135 (2006)
8. Krebs, V.E.: Uncloaking terrorist networks. First Monday **7**(4–1) (2002)
9. Enders, W., Su, X.: Rational terrorists and optimal network structure. J. Confl. Resolut. **51**(1), 33 (2007)
10. Latora, V., Marchiori, M.: How the science of complex networks can help developing strategies against terrorism. Chaos Solitons Fractals **20**(1), 69–75 (2004)
11. Baker, W.E., Faulkner, R.R.: The social organization of conspiracy: Illegal networks in the heavy electrical equipment industry. Am. Sociol. Rev. 837–860 (1993)
12. Memon, N.: Investigative Data Mining: Mathematical Models for Analyzing, Visualizing and Destabilizing Terrorist Networks. Ph.D. thesis, Aalborg University, Denmark (2007)
13. Rhodes, C.J.: Inference approaches to constructing covert social network topologies. Mathematical Methods in Counterterrorism, pp. 127–140. Springer, Berlin (2009)
14. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. **99**(12), 7821–7826 (2002)
15. Newman, M.E.J.: A measure of betweenness centrality based on random walks. Soc. Netw. **27**(1), 39–54 (2005)
16. Carley, K.M., Lee, J.S., Krackhardt, D.: Destabilizing networks. Connections **24**(3), 31–34 (2001)
17. Fischer, M.M., Lände, R.: GIS and network analysis. Handbook of Transport Geography and Spatial Systems, vol. 5, pp. 391-408. Pergamon Press, Amsterdam (2004)
18. Bell, M.G.H., Iida, Y.: Transportation Network Analysis. Wiley, New York (1997)
19. Jenelius, E., Petersen, T., Mattsson, L.G.: Importance and exposure in road network vulnerability analysis. Transport. Res. Part A **40**(7), 537–560 (2006)
20. Bienenstock, E.J., Bonacich, P.: Balancing efficiency and vulnerability in social networks. Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers, pp. 253–264. The National Academies Press (2003)
21. Wiil, U.K., Gniadek, J., Memon, N.: CrimeFighter Assistant: A knowledge management tool for terrorist network analysis. In: Proceedings of the International Conference on Knowledge Management and Information Sharing, pp. 15–24. INSTICC Press, Valencia, Spain Oct 2010

A Global Measure for Estimating the Degree of Organization and Effectiveness of Individual Actors with Application to Terrorist Networks

Sara Aghakhani, Khaled Dawoud, Reda Alhajj, and Jon Rokne

Abstract The motivation for the study described in this paper is realizing the fact that organizational structure of a group and critical members of the group are key indicators in determining its strengths and weaknesses. For instance, a general knowledge of the prevalent models of terrorist organizations leads to a better understanding of their capabilities. Though the framework is general, this chapter focuses more on terrorist networks; it is divided in two parts. The first part describes a novel approach for extracting structural patterns of terrorist networks with the help of social network analysis (SNA) measurements and techniques. A global organization measure (Org) is proposed in order to estimate the degree of organization of a social network. The second part contains a new approach which helps to find the group of the most influential people within the terrorist networks. To achieve this target, we utilize SNA measurements and techniques. The importance of such research comes from the fact that individuals in organized intellectual networks and especially terrorist networks tend to hide their individual roles and also, a general knowledge of the prevalent models of terrorist organizations leads to a better understanding of their capabilities. As a result, we argue the need to consider such networks as a whole and at the individuals' level for discovering the degree of organization with its strengths and weaknesses. The reported test results demonstrate the applicability and effectiveness of the analysis as depicted in the proposed framework.

S. Aghakhani · K. Dawoud · J. Rokne
Computer Science Department, University of Calgary, Calgary, AB, Canada

R. Alhajj (✉)
Computer Science Department, University of Calgary, Calgary, AB, Canada
and
Department of Computer Science, Global University, Beirut, Lebanon
and
Department of Information Technology, Hellenic American University, NH, USA
e-mail: alhajj@ucalgary.ca

1 Introduction

Networks can be used for representing various systems in diverse fields wherever it is possible to realize entities which could be connected by certain type of relationship. Networks consist of sets of nodes (interchangeably called vertices) linked together in pairs by edges (interchangeably called links) with nontrivial topological structures [30]. A network may be visualized as a graph and may be represented for processing using a matrix or list structure; the adjacency matrix is the most commonly used representation. Various techniques in graph theory and linear algebra are valuable for network analysis and manipulation. However, a network should be treated within a context in order to be analyzed for knowledge discovery within the specified context. In this respect, social network is one of the most popular network structures attracting considerable attention in the research community. It started as subfield of sociology and anthropology and recently expanded to serve different other applications, including public security by modeling and analyzing criminal and terrorist networks.

A Social network is made up of nodes and edges. In social network analysis (SNA) nodes are individual actors and links represent the relationship between actors. Building a model of a network structure helps in understanding the network and discovering prominent nodes and edges. Nodes may be relevant individually or in groups forming communities. A community is a set of nodes characterized by having more connections within the group and few connections outside the group.

Recently, organized terrorism expanded from the Middle and Far East to cover the developed western countries, e.g., East Europe and USA. Driven by this trend, research on the analysis of terrorist networks has attracted special attention in the last decade, especially after the events of September 11, 2001. This raised high interest in scientific methodologies that could help in systematically fighting terrorism, e.g.,[24–26, 28, 29, 31, 33]. Researchers mostly try to build a social network model of a terrorist organization in a way to track its activities. In this model the nodes of the network are the terrorists and links between them reflect their relationships which can be based on communication, being relatives, or even the places they lived-in, etc. The main target is to find out how members of the organization split into groups to plan for and conduct certain terrorist activities. The ultimate target of the governments, industry and the academia is the development of automated early warning systems with high prediction accuracy [25, 26]. The area is challenging and multidisciplinary, demanding expertise from anthropology, sociology, psychology, ecology, statistics, mathematics, computer science, among others.

Many researchers use graph and network analysis methods as measures to analyze how members of a terrorist network interact with each other and how they split into groups to plan for a terrorist activity [20, 21, 25, 28, 31, 33]. However, a graph model might not be the best representation of organizations such as terrorist organization and threat groups. In the recent work of Farley [9], he explains clearly

that modeling terrorist networks as graphs does not give us enough information to deal with the threat. Modeling a terrorist network as a hierarchy can be a good approach to give an idea about the subgroups present in the network and also how information flows from higher to lower ranks [28]. In other words, hierarchy is another perspective for analyzing a social network. We argue that it is necessary to consider all perspectives at once in order to produce a robust approach that could work as early warning system with high accuracy.

Carley et al. [3, 4] discussed that terrorist networks can be dynamic networks which are always changing. Also, Sparrow [29] suggested that instead of looking at the presence or absence of ties, it may be more informative to look at their strength based on task and timing. The methods proposed in this study are focused on static networks, and hence assume complete knowledge of the network. Here, we consider that all nodes with their ties are present, and then based on the network in hand, we discuss a new measurement. For our approach, it is impossible to consider the dynamic network as all the links and nodes in the network are essential for our measure.

As part of our effort to contribute to this essential field of research, we first introduce a new measure which reports how organized a network is; if the network is more organized, it is more important, and perhaps an organized terrorist network more dangerous. In another way, a well organized network structure leads to the discovery of leaders and important nodes. Concentrating on the structure would lead to vital discoveries that highlight such leaders and important nodes which should be well tracked in order to avoid as much as ever possible any future terrorist activities by their groups. The proposed measure is equally important to analyze other domains that could be modeled as social network; this is well demonstrated in the conducted testing. Second, we present a method which finds all important members of a terrorist network within a group. Here, we developed a method to show that all important members of a terrorist network can be found in one group and in this way the intelligent agencies do not need to search the whole network in order to find important people; here we also show how the network is changing after recursively excluding each important member of the terrorist network. The reported test results demonstrate the effectiveness and applicability of the proposed framework.

The remaining part of this chapter is organized as follows. Section 2 describes the concept of SNA measures utilized in our study; we also cover k-mean clustering. Section 3 includes the proposed measure for analyzing organized network structure (Org). In this section, the results of our proposed measure on three different data sets (two terrorist networks and one business related) are presented as well. In Sect. 4, we described the approach for eliminating key actors; we mainly explain the ranking and clustering of the three data sets. Finally, the summary and conclusion of our work are included in Sect. 5.

2   Social Network Analysis and Clustering

The main idea of SNA is to map and measure the relationships between people, groups, organizations or other information/knowledge processing entities [5, 7, 8, 13, 22, 27]. Entities (interchangeably called individuals or actors) are represented as nodes in the network along with links, with/without attribute weight ages. SNA attempts to provide both mathematical analysis and visual representation of relationships in a network. SNA is an interesting cross-fertilization of sociology, statistics, mathematics, computing, etc., to the benefit of all.

Previously, there have been other approaches applying SNA to understand the network [29]. In one form or another, network analysis has been used to uncover unlawful entities and activities. It has been used for evidence mapping in fraud and criminal conspiracy cases, e.g., [2, 6, 8, 19]. A suspect's network can be built through relational information, including internet blogs, emails, telephone logs, travel bookings, credit card transactions, etc. [21]. More recently, network methods have formed a useful part of intelligence work. As terrorists establish new relations or break existing relations with others, their position roles, and power may change accordingly. These node dynamics, resulting from changes in relationships, can be captured by a set of centrality measures from SNA. The major measures are covered briefly next in this section.

2.1   Betweenness Measure

Betweenness centrality measures the extent to which a particular node lies between other nodes in a network. The betweenness of a node say a, is defined as the number of geodesics (shortest paths between pairs of nodes) passing through a [10, 12, 14]:

$$C_{\mathrm{B}}(a) = \sum_{i}^{n} \sum_{j}^{n} g_{ij}(a) \tag{1}$$

where $g_{ij}(a)$ indicates whether the shortest path between two other nodes i and j passes through node a. A member with high betweenness centrality value may act as a gatekeeper or "broker" in a network for smooth communication (information passing) or flow of goods (e.g., drugs).

2.2   Degree Measure

Degree centrality measures how active a particular node is. It is defined as the number of direct links a node a has [10, 12, 14, 32]:

$$C_{\mathrm{D}}(a) = \sum_{i=1}^{n} c(i, a) \tag{2}$$

where n is the total number of nodes in a network, $c(i, a)$ is a binary variable indicating whether a link exists between nodes i and a. A network member with a high degree centrality value could be the leader or "hub" in a network.

In a directed graph, degree can be classified into in-degree and out-degree to differentiate between links going into and out of a node, respectively.

2.3 Closeness Measure

Closeness centrality measure is the sum of the length of geodesics between a particular node, say a, and all the other nodes in a network. It actually measures how far away one node is from other nodes; and sometimes it is called "farness" [2, 11]:

$$C_C(a) = \sum_{i=1}^{n} l(i, a) \tag{3}$$

where $l(i, a)$ is the length of the shortest path connecting nodes i and a. Some researchers measure closeness by considering the reciprocal of $C_C(a)$ in the above equation. This will lead to a value between zero and one to indicate actual closeness between two nodes.

2.4 Eigenvector Centrality Measure

Eigenvector centrality measures the centrality of an actor based on the centrality of the actors to whom it is connected [12, 32]. In other words, actors who have high eigenvector centrality are connected to many other actors who are connected to many others.

Adjacency matrix is used in order to find the eigenvector centrality measure values. For each node, centrality score is a proportion of the sum of all node's score.

$$X_i = \frac{1}{\lambda} \sum_{j \in M(i)} x_j = \frac{1}{\lambda} \sum_{j=1}^{n} A_{i,j} x_j \tag{4}$$

Here, x_i is the score of ith node and A is adjacency matrix of the whole network, which means that $A_{i,j} = 1$ if the ith node is adjacent to the jth node; and if they are not adjacent, the entry value is zero. Also, $M(i)$ is set of nodes which are connected to the ith node; N represents the total number of nodes and λ is a constant value. There are different eigenvalues λ which generate corresponding eigenvector solutions.

In reality, a person with high eigenvector centrality value can spread information much faster as he/she is well-connected to well-connected people. Actors with high value for eigenvector centrality are more important when rapid communication is needed.

2.5 Authority Measure

The authority and hub measures are related to each other. Entities that have many
other entities point to them are called authorities where entities which point to large
number of authorities are hubs. So, an authority which is referred to with the highest
hubs; it is the highest authority as well [15].

A node is authority central if its in-links are from nodes with many out-links [18].
Kleinberg [18] proposed an iterative algorithm in order to measure authority and hub
degree of each node in a network. In each iteration, it updates the authority value
for node i, denoted $a(i)$, to be the sum of the hub weights of all nodes in j.

$$a(i) = \sum_{j \to i} h(j) \tag{5}$$

In general, network member that acts as authority receives information from
other nodes that each sends information to many other nodes in the network.

2.6 Exclusivity Measure

Exclusivity measure identifies nodes that have links which few other nodes do
have [1].

$$KEI = \sum_{j=1}^{|K|} AK(i, j) \times \exp(1 - \text{sum}(AK(i, j))) \tag{6}$$

where KEI is knowledge exclusivity index for nodes i. Usually, network members
with high exclusivity for knowledge/resources are those ones that have knowl-
edge/resources which few other members do have.

2.7 K-Mean Clustering

K-mean is a partitional clustering algorithm which aims to partition data into k
number of clusters [16, 17, 23]. Clustering is generally unsupervised process in the
sense that no prior information is expected in order to progress towards the solution.
However, partitional clustering approaches like k-mean require the user to provide
the number of clusters as input; then, a centroid is defined for each cluster. The
locations of the centroids constitute very important factor in the partitioning; so, it is
better to place them as far away from each other as possible. Distances are computed
between each data point and all the centroids, and each data point is assigned to the
cluster with the nearest centroid. After all data points are distributed into the clusters,
the centroids should be recalculated. The last two steps (distance computation and

distribution of points) are repeated until the system stabilizes, i.e., data points stick to their clusters. The minimizing function is as follow.

$$\arg \min \sum_{i=1}^{k} \sum_{x_j \in S_i} ||x_j - c_j||^2 \tag{7}$$

where c_j is the centroid of points in S_j.

A good clustering method provides high quality clusters with high intra-class similarity and low-inter-class similarity. In other words, homogeneity (intra-class measurement) and separateness (inter-class measurement) are two factors which can be measured for each number of clusters k. The number of clusters with high value for homogeneity and low separateness is the best number for clustering.

3 Introducing a New Measure

The main idea behind this section is to focus on the whole structure of a network rather than the individual nodes. A terrorist network can be considered dangerous and hence important for further investigation if it reaches to certain level of organizational structure. Modern terrorist organizations have learned that they can effectively counter much larger and conventional enemies using dispersed and networked forms of warfare; striking when their target is least likely to expect it. However, planning a highly organized attack requires highly organized communications, which means, the fact that true leaders could be hiding themselves doesn't really affect the structure of a terrorist network.

3.1 Process Organizational Measure

In this section, we use five different measures on terrorist network structures in order to observe their characteristics. The five measures used in our study are some of the major measures of SNA; namely degree, betweenness, closeness, authority, and exclusivity.

Our proposed method follows the following steps:

1. Rank each node of the network based on each of the mentioned five measures.
2. Start grouping nodes from the lowest values of each measure. The calculated values for each measure tend to be categorized in groups of close values. As using the normalized values are preferred here, and in order to keep the range of the data limited, this approach is considered:
   If $0 < data\text{-}range < 0.1$ then (if $|x_{i+1} - x_i| < 0.01$) put x_{i+1}, x_i in one group
   If $0 < data\text{-}range < 1$ then (if $|x_{i+1} - x_i| < 0.1$) put x_{i+1}, x_i in one group

where x_i is the ith calculated value for the target measure. Also, A is used for set of groups.

3. Each group is given a weight. These weights reflect the influence of each group. The group with the higher weight has more influence. Weights can be given in many ways depending on to what extent high SNA values affect the network. In our study, we assign sequential numbers, where the first group created has the lowest weight and the last one has the highest weight; weights are integer numbers starting from one.
   For each (A_i), (W_i) *denotes the weight of the group* (A_i)

4. Multiply the weight of each group with the number of nodes within the group, which accordingly expresses the influence of the group. Then, add all values together. The result is one single value which is relevant to the size of the whole network.

$$Org(measure) = \frac{\sum_{i=1}^{n} \#nodes\_within\_A_i \times W_i}{N} \qquad (8)$$

Here, n is number of groups and, N is the total number of nodes within the network.

At the end, apply the *Org* function on each of the five previously mentioned SNA measures. Then, sum all Org values for each measurement and find the total measure value for the whole network.

$$Org(network) = Org(betweenness) + Org(degree) + Org(closeness) \qquad (9)$$

$$+ Org(authority) + Org(exclusivity) \qquad (10)$$

A higher value of $Org(network)$ measure reflects higher organized network structure, which represents more important and more dangerous terrorist networks. In the next section, we demonstrate the effectiveness of the proposed method for two different terrorist network data sets. We also computed the new measure for a business data set.

3.2 Testing the Effectiveness of the New Measure

For the testing, we used three different data sets. The first two are terrorist networks data sets, namely 9/11 data set, and Madrid bombing data set. The 9/11 data set consists of 63 nodes of actors suspected to be involved in 9/11 attacks, the 153 links between nodes constructed based on (communication, relatives, places belonging to, and roommates) relationships. Madrid Bombing data set consists of 67 nodes and 89 links, which were constructed by the same way as 9/11 data set. The third data set is related to World Trade of 80 countries and 998 interactions of trading activities between countries. The latter data set is intended to demonstrate the general scope of the proposed measure.

Table 1 9/11 data set measures

	Total degree		Betweenness		Closeness		Authority		Exclusivity	
Rank	Agent	Value	Agent	Value	Agent	Value	Agent	Value	Agent	Value
1	33	1	33	0.141	59	0.059	40	1	21	0.067
2	40	0.857	21	0.074	54	0.057	12	1	5	0.048
3	46	0.619	5	0.043	51	0.049	6	1	46	0.042
4	21	0.524	46	0.043	53	0.047	7	1	54	0.029
5	55	0.476	15	0.041	52	0.047	55	0.872	53	0.023
6	25	0.476	40	0.033	61	0.046	11	0.847	59	0.022
7	41	0.476	14	0.025	62	0.044	17	0.847	61	0.022
8	49	0.381	19	0.024	56	0.043	13	0.847	7	0.016
9	43	0.381	56	0.021	50	0.043	41	0.499	43	0.014
10	31	0.381	43	0.019	49	0.041	57	0.443	9	0.012

First, each of the five selected SNA measures is applied on each data set separately. Then, the organization measure for each of the five measures is calculated; and finally, the overall organization measure is determined. Table 1 represents the first ten top ranked nodes for each of the selected measures for 9/11 data set. Tables 2 represent the same five measures calculated for Madrid data set for the eight top rank nodes.

To avoid any bias in the results and to demonstrate the effectiveness of the proposed measure, we have also applied the same set of experiments on a non-terrorist network (namely World Trade) data set. The results of the top ranking nodes for the five measures are listed in Table 3; values of SNA measures of World Trade data set are observed to gradually decrease from top ranked to least ranked values. Comparing the distribution curves of SNA measures for terrorist networks (highly organized networks) to the World Trade network demonstrates the grouping distribution of terrorist networks; here, it is worth noting that the distribution of the SNA measures in World Trade data set doesn't reflect any grouping factors; this supports the idea of non-organizational structure.

We can notice from the top ranked nodes of each measure, the way nodes are grouped in close values. We believe this trend shows in well-organized social networks, where actors tend to have responsibilities according to their importance within the hierarchy of the network. In Tables 1, 2 and 3, SNA measures of a sample of top ranked actors are not gradually decreased from the top ranked to the least, they tend more to decrease in gaps leaving group of actors having close values together.

As shown in Figs. 1, 2 and 3, we plotted values of each node of the network for total degree measure as compared to normal distribution using the ORA tool in order to show the non-normal distribution. Values plotted in this chart correspond to all actors in the network. Here, gaps are visually observed and the trend is repeated all over the network. The gaps between values of 9/11 data set are less than Madrid data set, and both of them are better than World trade data set. The graph of 9/11 is closer to the line of normal distribution.

Table 2 Madrid data set measures

Rank	Total degree Agent	Value	Betweenness Agent	Value	Closeness Agent	Value	Authority Agent	Value	Exclusivity Agent	Value
1	Madrid Bombings	1	Jamal Zougam Serhane ben Abdelmajid	0.008	Madrid Bombings	0.038	Iraq	1	Madrid Bombings	0.344
2	Jamal Zougam Serhane ben Abdelmajid	0.44	Abdelmajid Fakhet	0.005	Lavapies Madrid	0.0166	Syrian Arab Republic	1	Zarqa Jordan Assasination	0.045
									Plot	0.045
3	Abdelmajid Fakhet	0.19	Jamal Ahmidan	0.002	Madrid Store	0.0164	Jordan Al Qaeda Cell	1	Takfir wal-Hijra	0.03
4	Mustafa Setmariam Nasar	0.19	Spain Abdennabi	0.001	Spain	0.0164	Jamal Zougam	1	Lavapies Madrid	0.015
5	Madrid Store	0.13	Kounjaa	0.001	Belgium	0.0161	Redouan al-Issar	1	Madrid Store	0.013
6	Jamal Ahmidan		Zarqa Jordan Assasination	7E-04	Tunisia	0.0161	Mohommad Bouyeri	1	Moroccan Islamic Combatant Group	0.011
7	Zarqa Jordan Assasination	0.11	Plot Moroccan Islamic Combatant Group	5E-04	Tetouan Morocco	0.0161	Mohamed Bekkali	0.802	al qaeda	0.006
8	Mohamed Bekkali	0.11	Takfir wal-Hijra	5E-04	Morocco	0.0159	Mohamed Chaoui	0.802	Afghanistan	0.006

Table 3 World trade data set measures

	Total degree		Betweenness		Closeness		Authority		Exclusivity	
Rank	Agent	Value	Agent	Value	Agent	Value	Agent	Value	Agent	Value
1	Finland	1	Rep.	0.1124	Finland	0.9744	Finland	0.1159	Slovenia	1
2	Slovenia	0.8817	Iceland	0.0784	Hungary	0.9268	Hungary	0.1155	Brazil	0.5995
3	Iceland	0.4545	Finland	0.0649	Slovenia	0.9157	Salvador	0.095	Finland	0.4778
4	Singapore	0.3861	Mexico	0.061	Singapore	0.8444	Slovenia	0.0905	Singapore	0.3743
5	Hungary	0.3662	Ecuador	0.0501	Chile	0.7917	Belgium	0.0441	Kuwait	0.3092
6	Brazil	0.315	Slovenia	0.0477	Salvador	0.7835	Austria	0.0191	Belgium	0.2623
7	Chile	0.2793	Of	0.04	Iceland	0.7755	Greece	0.0183	Germany	0.2538
8	Kuwait	0.2664	Singapore	0.0356	Belgium	0.7037	Pakistan	0.0183	Austria	0.2411
9	Salvador	0.2481	Moldava.	0.0322	Rep.	0.7037	Chile	0.0092	Rep.	0.2411
10	Belgium	0.2285	Hungary	0.0279	Kuwait	0.7037	Philippines	0.0068	Iceland	0.2165
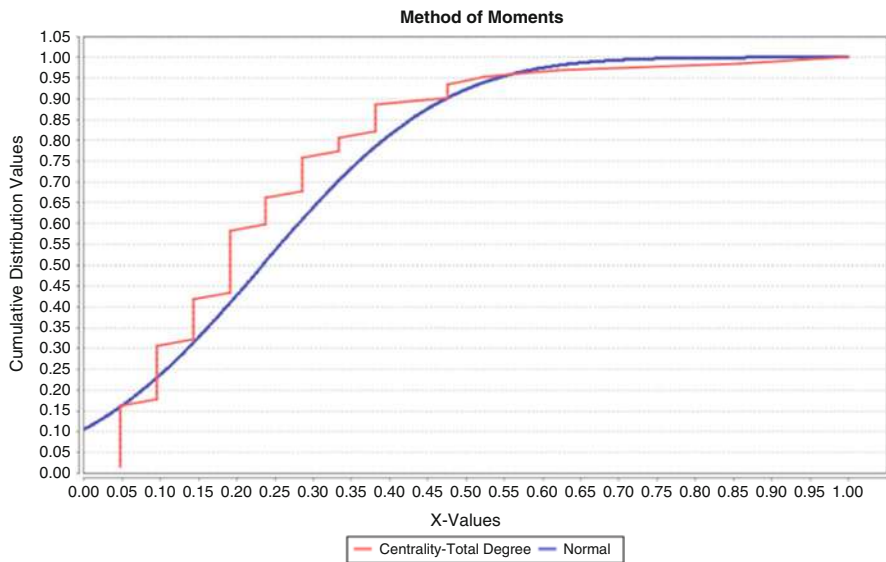
Fig. 1 Total degree distribution of 9/11 data set

Our method tends to transform the degree of organization into a numerical value as demonstrated by the values reflected in the tables and the curves plotted in the figures.

The values of the proposed organization measure are shown in Tables 4, 5 and 6 for the three networks used in the tests. The results show 9/11 total value is significantly higher; this reflects the degree of organization the network did reach to, and as a result became more and more seriously dangerous.

As shown in Tables 4 and 5, generally all measure values of 9/11 are more than Madrid, but the main difference is related to the authority. Based on our method definition, this means the existence of many organized groups within the whole network with different levels of authority, and possibly a s specific task is assigned
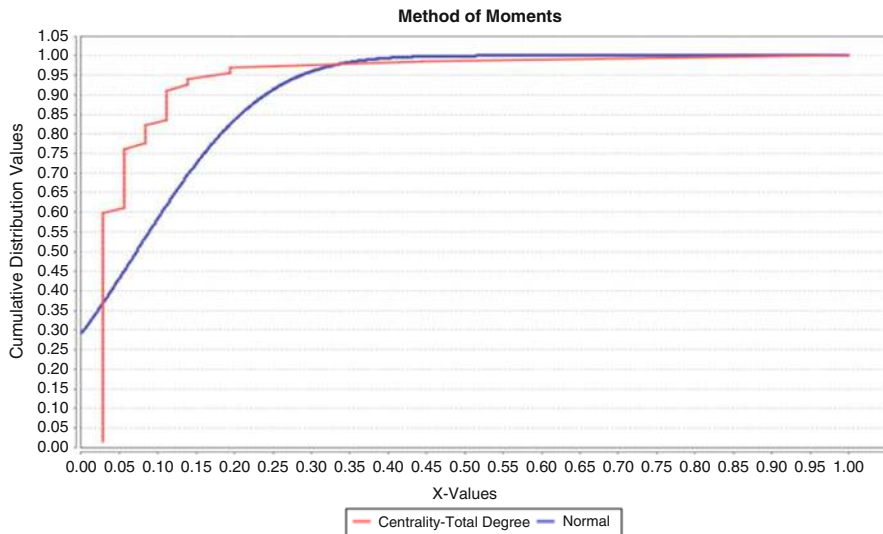
Fig. 2 Total degree distribution of Madrid data set
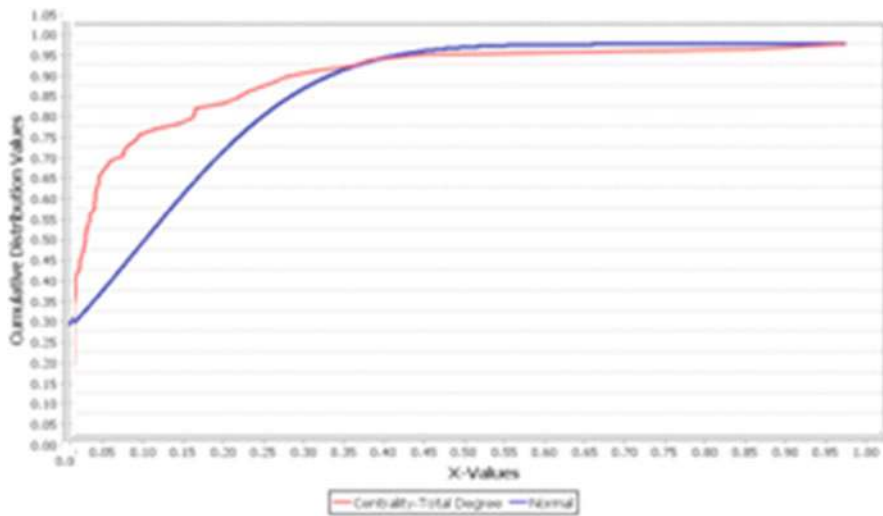
Fig. 3 Total degree distribution of World Trade data set

to each of these groups. This shows how organized and dangerous that network is. Also, value of total degree has more effect on the value of organization; this can make it a dominating factor as well. Finally, the organizational measures for the World Trade data are reported in Table 6; it has the least value in comparison with the two terrorist networks.

Table 4 9/11 data set organization values

	Total degree	Betweenness	Closeness	Authority	Exclusivity	Total
Organization value	4.58	2.171	1.951	10.952	1.482	21.136

Table 5 Madrid data set organization values

	Total degree	Betweenness	Closeness	Authority	Exclusivity	Total
Organization value	2.014	1	1.014	4.243	1.342	9.613

Table 6 World trade data set organization values

	Total degree	Betweenness	Closeness	Authority	Exclusivity	Total
Organization value	1.55	0.757	1.67	1.246	1.636	6.859

Generally, our measure would be helpful in detecting highly organized terrorist networks and as a result preventing possible terror attacks. Further, our approach is different than other approaches studying the structure of networks; we focus on terrorist networks' structure in which we believe the redundancy of connectivity and communication levels play a vital role in composing serious well organized terrorist networks.

4 Effect of Individual Actors on the Network

In this section, we investigate the effect of removing the most important actor(s) from the network. In other words, we want to know if the most important actor disappears (like an important individual in a terrorist network dies) who would be his/her replacement. Even, when the removal is repeated (the disappearing situation happens again), how can we detect this group of actors. Our experimental evaluation has two different parts. The first part follows these four steps:

1. Apply the two selected SNA measures; authority and eigenvector centrality on each of the data sets.
2. Find the actor with the highest value for each of the two measures.
3. Exclude from the network the identified actor with its entire links.
4. Repeat steps 2 and 3 for 4 iterations or until the size of network becomes less than half the size of the whole network (based on each of the measures).

For the second part of our experimental evaluation, we used k-mean clustering on the whole network based on each of the measures separately. The key here is to find the best k for the number of clusters based on homogeneity and separateness of the clusters. Finally, we tried to realize what percentage of these excluded actors are within the same cluster.

4.1   Testing the Influence of Individual Actors

In this section, we used two terrorist network data sets for our experimental evaluations. The first is 9/11 data set and the second one is Madrid bombing data set.

4.2   The 9/11 Data Set

The next subsections represent our method evaluation on 9/11 data set both for authority and eigenvector centrality measures.

4.2.1   Eigenvector Centrality

Here, first we applied the eigenvector centrality measure to the whole network. After excluding the actor with the highest value, we repeat the actor excluding process for four times. Table 7 shows the eigenvector centrality values for the original data set and for four different levels for the top ten actors.

Figure 4 shows how the 9/11 network is changing after excluding the nodes in each level based on the eigenvector centrality measure. It can be easily seen that the size of the network is becoming small and it means that in this way we can lead to elimination of the 9/11 network.

In the second step, we applied k-mean clustering for three values of k on the whole data set based on the eigenvector centrality values. The homogeneity and separateness values for different values of k are shown in Table 8.

As shown in Table 8, the highest homogeneity value is obtained for $k = 2$ clusters, but the separateness is the largest when $k = 3$ clusters; we decided to go for $k = 3$ because the corresponding homogeneity is acceptable. After deciding on the value for k, we searched the excluded actors in each clusters. As a result, we find out that all the excluded actors are within one cluster; this means that if an actor with the highest value based on eigenvector centrality is found; then after omitting that actor, other actors with highest values are within the same cluster in which the first actor was.

Table 7   Eigenvector centrality values for the original network and for four different levels

Rank	Original Agent	Level 1 Agent	Level 2 Agent	Level 3 Agent	Level 4 Agent
1	*33*	*40*	*46*	*55*	*15*
2	40	46	55	49	25
3	41	41	49	53	5
4	46	49	56	54	41
5	25	39	41	44	14
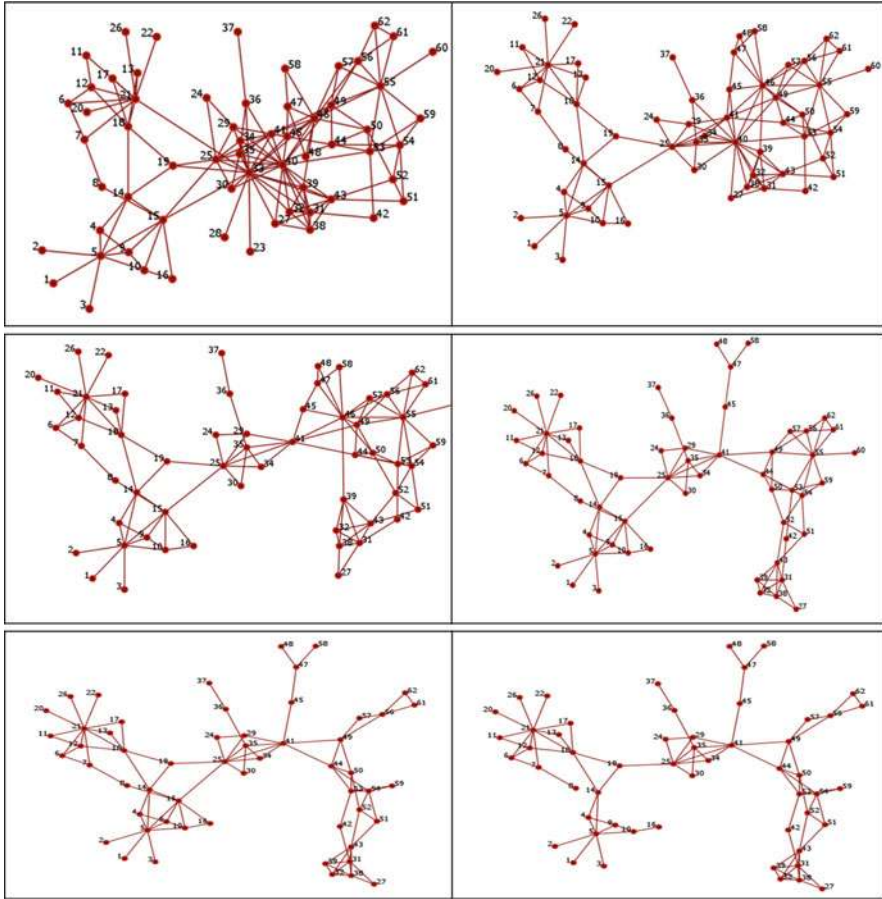
Fig. 4 9/11 network changes after each level node excluding for eigenvector centrality measure

Table 8 k-Mean clustering for eigenvector centrality values

Number of clusters	Homogeneity	Average separateness
2	0.89	0.427
3	*0.523*	*0.613*
4	0.2006	0.52

4.2.2 Authority

In this section, we applied authority measure on the 9/11 data set. The results for the first part of the experiments are reported in Table 9.

Figure 5 shows how the 9/11 network is changing after excluding in each step the node with the highest value for the authority measure. It can be seen that the

Table 9 Authority values for the original network and different levels

	Original	Level 1	Level 2	Level 3	Level 4
Rank	Agent	Agent	Agent	Agent	Agent
1	*33*	*40*	*55*	56	*31*
2	18	12	15	57	34
3	1	6	11	25	4
4	2	7	17	5	10
5	3	55	13	41	32
6	40	11	5	46	29
7	6	17	4	44	36
8	12	13	57	4	38
9	7	41	44	39	35
10	11	57	10	62	30
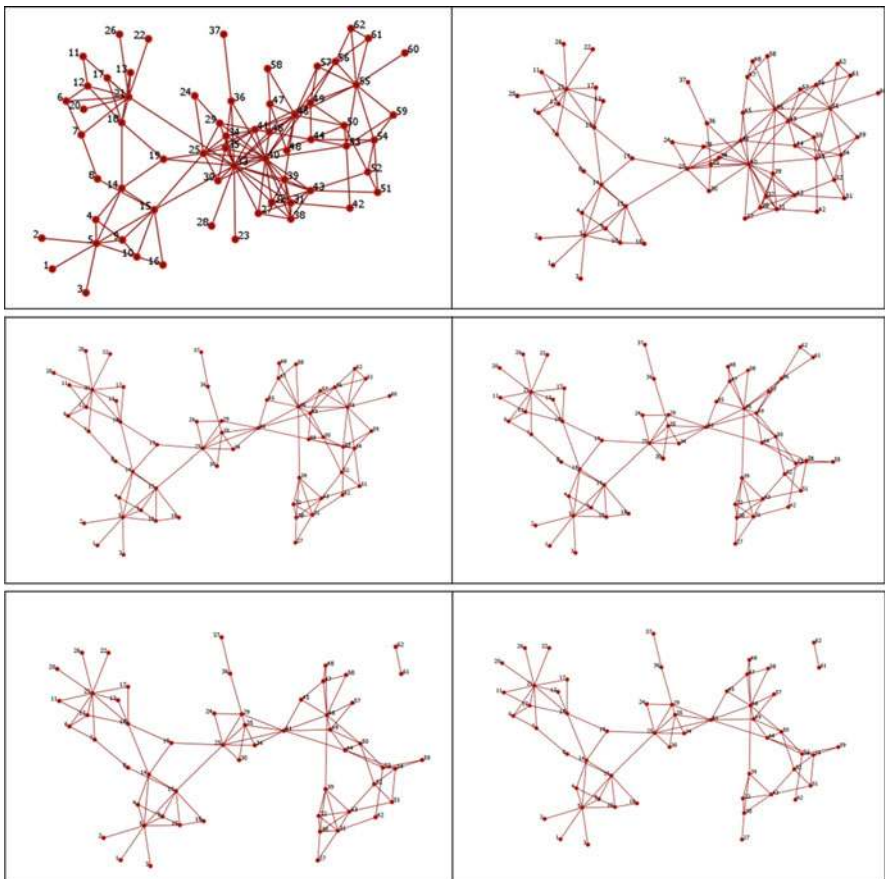
Fig. 5 9/11 network changes after each level node excluding for Authority measure

Table 10 k-Mean clustering for authority values

Number of k-mean clusters	Homogeneity	Average separateness
2	1.316	0.4801
3	*0.516*	*0.5534*
4	0.2201	0.4365

size of the network is becoming small; this means that in this way we can lead to elimination of the 9/11 network.

Then in the second part of our method, we applied k-mean clustering on the whole data set based on the authority measure; we used three different values for k, namely 2, 3 and 4. Table 10 includes the homogeneity and average separateness for different values of k. Here, we also decided on using $k = 3$ clusters because it has the best separateness and acceptable homogeneity as reported in Table 10. Investigating the three clusters, we observed that all the excluded values are within one of the three clusters. This means that those agents who have the most authority in each level are all in one cluster; this helps us to find the group of suspects within one cluster.

4.3   Madrid Data Set

Our method has also been applied on Madrid bombing network based on both eigenvector centrality and authority measures.

4.3.1   Eigenvector Centrality

Table 11 represents the results for the first part of the experimental evaluation for Madrid bombing data set. This evaluation is done for the original data set and three different steps until the size of the data set becomes small based on eigenvector centrality.

In Fig. 6, Madrid bombing network is represented with its changes after excluding the nodes in each level based on eigenvector centrality measure. Here, again the size of the network is reduced significantly.

Then k-mean clustering is applied on Madrid bombing data set based on eigenvector centrality values. As shown in Table 12, the homogeneity of twofold and threefold are very close, also the average separateness for twofold is superior to the rest. Therefore, twofold is considered as the preferred number of clusters.

After exploring each cluster, we realized that all the excluded actors in the first part are within the same cluster among the clusters chosen in the second part.

Table 11 Eigenvector centrality values for the original network and three different levels

Rank	Original Agent	Level 1 Agent	Level 2 Agent	Level 3 Agent
1	*Madrid Bombings*	*Zarqa Jordan Ass.Pilot*	*Madrid Store*	*Morocco*
2	Jamal Zougam	Jamal Zougam	Serhane ben Abdelmajid	Jamal Ahmidan
3	Rachid Adli	Mohamed Moussatten	Spain	Rachid Bendouda
4	Mohamed Bekkali	Brahim Moussatten	Mustafa S. Nasar	Abdennabi Kounjaa
5	Mohamed Chaoui	Takfir wal-Hijra	Morocco	Abu Hafs al-Masri Brigade
6	Serhane ben Abdelmajid Fakhet	Madrid Store	Jamal Ahmidan	Mustafa S. Nasar
7	Jamal Ahmidan	Mohamed Bekkali	Abdennabi Kounjaa	al qaeda
8	Abdennabi Kounjaa	Mohamed Chaoui	Belgium	Rachid Adli
9	Moroccan Islamic Combatant Group	Morocco	Tetouan Morocco	Morata de Tajuna Spain
10	Madrid Store	Rachid Adli	Tunisia	Afghanistan

4.3.2 Authority

Authority is another measure considered for evaluation of our method for Madrid bombing data set. Table 13 represents the results for the first part of our proposed method.

Figure 7 shows Madrid bombing network with its changes after excluding the nodes in each level based on authority measure.

Different numbers of folds for k-mean clustering are shown in Table 14. Here also if the homogeneity is more important, then twofold is suitable. Otherwise, as the average separateness of threefold has the most value, it can be considered as well. In this situation, we searched the excluded nodes within the clusters for both twofold and threefold. The result was really impressing as in both cases all the excluded actors were within the same clusters (Figs. 8 and 9).

4.4 World Data Set

We tried to apply our approach on a non-terrorist network which is world-trade network.

Fig. 6 Madrid network changes after each level node excluding for eigenvector centrality measure (part I of III)

Fig. 6 (continued) (part II of III)

Fig. 6 (continued) (part III of III)

Table 12 K-mean clustering for eigenvector centrality values

Number of clusters	Homogeneity average separateness	
2	0.7637	0.1607
3	0.755	0.1432
4	0.0198	0.506

4.4.1 Eigenvector Centrality and Authority

In this section, we tried to compare the result of a non-terrorist network (World trade network) with two terrorist networks, namely 9/11 and Madrid networks. First, we applied eigenvector centrality and then authority measure on the whole network. Finally, we used k-mean clustering to find the members within one cluster (Table 15).

We realized that our approach is not applicable to non-terrorist networks. We realize that after applying k-mean clustering, mainly most of the nodes are put in one cluster and even though all the excluded nodes are not in the same cluster (Table 16). The main reason is that terrorist networks are more organized as we discussed before and each of their important nodes are real keys of the network; they are more central and our approach mainly was successful for eigenvector centrality which put more focus on the centrality measure of the nodes. The World Trade network is similar to complete graph which means that all nodes are central, and therefore our approach is not suitable for this network (Tables 17 and 18).

Table 13 Authority values for the original network and different levels

Rank	Original Agent	Level 1 Agent	Level 2 Agent	Level 3 Agent	Level 4 Agent
1	Iraq Syrian-Arab Republic	Mustafa S. Nasar	Mohamed Bekkali	Jamal Ahmidan	Serhane ben Abdelmajid
2	Jordan-Al Qaeda Cell	Mohamed Bekkali	Mohamed Chaoui	Serhane ben Abdelmajid	Rachid Bendouda
3	Jamal Zougam	Mohamed Chaoui	Jamal Ahmidan	Abdennabi Kounjaa	Abdennabi Kounjaa
4	Redouan-al-Issar	Serhane ben Abdelmajid	Vinay Kohly	Bendouda	Vinay Kohly
5	Mohomad Bouyeri	Jamal Ahmidan	Suresh Kumar	Rabei Osman el Sayed	Suresh Kumar
6	Mohamed Bekkali	Vinay Kohly	Serhane-ben Abdelmajid	Rachid Adli	Hasan-El Haski
7	Mohamed Chaoui	Suresh Kumar	Abdennabi Kounjaa	Hasan El Haski	Youssef Belhadj
8	Mustafa Setmariam Nasar	Abdennabi Kounjaa	Rachid Bendouda	Youssef Belhadj	Moroccan Islamic C. Group
9		Rachid Bendouda	Rabei Osman el Sayed	Vinay Kohly	Rachid Adli
10	Vinay Kohly	Moroccan Islamic C. Group	Rachid Adli	Suresh Kumar	Rabei Osman el Sayed

5 Conclusions

Generally, a clear understanding of network structures and individual roles can help law enforcement and intelligence agencies to develop effective strategies in order to prevent future terrorist attacks. More organized terrorist network structure is considered to be more important and more dangerous. SNA is widely used for discovering important agents; the growing awareness of terrorists makes it more difficult to discover important agents, since they tend to hide themselves. We argue that a measure of degree of organization of the whole network without concentrating on discovering certain individuals would help in discovering and predicting terrorist networks. Using SNA measures to extract a measure referring to the degree of organization of the whole network is the technique we used. We tested our measure on three different networks to validate the proposed approach for predicting the organizational structure of a network. To avoid the biased, we applied the organization measure on world trade network as well. The measure had higher value for 9/11 network; the result shows how organized this network is.

Fig. 7 Madrid network changes after each level node excluding for authority measure (part I of III)

Fig. 7 (continued) (part II of III)

Fig. 7 (continued) (part III of III)

Number of clusters	Homogeneity	Average separateness
2	0.11103	0.1603
3	0.03449	0.2765
4	0.0104	0.2286

Table 14 K-mean clustering for authority values
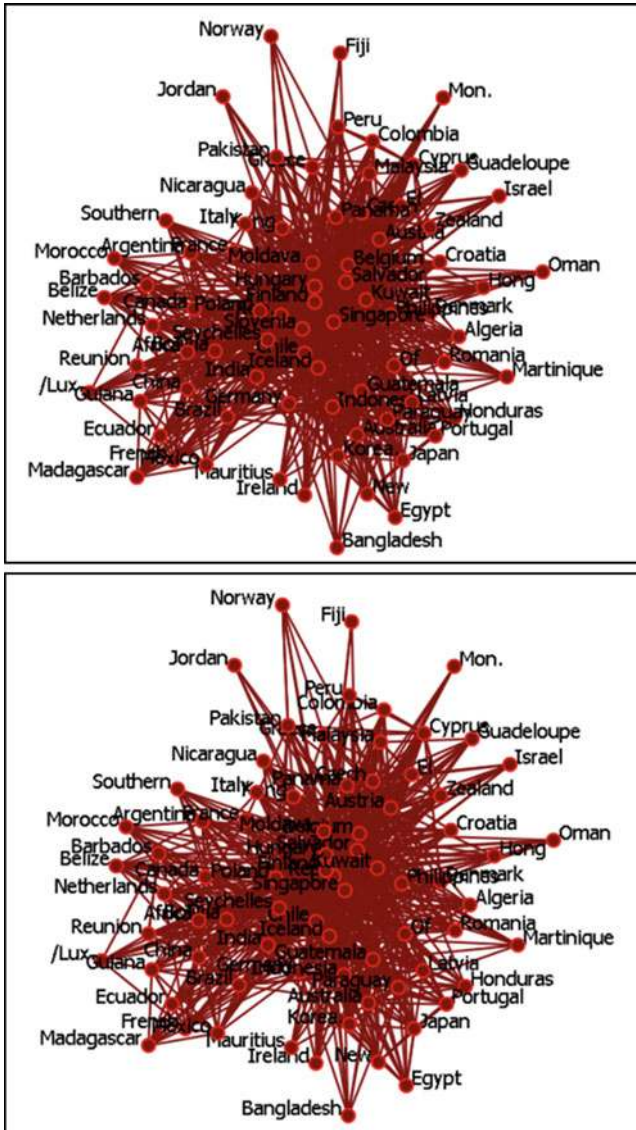
Fig. 8 World trade network changes after each level node excluding for authority measure (part I of III)
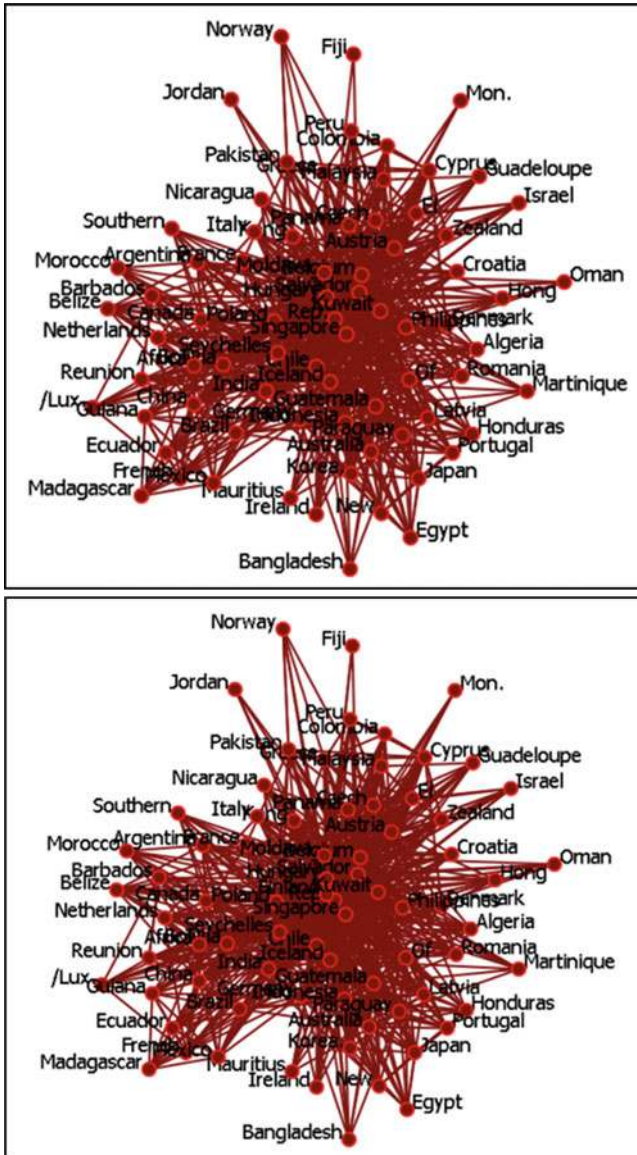
Fig. 8 (continued) (part II of III)
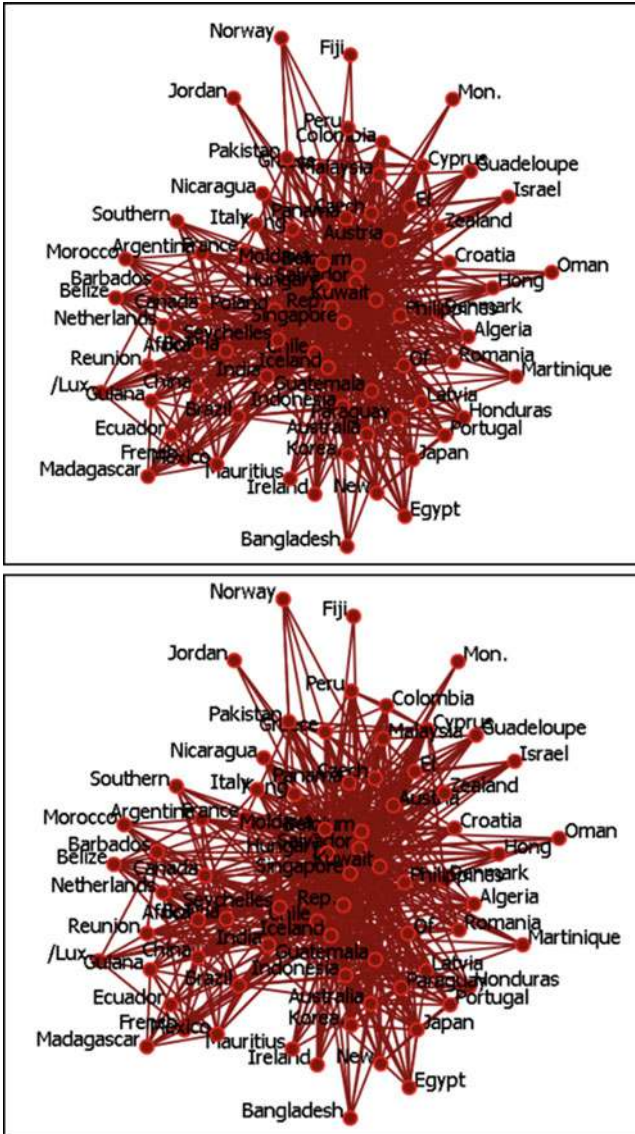
Fig. 8 (continued) (part III of III)
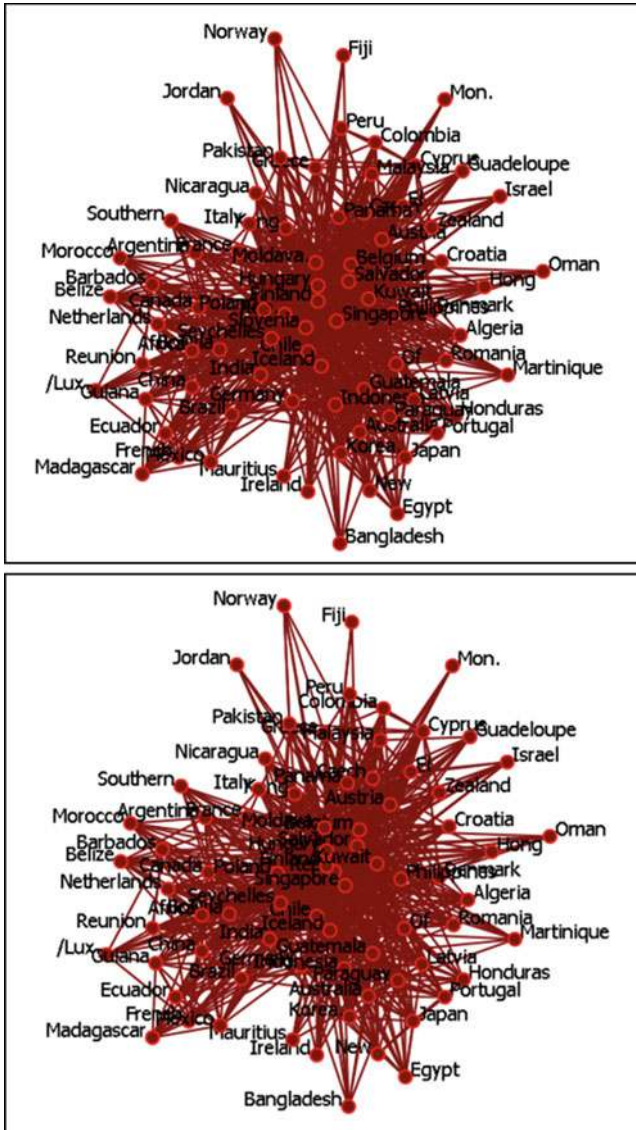
Fig. 9 World trade network changes after each level node excluding for eigenvector centrality measure (part I of III)
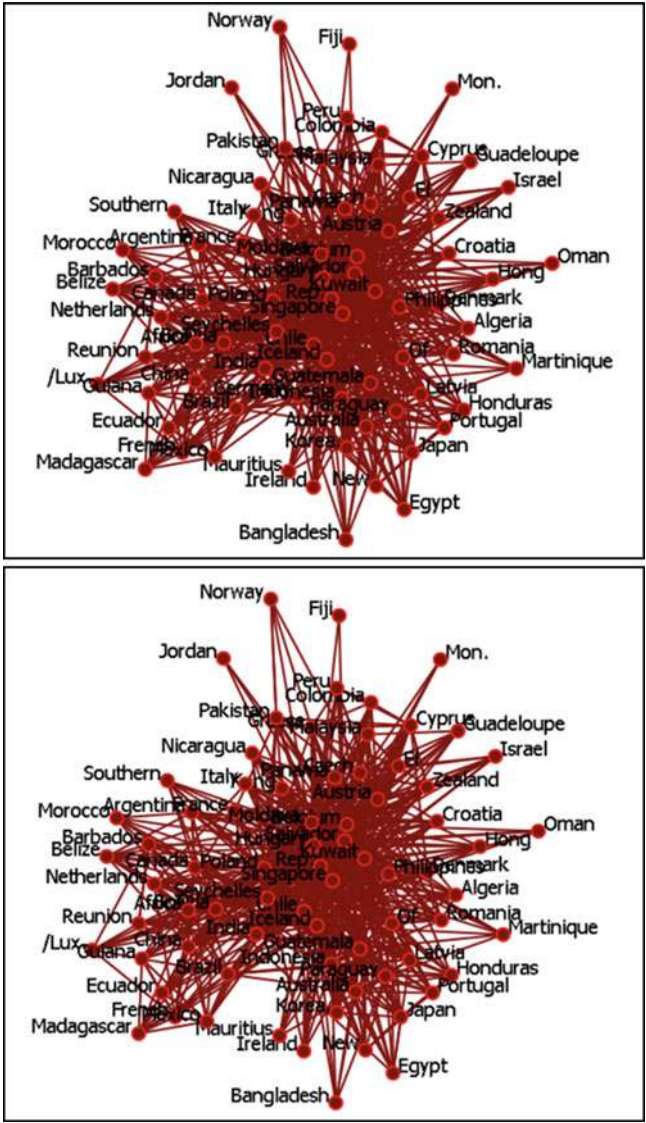
Fig. 9 (continued) (part II of III)
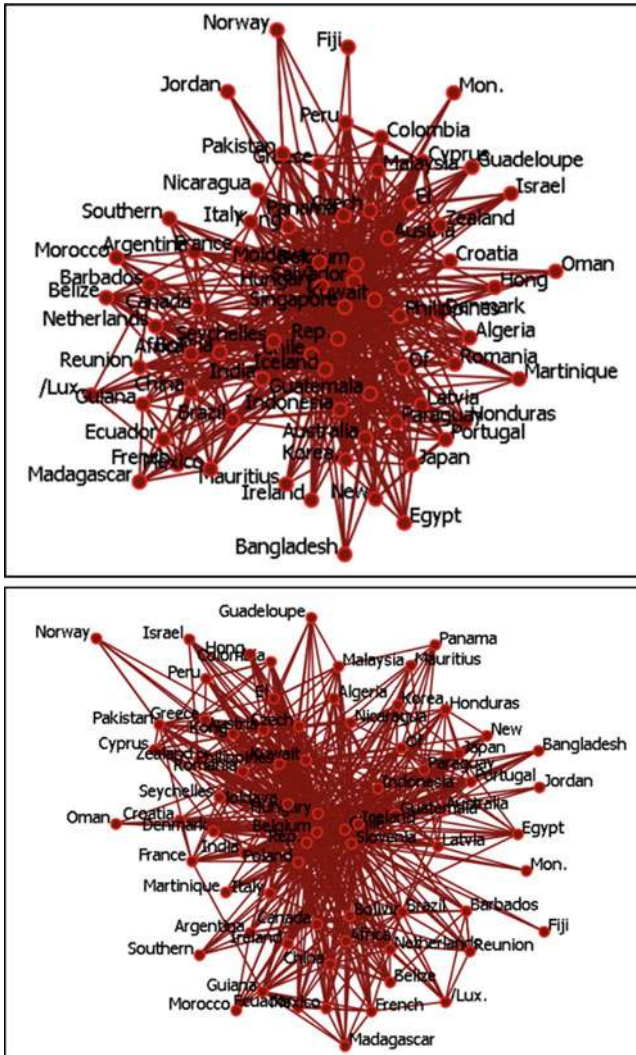
Fig. 9 (continued) (part III of III)

Also, having a good recognition of the terrorist networks, help intelligence agencies to develop strategies to prevent future attacks. In this paper, we presented and evaluated a novel approach to find the most important actors in a terrorist network. Two SNA measures are used to rank the actors. Then, k-mean clustering is applied to group the actors in the network based on each of the SNA measures. We found that the important actors in each level of ranking are within the same cluster. This means that if an important actor is found, then knowing which cluster

Table 15 Authority values for the original network and for four different levels

	Original	Level 1	Level 2	Level 3	Level 4
Rank	Agent	Agent	Agent	Agent	Agent
1	Finland	Finland	Germany	Singapore	Of
2	Hungary	Kuwait	Of	Belgium	Belgium
3	Salvador	Singapore	Singapore	Of	Indonesia
4	Slovenia	Belgium	Belgium	Poland	Japan
5	Belgium	Austria	Indonesia	Kuwait	Poland
6	Austria	Rep.	Paraguay	Indonesia	Paraguay
7	Greece	Hungary	Poland	Hungary	Kuwait
8	Pakistan	Poland	Kuwait	Paraguay	Chile
9	Chile	Moldava	Japan	Japan	Hungary
10	Philippines	Germany	Iceland	Rep.	Rep.

Table 16 Eigenvector centrality values for the original network and four different levels

	Original	Level 1	Level 2	Level 3	Level 4
Rank	Agent	Agent	Agent	Agent	Agent
1	Finland	Germany	Singapore	Of	
2	Hungary	Kuwait	Of	Belgium	Belgium
3	Salvador	Singapore	Singapore	Of	Indonesia
4	Slovenia	Belgium	Belgium	Poland	Japan
5	Belgium	Austria	Indonesia	Kuwait	Poland
6	Austria	Rep.	Paraguay	Indonesia	Paraguay
7	Greece	Hungary	Poland	Hungary	Kuwait
8	Pakistan	Poland	Kuwait	Paraguay	Chile
9	Chile	Moldava	Japan	Japan	Hungary
10	Philippines	Germany	Iceland	Rep.	Rep.

Table 17 k-Mean clustering for eigenvector centrality values

Number of clusters	Homogeneity	Average separateness
2	*0.666*	*0.765*
3	0.472	0.170
4	0.413	0.183

Table 18 k-Mean clustering for authority values

Number of clusters	Homogeneity	Average separateness
2	*0.733*	*0.795*
3	0.476	0.173
4	0.405	0.183

he belongs to, leads us to find other important actors as well. This method helps intelligence agencies to find the important actors more easily. Instead of searching the whole network, they can put their focus on just a part of the network. This significantly reduces the time and the money needed to be spent in order to find these actors. Also, our results show that applying eigenvector centrality measure is more useful in reducing the size of terrorist networks, and finally eliminating them.

References

1. Ashworth, M.J.: Identifying key contributors to performance in organizations: The case for knowledge-based measures. In: NAACSOS Conference Proceedings (2003)
2. Baker, W.E., Faulkner, R.R.: The social organization of conspiracy: Illegal networks in the heavy electrical equipment industry. Am. Sociol. Rev. 837–860 (1993)
3. Carley, K.M.: Dynamic network analysis. In: Breiger, R., Carely, K.M. (eds.) In the Summary of the NRC Workshop on Social Network Modeling and Analysis. National Research Council (2003)
4. Carley, K.M., Reminga, J., Kamneva, N.: Destabilizing terrorist networks. In: NAACSOS Conference Proceedings. Pittsburgh, PA (2003)
5. Carrington P.J., Scott J., Wasserman S.: Models and Methods in Social Network Analysis. Cambridge university press, Cambridge (2005)
6. Criminal Network Analysis Training Course. Defense Intelligence Agency (2000)
7. Degenne, A., Forsé, M.: Introducing social networks. Sage, London (1999)
8. Doreian, P., Stokman, F.N.: Evolution of Social Network. Gordon and Breach Publishers, Amsterdam (1997)
9. Farely, D.J.: Breaking Al Qaeda cells: A mathematical analysis of counterterrorism operations. Stud. Confl. Terrorism **26**, 399–411 (2003)
10. Freeman, L.C.: A set of measures of centrality based on betweenness. Sociometry **40**, 35–41 (1977)
11. Freeman, L.C.: Centrality in social networks I: Conceptual clarification. Soc. Netw. **1**, 215–239 (1979)
12. Freeman, L.C.: The gatekeeper, pair-dependency and structural centrality. Qual. Quant. 585–592 (1980)
13. Freeman, L.C., White, D.R., Romney, A.K.: Research Methods in Social Network Analysis. Transaction Publishers, New Brunswick, N.J. (1992)
14. Jialun, Q., Xu, J.J., Daning, H., Sageman, M., Chen, H.: Analyzing terrorist networks: A case study of the global Salafi Jihad network. In: Proceedings ofIEEE International Conference on Intelligence and Security Informatics, pp. 287–304. Atlanta, GA (2005)
15. Jung, J.J.: Query transformation based on semantic centrality in semantic social network. J. Univers. Comput. Sci. **14**(7) (2008)
16. Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C., Silverman, R., Wu, A.Y.: The analysis of a simple k-mean clustering algorithm. In: Proceedings of 16th Annual Symposium of Computational Geometry, pp. 100–109. (2000)
17. Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R., Wu, A.Y.: An efficient k-mean clustering algorithm: Analysis and implementation. IEEE Trans. Pattern Anal. Mach. Intell. **24**(7), 881–892 (2002)
18. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. In: Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 668–677. (1998)
19. Klerks, P.: The network paradigm applied to criminal organizations. Connections **24**(3) (2001)
20. Klerks, P.: The network paradigm applied to criminal organizations. Connections **24**(3) (2001)
21. Krebs, V.: Mapping terrorist networks. Connections **24**(3) (2002)
22. Knoke, D., Yang, S.: Social network analysis. Series: Quantitative applications in Social Sciences. Sage publication, London (2008)
23. MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297. Berkeley, University of California Press (1967)
24. Memon, N., Larsen, L.H.: Practical algorithm for destabilizing terrorist network. Intelligence and Security Informatics. Lecture Notes in Computer Science, vol. 3975, pp. 389–400. Springer, Berlin (2006)

25. Memon, N., Wiil, U.K., Qureshi, A.R.: Design and development of an early warning system to prevent terrorist attacks. In: Proceedings of the International Conference on Artificial Intelligence and Neural Networks, pp. 222–226. (2009)
26. Memon, N., Wiil, U.K., Alhajj, R., Atzenbeck, C., Harkiolakis, N.: Harvesting covert networks: The case study of the iMiner database. J. Netw. Virtual Organ. **8**(1/2), 52–74 (2011)
27. Scott, J.: Trend report: Social network analysis. Sociology 109–27 (1998)
28. Shaikh, M.A., Wang, J.: Discovering hierarchical structure in terrorist networks. In: Proceedings of the International Conference on Emerging Technologies, pp. 238–244. (2006)
29. Sparrow, M.K.: The application of network analysis to criminal intelligence: An assessment of the prospects. Soc. Netw. **13**(3), 251–274 (1991)
30. Strogatz, S.H.: Exploring complex networks. Nature **410**, 268–276 (2002)
31. Tsvetovat, M., Carley, K.M.: Structural knowledge and success of anti-terrorist activity: The downside of structural equivalence. J. Soc. Struct. **6**(2) (2005)
32. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Cambridge University Press, Cambridge (1994)
33. Xu, J., Chen, H.: CrimeNet explorer: A framework for criminal network knowledge discovery. ACM Trans. Inform. Syst. **23**(2), 201–226 (2005)

Counterterrorism Mining for Individuals Semantically-Similar to Watchlist Members

James A. Danowski

Abstract A key counterterrorism problem is how to identify people that should be added to a watchlist even though they have no direct communication with its members. One of the main ways a watchlist is expanded is by monitoring the emergence of new persons who establish contact with those on the list. Unfortunately, this severely limits the time horizon for managing risks of dark network behaviors because individuals are already actively involved with one another and more likely to be discussing and planning terrorist actions. In contrast, a wider time horizon results from identifying individuals who do not yet have communication with watchlist members, while they have highly similar semantic networks. Discussion forums are considered a primary source of intelligence about plans for dark behaviors. The research reported here develops a method for locating individuals in discussion forums who have highly similar semantic networks to some reference network, either based on watchlist members' observed message content or based on other standards such as radical jihadists' semantic networks extracted from messages they disseminate on the internet. This research demonstrates such methods using a Pakistani discussion forum with diverse content. Of those pairs of individuals with highly-similar semantic networks, 61% have no direct contact in the forum. It is likely that adding to watchlists individuals who have a high match to a reference semantic network lengthens the time horizon for identifying high risk dark behaviors.

J.A. Danowski (✉)
Department of Communication, University of Illinois at Chicago, Chicago, IL, USA
e-mail: jimd@uic.edu

1 Introduction

1.1 Chapter Focus: Counterterrorism Text Mining to Find Similar Semantic Networks

The focus of this chapter is on text analysis using methods of social network analysis. Before more fully conceptualizing semantic networks, I highlight the counterterrorism strategy to which semantic network analysis will be adapted. One probably has a good sense of network analysis to facilitate comprehension of basic ideas framing this chapter.

The motivation for the research reported here came from a talk at the European Intelligence and Security Informatics 2008 conference in Esbjerg, Denmark. It was not my talk but an intelligence analyst's. Later in this chapter I, will tie together the nature of my counterterrorism paper with the more recent evolution of my research. I was there to report on research applying my optimal message generation software, OptiComm [20] to a corpus of approximately 8,000 international news stories about Al Qaeda. The objective was to explore the formulation of messages that might be appropriate for a counterterrorism public information campaign. In the aggregate word network I identified the shortest paths across the stories between the words 'Al Qaeda' and 'bad.' [19]. Another goal of this study was to automatically create optimal messages as a basis for a subsequent experiment on how well these messages worked compared to control messages.

Because in the international news there were many positive stories about Al Qaeda along with the negative ones, the words 'good' and 'bad' were directly connected. As a result every potential optimal message to move Al Qaeda closer to 'bad' included the word 'good.' An example is, "Al Qaeda [is] good [because they] kill Iraqi soldiers." This suggested modifying the shortest path algorithm to include obstacles to be skirted.

The prime motivator for the current chapter derived from a talk by another conference speaker, a counterterrorism specialist, who asserted that online discussion forums were the best sources of counterterrorism intelligence. Despite the value of forums, he indicated that there was a key problem: how to effectively find individuals to add to a watchlist when these potential new members had no direct communication with watchlist persons. At the same time another problem was that watchlist individuals who were in frequent contact frequently changed their login IDs by switching them with other participants for a day or two, and also using an array of aliases, to make surveillance more difficult.

These problems created a challenge that linked to work I had previously done [16] whose purpose was to develop a method to measure the semantic similarity of potential new hires to a reference group of successful current employees. This enables hiring people who already encode messages like the reference cohort does. The summer prior to the Esbjerg conference I was revisiting that work and planning on extending it to the online discussion forum domain. Not quite sure how I was

going to structure that analysis, the talk at the conference lit a bright conceptual light bulb for me, powered by the realization that I could tailor my analysis to address the challenge about expanding watchlists to include individuals who were not in direct contact. The potential to further contribute to counterterrorism measures energized my plans to do semantic similarity research with discussion forum posters. This chapter resulted.

1.2 Benefits of Semantic Network Analysis to Counterterrorism Intelligence

It has become well established that open-source intelligence content, when processed with tools based on social network analysis, provides opportunities for analysts to extend their predictive time horizons further into the future, to increase the probabilities of spotting appropriate individuals to watch more closely, and to better assess and manage risks. We know from education and psychological research [48] that the act of encoding information, particularly writing messages, leads to substantial changes in the conceptual structure because of reflection leading to self-reinforcement, and the strengthening of conceptual and affective associations in the person's internal information network.

Moreover, public expressions of opinion increase the likelihood of more active subsequent behaviors consistent with the opinion. This is because the individual knows he has made a commitment visible to others. He feels a tacit social influence to behave in a manner consistent with these proclamations.

The persuasion research of the World War II and post-war period well documented these effects of social comparison and reference group influence tied to strengthening the underlying attitudes on which the public expressions were based [35]. An example of Cold War American domestic propaganda that utilized these processes was the television public service advertising campaign "Stand Up and Be Counted for America" in which an initially seated audience was challenged to stand up and be counted as a large American flag waved in a back frame layer. This advertisement appeared to be based on the assumption that through vicarious means of parasocial interaction [30], the home audience that identified with the on-screen audience members who rose from their chairs was feeling some degree of social pressure to conform to the patriotic theme. Mediated public opinion displays probably generally have less strong effects than live face time with others. Nevertheless, identification with others on the screen engaging in public opinion expression is probably somewhat effective. Most of the persuasion research of the time presented mediated stimuli to participants and there were numerous findings of factors influencing persuasive effects.

To consider the essence of social influence processes fostered by public communication, recall your experiences in a live audience for performances when some audience members begin to stand and clap for an encore. Those closest to

the performers typically stand first. This creates social pressures for those close to the initial standers to also stand. As the wave moves through the audience coming closest to your seated position, you most likely also rise. Research on such intra-audience effects shows that audience members are more likely to think the performance was better, in other words feel more favorable attitudes toward it, because of this positive public expression [28]. Based on these findings one can also expect that if members of a live audience exhibit some public displeasure toward some object, others closest to them feel the urge to join in the expression of negative opinion. A field experiment was headed by my graduate school office mate [28], whose treatment conditions were created through use of confederates. In one treatment condition, the confederates in the audience stood and cheered for the bar band with a former Top 40 record achievement, the Woolies. In the control condition, the confederates sat quietly. My colleagues found significantly more positive feelings toward the band performance among other audience members on nights when the confederates publically expressed their pleasure than when they sat unmoved.

Attitudes are typically conceptualized as predispositions to respond positively or negatively to some object. Online public writing and self-reflection increases the strength of attitudes underlying the encoded messages. An important form of open-source intelligence, therefore, is to monitor the message content of online forums. The most readily observable online behaviors are writing posts and replying to others' messages, forming a direct communication link. Nevertheless, these actions result in the shortest forward time horizons.

In contrast, the main premise of this chapter is that the counterterrorism analyst gains a longer predictive time horizon when studying individuals who post similar messages to those of other individuals of interest, yet who are currently not in direct communication with one another. There are no post-reply cycles between these individuals. Given the conceptual and attitudinal principles discussed for public communication made in direct contact with others, it is likely that even online public posts made in relative isolation have some degree of social effects on the poster. This is likely because of the poster's imagined others perceiving his online behavior. Internal network crystallization will likely still take place to an extent worth monitoring. Moreover, there is an increased probability that such individuals will eventually come into direct contact with others of interest who espouse similar opinions. A key reason is that individuals' identities are social constructed. This is why monitoring semantically-similar individuals who are not in direct contact with one another lengthens the analysts' predictive time horizon for a dark behavior of interest. These apparently isolated individuals serve as early warning signposts for their more likely future direct coordinated social actions. In addition, the extent to which there are greater numbers of individuals in this incubated state, the greater the likelihood of larger waves of later social mobilization.

In short, individuals with more similar semantic networks based on their encoding of messages are likely to perceive the world more similarly and behave more similarly. Nevertheless, to assume semantic network similarity from frequency of direct communication exchange is a pursuit fraught with potential error because at

a single point in time two communicating individuals may, for example, be arguing different points of view using different message constructions. More appropriate to identifying networks of similar individuals for improving counterterrorism effectiveness is analyzing the similarity of their semantic networks without their direct interaction.

These benefits from identifying semantically-similar individuals who are not in direct contact, suggests that this kind of observation is more likely to be cost-effective. This is because the collection and processing of information to identify such individuals is readily automated on a large scale, leaving only the final interpretation to the analyst. Moreover, these semantically-similar persons can be automatically added to automated watchlists.

This kind of counterterrorism intelligence process is virtually open-source in nature. Practically, it is not fully open-source practice because to efficiently obtain large corpora of discussion posts the analyst deploys special, although simple, software to crawl the discussion forum sites, most of which do not block bots because Google uses them to update page-ranks. This can be done without the knowledge and consent of the forum participants because even if a forum requires creation of a login ID and password, this does not render the information private. Such forums are still widely considered to be public. Nevertheless, some academic research communities, for example, the Association of Internet Researchers, considers the extraction of such public information without the consent of the participants to possibly be unethical.

Do participants in this environment assume/believe that their communication is private? If so – and if this assumption is warranted – then there may be a greater obligation on the part of the researcher to protect individual privacy in the ways outlined in human subjects research (i.e., protection of confidentiality, exercise of informed consent, assurance of anonymity – or at least pseudonymity – in any publication of the research, etc.) [23], p. 7

Such a view, however, is not, for example, shared by most U.S. university Institutional Review Boards (IRBs) for the protection of human subjects, although this is not the case in all countries. As an American researcher, whose own IRB considers extraction of discussion forum posts as exempt from human subjects protections, I therefore report in this research the forum participants IDs because it adds face validity to the findings.

In contrast, all persons who are citizens of the European Union enjoy strong privacy rights by law as established in the European Union Data Protection Directive (1995), according to which data-subjects must:

- Unambiguously give consent for personal information to be gathered online
- Be given notice as to why data is being collected about them
- Be able to correct erroneous data
- Be able to opt-out of data collection
- Be protected from having their data transferred to countries with less stringent privacy protections. (See http://www.privacy.org/pi) [23], p. 7

A side benefit of the protocols of many forums that require registration and use of a login ID and password is that it probably fosters a sense for authors of posting in a protected space where one can express opinions that presumably will be seen only by other forum members. This is likely to increase the propensity to write more impulsively with less self-censorship. This probably increases the intelligence value of the information because it is therefore more likely to be valid in representing the underlying attitudes of the participants.

The volume of information that can be extracted from forum sites is considerable. Researchers at the University of Arizona [11] report having obtained nearly 13 million posts from 28 jihadist portals, an average of nearly half a million posts per portal. Despite these volumes of messages, it is not difficult to quickly write software to group all of the posts made by an individual into one file, conduct word-cooccurrence network analysis at the individual level, then compare the semantic networks of pairs of individuals using a well-established similarity index, such as the Pearson correlation produced by the Quadratic Assignment Procedure (QAP) [32, 37]. At the same time, one can automatically identify which individuals are in direct contact through replying to another's post. This enables a network analysis of persons in addition to of words. When used together, one then knows who the most similar individuals are to others of interest, even though these individuals are not in direct contact.

When one has a predefined list of individuals of interest from a forum, the computation resources required may be relatively easily provided by a standard PC in a short compute time. Nevertheless, to demonstrate the case where one has not yet created a watchlist, and therefore must compare many more pairs of authors to one another, the research reported here compared each of the 3,272 other forum participants, resulting in 5.35 million paired comparisons of authors' semantic networks. This took 11 days to process on a common PC. While feasible and producing useful results whose shelf-life is probably much longer than the compute time, the unacceptable run time immediately motivated me to try to cut it by at least two orders of magnitude. After more systematically conceptually defining semantic network analysis, and reviewing relevant research on semantic similarity I present the research methods, results, and discussion.

1.3 Conceptualizing Semantic Networks

Imagine a large group of people using the same language over time [17]. Assume that the full text of their messages is available to you in natural language form. How would you come to some representation of what they are talking about? Your first thought may be to use traditional content analysis methods that categorize text, either manual procedures or computerized ones like the General Inquirer [49] or it's more recent cousin LIWC [43], or topic-modeling software based on Bayesian statistical procedures [2]. These automated procedures, while computationally

sophisticated, are relatively crude at the conceptual level. They merely assign message elements or individuals to a limited number of nominal categories.

Instead of categorizing messages, a network perspective captures the relationships among words within the messages. Word-pair link strength is the number of times each word occurs closely in text with another. All possible word pairs have an occurrence distribution whose values range from zero on up. This ratio scale of measurement allows the use of sophisticated statistical tools from social network analysis toolkits. These enable the mapping of the structure of the word network. They can also identify word groups, or clusters, and quantify the structure of the network at different levels. Using these word-pair data as input to network analysis tools, one maps the language landscape. On the map, instead of cities as nodes there are other kinds of words. Rather than roads, there are links or edges among words.

Travelling through the word network are fleets of social objects. These communication vehicles are the concepts, ideas, or physical things that people linguistically describe. They link words to these vehicles in the course of their everyday informal and formal communication. This creates energy that propels them through the network. Sometimes these movements are unplanned. At other times, groups or organizations try to manage vehicular traffic. By means of optimal messages, they try to steer vehicle traffic away from certain words or toward them.

Mathematically-based procedures have been developed to create optimal messages [13, 20] constructed through systematic analysis of the paths connecting word nodes of interest. The procedures identify the optimal association network across the aggregated messages of the social community. The underlying assumption is that stimulating associations across it is more effective as the shortest effective sequence of words based on particular constraints is selected for the message. This is because people process strings of words linearly over time, encoding and decoding them in ordered sequences. Furthermore, the triggering of associations to words in context takes cognitive time. The most effective and efficient messages, therefore, optimize the association networks in the receivers' minds as they read or hear these messages.

In short, this paper focuses on the similarity of messages encoded by individuals, in other words, individuals' semantic network similarity. This constitutes a new kind of network variable, in addition to cohesion based on actual communication exchange, and in addition to structural equivalence based on similarity of network position. This new network variable is semantic equivalence. Some may think this construct means entities have the same linguistic tags or keywords from a controlled vocabulary, such as the same name for a person, organization, or object, or the same keywords produced manually that describe a document. These tags are not semantic network characteristics but instead are semantic attributes of some entity. They are akin to the words in a dictionary or elements of an ontology. In contrast, one can consider the networks among semantic elements encoded by persons or other social units. Our interest is in semantic encoding similarity.

People who string words together in a similar manner are similar to one another in speech/act behaviors. We assume that language reflects perception [34] and behavior to a considerable extent. The people who write more like each other are likely to behave more similarity given similar contextual circumstances. This is probably

because they perceive their environments and their choices for behavior within them more similarly.

Time frame is also relevant. If I write like a terrorist, I may currently be a terrorist, I may become a terrorist, or I previously was a terrorist. Because message strength decays over time in relation to inner and outer states of contextual structures, we are less interested in former members of a class of individuals based solely on their communicative similarity. If we are tracking semantically similar people, such as for expanding a watchlist, we want to identify people who currently encode messages like our exemplars do. Or, we may want to become more active and covertly connect with these individuals to see who has ill intent and whether they are willing to fully implement their intentions, then thwarting their accomplishment of their desired dark behaviors just short of the terrorist's anticipated end effects.

1.4 Related Work

When computer bulletin boards emerged as dialup information servers in the late 1970s I computed the similarity among discussion forum messages based on the extent to which they shared similar concepts, coding messages posted to a forum and the replies to them as constituting message pairs for comparison [12, 13, 15].

Monge and Eisenberg [42] subsequently network analyzed organizational members based on how similar their perceptions of the corporate mission were. They used a categorical form of manual content analysis of brief open-ended statements respondents wrote in a self-administered survey.

Latent Semantic Indexing (LSI) uses semantic similarity to index documents, but uses a highly reduced dimensionality in the network [22]. Similar in scope, but using tensors rather than linear dimensions is an approach demonstrated in the context of indexing the similarity of blogs, based on both direct links and on indirect semantic similarities [40]. An entirely different level of semantic analysis was the focus of Resnik [45] who defined semantic similarity as the extent to which two words share links to others words in common, as he explicated semantic similarity as a property of a taxonomy. In the social network analysis literature, Burt earlier refers to an approach of indexing the extent to which individuals occupy similar network positions, but do not directly communicate, as "structural equivalence" whereas direct contact between nodes is called "cohesion" [8]. Budanitsky and Hirst [5] evaluate five categories of semantic similarity approaches, finding as superior a method of semantic similarity conceptualization and measurement by Jiang and Conrath's [33] focusing on network edges. They focus on similarities of pairs of concepts, measuring similarity as the shortest path between them in a network of words based on parent-child relations, therefore focusing on hierarchical relations.

Our approach takes into account the entire semantic network based on all of the word pairs that cooccur within three word positions on either side of each word in the text file, and indexes the similarity of two semantic networks, not just for a pair of words, but for two entire networks built from all word-pair cooccurrences

in bodies of text. Our basic focus in the current research is on overall semantic similarity of two authors' texts.

Analyzing the similarity of bodies of text has been pursued for purposes of resolving controversies over authoring of texts, particularly in the humanities. Forensic analyses of texts to determine authorship when unknown for a document include a variety of non-network semantic similarity indicators [10, 38] including syntactic similarity based on punctuation, phrase structure, and stylistic errors.

People with similar semantic networks can be considered to form a discourse community, a set of individuals who encode similar information content in a similar structure. This is perhaps a more useful method for identifying community than the currently predominant approach to measures of community based on within to between community cohesion resulting from direct exchange of messages or other objects such as web hyperlinks [51]. I do not prefer such a restricted approach. Just because individuals talk to one another does not necessarily imply that they represent the same community. Take an example in which there are two or more different communities engaged in discourse who have conflicting points of view and are communicating these points with one another either in preparation for more overt hostilities or in an attempt to more peacefully resolve their differences. Under these conditions, merely network analyzing who communicates with whom, without regard for message content, can lead to the misleading interpretation that all the individuals are members of one community. Nevertheless, some of these members may be communicating in polar opposite ways, with their semantic networks highly negatively correlated. Likewise, one encounters situations in which individuals may invoke their own membership in a community based on subjective identification with it, even though they may not communicate with one another. For example, Danowski and Ruchinskas [21] define "content cohorts" based on their sharing of interests in the same media content. More specifically, one can posit that individuals who process some media content in common are members of a community even though they may not perceive themselves as such. Moreover, the processors of mediated information may subjectively define themselves as belonging to a community as framed in the media messages without the requirement that they send messages about this content or their identification with it to other individuals in their social network.

To identify communities based on some direct message exchange among members may render a net loss of information in terms of what one would consider a community in terms of a set of individuals who share the same language, whether or not they directly communicate with one another. Similarly, sharing certain beliefs that others have previously reproduced and formalized as an ideology may define a community where members can be considered as belonging to it without regard for their actual communication behavior with other individuals. For example, members of a religious denomination may only communicate locally with a limited number of individuals in a congregation and not have any communication outside the locale, yet can be considered, as well as consider themselves, to be members of the larger global religious community for the denomination. Particularly, when one wants to mask one's membership in a community, such as a radical jihadist is likely to do

when communicating with counterterrorism agents, actual communication among community members sharing radical jihadist beliefs may be quite constrained.

Placed in the context of community research, my interest in the work reported here is testing a method for identifying communities of linguistic practice that contain members of relatively high semantic network similarity. This can be considered more fundamental community identification in that members of these elemental similar semantic communities may be observed to communicate with one another across different community boundaries defined by direct cohesion properties.

Over the years, when attempting to speed processing of word-cooccurrence networks, researchers have used procedures to reduce the dimensionality of the matrix. The first such attempts were labeled LSI [22]. This approach takes the document-by-word matrix (in our case the author_file-by-word matrix), and runs a singular value decomposition procedure to reduce the number of word dimensions, where originally each word is a dimension, and finds the set of underlying dimensions that explain most of the variation in the matrix. The basic approach is to do a factor analysis of the document/term matrix. This is somewhat analogous to finding the bedrock bases of the highest linear mountain ranges in a landscape and using only these large structures in subsequent analysis, ignoring all of the lower hills and valleys as well as non-linear mountain ranges. In the computations, these high mountain ranges are at 90° angles, and there can be many more than three dimensions. The eigenvectors identified, the bedrock bases of the mountain ranges, are the latent dimensions that underlie similarly used words. This latent dimension includes words and their synonyms, among a relatively large subset of all words that frequently cooccur. The original full manifold of dimensions is considerably reduced as these latent dimensions replace them in the matrix for further analysis. This speeds up subsequent computation, such as indexing the similarity of documents or authors, because of this substantial reduction of the dimensionality of the matrix. For purposes such as information retrieval of documents, the application for which this method was developed, it works very well compared to using the full matrix. Nevertheless, if the goal is precise extraction of social meanings from aggregations of individual-level word cooccurrence matrices, such reduction techniques introduce considerable error because they remove nuanced language with particular inflections. This is discarded in favor of the prominent linear ridges of a document by word matrix. An analogy would be to first have data on individual vehicle traffic patterns such that for every driver for every trip one had recorded in a matrix each town passed through in a metropolitan area, including travel on every surface street. The reduced form of this matrix would be to find at the aggregate level the most traveled routes. As a result, one would probably be left mainly with the freeway or toll way routes among towns. After such reduction, no longer could one recover trips that involved use of surface streets, or unique "back road" routes that some individuals may have taken. Moreover, any particular individual may be represented as using a set of main freeways or toll ways even though on any of their trips they may use only a subset of these thoroughfares.

One of the major limitations of these latent dimension analyses, ignored by most researchers, is that they dump all of the words in a document or file into what is called a "bag of words." Words are considered equally related no matter how far apart they may be in a document or text file. Nevertheless, the methods we use in this research are not based on bag-of-words approaches. Instead we take into account the proximity of words and count as related those words that appear within three word positions on either side of each word, a window size that was empirically validated [18]. In effect, a sliding window moves through the text as it advances from one word to the next and word pairs within the window are indexed as occurrences of network edges among the words [17].

Because of the use successful use of LSI for information retrieval, researchers have sought to improve on the Gaussian model underlying it by developing Bayesian probability models. The first widely available such application was Probabilistic LSI (PLSI), implemented in a Matlab script [29] This method is based on a log-likelihood principle and on Bayesian statistics. Its chief advantage over LSI is that documents are not assigned to a single factor. Rather, each document is a mixture of factors with weights. This enables the effective handling of both synonymy and polysemy of words. The method also enables computation of document similarity coefficients [25]. Subsequent research, however, has improved upon the PLSI results with Linear Dirichlet Approximation (LDA) [2].

LDA is the most recent development of dimensionality reduction methods and more fully uses Bayesian models. LDA has been characterized by its initial developers [2] as a three-level hierarchical Bayesian model. LDA represents each item of a textual collection as a finite mixture over an underlying set of topics. The method then models each topic as an infinite mixture over an underlying set of topic probabilities. A topic is essentially a latent dimension. The method allows for a fast and seamless computation of both the composition of topics comprised by units of the corpus, typically words, and the similarities of the higher order units, most often documents comprised of words.

Two most popular LDA programs are MALLET [40] and Chang's implementation of LDA in the statistical programming language R [9]. I used all three of these dimensionality reduction methods and both LDA programs in search of the best one for the purpose of finding semantically-similar individuals who do not directly communicate with one another. Chang's LDA in R code was most readily adapted to our purposes. The code enabled the computation of the cosine across the topics for each pair of authors. Because the cosign is equivalent to the correlation coefficient, this offered the best means for comparing the dimension-reduction methods to the QAP correlations computed for pairs of authors using the unreduced full-text word-cooccurrence network analysis.

Despite the advancement of these latent dimension methods, available LDA software also uses the crude bag-of-words approach. Nevertheless, one study has been reported that demonstrated that LDA models that incorporate word proximity information are more valid representations [53]. Unfortunately, the researcher has not made the proximity-based LDA software available.

Despite the bag-of-words limitation, my primary motive of finding a way to reduce computation time by two or more orders of magnitude justifies the exploration of the LDA method. I found that an LDA analysis was completed in less than four minutes compared to the 11 days it took for the full-network method. This encouraged me to further explore the method to see how well it performed in identifying semantically-similar authors who did not directly communicate with one another.

2 Methods

2.1 Population Studied

To develop and test this method of semantic similarity identification I chose a diverse linguistic community. This was a discussion forum of a Pakistani accountancy organization (http://accountancy.com.pk). Initial manual examination of the posts revealed that the diversity of topics was much broader than accounting, rather more like a general discussion forum. Posts ranged from political statements in English to Urdu poetry. At the same time the languages used were diverse. Most content was in English but some was in different ethnic dialects. The diversity of the forum would challenge the methodological approach. As a contrasting forum we selected http://interface.edu.pk's ISSB (Inter Services Selection Board) Questions and Answers Forum, mainly focused on the selection test for the Pakistani armed forces. The modal content was brief questions and answers in English.

The reason why I chose Pakistani discussion forums is because that society is considered to contains numerous ideologies. Moreover, it has multiple radical jihadist organizations. Studying a diverse discussion forum provide a stronger basis for detecting semantic similarity than in a forum in which members were less diverse. "... Pakistani cyberspace includes English, Punjabi, Urdu, and Arabic blogs." [6, p. 170]. This diversity is reflected widely. One example is All Things Pakistan, a blog founded in 2006, intended "to embrace Pakistan in all its dimensions ... including its politics, its culture, its minutia, its beauty, its warts, its potential, its pitfalls, its facial hair, its turbaned heads, its shuttlecock burqas, its jet-setting supermodels, its high-flying bankers, its rock bands, its qawalls, its poets, its street vendors, its swindling politicians, its scheming bureaucrats, its resolute people-in essence, all things Pakistani." [6, p. 171] The accountancy discussion forum, while supposedly focusing on accounting, contained posts of a wide scope, not unlike the ATP blog scope. This diversity would challenge our method.

2.2 Post Extraction

We extracted all posts from the forums on February 1, 2010 using a python script. For the accountancy forum the total number of posts was 10, 001 and number of

unique persons posting content was 3,272. For the ISSB forum there were 530 posts and 503 authors.

2.3   Partitioning by Author

A python script was used to aggregate the posts into a separate file for each poster in the respective forums.

2.4   Extracting Word Proximity Pairs

Next, for each author, WORDij's [20] WordLink program implemented in C++, was used to identify all word pairs occurring within three word positions on either side of each word in the text. All frequencies were retained, even frequencies of 1, which are typically discarded in natural language computational linguistics. In setting the program parameters we dropped only punctuation, numbers, and ignored case. No stop list was used and stemming was not performed. Word order in pairs was retained.

Rather than computing semantic similarity based on the presence of similar concepts we take a micro-level semantic similarity approach. We code all messages produced by individuals and represent the message content as semantic networks. For maximum internal validity we do not use a stop word list nor do we stem words, two techniques common in natural language processing for applications such as document retrieval and in most contemporary automated content analysis of natural language. We assume that that manner in which people use pronouns and grammatical function words lends form to their semantic signatures. Their entire semantic network using all words and all cooccurrence frequencies within a particular proximity provides the most high-fidelity representation of their message encoding. If at this level, two individuals have similar semantic structures, we assume that indeed they are semantically similar with a high degree of validity.

An adjacency list of all such word pairs and their frequencies of occurrence was retained for subsequent analysis. Word order within pairs was maintained because it imbeds considerable syntactic information that enables relatively linguistically valid optimal messages with the OptiComm software in WORDij, although this is an analysis awaiting future research.

2.5   Computing Network Similarity

The QAP in C++ [20, 32, 37] was performed on the semantic network adjacency lists for each pair of authors, numbering 5.35 million paired comparisons for the

accountancy forum and 126,504 paired comparisons for the ISSB forum. QAP produces a Pearson correlation coefficient of the network similarity of a pair of authors. Because identical rows and columns of two matrices are required, we created the union of the two semantic networks and entered zeros for the non-occurring pairs in one adjacency list compared to the other. Because we did not use a dimension reduction, so that we would retain the actual word pair network structures throughout the analysis, computational time was lengthy for the larger forum, averaging close to 0.17 s for each pair of authors. Computing all 5.35 million paired comparisons (3272*3272 authors/2) was done on a dual core PC running Ubuntu Linux, which ran for 11 days. The ISSB forum ran for less than 1 h because the posts were shorter and there were few authors.

Based on the preliminary findings, the current research explored LDA as a dimensionality reduction approach that could cut run time by several orders of magnitude or more. In implementing LDA topic modeling one of the problems is that there is insufficient empirical evidence to suggest how many topics to specify for the analysis. The researcher must stipulate the number of topics in advance to generate a model. There is no code that automatically sets an optimal number of topics for a particular analysis. I asked three LDA developers how many topics they would use for the data at hand. The answers were: 400, 400, and 300.

3 Results

Generally correlations in the 0.60 range are considered moderately strong. Accordingly, a threshold of 0.60 for the Pearson correlation coefficients from the QAP analysis filtered the overall similarity of pairs of authors' semantic networks. Figure 1 shows the network of such authors for the Pakistani accountancy forum, regardless of whether or not the pair of individuals actually communicates directly. The network has 70 individuals forming 315 author pairs.

These author pairs were analyzed using standard social network analysis procedures. Pairs files output from WORDij's WordLink program in the .net Pajek format [1] were imported into UCINet [3] to create a system file that was then entered into NetDraw [4] to render Fig. 1. The layout is based on the spring-embedding algorithm. The size of the nodes is in relation to their betweenness centrality [24].

To see whether this network had any subgroups we used NEGOPY, which the benefit of a fixed set of rules for identifying groups: (1) all nodes are directly or indirectly reachable within the group boundary, (2) nodes have at least 51% of their links with others in a tentative group or cluster, (3) and the removal of any single node does not result in two disconnected components [46]. This analysis revealed a single group network structure with no subgroups. Likewise, Girvan–Newman computations in UCINET revealed evidence for a single group [26].

The set of nodes with high similarity is 0.7% of the total number of discussion forum participants. Figure 1 shows a network structure with varied features. There is one group and there are a number of pendants (single links to the group members) and isolated dyads.
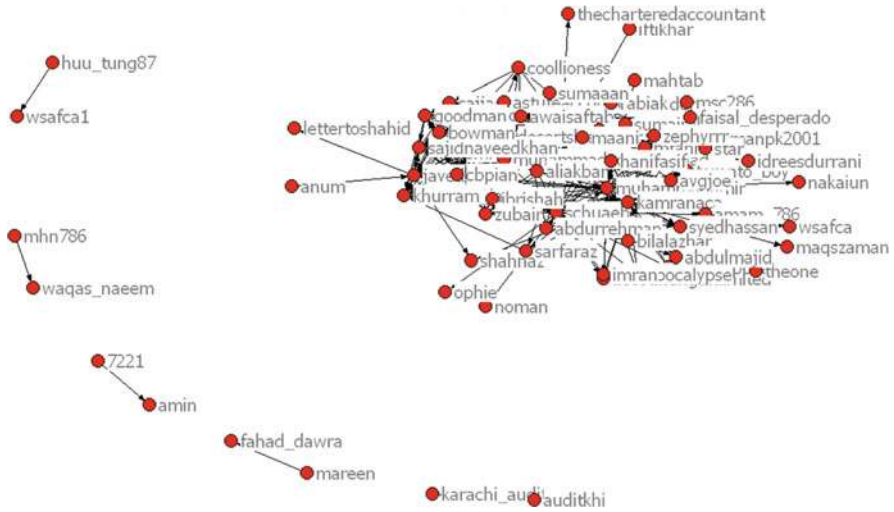
Fig. 1 Accountancy forum authors with correlations of 0.60 or higher
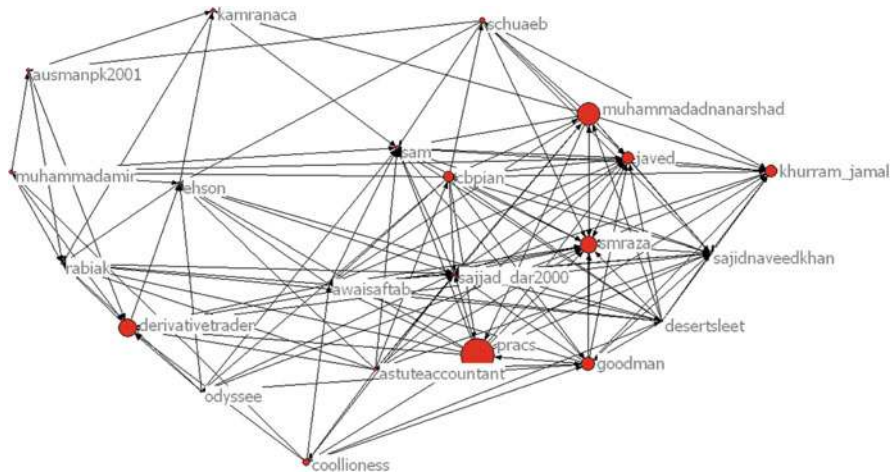
Fig. 2 Accountancy forum authors with highest semantic network similarity ($r > 0.68$)

The highest Pearson correlation coefficient for this forum was 0.71 and there were 22 authors paired with each other with correlations of 0.69–0.71. The number of such pairs was 130. Thus, this similarity threshold is approximately 50% shared variance in the semantic networks.

The higher correlation range was chosen for graphical communicability in that it limited the number of nodes shown in the network so that the pattern of links could be readily observed. Consider the differences in Figs. 2 and 3. Figure 2 is based on semantic similarity edges, while Fig. 3 is based on direct communication
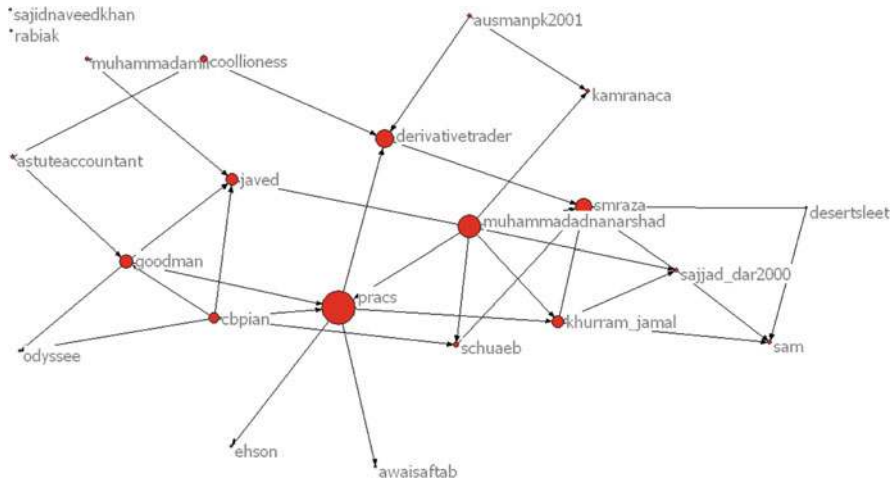
Fig. 3 Accountancy forum high-similarity authors with direct communication

among those with semantically-similar edges. When considering all pairs at this higher similarity level there are 130 among the 22 individuals. When restricting the network to only those similar individuals who directly communicate, the 22 individuals constitute only 35 directly connected pairs. The direct communication network among these individuals is only 39% as linked as the Fig. 2 network, or put the other way, 61% of highly-semantically similar author pairs' do not have direct communication.

Figure 4 is from the Pakistani ISSB military forum. It has a narrow scope, primarily a question and answer venue. This is reflected in the set of isolated dyads found with correlations ranging from $r = 0.30$ to $r = 0.96$. Arc thickness represents correlation magnitude

3.1 Linear Dirichlet Approximation

Based on the advice from developers to use 300–400 topics I systematically ran a series of analyses starting at 400 topics, decrementing the topic number at 20 fewer for each subsequent analysis, down to 20 topics. No correlations greater than 0.40, explaining only 16% shared variance between authors, were found. As a result, I proceeded in the opposite direction, starting with three topics and incrementing each subsequent run by one topic up through 10 topics. After 8 topics there was a large drop off in the number of highly similar author pairs for topic numbers of 9 and 10 so the analysis was stopped after that point. It was found that at this lower end of the topic distribution there were occurrences of correlations higher than 0.60. In the LDA analysis, at topic numbers 8 and lower nearly all of the correlations were in the 0.60–0.69 range with few above that.
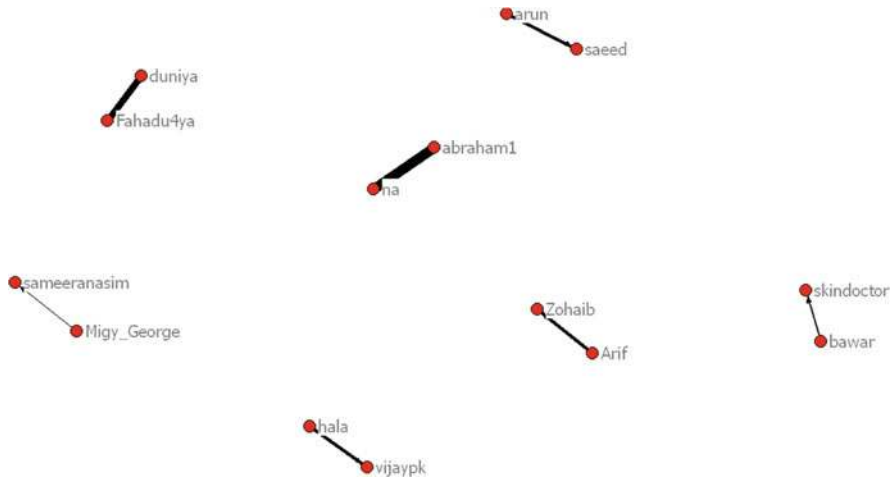
Fig. 4 Pakistani military test forum authors with correlations of 0.30 or higher

	K3	K4	K5	K6	K7	K8	K9	K10
■ Series1	113	2367	132	92	21	134	0	6
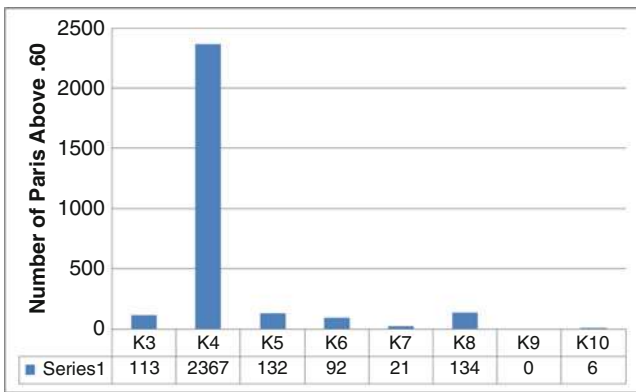
Fig. 5 LDA Analysis: Similar pairs of authors for different numbers of topics

As seen in Fig. 5, the topic number models with topics equal to 5 and 8 produced the number of similar author pairs above 0.60 closest to the full-matrix analysis. The next issue of importance is what percentage of the same author pairs were identified by the full-network and the reduced network methods. In none of these comparisons for 3–8 topics were any of the similar author-pairs from the LDA found to be the same as in the full-network method. Because the full-network method is based on representing the complete textual data, it is assumed to be a valid benchmark against which to compare the LDA method.

4 Discussion

The results of this research demonstrate the feasibility of identifying semantically-similar individuals from network analysis of their forum message content. The current research analyzed all 5.35 million pairs of individuals' semantic networks to index similarity. The U.S. Federal Bureau of Investigation's Terrorism Screening Center is reported to have approximately 400,000 unique names on its watchlist as of September 9, 2008 and it is estimated that approximately 1,600 names are qualified for the list daily [52]. Extrapolating that figure to the present, January 2011, would indicate that the current list numbers nearly 2,000,000 names. Although there is unlikely to be discussion post content in the FBI database for most watchlist individuals, for the sake of discussion, let's assume that there was, as well as discussion list content for the 1,600 names added daily. This would result in approximately 600 times more computations than we demonstrated on a dual-core PC with 4 gb RAM. A more powerful parallel machine such as the Blue Waters machine at the National Center for Supercomputer Applications at the University of Illinois contains more than 2.4 million processor cores, compared to the single 2-core processor we used. The compute time would be approximately 1.2 million times faster, so processing 3.2 billion comparisons for the estimated complete watchlist would be completed in approximately 4 min.

A recent development is parallelized implementation of network analysis algorithms using the compute power of the PC's Graphical Processing Unit (GPU) [27] This unit already has parallelized architecture necessary to process complex game graphics at high speed. Nvidia, the dominant supplier of graphics processors for PCs, has recently released to the open-source community a developer programming package to enable efficient parallel programming using their GPU. This is likely to be an adequate solution for optimal compute time for corpora in the range of thousands of authors and posts. If the corpora size were in the millions of authors and posts then one would probably use a massively parallel computer platform such as the Blue Waters machine at the National Center for Supercomputer Applications at the University of Illinois at Urbana-Champaign (http://www.ncsa.illinois.edu/BlueWaters/). It offers sustained petascale computing using 300,000 8-core IBM System 7 processors.

In addition to use of a watchlist's authors' posts for the reference semantic network for similarity assessment, analysts could take a valid corpus of radical jihadist messages posted on the web, such as collected daily by Gabriel Weimann, University of Haifa, Israel and colleagues at the University of Illinois at Chicago, and use those messages to create the semantic network to use as a reference network, comparing large numbers of discussion posters' semantic networks to this standard. I demonstrated the feasibility of such a semantic network analysis by analysis of only 12 web sites of radical jihadists reacting to Obama's "A New Beginning" speech in Cairo, Egypt in June, 2009, obtained fron Gabriel Weimann. Figure 6 shows the network including links up to 5 steps away from the word 'obama.' In contrast, the University of Arizona's Dark Web Forum Portal (http://ai.arizona.edu/research/terror/) is said to contain nearly 13 million messages
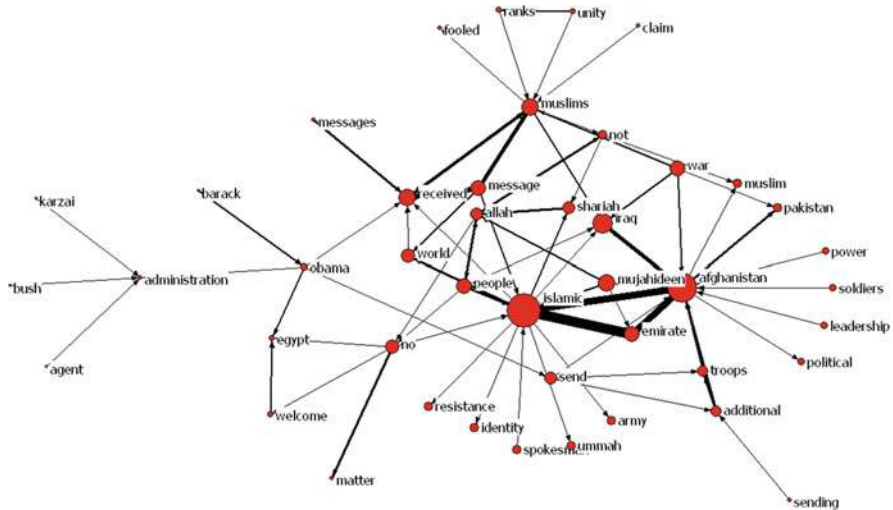
Fig. 6 Radical Jihadist responses to Obama's June 4, 2009 Cairo speech translated into English

from 28 radical jihadist web portals. Such data would provide a more externally valid aggregate semantic network to serve as the similarity reference network in finding authors with high semantic similarity on forums not currently included in the Dark Web Forum Portal project.

4.1   On Finding Needles in Haystacks

Because we wished to use only open-source information to demonstrate the counterterrorism applications of the semantic network content analysis methods, we did not have a predefined watchlist to provide for a reference network for indexing similarity of other forum members. Instead, we compared all possible pairs of forum authors. In the accountancy forum, among 3,272 forum post authors only 22 shared at least approximately 50% similarity in their semantic networks, an *r* greater than or equal to 0.69. The analysis does demonstrate that through semantic network text mining one can locate a group of authors with relatively high post content similarity. For the second forum we analyzed, the similarity was much lower with only 2 of 503 authors sharing at least 50% semantic network similarity. Figure 4 shows the series of isolated dyads with correlations greater than 0.30.

4.2   Semantic Scope Paradox

Not only was this ISSB forum more content focused, it tended to have shorter posts asking for factual details about the armed services entry test and its administration. It may be that that the more focused the purpose of a forum, the less diverse its

posts, because there is little scope within which to construct similar messages. As well, authors would not be inclined to repeat the same queries and comments that others had already made in the forum. Threadedness would be minimal. On the other hand, when the scope of discussion is broader, authors may perceive more latitude to develop threads of posts. Threadedness is probably positively associated with similarity of posts within threads. The seeming "semantic paradox" that the more general the content focus, the more likely the occurrence of similar posts, may be something for which future research finds additional evidence.

4.3 Linear Dirichlet Approximation Versus Full-Network Results

A likely explanation for the complete lack of overlap in highly similar pairs of authors in the LDA versus the full-network results may be the fundamental nature of LDA. In its dimension reduction to find major topics in the corpus, it eliminates much of the individual-level variation in language use. The author similarity approach it fosters, that of seeing to what extent an author writes posts containing the same mix of topics as another author, may produce a level of granularity that is too gross. The vehicle traffic metaphor of replacing the individual commuter routes at the street level with the inter-town freeway/toll way flows may be apt.

It is conceivable that the LDA approach may produce opposite results to the full-network method as it shifts the level of analysis. Two authors are more similar to the extent that they use a similar mix of topics in their messages. Because the similarity coefficient is the cosine across all of the topics, individuals are more likely to be similar to the extent they produce messages across more topics in common. In contrast, individuals who communicate about a few topics or perhaps only one topic would be less similar to most others based on this use of all topics for the cosine computation. Testing this assumption is a worthwhile future research effort. If it is indeed valid then the efficiency of the LDA computation is a benefit.

When we examined the words comprising each topic we observed that the use of all words may have reduced the internal validity of the LDA results, unlike the case with the full-network method. Accordingly, we further repeated analysis using a stop list removing 761 words and characters. The LDA analyses from 3 to 10 topics were rerun and similar pairs of authors were compared to the full-network results. Again, no overlap was found at any topic size.

Our assumption that the full-network representations are a valid benchmark should be empirically tested. One way to do this is through predictive validity assessment. One could observe to what extent currently highly similar yet unconnected individuals eventually directly communicate, comparing the difference between the full-network similar individuals and the LDA topically-similar individuals.

4.4   Testing External Validity

Analysis of many more forums is required to increase the probability of externally valid findings. These current results based on a single forum should be considered only exemplary and should not be generalized.

4.5   Research Questions for Future Research Using Correlational Designs

RQ1: Are full-network highly-similar non-communicating individuals (FNHSNCIs) less central in the who-to-whom communication network, while highly similar communicating pairs are more central?

RQ2: Do FNHSNCIs read each other's posts yet do not reply for some personality-related reasons? Are individuals more anti-social? . . . less extroverted? . . . more introverted? . . . more narcissistic?

Research on narcissism has found that a high ratio of the word "I" to the word "we" is significantly correlated with the validated and reliability Narcissism Scale [44]. Other research has found that individuals more central in an online network are more likely to use the "you" relatively more than the word "I," [14] evidence of an other-orientation and perhaps greater empathy.

RQ3: Although FNHSNCIs are homophilous [48] considering semantic networks, do less similar individuals respond to them?

The homophily principle that more similar individuals are more likely to communicate suggests not.

RQ4: Why are FNHSNCIs outliers on the "birds of a feather flock together" homophily and communication relationship [36]?

RQ5: Why do FNHSNCIs appear to have lower need for mainstream social approval within the discourse community?

RQ6: Are LDA-based topically-similar individuals, indicative of higher shared topical mixtures, more central in the who-to-whom communication network, while full-network-based highly-similar individuals are less central?

RQ7: FNHSNCIs read each other's post yet choose not to reply due to some linguistic features of the posts that cause an avoidance reaction?

RQ8: Is it the case that the posts of FNHSNCIs are topically unusual to the extent that others do not feel they have adequate basis for replying?

RQ9: Are FNHSNCIs more likely to post extreme opinions, which are avoided by the more central mainstream individuals in the network?

RQ10: Have FNHSNCIs been more likely to once been more central but later ostracized from the central groups?

Small group research has demonstrated that groups that are more cohesive, hence having more dense interlocking networks, when faced with a member expressing non-normative messages, are more likely to first try to persuade the individual to

abandon the deviant viewpoints. This attempt to rein in the deviant is relatively short-lived, after which the deviant is excluded from the group [39].

RQ11: Is LDA-based topical dissimilarity a precursor to FNHSNCIs status?

If this is supported by empirical evidence, then the LDA-based analysis would provide an earlier "early warning" signpost for future dark behaviors than full-network highly-similar non-communicating pairs of individuals.

RQ12: To what extent are FNHSNCIs actually the same individual using an alias?

At a 2008 European Intelligence and Security Informatics conference, it was reported that discussion forums are the richest intelligence source for dark network behaviors yet it is difficult to track specific individuals reliably, because as a diversionary tactic to counter expected surveillance, they frequently change or trade IDs with others. It would be useful to compare the "writeprint" [38] methods to the current semantic network approach in addressing this kind of question.

Answering such research questions calls for multi-method approaches ranging from automated content analysis to network ethnography [31].

4.6 Research Questions Requiring Experimental Interventions

RQEx1: If a third party replies positively to a post from FNHSNCI, what effects does this have on their subsequent communication behaviors?

RQEx2: If a third party replies negatively to a post from FNHSNCIs, what effects does this have on their subsequent communication behaviors?

Research on organizational email networks and sentiment [47] has found that more peripheral individuals in the who-to-whom network express more negative sentiment while more central individuals express more positive sentiment.

RQEx3: If a third party posts a message that states that some set of FNHSNCIs are perceived as having similar messages, what subsequent behaviors of these individuals occur?

RQEx4: If a third party invites FNHSNCIs to a dark web site, are they more likely to spend more time on the site than other individuals?

RQEx4: If a third party invites FNHSNCIs to a dark high-risk web site, are they more likely engage in an advocated behavior on the site than other individuals?

RQEx5: Can FNHSNCIs be inoculated [7] by messages from a third party to subsequent requests for the individuals to go to a dark web site?

RQEx6: What kinds of message features motivate FNHSNCIs to reply to a message?

5 Conclusion

In counterterrorism work, it has been difficult to identify individuals to add to watchlists who are not in direct contact with one another, yet have high semantic similarity to watchlist members or to other terrorist messages, increasing the

probability that they may at a future time be the most likely individuals to enter into direct contact with known dark actors. This research has demonstrated that it is feasible to match large numbers of form posters' semantic networks to a reference network, as a basis for expanding watchlists. This lengthens the future time horizon for anticipating and managing individuals with a propensity to eventually join terrorists in their pursuits. Individuals semantically-similar to dark actors, yet who do not yet communicate with them, offer an earlier warning signpost for reducing risks of terrorist activity. Interventions can be launched to attempt inoculation of vulnerable individuals against terrorists' persuasive messages. As well, individuals indicating intent to engage in terrorist acts, yet have no direct contact with terrorists, can be covertly tested to see if they are willing to carry out terrorist plans to their ends, when covert operatives then thwart these actions prior to their assumed execution by targeted individuals. This chapter proposed a number of research questions for future research to refine these identification methods and to better predict their implications for counterterrorism activities.

References

1. Batageli, V., Mrvar, A.: Pajek program for analysis and visualization of large networks. Version 2.0 Reference Manual. Ljubljana, Slovenia, University of Ljubljana (2010)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
3. Borgatti, S.P., Everett, M.G., Freeman, L.C.: UCINet for Windows: Software for Social Network Analysis [computer program]. Analytic Technologies, Harvard, MA (2002)
4. Borgatti, S.P.: NetDraw: Graph visualization software [computer program]. Analytic Technologies, Harvard, MA (2002)
5. Budanitsky, A., Hirst, G.: Evaluating WordNet-based measures of lexical semantic relatedness. Comput. Linguist. **32**(1), 13–53 (2006)
6. Bunt, G.R.: iMuslims: Rewiring the House of Islam. University of North Carolina Press, Chapel Hill, NC (2009)
7. Burgoon, M., Miller, M.D., Cohen, M., Montgomery, C.L.: An empirical test of a model of resistance to persuasion. Human Commun. Res. **5**(1), 27–39 (1978)
8. Burt, R.S.: Cohesion versus structural equivalence as a basis for network subgroups. Sociol. Methods Research, 7; 189–211 (1978)
9. Chang, J.: Package 'lda': Collapsed Gibbs sampling methods for topic models [computer program.] Princeton, NJ: Princeton University. Retrieved August 30, 2010 from http:// cran.r-project.org/web/packages/lda/ (2010)
10. Chaski, C.E.: Empirical evaluations of language-based author identification techniques. Forensic Linguistics **8**(1), 1350–1771 (2001)

11. Chen, H., Yang, C. (eds.): Terrorism informatics: Knowledge management and data mining for homeland security. Springer, New York, NY (2008)
12. Danowski, J.A., Martin, T.H.: Evaluating the health of information science: Research community and user contexts. Final report to the Division of Information Science of the National Science Foundation, no. IST78-21130 (1979)
13. Danowski, J.A.: A network-based content analysis methodology for computer-mediated communication: An illustration with a computer bulletin board. In: Burgoon, M. (ed.) Communication Yearbook 5, pp. 904–925. Transaction Books, New Brunswick, NJ (1982)
14. Danowski, J.A.: Interpersonal network radiality and mass and non-mass media behaviors. In: Gumpert, G., Cathcart, R. (eds.) Inter-Media, 3rd edn., pp. 168–175. Oxford University Press, New York (1986)
15. Danowski, J.A.: Organizational infographics and automated auditing: Using computers to unobtrusively gather and analyze communication. In: Goldhaber, G.,Barnett, G. (eds.) Handbook of organizational communication. pp. 335–384. Ablex, Norwood, NJ (1988)
16. Danowski, J.A.: A mathematical model for selection based on individuals' semantic fit with the organization's aggregate semantic network in high performance units. Presented to the Speech Communication Association, Chicago (1990)
17. Danowski, J.A.: WORDij: A word-pair approach to information retrieval. Proceedings of the DARPA/NIST TREC Conference, pp. 131–136. National Institute of Standards and Technology, Washington, DC (1993a)
18. Danowski, J.A.: Network analysis of message content. In: Barnett, G., Richards, W. (eds.). Progress in communication sciences XII, pp. 197–222. Ablex, Norwood, NJ (1993b)
19. Danowski, J.A.: Evaluative word locations in semantic networks from news stories about Al Qaeda and implications for optimal communication messages in anti-terrorism campaigns. Paper presented to the conference: EuroISI2008: European Conference on Intelligence and Security Informatics, Esbjerg, Denmark (2008)
20. Danowski, J.A. (2010). WORDij 3.0 [computer program]. Chicago: University of Illinois at Chicago. Available at http://wordij.net.
21. Danowski, J.A., Ruchinskas, J.E.: Period, cohort, and age effects: A study of television exposure in presidential election campaigns, 1952–1980. Communic. Res. **10**(1), 77–96 (1983)
22. Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W., Harshman R.A.: Indexing by latent semantic analysis. J. Am. Soc. Inform. Sci. **41**(6), 391–407 (1990)
23. Ess, C., and the AoIR ethics working committee: Ethical decision-making and Internet research: Recommendations from the aoir ethics working committee (2002). Retrieved on December 1, 2010 from http://www.aoir.org/reports/ethics.pdf
24. Freeman, L.C.: Centrality in social networks: Conceptual clarification. Soc. Netw. **1**, 215–239 (1979)
25. Gehler, P.V. (n.d.). pLSA: Probabilistic Latent Semantic Analysis. [computer program.]. Tubingen, Germany: Max Planck Institute for Biological Cybernetics. Retrieved August 31, 2010 from http://www.kyb.mpg.de/bs/people/pgehler/code/index.html
26. Girvan, M., Newman, M.E. J.: Community structure in social and biological networks. PNAS **99**(12), 7821–7826 (2002)
27. Harish, P., Vineet, V., Narayanan, P.J.: Large graph algorithms for massively multithreaded architectures. Report No: IIIT/TR/2009/74. Hyderabad, India: Center for Visual Information Technology, International Institute of Information Technology (2009). Retrieved September 10, 2010 from http://iiit.ac.in
28. Hocking, J.E., Margreiter, D.G., Hylton, C.: Intra-audience effects: A field test. Hum. Commun. Res. **3**(3), 243–249 (1977)
29. Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of the Twenty-Second Annual International SIGIR Conference (1999)
30. Horton, D., Wohl, R.: Mass communication and parasocial interaction: Observations on intimacy at a distance. Psychiatry **19**, 215–29 (1956)
31. Howard, P.N.: Network ethnography and the hypermedia organization: New media, new organizations, new methods. New Media Soc. **4**(4), 550–574 (2002)

32. Hubert, L., Schultz, J.: Quadratic Assignment as a general data analysis strategy. Br. J. Math. Stat. Psychol. **29**, 190–241 (1976)
33. Jiang, J.J., Conrath, D.W.: Semantic similarity based on corpus statistics and lexical taxonomy. Proceedings of International Conference on Research in Computational Linguistics, Taiwan (1997)
34. Kay, P., Kempton, W.: What is the Sapir-Whorf Hypothesis? Am. Anthropol. **86**(1), 65–79 (1984)
35. Kelman H.C.: Compliance, identification, and internalization: Three processes of attitude change. J. Conflict. Resolut. **2**(1), 51–60 (1958)
36. Kincaid, D.L.: The convergence theory of communication, self-organization and cultural evolution. In: Kincaid, D.L. (ed.), Communication theory: Eastern and Western perspectives, pp. 209–221. Academic, New York (1987)
37. Krackhardt, D.: Predicting with social networks: Nonparametric multiple regression analysis of dyadic data. Soc. Netw. **10**, 359–382 (1988)
38. Li, J., Zheng, R., Chen, H.: From fingerprint to writeprint. Commun. ACM **49**(4), 76–82 (2006)
39. Marques, J.M., Abrams, D., Paez, D., Hogg, M.A.: Social categorization, social identification, and rejection of deviant group members. In: Blackwell Handbook of Social Psychology: Group Processes, vol. 3, pp. 400–420 (2001)
40. McCallum, A.K.: MALLET: A machine learning for language toolkit (2002). Retrieved August 29, 2010 from http://mallet.cs.umass.edu
41. Mirzal, A., Furukawa, M.: Node-context network clustering using parafac tensor decomposition. Proceedings of the 4th International Conference Information & Communication Technology and Systems, pp. 283–388 (2010)
42. Monge, P.R., Eisenberg, E.M.: Emergent communication networks. In: Jablin, F., Putnam, L.L., Roberts, K., Porter, L. (eds.). Handbook of organizational communication, pp. 304–342. Sage, Newbury Park, CA (1987)
43. Pennebaker, J.W., Booth, R.J., Francis, M.E.: Linguistic inquiry and word count (LIWC). [computer program]. Liwc, Austin, TX (2007)
44. Raskin, R., Shaw, R.: Narcissism and the use of personal pronouns. J Pers. **56**(2), 393–404 (1988)
45. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, pp. 448–453, Montreal (1995)
46. Richards, W.D., Jr.: A manual for network analysis: Using the NEGOPY network analysis program. ERIC, ED #114110 (1975)
47. Riopelle, K., Danowski, J.A., Bishop, A.: Expression of sentiment by different node positions in email networks. Paper presented to annual meetings of the International Network for Social Network Analysts, Riva Del Garda, Italy, 29 June–4 July 2010
48. Rogers, E.M., Bhowmik, D.K.: Homophily-heterophily: Relational concepts for communication research. The Public Opinion Quarterly **34**(4), 523–538 (1970)
49. Rogers, T.B., Kuiper, N.A., Kirker, W.S.: Self-reference and the encoding of personal information. J. Pers. Soc. Psychol. **35**(9), 677–688 (1977)
50. Stone, P.J., Bales, R.F., Namewrith, Z., Ogilvie, D.M.: The General Inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information. Behav. Sci. **7**(4), 484–498 (1962)
51. Thelwall, M.: Extracting macroscopic information from Web links. J. Am. Soc. Inf. Sci. Technol. **52**(13), 1157–1168 (2001)
52. Office of the Inspector General: The Federal Bureau of Investigation's Terrorist Watchlist nomination practices. U.S. Department of Justice, Audit Division, Audi Report 09-25, May (2009)
53. Wallach, H.M.: Topic modeling: Beyond bag-of-words. In: Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA (2006)

Detection of Illegitimate Emails Using Boosting Algorithm

Sarwat Nizamani, Nasrullah Memon, and Uffe Kock Wiil

Abstract In this paper, we report on experiments to detect illegitimate emails using boosting algorithm. We call an email illegitimate if it is not useful for the receiver or for the society. We have divided the problem into two major areas of illegitimate email detection: suspicious email detection and spam email detection. For our desired task, we have applied a boosting technique. With the use of boosting we can achieve high accuracy of traditional classification algorithms. When using boosting one has to choose a suitable weak learner as well as the number of boosting iterations. In this paper, we propose suitable weak learners and parameter settings for the boosting algorithm for the desired task. We have initially analyzed the problem using base learners. Then we have applied boosting algorithm with suitable weak learners and parameter settings such as the number of boosting iterations. We propose a Naive Bayes classifier as a suitable weak learner for the boosting algorithm. It achieves maximum performance with very few boosting iterations.

S. Nizamani (✉)
Counterterrorism Research Lab, The Maersk Mc-Kinney Moller Institute, University of Southern Denmark, Odense, Denmark
and
University of Sindh, Jamshoro, Pakistan
e-mail: saniz@mmmi.sdu.dk

N. Memon
Counterterrorism Research Lab, The Maersk Mc-Kinney Moller Institute, University of Southern Denmark, Odense, Denmark
and
Hellenic American University, NH, USA
e-mail: memon@mmmi.sdu.dk

U.K. Wiil
Counterterrorism Research Lab, The Maersk Mc-Kinney Moller Institute, University of Southern Denmark, Odense, Denmark
e-mail: ukwiil@mmmi.sdu.dk

1   Introduction

As the usage of the Internet grows over the time, the number of email users also grows. For many people, the main purpose of using the Internet is email and other communication. Emails are used by a variety of individuals in a variety of ways. Many businesses receive their orders by emails. Ordinary persons communicate through email with friends. Researchers communicate with their colleagues and supervisors through email. Teachers send assignments to their students and students submit their assignments to the teacher by email. Researchers submit their papers to conferences and journals and get feedback through emails. These and many other examples relate to positive and legitimate usage of emails. Like other inventions, emails are also used for negative and illegitimate purposes as well. We call an email illegitimate if it is not useful for the receiver or for the society. Illegitimate emails may contain unwanted messages, phishing emails [11], threatening messages, or planning messages for some disastrous events such as terrorist plots. Emails are the favorite medium of communication for criminals and terrorists for planning their plots [20]. They can send emails to a number of recipients without revealing their identity (if they wish to stay anonymous). By taking advantage of the characteristics of emails, terrorist communicate with each other as well as sending threats using emails.

1.1   Motivation

After the events of 9/11, a report was published by the "National Commission on Terrorist Attacks Upon the United States" [20]. The report shows the use of emails in terror plots. It is therefore important to identify the best approach to detect emails containing clues for such events in advance to help preventing them from happening. In the paper, we have addressed this problem as an email classification problem. The problem has many dimensions; we focus on emails containing text thus considering the problem from a text classification perspective. In relation to emails, the major problem faced by most users is the arrival of spam emails. A spam email may have many characteristics but in simple words we can say that any unexpected or unwanted email can be defined as a spam email. According to the Ferris Research Report [8], businesses in the United States spend $10 Billion in 2003 on spam costs (solutions for handling spam emails, network load, and the valuable time spent by employees on spam emails). In this paper, we have applied a boosting algorithm [9] to address these two email classification problems. By email classification we mean classification of emails as either spam or non-spam (to address the problems identified by Ferris [8]) and as either suspicious or non-suspicious (to address the problems identified in the 9/11 report [20]).

1.2 Methodology

Our goal is to focus on these two major areas of email content analysis and propose an optimal solution. For achieving the desired goal, we have performed various experiments on two email datasets. For detecting spam emails the publicly available Enron Email Dataset [6] is used. For detecting suspicious emails we have developed our own dataset due to the unavailability of such public datasets.

We have applied a machine learning classification technique for our desired task. Machine learning is a technique by which we train the machine by providing some training dataset and a machine learning algorithm (classification/clustering). A machine learning algorithm learns from the training dataset and applies its learning on the test dataset. Machine learning can be either supervised or unsupervised. In supervised learning, the training examples are labeled and the machine learns the behavior of the training example for which it has been assigned such label (i.e., class); this task is known as classification. In unsupervised learning, the machine identifies the common characteristics of the training examples and organizes them into groups (i.e., clusters); this task is known as clustering.

Firstly, we have conducted experiments on the two email datasets, initially applying the various classification algorithms as base learners. The algorithms that we have used include Decision Tree, Naive Bayes (NB), and Support Vector Machine (SVM). Secondly, we have conducted the experiments with the Adaptive Boosting Algorithm using all of these three algorithms as weak learners of the boosting algorithm. Finally, we have compared the performance of all the algorithms as base learners. Then we compared the performance of various weak learners of the boosting algorithm on both datasets.

The rest of the paper is organized as follows: Sect. 2 discusses related work whereas Sect. 3 describes classification techniques. Section 4 illustrates the boosting algorithm, while section 5 explains email preprocessing. Section 6 gives experimental results and Sect. 7 portrays the results and discussions. Finally, Sect. 8 concludes the paper and describes future work. Some emails from the datasets are provided in Appendices 1 and 2.

2 Related Work

In the domain of email analysis, the research is mainly focused on email traffic analysis and email content analysis. An email traffic analysis system applies computational intelligence techniques to investigate the suspicious individuals based on their traffic behavior. Email traffic analysis [7, 15] is actually the analysis of email headers, meaning only the structured part of the email such as the To field, the Carbon copy (Cc) field, and the Blind carbon copy (Bcc) field. For this purpose, many email datasets have been developed by the research community. The most famous among them is the Enron Email Dataset [6], which was developed by the

Cognitive Assistant that Learns and Organizes (CALO) project. This email dataset contains about a half million messages and made publicly available on the web by the Federal Regulatory Commission [12] during its investigation. The dataset has been processed by researchers according to their requirements. As for purpose of spam email filtration the dataset was processed by Metsis, Androutsopoulos and Paliouras [19]. Spambase [25] is another dataset created especially for spam email detection.

2.1 Spam Email Detection Research

Youn and McLeod [33] have presented a comparative study of various classification techniques for the task of spam email classification. The classification methods that the authors have used are NB, SVM, J48, and Neural Networks. They proposed the J48 decision tree for the required task.

Youn and McLeod [34] have also proposed an ontology based spam filtering method. For ontology creation they have used the J48 decision tree for creating the concepts of the ontology. Initially, J48 is run on the email dataset; the rules generated by J48 are used to create ontologies instead of classification.

Renuka and Hamsapriya [23] have proposed a spam email detection method using word stemming instead of simple content based classification and the authors have showed the efficiency of their method over simple content based classification. The authors [23] have emphasized the use of stemming because spammers may use misspellings deliberately instead of correct spellings of spam keywords to avoid content based spam filters to detect the keywords.

Graham [13] discusses the Spambayes email classification method used as a plug-in for Microsoft Outlook. The Spambayes system uses Bayes' theorem for classification of incoming email based on a keyword approach.

2.2 Suspicious Email Detection Research

After the tragic events of 9/11, a number of researchers [1–4] started their research on analyzing emails related to the area of counterterrorism. Our research is also a contribution in this area.

Appavu and Rajaram [1] have applied the decision tree algorithm [21] to detect suspiciousness. The detection system described by the authors [1] applied simple rules to detect suspicious emails by using a simple strategy based on keywords. This system works independent of the context in which those keywords appear in the body of the email.

Association rule mining is also used for detecting suspicious emails with additional benefits of classifying the (suspicious) emails further into specialized

classes [2]. This system is capable of alerting the users, if it encounters presence of keywords along with future tense verbs.

Weber et al. [32] have developed an integrated approach to detect Inconspicuous Contents from terror related web pages using a small training sample.

3 Classification Techniques

Classification is one of the data mining techniques aimed at finding hidden relationships in the data items that belong to the same class or category. It learns relationships from the training instances and then applies that learned knowledge to predict the classes of the instances whose class is unknown.

$$D = \{t_1, t_2, \ldots, t_n\} \tag{1}$$

$$t_i = \{a_1, a_2, \ldots a_m c_i\} \tag{2}$$

$$C = \{c_1, c_2, \ldots, c_k\} \tag{3}$$

Where D is a dataset and each t_i is a training instance. Each training instance is defined by set of attributes $a_{i(i=1..m)}$, m is number of attributes and a target variable class name $c_{(i=1..k)}$, where k is total number of classes to whom an instance can belong.

Several classification techniques have been developed over time. Here we will discuss the techniques that we have used and also have been used by many other researchers due to their prominent features.

3.1 Decision Tree

A decision tree is a collection of nodes. Each node is either an internal node that contains a test attribute a_i selected from the set of attributes by some measure from information theory or an external node (leaf node) labeled by a class label (target variable). A decision tree has some advantages:

1. It is simple to construct.
2. It generates rules that are easily understandable by humans.
3. It is inductive in nature.
4. If provided with sufficient training examples, its accuracy is high.

Many decision tree algorithms have been developed over time; the most famous of them is ID3 (Iterative Dichotomiser 3) [21]. It was introduced by Quinlan in 1986. After its development, ID3 evolved in different ways. ID3 can only work with categorical attributes and cannot handle missing values and it is non-incremental. In 1993 Quinlan presented improvements to ID3 in the form of the C4.5 algorithm

[22] that can handle missing values and categorical attributes. A number of other variations exist that deal with the non-incremental problem of ID3 such as ID'3 [24], ID4 [24], ID5 [28], ID5R [29], and ITI [30].

3.2 *Naive Bayes*

Naive Bayes [17] is a classification technique that uses Bayes' theorem to calculate the probabilities of the attribute values corresponding to the class label (target variable) in the training phase. In the prediction phase, it uses the prior probabilities of each class variable from observed attributes to predict the maximum priority class for that instance.

3.3 *Support Vector Machine*

SVM is a supervised machine learning technique used for classification. SVM is based on Vapnik's statistical learning theory [31]. SVM has some unique features due to which it is considered as state-of-the-art in classification. It is considered well suitable for the task of text classification and hand written digit recognition. Its unique features are as follows:

1. It works well with high dimensional data.
2. It can make a decision boundary by using only a subset of training examples called support vectors.
3. It can also work well on non-linearly separable data by transforming the original feature space into a new feature space that is linearly separable by using the kernel trick.

Joachims [14] has defined some properties of text classification for which SVM is the ideal choice of solution.

4 Boosting Algorithm

"Boosting" is a general method for improving the performance of any learning algorithm [9]. Boosting can improve the performance of any weak learner by reducing the error rate by consistently generating classifiers on different samples with varying weights of the samples [27]. Weak learning means learning with a traditional classification algorithm. Boosting takes as input T (number of rounds) and any base classifier (weak learner) that runs on the samples from the original dataset by attaching calculated weights to each instance of the sample. The instances with higher weights get higher chance to be selected by the weak classifiers. After

running the base classifier on the sample it returns a hypothesis h(x) = y. An error rate is calculated for each hypothesis. Depending on the error rate, a weight is associated with the hypothesis. The smaller the error rate, the higher weight the hypothesis will receive. A number of hypotheses are generated in this way, say for T times. Each time, a weak learner will run on a distribution of samples and emphasizing on the examples that were misclassified during previous rounds that have received higher weights. Finally, a strong hypothesis is generated from T weak hypothesis that will have low error rate as shown in Fig. 1. It is the basic idea of the AdaBoost algorithm, that Freund and Schapire have used in [9] and [10]. Since its development there has been a number of extensions and variations but still most of the applications refer to its basic form defined as AdaBoostM1 [9, 10] as given in Algorithm 6. A well known Machine Learning Tool WEKA (Waikito Environment for Knowledge Analysis) [16] has implemented this version of the boosting algorithm.

Algorithm 6 AdaBoostM1

AdaBoostM1(WeakLearner,T,D_t, $\{(x_1, y_1)...(x_m, y_m)\}$)Initialize $D_1(i) = 1/m\{i = 1...m\}$
[initial weights are uniformly distributed weights]
Following are the major steps that are performed by each of T rounds
Step1: [Calculate error rate of each weak learner t]
$\varepsilon_t = \sum_{t=1}^{m} D_t(i).(h_t(x_i) \neq y_i)$
if ($\varepsilon_t > \frac{1}{2}$) then discard the hypothesis (weak learner)
Step2: [Calculate weight for the weak learner t depending on the error rate]
$\beta_t = \frac{\varepsilon_t}{1-\varepsilon_t}$
Step3: [Update the weight of instances for the next weak learner]
$D_{t+1}(i) = \frac{D_t(i).\beta_t}{Z_t}$. if ($h_t(x_i) = y_i$) otherwise $\frac{D_t(i)}{Z_t}$
[where Z_t is a normalization factor]
Step4: [Output the final classifier (strong learner)]
arg max $\sum_{t=1}^{T} log \frac{1}{\beta_t}.(h_t(x) = y)$

AdaBoostM1 takes as input a weak learner, T number of boosting rounds, and an initial distribution of dataset with uniform weights assigned to each instance in the distribution. D_1 is the initial distribution and 1/m is the uniform weights assigned to each instance. In each round, a hypothesis is generated and the instance that are correctly classified by the hypothesis will decrease their weights as defined in step 3 and the weight of incorrectly classified instances will remain unchanged. In step 1 the error rate is calculated for each hypothesis and any hypothesis whose error rate is greater than 1/2 will be discarded. Step 2 calculates the weights for the instances and hypotheses. In step 4, the final hypothesis is generated as output. The hypothesis that have lower error rate, will have higher weights. Final decision for classifying an instance is made by combining all the hypotheses with a weighted voting scheme.

The concept of boosting illustrated in Algorithm 6 was originally given by Freund and Schapire in [9] and [10]. Ron and Ghunnar [18] have defined AdaBoost with a slight variation in step 2 of the algorithm. The stopping condition that has been used [18] is (if $\varepsilon_t = 0$ or $\varepsilon_t > 1/2$). They have added an upper bound of the
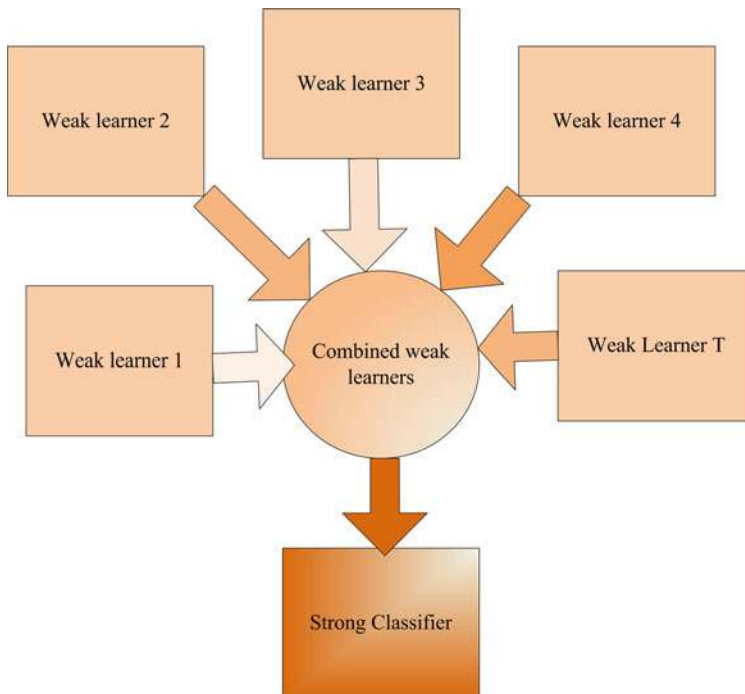
Fig. 1 Illustration of boosting

error $\varepsilon_t = 0$ to stop the loop. However, when $\varepsilon_t = 0$ it means that the classifier has correctly classified all the instances. It is very rare that a classifier results in 100% accuracy having zero error rate. There will be no example to be chosen for the next round because each example will receive zero weight. The hypothesis will receive the weight infinity (step 4). In this special case, a solution can be achieved by adding some constant to the weight of a hypothesis with an error rate of zero. 100% accuracy can also be caused by overfitting; to avoid this, a validation technique may be applied as in [5]. Bylander and Tate [5] have provided a mechanism for overfitting that still did not consider this special case. They divide the dataset in training and validation set. Initially the solution in [5] generates the same hypothesis as [9] and [10] on the training instances. Then the hypothesis is applied to the validation set. If the average error rate of the hypothesis is still greater then 1/2 the hypothesis is discarded, otherwise it is not.

Choosing the right number of boosting iterations is as important as the boosting algorithm itself. By using an improper number of boosting iterations, the boosting performance can decrease from any of the single weak learners. The AdaBoost algorithm and its variants run for a fixed number of rounds chosen at run time. When running boosting, the number of iterations must be specified even though at that time it is unknown for how many iterations the best classifier will be generated. The number of iterations must be randomly guessed. Thus, to get the maximum advantage of the boosting algorithm, we need to run it a number of times. Sometimes
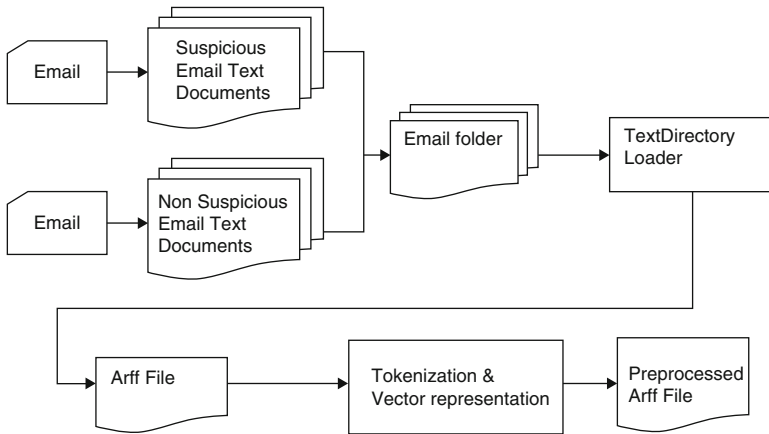
Fig. 2 Illustration of email preprocessing

boosting may yield maximum accuracy in very few rounds, say 4 or 5. But all possibilities must be checked; this will potentially waste many hours for fine tuning the number of rounds of the boosting algorithm. We propose to include a stopping condition for the number of rounds possibly depending on the variation of the error rate or weight of a classifier instead of a fixed number of iterations.

5 Email Preprocessing

Preprocessing is the initial step needed to work with machine learning techniques. The dataset must be in a form understandable by the machine learning scheme. Before performing the experiments, the emails need to be preprocessed. Initially emails are saved as text files and stored in a folder. TextDirectoryLoader is a utility in WEKA that extracts the files from text directory and stores each of text file of email as a row in Arff file. After the Arff file is created further preprocessing is performed by tokenization and vector representation. Figure 2 illustrates the discussed preprocessing step. We have performed experiments using the WEKA machine learning tool. WEKA takes an Arff (Attribute Relation File Format) file as input.

6 Experiments

We have performed experiments on two datasets one related to spam email detection and one related to suspicious email detection. The evaluation criteria we have used is the accuracy percentage A measured as

$$A = \frac{N_c}{N}.100 \tag{4}$$

Table 1 Base learner tenfold cross validation terror email dataset

NB	ID3	SVM
86.7	77.5	89.8

Table 2 Boosting with various weak learners tenfold cross validation default iterations on terror email dataset

Weak learner	NB	ID3	SVM
Accuracy	86.7	83.6	87.8

Table 3 Boosting with various weak learners tenfold cross validation using varying iterations on terror email dataset

Weak learner	NB	ID3	SVM
Accuracy	88.8	86.4	87.8
Iterations	3	12	4

Table 4 Base Learners tenfold cross validation spam email

NB	ID3	SVM
91.9	95.0	97.9

Table 5 Boosting with base learners tenfold cross validation on spam email dataset

NB	ID3	SVM
98.1	97.4	98.9

where N_c is the number of correctly classified instances and N is the total number of instances. We have performed two sets of experiments on both datasets.

First, we performed experiments using base learners. We have evaluated each base learner on both datasets. Table 1 gives the comparison of the three base learners on the terror email dataset using tenfold cross validation. It is clear that SVM as base learner outperforms the NB and ID3. Table 2 gives the effect of boosting on each of these base learners on default number of rounds. Table 3 shows the varying boosting rounds on base learners and NB outperforms other learners only in 3 boosting rounds.

In the second experiment, we have evaluated each base learner as a boosted weak learner for the AdaBoostM1 algorithm on spam email dataset. The evaluation method is the same for both experiments that is tenfold cross validation. Table 4 shows the comparison of base learners on spam email dataset and SVM outperformed NB and ID3. Table 5 illustrates the boosting each of the base learners, we can see that the performance of NB is increased by boosting from 91 to 98% and the accuracy of ID3 is slightly improved from 95 to 97%, where SVM raised its accuracy just from 97 to 98%. SVM could not be recommended as weak learner of boosting due to its long running time. We have showed the base learner and boosted weak learner's accuracy on terror email dataset on default rounds in Fig. 3. Figure 4
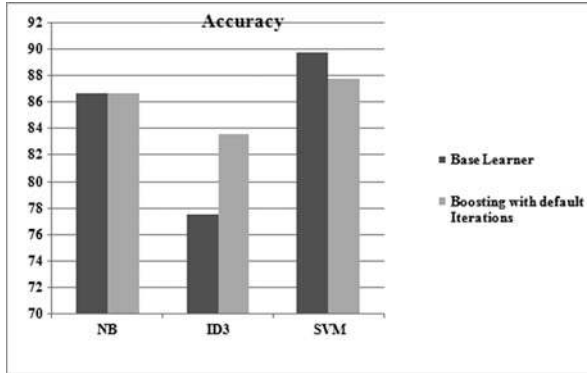
Fig. 3 Boosting using various base learners tenfold cross validation on terror email dataset using default number of iterations
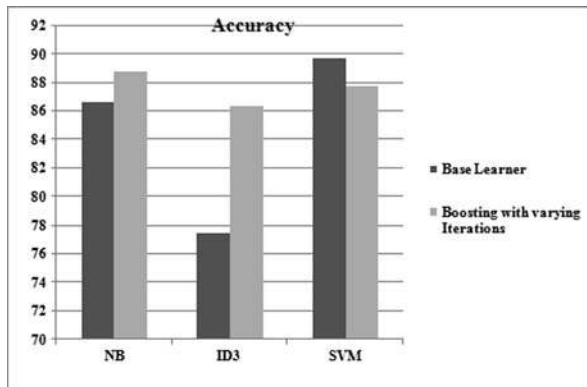
Fig. 4 Boosting using various weak learners tenfold cross validation on terror email using varying number of iterations

illustrates the varying boosting rounds along with weak learners on terror email dataset. Figure 5 shows the comparisons of base learners and boosted base learners on spam email dataset.

7 Results and Discussions

From the experiments, we can easily understand the effect of boosting using various weak learners. We propose NB as a weak learner for the task of email classification. During our experiments we observed that SVM cannot be recommended for the boosting because in some cases the boosting with SVM reduced the accuracy of
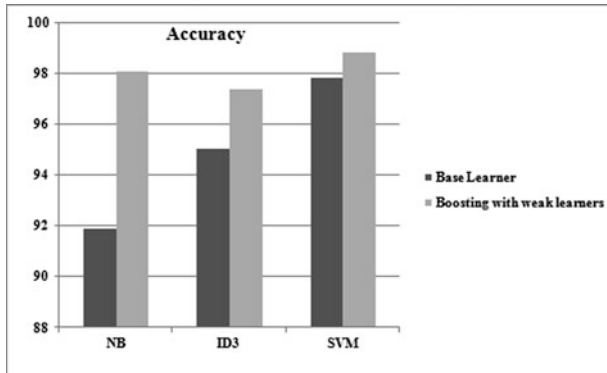
Fig. 5 Effect of boosting on each learner on spam email dataset using tenfold cross validation

the task. In some cases, SVM without boosting outperformed boosted ID3 and NB. An important aspect of the boosting is the number of boosting iterations (rounds). During our experiments, we analyzed that to get the maximum benefits of boosting we have to perform a number of boosting experiments with varying number of iterations. Performing the boosting experiments a number of times is quite time consuming (i.e., several hours). The default number of iterations of the AdaBoostM1 in WEKA is 10. In the experiments using NB as weak learners, we got high accuracy in just 3 iterations as compared to the default number of iterations. In some cases we found that when using SVM as weak learner, the accuracy of the base learner is 90%, while using boosted SVM we got the maximum accuracy 87.5% in 5 iterations as well as in default 10 iterations. We must have some alternative way to stop the boosting rounds, when an optimal solution is achieved, instead of running it up to the fixed number of iterations to save time.

8 Conclusion and Future Work

We conclude our paper by proposing suitable weak learners for the task of detecting illegitimate email. We propose NB as a suitable weak learner for the boosting algorithm. It is also concluded that when NB is used as weak learner one must go for fewer iterations, where as when ID3 is used as weak learner one must look for more iterations. By increasing the number of iterations the accuracy of NB as weak learner decreased in some cases. The accuracy of ID3 was increased greatly by boosting, but overall it was less compared to NB. SVM as weak learner has little or no effect of boosting. SVM itself is a good base learner without boosting for the task. In some experiments, the accuracy of boosting was reduced when using SVM as weak learner as compared to SVM as base classifier. In the future, we propose to add a stopping condition in the boosting algorithm to avoid the repeated experiments of estimating the exact number of boosting iterations.

Appendix 1

Some Examples of Suspicious Emails Used in Terror Email Dataset

Subject: Reaction

Government of India

This attack is a reaction to those actions that have been taken against us since 1947 and onwards. Now, there shall be no actions. There shall only be reactions, again and again. These shall continue until we have avenged each and every atrocity.

Suspects of mumbai attack

Subject It is our innings

Attention Government of India

It is our innings now, We shall not, allow this innings to go waste.

We shall play this innings with that style which was taught to us by our leaders. We know that the Government of India will not listen to our warnings seriously. Therefore, we have decided that this warning should remain not just a warning, but demonstrated through actions – actions of which you have seen a living example in Mumbai.

Suspects of mumbai attacks

Subject: US embassy

Get ready everybody. We are a band of patriots who has currently captured a Brahmas and a nuclear warhead. We threaten to destroy the parliament building in session if any is held. With captures technology and our expertise. We have started mass production we can destroy any place, anytime since we have many strategically placed base stations, throughout the planet.
Enemy

Subject: Government of India

We, the terrorist society, have planned to do a bomb blast in the Indian airline plane today. We don't like this government. We do this act as criticism to this government.
Enemy

Subject: Threat

This is a terrorist threat! Take this seriously. I hate the way you people are spending money you don't have ... I'm assigning myself to be judge, jury and executioner. Since you folks have spent what you don't have, it's time to pay the ultimate price.

Subject: Financial Support

You will receive all kind of moral and financial support there. You only have to be careful about the target. This time you must not lose any opportunity for the loss of enemy.

Hope you will do best.

Best Regards

John

Appendix 2

Some Examples of Spam Emails

Subject: major stock play

amnis systems , inc. (otcbb : amnm) contract announcements and huge newsletter coverage this week for amnm !!! this thursday amnm will be profiled by some major newsletters. there will be huge volume and a strong increase in price for several days . these are the same newsletters that profiled clks two weeks ago. they brought clks from $ 1.50 to $4.35 in ten days . we know for certain that the same groups are going to profile amnm starting on thursday.

we are very proud that we can share this information with you so that you can make a profit out of it .it is highly advisable to take a position in amnm as soon as possible , today before the market closes , or tomorrow . the stock is trading near its 52 week low , and will start moving up immediately . we believe the stock could easiely reach $ 4 in less than a month.

good luck and watch amnm fly this week ! !

Subject: save now

* * * * * write down * * * * *

hello ,

it is time to refinance !your credit does not matter , we can approve anyone . now is the time to let some of the top mortgage companies in the country compete for your

business . if you have good credit we will give you the most amazing rates available anywhere ! if you have poor credit , don ' t worry ! we can still refinance you with the most competitive rates in the industry ! let us put our expertise to work for you ! guaranteed !http : / / 21377 www.top-lenders.com/app

best regards ,

top – lenders

References

1. Appavu, S., Rajaram, R.: Suspicious email detection via decision tree: A data mining approach. J. Comput. Inform. Technol. **15**, 161–169 (2007)
2. Appavu, S., Rajaram, R.: Association rule mining for suspicious email detection: A data mining approach. IEEE International Conference on Intelligence and Security Informatics, pp. 316–323. (2007)
3. Appavu, S., Rajaram, R.: Learning to Classify threatening e-mail. Int. J. Artif. Intell. Soft Comput. **1**, 39–51 (2008)
4. Allanach, J., Tu, H., Singh, S., Willet, P., Pattipati, K.: Detecting, Tracking and Counteracting Terrorist Networks Via Hidden Markov Model. In: IEEE Aerospace Conference, pp. 3246–3257 (2004)
5. Bylander, T., Tate, L.: Using Validation Sets to Avoid Overfitting in AdaBoost. In: 19th International Florida Artificial Intelligence Research Society Conference, pp. 544–549. (2006)
6. Carnegie Mellom Universiy. http://www.cs.cmu.edu/ēnron/. Accessed on 12-08-2010
7. Clayton, R.: Email traffic: A quantitative snapshot. In: CEAS 2007-Fourth Conference on Email and Anti-Spam, Mountain View, California USA (2007)
8. Ferris Research Report: Spam Control: Problems and opportunities", http://www.ferris.com. Accessed on 25-08-2010
9. Freund, Y., Schapire, R.E.: Experiments with a New Boosting Algorithm. In: Machine Learning: 13th International Conference on Machine Learning, pp. 148–156. (1996)
10. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. **55**, 119–139 (1997)
11. Fette, I., Sadeh, N., Tomasic, A.: Learning to Detect Phishing Emails. Technical Report. Carnegie Mellon Cyber Laboratory (2006)
12. Federal Energy Regulatory Commission. A report downloaded from http://www.ferc.gov/. Accessed on 20-08-2010
13. Graham, P.: A plan for Spam. http://www.paulgraham.com/spam.html. An Internet article. Accessed on 23-08-2010
14. Joachims, T: A Statistical Learning Model of Text Classification for Support Vector Machines. In: 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. (2001)
15. Lim, M.J.H.: Computational Intelligence in Email Traffic Analysis. Ph.D. Dissertation, University of Tasmania. (2008)
16. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Ian H. Witten, I. H.: The WEKA Data Mining Software: An Update; SIGKDD Explorations, vol. 11(1). (2009)
17. McCallum, A., Nigam, K.: A Comparison of Event Models for Naive Bayes Text Classification. Technical Report . Workshop on Learning for Text Categorization, pp. 41–48. (1998)
18. Meir, R., Rastch, G.: An Introduction to Boosting and Leveraging. Advanced lectures on Machine Learning, pp. 118–183. Springer, New York (2003)

19. Metsis, V., Androutsopoulos, I., Paliouras, G.: Spam Filtering with Naive Bayes – Which Naive Bayes. In: 3rd Conference on Email and Anti-Spam, pp. 1702–1761. (2006)
20. National Commission on Terrorist Attacks Upon the United States. http://govinfo.library.unt. edu/911/report/911Report.pdf, (2004). Accessed on 25-08-2010
21. Quinlan, J.R.: Induction of Decision Trees. J. Mach. Learn. **1**, 81–106 (1986)
22. Quinlan, J.R.: C4.5: Programs for machine learning. Machine Learning, vol. 16, pp. 235–240. Springer, Berlin (1993)
23. Renuka, D.K., Hamsapriya, T.: Email Classification for Spam Detection using Word Stemming. Int. J. Comput. Appl. **1**, 45–47 (2010)
24. Schlimmer, J.C., Fisher, D.: A case study of incremental concept induction. In: 5th National Conference on Artificial Intelligence, pp. 496–501. (1986)
25. Spambase dataset. Downloaded from UCI Machine Learning Repository. http://archive.ics.uci.edu/ml/datasets/Spambase. Accessed on 15-08-2010
26. Shawkat, A., S., Xiang, Y.: Spam classification using adaptive boosting algorithm. In: IEEE 6th Conference on Computer and Information Science, pp. 972–976. (2007)
27. Tan, P.N., Michael Steinbach, M., Kumar, V.: Introduction to Data Mining. pp. 285–290. (2006)
28. Utgoff, P.E.: ID5: An incremental ID3. In: 5th International Conference on Machine Learning, pp. 107–120. (1988)
29. Utgoff, P.E.: Incremental induction of decision trees. Mach. Learn. **4**, 161–186. (1989)
30. Utgoff, P.E., Berkman, N.C., Clouse, J.A.: Decision tree induction based on efficient tree restructuring. Mach. Learn. **29**, 5–44 (1997)
31. Vapnik, V.: The Nature of Statistical Theory. Springer, New York (1995)
32. Weber, R., Waldstein, I., Deshpande, A., Proctor, M.J.: Integrated approach to detect inconspicuous contents. LNAI. 304–315. (2005)
33. Youn, S., Dennis, M.: A comparative study for email classification. Advances and Innovations in Systems, Computing Sciences and Software Engineering, pp. 387–391. Springer, Berlin (2007)
34. Youn, S., Dennis, M.: Efficient spam email filtering using an adaptive ontology. In: IEEE 4th International Conference on Information Technology: New Generations (ITNG), pp. 249–254. (2007)

Cluster Based Text Classification Model

Sarwat Nizamani, Nasrullah Memon, and Uffe Kock Wiil

Abstract We propose a cluster based classification model for suspicious email detection and other text classification tasks. The text classification tasks comprise many training examples that require a complex classification model. Using clusters for classification makes the model simpler and increases the accuracy at the same time. The test example is classified using simpler and smaller model. The training examples in a particular cluster share the common vocabulary. At the time of clustering, we do not take into account the labels of the training examples. After the clusters have been created, the classifier is trained on each cluster having reduced dimensionality and less number of examples. The experimental results show that the proposed model outperforms the existing classification models for the task of suspicious email detection and topic categorization on the Reuters-21578 and 20 Newsgroups datasets. Our model also outperforms Automatic Decision Cluster Classification (ADCC) and the Decision Cluster Forest Classification (DCFC) models on the Reuters-21578 dataset.

S. Nizamani (✉)
Counterterrorism Research Lab, The Maersk Mc-Kinney Moller Institute, University of Southern Denmark, Odense, Denmark
and
University of Sindh, Jamshoro, Pakistan
e-mail: saniz@mmmi.sdu.dk

N. Memon
Counterterrorism Research Lab, The Maersk Mc-Kinney Moller Institute, University of Southern Denmark, Odense, Denmark
and
Hellenic American University, NH, USA
e-mail: memon@mmmi.sdu.dk

U.K. Wiil
Counterterrorism Research Lab, The Maersk Mc-Kinney Moller Institute, University of Southern Denmark, Odense, Denmark
e-mail: ukwiil@mmmi.sdu.dk

1   Introduction

In this paper, we focus on the text classification problem in the perspective of suspicious email detection and topic categorization on the Reuters-21578 and 20 Newsgroups datasets. A suspicious email is one that contains clues regarding some future terrorism events, or threatening emails by terrorists. The motivation towards this area of research is the tragic events of 9/11. The report [19] shows that the terrorists used emails for their terrorism plot. Our motive in this regard is to identify such emails to help prevent such events from happening in the future.

Many researchers have focused on text classification problems over the last two decades. Sebastiani described text categorization problem in a survey report [24].[1] Sebastiani has also discussed various issues related to text categorization. There are many issues in text classification ranging from text representation to high feature dimensionality. Text classification models are usually comprised of a high number of text examples resulting in complex classification models. The text is normally represented in a vector space model and each text is represented as a vector of size equal to the number of terms (words) in the vocabulary. The approach used in this paper addresses some of these issues. We divide the dataset into clusters based on their word similarities. Thus, we create smaller and less complex models – each having less dimensionality. Our model classifies the test example with the classifier that is most suitable for that example. This greatly increases the chance of correctly classifying that test example. Experimental results show that our model achieves higher accuracy than traditional classification models on three datasets i.e., a terrorism email dataset, Reuters-21578, and 20 Newsgroups. Our model also achieves higher accuracy than the reported accuracy of the ADCC [16] and DCFC [15] models on the Reuters-21578 dataset.

We propose a model of clustering based classification, using K clusters in the task of suspicious email detection and news text categorization. Our model suggests to initially create K clusters from the original training dataset using the K-means clustering algorithm. The number K is determined by visualization of the training data. When K is determined, the K-means clustering algorithm is used to form K clusters of the original dataset. The training examples in the clusters are organized by their similarity of word features. The original number of features of the entire dataset is very high and in vector representation of the text documents (emails/news) many of the features are sparse. Therefore, we propose a classification model based on clusters that works well in this situation. The unseen test examples are classified by the classification model in the cluster that has highest similarity score to that test example. In each cluster, a classification model is created by applying a decision tree algorithm. The reason for using the decision tree classification model is due to its simplicity and its easy understanding. When the model is fully constructed, we have M optimized decision tree classifiers. The training examples are assigned to a

---

[1]The terms categorization and classification will be used interchangeably throughout the paper.

particular cluster, not because of the same class labels but because of various other common features. In each cluster, we have fewer sparse features. The decision tree model that is applied on each cluster has less depth.

The survey report [32] presents the top ten algorithms used by most of the data mining research community. According to the report, decision tree is one among the classification algorithms and K-means is the leading clustering algorithm. This motivates us to use these algorithms for our model building.

In the literature, boosting [7] is another technique that also constructs M ensemble classifier of the dataset. In boosting, M ensemble classifiers are constructed, then in the first boosting iteration training examples are randomly selected from the dataset. The randomly selected training examples have no relationship at all. In the next iteration, all those examples are taken for model building which were wrongly classified by the previous model and so on. In boosting, one cannot determine the number of boosting rounds in advance. To exactly know the number of boosting rounds, one has to perform a series of experiments with varying boosting rounds.

In the literature [9, 11, 15, 16, 35] a number of classification by clustering models have been proposed, but these work in different way. Jing et al. [9] and Li and Hung [15] have proposed a classification by clustering model. The approach used by these authors [9, 15] divides the dataset into a certain number of clusters and assigns the class label of the dominant class of the training examples to the cluster. They do not use any classification algorithm within the clusters. At classification time, the similarity of the test example is measured with all the cluster centroids and the test example is assigned the label of the cluster having the highest similarity score. Our model divides the dataset into clusters, because each cluster contains training examples from various classes, so the classification algorithm is applied on each of the clusters. Each test example is classified by the classification model that is trained on a similar training example of that test example.

We have performed our experiments using Waikato Environment for Knowledge Analysis (WEKA) [8]. It is an open source and freely available machine learning and data mining tool that is used by a number of researchers in the field.

The rest of the paper is organized as follows: Sect. 2 describes related work, whereas Sect. 3 elaborates on clustering. Section 4 describes classification, while Sect. 5 discusses the proposed model. Section 6 illustrates the data preprocessing, and Sect. 7 demonstrates the experimental results. Section 8 concludes the paper.

2 Related Work

Text classification remains a field of research interest in many different directions. Few researchers have focused on the problem of classifying threatening emails [1], while others have studied the problem for spam email detection using word stemming [23]. Moore et al. have applied text classification for web page categorization and feature selection using association rule and principal component clustering [18]. Dumais et al. [6] have applied inductive learning algorithm for text

categorization and representations on the Reuters-21578 dataset. Dumais et al. have achieved high accuracy using traditional classification on Reuters-21578 dataset. They trained a separate classifier for each category. In our approach, we have applied a classification model on all categories at once. It has the capability to distinguish a document from other categories. Backer and McCallum [2] present distributional word clustering for text categorization task. Hence, there are many applications of this major area of research. Kyriakopoulou [13] describes the major categories of text classification aided by clustering. There are three main categories:

1. Clustered based classification for feature reduction.
2. Clustering unlabeled data in semi-supervised classification.
3. Clustering large problem space to reduce the problem in smaller space.

The study by Backer and McCallum [2] belongs to category 1, and in this study the features having the similar class distribution are clustered together. All the features in a cluster play the same role in predicting the class label of an example, only the cluster representative features are used to the reduced feature space. The authors have compared their approach to other methods and their approach outperform other methods as mutual information [33] and class based clustering methods [4]. The approaches used by the authors [11, 35] fall into category 2. The models proposed by the Li and Hung [15, 16] are classification models using a number of clusters; these model fall into category 3. Our model belongs to categories 1 and 3 of classification based clustering.

The DCFC [15] model builds the classification model using the forest of decision cluster trees. This model is based on the ADCC [16] model, which creates the K clusters of the dataset by creating decision clusters. The whole dataset is divided in K clusters, where K is the number of classes. Each cluster is labeled with the majority class. The centroids of the clusters are retained. When classifying an instance, the instance is compared with all centroids and is classified with the class label of the cluster having maximum similarity with that instance. The model ADCC [16] suffers from the problem when there is no majority class in the cluster. This problem was addressed by authors using DCFC [15]. The authors [15] propose the DCFC model. In DCFC [15] the dataset is divided in the K sub-datasets (k is the number of classes). For each class, a decision cluster tree is constructed. In the decision cluster tree, each cluster is labeled with the label of its class, because in the tree there are the objects belonging to same class (thus no conflict for the dominant class). Finally, all the decision cluster trees are combined to constitute the decision cluster forest. The DCFC model works fine when the centers of each class clusters have large distance. In the real world, we may have the situation in which objects from varying classes may share a lot of vocabulary. In this situation, classifying an object just by the centroid distance may misclassify it. In the DCFC model, each cluster is built from a single class; it does not consider objects from different classes for cluster building. In this situation, if the vocabularies of all the classes vary a lot, then the Naive Bayes (NB) classification model and other classification models work fine. Both these models do not achieve high accuracy on the Reuters-21578 dataset.

The authors [11, 35] and many others, developed clustered based classification models and most of them use clustering only when they have very few labeled examples. Clustering is used only to label the unlabeled examples. Then, classification models are applied.

Kyriakopoulou and Kalamboukis [12] proposed to combine clustering with classification for spam detection. The approach used by these authors assumes that the number of clusters are equal to the number of classes. After the desired number of clusters has been created then a meta-feature set for each clusters is extracted. The classification algorithm Support Vector Machine (SVM) is then applied to each cluster. In our approach, we do not assume that the number of clusters are equal to the number of classes. We assume that the examples in the different classes may share common vocabulary. There may be a varying number of clusters depending on the training examples of the different domains not only depending on the number of classes.

The work presented in the paper [34] proposes an improved kNN classification model using clustering. The approach used by these authors, speeds up the traditional kNN algorithm. It reduces the training samples to the number of clusters from the actual number of training examples. It takes only the centroid of each cluster as training sample for classification of new unseen samples. The number of the clusters is equal to the number of categories (classes). It assumes that the centroid samples are representative of all examples in the cluster.

Many authors [3, 10, 14, 17, 25] have studied the text categorization problem. A number of approaches have been proposed. Lewis [14] has evaluated clustered phrasal representation of text for text categorization. Bekkerman and Allan [3] have evaluated the use of bigrams for text categorization.

3 Clustering

The Clustering [27] is an unsupervised machine learning technique. It organizes the data objects into groups called clusters. The organization of the objects into the similar groups is based on their similarity. When talking about text clustering, the text documents are clustered by their word similarities. The clustering algorithms can roughly be divided in three main categories:

- K-means (partitioning)
- Hierarchical clustering
- DBSCAN

The approach that we have used in our proposed model is a hybrid of partitioning and hierarchical clustering. In each level of the hierarchy, we have used the K-means clustering algorithm. It organizes the text into K clusters. It finds the K centroids, then for any object, it calculates its Euclidean distance with all the centroids using (1). The object is put into the cluster having the smaller distance with the cluster center.

$$dist(X, C) = \sqrt{(x_1 - c_1)^2 + (x_2 - c_2)^2 + \dots (x_n - c_n)^2} \qquad (1)$$

where X is the object to be classified and C is the centroid of the cluster.

The clustering algorithm that we have used for our experiments is SimpleK-means, that is WEKA's [8] implementation of K-means algorithm.

3.1 K-Means Algorithm

Following are the main steps that are performed in the K-means algorithm.

1. First we need to provide the parameter K (i.e., the number of clusters).
2. The algorithm selects K points as the initial centroids.
3. Form K clusters by assigning each training example to its nearest centroid.
4. Re-compute the centroids of each cluster.
5. Steps (3) and (4) are repeated until the best cluster centroids are computed.

4 Classification

Classification is a supervised machine learning technique. Classification algorithms require labeled training data. They learn from training examples and create the rules from labeled data. Then they apply these learned rules on unlabeled data to classify them. The classification techniques that we have used for our experiments are described below:

4.1 Decision Tree

A decision tree is a collection of nodes. Each node is either an internal node that contains a test attribute selected from the set of attributes by some measure from information theory or an external node (leaf node) labeled by a class label (target variable). A decision tree has some advantages:

1. It is simple to construct.
2. It generates rules that are easily understandable by humans.
3. It is inductive in nature.
4. If provided with sufficient training examples, its accuracy is high.

Many decision tree algorithms have been developed over time; the most famous of them is ID3 (Iterative Dichotomiser 3) [21]. It was introduced by Quinlan in 1986. After its development, ID3 evolved in different ways. ID3 can only work with categorical attributes and cannot handle missing values and it is non-incremental.

In 1993, Quinlan presented improvements to ID3 in the form of the C4.5 algorithm [22] that can handle missing values and categorical attributes. A number of other variations exist that deal with the non-incremental problem of ID3 such as ID'3 [26], ID4 [26], ID5 [28], ID5R [29], and ITI [30].

4.2 Naive Bayes

Naive Bayes [17] is a classification technique that uses Bayes' theorem to calculate the probabilities of the attribute values corresponding to the class label (target variable) in the training phase. In the prediction phase, it uses the prior probabilities of each class variable from observed attributes to predict the maximum priority class for that instance.

4.3 Support Vector Machine

SVM is a supervised machine learning technique used for classification. SVM is based on Vapnik's statistical learning theory [31]. SVM has some unique features due to which it is considered as state-of-the-art in classification. It is considered well suitable for the task of text classification and hand written digit recognition. Its unique features are as follows:

1. It works well with high dimensional data.
2. It can make a decision boundary by using only a subset of training examples called support vectors.
3. It can also work well on non-linearly separable data by transforming the original feature space into a new feature space that is linearly separable by using the kernel trick.

Joachims [10] has defined some properties of text classification for which SVM is the ideal choice of solution.

5 Proposed Approach

We propose a cluster-based classification approach to suspicious email detection and news topic categorization problems. Initially we take a full problem space as universe (whole dataset), then we extract useful features using the approach proposed by the authors [20] and apply a classification model. The accuracy of the classification model is saved and then visualization is performed. We used the visualization to find densities of the clusters and the number of clusters in the current level of hierarchy of clusters. When a cluster is obtained that contains examples from
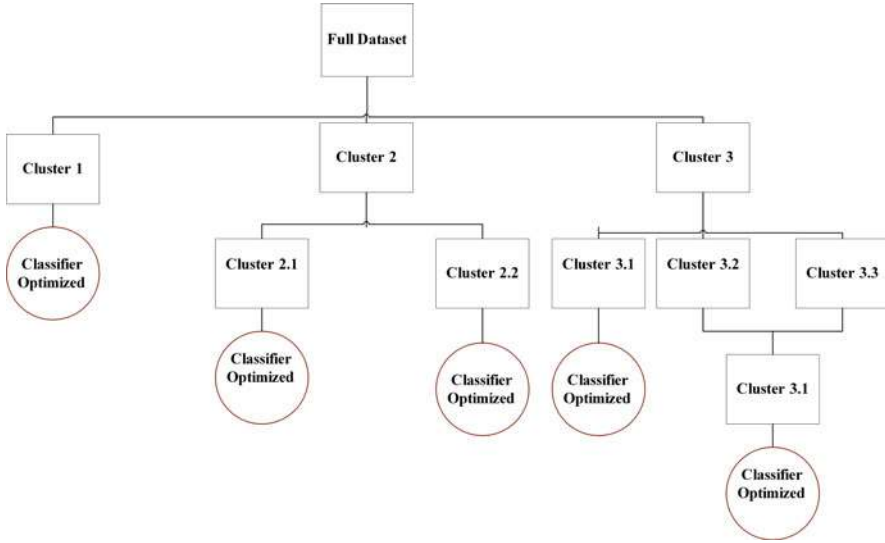
Fig. 1 Illustration of cluster-based classification model

the single class, it is labeled with the label of training examples in that cluster. When clusters are created, the classification algorithm is applied on each of the clusters, because in the cluster there are the examples from various classes. Before applying the classification model, features relevant to the clusters are extracted with reduced dimensionality.

The proposed model is depicted in Fig. 1. The root in the tree comprises the universe (entire dataset). At the root the classification algorithm is applied on the entire dataset considered as the universal cluster. The accuracy of the classifier on the universal cluster is saved. The clusters are then created from the universal cluster. The classification algorithm is then applied on each of the clusters created and the accuracy of the classifier in each of the clusters is saved. When the accuracy of a cluster is higher than the root cluster, the classifier is accepted and labeled as the optimized classifier. If the classifier in the cluster is not optimized, it means its accuracy is less than the accuracy of the root cluster; then that cluster is further divided in sub-clusters. If a cluster is created whose size is less than the threshold min (minimum size), then if all the examples in the cluster that belong to the same class, the cluster is labeled with that class label otherwise that cluster is merged with its adjacent cluster. After merging, the classification algorithm is applied on the merged cluster. The optimization and merging criteria is applied for each node in the tree.

There are some concepts used in our model, that are defined the below:

Definition1: The set of all training examples is considered as U or Universal set. $U = C_1 \cup C_2 \cup \ldots C_k$. Where C_i is a i th cluster.

$C_1 \cap C_2 \cap \ldots C_k = \emptyset$. All the clusters are non-overlapping.

Definition2: Visualization is a way to observe the training examples in each cluster and Universal set to decide the number of clusters in which the current set of instances can be divided.

Definition3: Optimization is the way to decide the acceptance of the classification model in the current cluster.

Definition4: Merging is the union of the two clusters when the size of one cluster is less than the minimum cluster size.

Definition5: Tenfold cross validation is a method of model evaluation. The dataset is divided into 10 subsets and 9 of the subsets are used for training and 1 of the subset is used for testing. In this way, 10 rounds are performed and the average accuracy of all the rounds is returned finally.

Definition6: The original dataset is characterized by set of feature vectors whose size is for example d. $U = x_1, x_2, \ldots .. x_n$, where n is the number of training examples. In U, each $x_i = w_1, w_2, \ldots . w_d$, where w_i is the ith feature of the training example x_i. Each cluster C derived from U, $C = x_1, x_2, \ldots .. x_t$, each x_i in C is $x_i = w_1, w_2, \ldots . w_k$ where k < d and it is the number of features in C and w_i is the ith feature in x_i in cluster C .

In the following section, we describe the algorithms used in the proposed approach.

5.1 Algorithms

In the proposed model, we have developed four algorithms. The main algorithm of our model is *Clus_Classification* (Algorithm 7). The algorithm takes two parameters as input U and Cl_{al}. Parameter U is the original dataset and Cl_{al} is the classification algorithm.The output of the algorithm is the set of optimized classifiers.

The *isOptimized* (Algorithm 8) takes three parameters as input. It checks for the optimality for the current cluster. In Algorithm 8 A_u is the accuracy of universal cluster, C_t is the set of training examples in current cluster, Cl_{al} is the classification algorithm. Accuracy is measured as tenfold cross validation.

The algorithm *Clustify* (Algorithm 9) is used to divide the current cluster into the further sub-clusters using K-means algorithm.

The algorithm merge (Algorithm 10) is used for merging the clusters. The merge algorithm is called when the size of a cluster is less than the min. It takes three parameters, the current cluster and its two adjacent clusters. The current cluster is merged with an adjacent cluster that has more similarity with the current cluster.

5.2 Workflow of the Proposed Model

Our proposed model consists of three main blocks. The first block is responsible for building classification models using clusters. The output of this block is the

---

Algorithm 7 Clus_Classification(U, Cl$_{al}$)

1: Fs←Apply filtered feature selection //Fs selected features
2: CModel←Apply Cl$_{al}$ to U using Fs
3: A_u←Accuracy by 10 fold cross validation
4: Clustify(U)
5: **for** (Each C_k in U) **do**
6:     **if** isOptimized(A_u, C$_k$, Cl$_{al}$) **then**
7:         Return classification model
8:     **else if** (size(C$_k$) < min) **then**
9:         **if** ($\forall$x in $C_k \in$ y)
10:             Label the C_k with y
11:         **else**
12:             Merge(C_k, C_{k+1}, C_{k-1})
13:         **end if**
14:     **else**
15:         Clustify(C$_k$)
16:     **end if**
17: **end for**

Algorithm 8 isOptimized (A_u, C_t, Cl_{al})

1: FsC_t ←Apply filtered feature selection
2: CModelC$_t$ ←Apply Cl$_{al}$ to C_t using FsC$_t$
3: A$_{ct}$←Accuracy //Apply 10 fold Cross Validation on C$_t$ using Cl$_{al}$
4: **if** $A_{ct} > A_u$ **then**
5:     return True
6: **else**
7:     return False
8: **end if**

Algorithm 9 Clustify(C$_t$)

1: Visualize the C_t to pass the parameter K to kMeans clustering algorithm
2: kMeans(k, Eucliduan) // Euclidian is a distance function
3: **for** (i = 1 to k) **do**
4:     return C_i
5: **end for**

Algorithm 10 Merge(C$_k$, C$_{k+1}$, C$_{k-1}$)

1: **if** (sim(centroid(C$_{k+1}$),centroid(C$_k$)) >
    sim(centroid(C$_{k-1}$),centroid(C$_k$)) then **then**
2:     C$_{new}$←C$_{k+1}$ ∪ C_k
3:     Remove cluster C$_{k+1}$
4: **else**
5:     C$_{new}$←C$_{k-1}$ ∪ C_k
6:     Remove cluster C$_{k-1}$
7: **end if**
8: Remove C_k
9: Return C_{new}
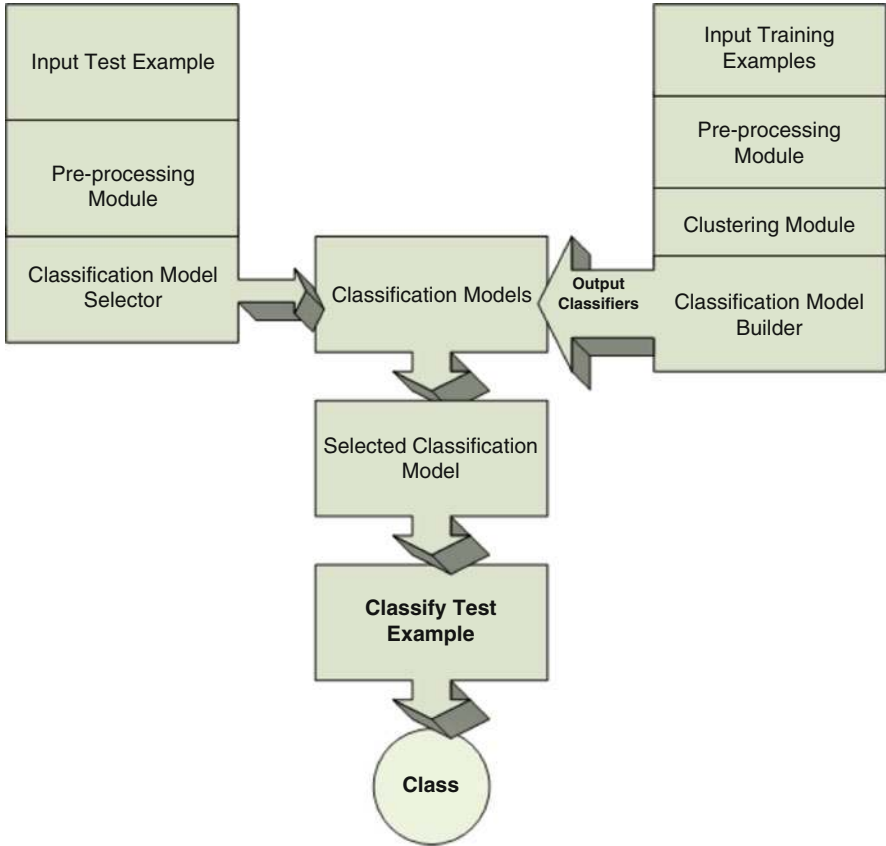
Fig. 2 Workflow of the proposed model

number of optimized classifiers. The second block is responsible for selecting one of the available classifier models for the test example. This block receives the test example and the set of classifiers in the system as input. The output of the second block is a selected classifier model for classifying the test example. The third block is responsible for classifying the test example using the selected classifier by the second block. The classifier outputs the class of the test example. The working of the proposed model is depicted in Fig. 2.

6 Data Preprocessing

Before performing experiments, we need to preprocess the datasets. All the datasets in raw form are available as text. Each instance of all the datasets is an individual text file. We need to transform all the datasets in the Attribute Relation File Format (ARFF) form i.e., the input form for the WEKA. We used WEKA utility
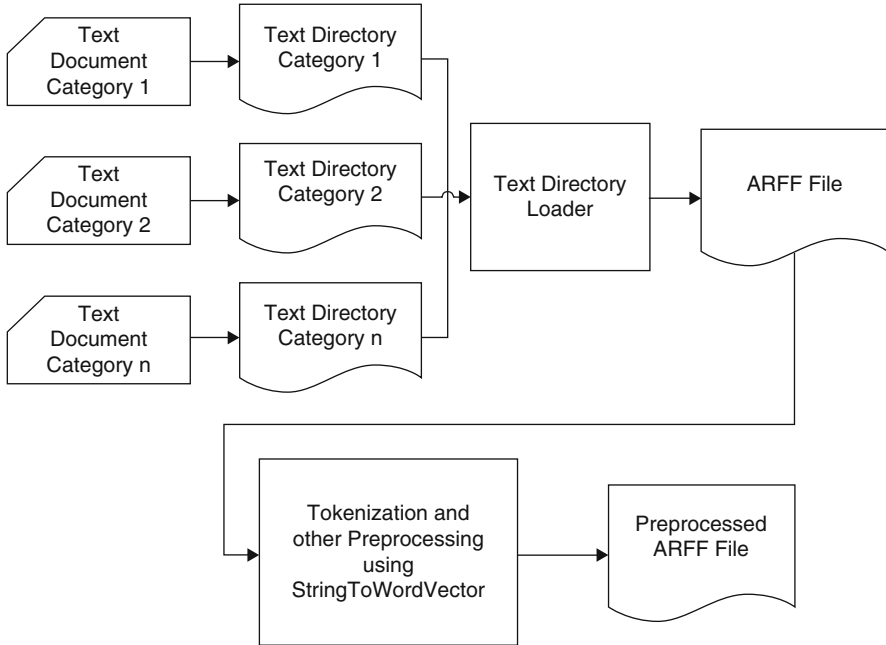
Fig. 3 Dataset preprocessing

TextDirectoryLoader that converts each of the text document/email in an ARFF instance. This utility stores the each document with two attributes one class name and the other is text of the document. After using this utility, we need to apply the filter StringToWordVector that creates a vector for each instance. In this filter, we apply further preprocessing tasks like stemming, stop word removal, TFIDF transformation, setting frequency threshold, and so on. Once the datasets have been preprocessed, the datasets are ready for experiments. The preprocessing step is illustrated in Fig. 3.

7 Experimental Results

We have performed experiments on three datasets. The first dataset is terrorism domain email dataset. This dataset is used to detect suspicious emails and it contains emails which provide clues regarding future terrorism. The second dataset is a subset of the Reuters-21578 dataset downloaded from the site [5]. We used this dataset for text categorization. The third dataset we used is a subset of 20 Newsgroups dataset. We also used 20 Newsgroups dataset for text categorization purposes. We performed two sets of experiments on all three datasets. In the first experiment, we performed suspicious email detection and text categorization using a traditional classification algorithms. The traditional classification algorithms that we used for

the experiments are ID3, NB, and SVM. For performing experiments on SVM, we need to pass parameter kernel type and we used the linear kernel type and rest of the parameter are set to default WEKA parameters. In the second experiments, we performed experiments using the proposed cluster-based classification. The clusters created in each dataset are based on similarity of features in training examples. In the cluster-based classification model, we used ID3 as a classification algorithm due to its simplicity and its easy understanding. The evaluation criteria used for all the experiments, is the accuracy of correctly classified examples using tenfold cross validation method. The average cluster accuracy is measured by the equation given below:

$$\text{Avg-cluster-Ac} = \sum_{i=1}^{k}(c_i * Ac_i)/100, \tag{2}$$

where c_i is the percentage of the total examples of dataset in cluster c_i, A_{ci} is the accuracy of correctly classified examples in cluster c_i, and k is the total number of optimized clusters.

7.1   Suspicious Email Detection Experiment

The terrorism domain email dataset is developed by us because there is no such publicly available dataset. Some emails in the dataset are threatening emails by the suspects of the Mumbai attacks. Each email in the dataset is simply treated as a text document. We used the contents of the emails for suspicious email detection. We present a few examples from the dataset in Appendix 1. First, we performed the experiments using a traditional classification algorithms. We created two clusters of the datasets using the simple K-means clustering algorithm. The first cluster contained 20% of the whole dataset and the other cluster contained the 80% of the whole dataset as show in Fig. 4. The different colors of the instances show the different classes. The first cluster contains all the examples from suspicious emails so we labeled that cluster as the suspicious one. We applied the ID3 classification on the second cluster that classified the emails with 89% accuracy. The average cluster accuracy obtained using (2) is 91.2% . We did not apply the classification model on the first cluster because that contained all the examples from the same class. The average accuracy of the two clusters is higher than the traditional classifiers. The experimental results for suspicious email detection are depicted in Fig. 5.

7.2   Text Categorization on 20 Newsgroups

This dataset consists of the four most famous categories of the 20 Newsgroups dataset. We divided the dataset into four clusters. The first cluster contains 43.6% instances and is classified with 73% accuracy. The second cluster contains 20.7% instances and is classified with 89.2% accuracy. The third cluster contains 28.9% instances and is classified with 78.07% accuracy. The fourth cluster contains
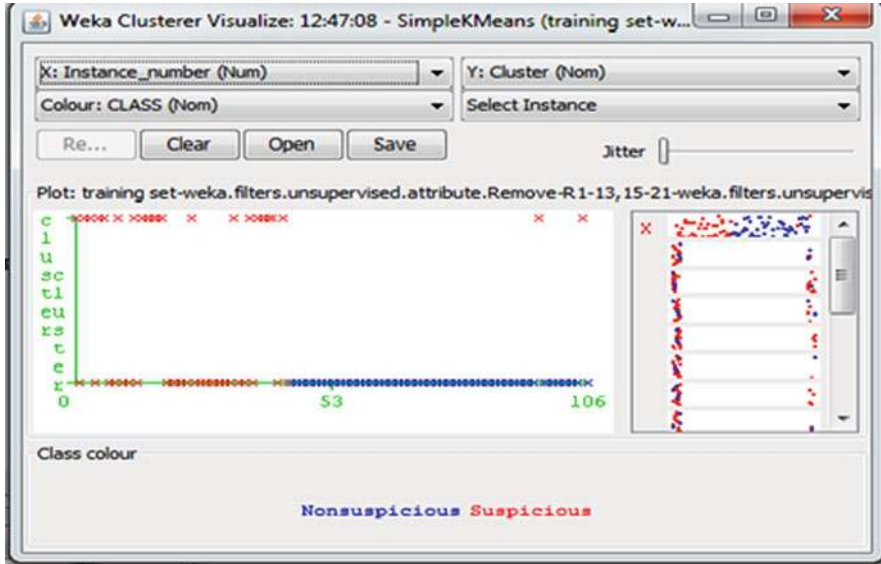
Fig. 4 Initial cluster assignment on terrorism domain email dataset
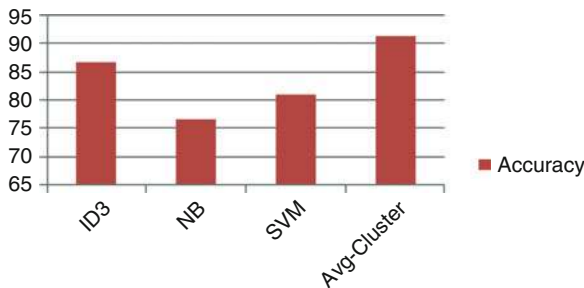
Fig. 5 Experimental results on terrorism domain email dataset

6.6% instances, all the instances in this cluster belong to same class so we do not need to apply classification model on that all the examples in this cluster are classified with 100% accuracy. The cluster assignment is presented in Fig. 6. The average accuracy of all the clusters is 79.6%. Using the traditional classifiers, we obtained a maximum accuracy of 70.8% with SVM classifiers, where the avg-cluster-ac obtained an accuracy of 79.6%. The experimental results are illustrated in Fig. 7.

7.3 Text Categorization on Reuters-21578 Dataset

The dataset is initially divided into two clusters. The first cluster contains 38.1% instances and this is classified with 98% accuracy. The cluster assignments are given in Fig. 8. According to the process shown in Fig. 1, the first cluster is optimized, so
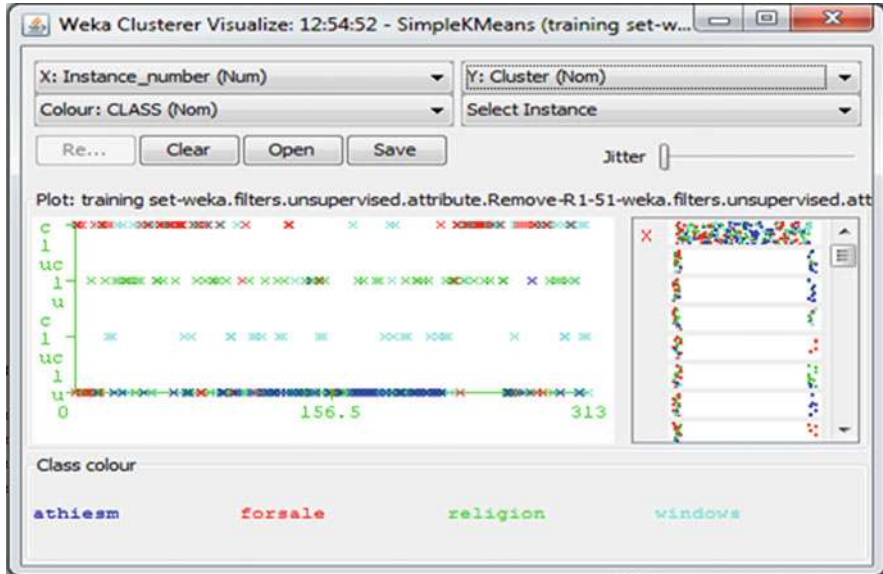
Fig. 6 Initial cluster assignment in 20 Newsgroups dataset
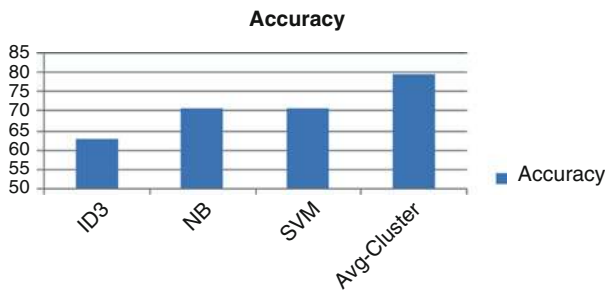
Fig. 7 Experimental results on 20 Newsgroups

it is not further divided it into more clusters. The second cluster contains 61.9% instances and does not provide satisfactory results, so according to the algorithm this cluster is further divided into three clusters and retained the classification model on a cluster when optimized model is obtained. In this way, we performed the classification task using our proposed approach with an accuracy of 89.4% calculated using (2). The Reuters-21578 dataset is also used as benchmark dataset by many of the researchers. We have also compared results on Reuters-21578 dataset with the other cluster based classification models. We have not performed experiments by ourselves on the models ADCC [16] and DCFC [15]; however, the reported accuracy of ADCC [16] is 69% and the reported accuracy DCFC [15] is 68%. The experimental results are shown in Fig. 9.
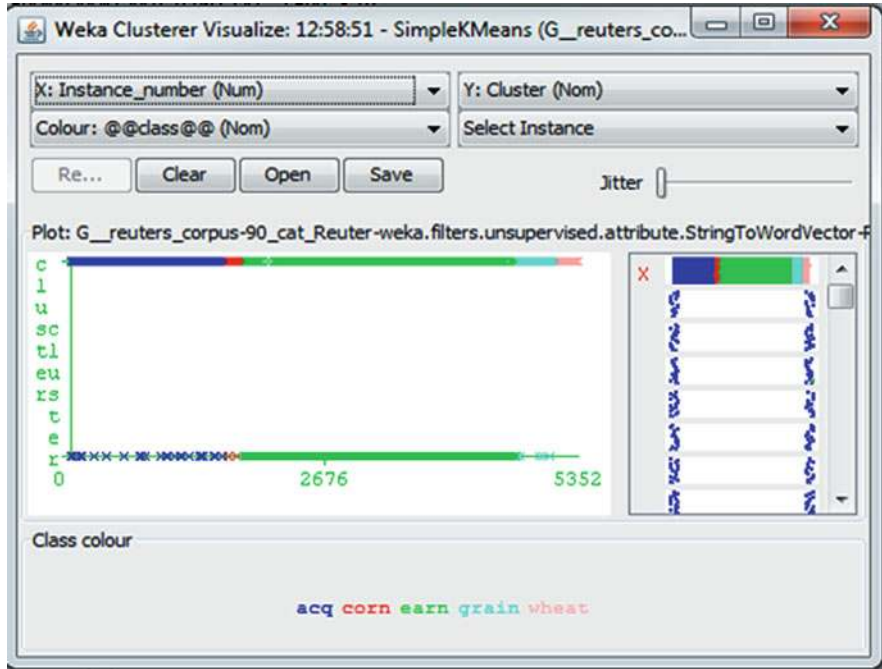
Fig. 8 Initial cluster assignment on Reuters-21578 dataset
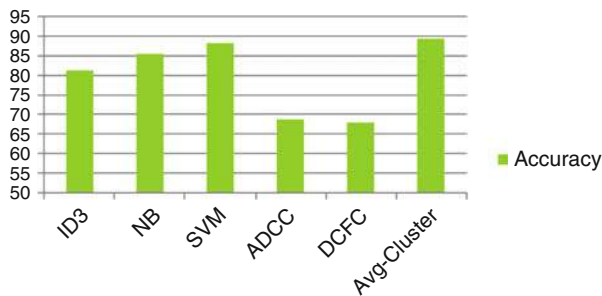
Fig. 9 Experimental results on Reuters-21578 dataset

8 Conclusion

We have performed experiments on three datasets (terrorism domain emails, Reuter21578, and 20 Newsgroups) using our approach and baseline classifiers. The Reuters-21578 dataset is used by the many other researchers, so we have compared the results on Reuters-21578 dataset not only with the baseline algorithms but also the ADCC [16] and DCFC [15] models. It is clearly shown that on this dataset, our proposed model not only outperforms the baseline models but also

the other approaches as well. Although we have used the simple ID3 algorithm for classification and K-means algorithm for clustering, our model results in high accuracy because of the small, simple, and less sparse models.

We propose a text classification model based on clustering. We build a model by first applying a decision tree classification model on the dataset. Then we cluster the dataset. We applied a classification algorithm on each cluster and then applied an optimization test. When a model is optimized in the cluster, that model is retained, otherwise the cluster is further divided into sub-clusters and so on. We performed experiments for suspicious email detection and two text classification tasks. We observed that in all the three tasks the accuracy of our model is higher than the existing classification models such as ID3, NB, and SVM. On the Reuters-21578 dataset, our model also outperforms the model DCFC [15] and ADCC [16]. We conclude that the simple and less sparse model achieves higher accuracy than more complex and highly sparse data.

Appendix 1

Examples of Suspicious Emails from the Terrorism Email Dataset

Subject: Reaction

We warn the international community not to send their people to the 2010 Hockey World Cup, the Indian Premier League and Commonwealth Games. Nor should their people visit India, if they do, they will be responsible for the consequences.

wanted terrorist Ilyas Kashmiri

Subject: Alert everybody

We have planned to attack trade center on the coming week. This attack is been planned by us to insist on the freedom of our people. We are ready to loss our lives for the sake of Our people if possible, try your level best to save your lives.

Your enemies

Subject: Map

Attached following is the map of the target place. Follow it carefully before the day of attack.

regards

Subject: Government of India

This attack is a reaction to those actions which Hindus have taken since 1947 onwards. Now, there shall be no actions. There shall only be reactions, again and again. These shall continue until we have avenged each and every atrocity."

suspects of mumbai attack

Subject: Threat

This is a terrorist threat! Take this seriously. I hate the way you people are spending money you don't have ... I'm assigning myself to be judge, jury and executioner. Since you folks have spent what you don't have, it's time to pay the ultimate price.

Subject: Flight

Flight 88 has been cancelled. Our today's hijacking plan is cancelled because the plan has been leaked and now the security is high alert. Now for few days you must go anywhere away from the security.

Best Regards

References

1. Appavu, S., Rajaram, R.: Learning to classify threatening e-mail. Int. J. Artif. Intell. Soft Comput. **1**, 39–51 (2008)
2. Backer, L.D., McCallum, A.K.: Distributional Clustering of Words for Text Classification. In: 21st annual international ACM SIGIR conference on Research and development in information retrieval. SIGIR'98. ACM (1998)
3. Bekkerman, R., Allan, J.: Using Bigrams in Text Categorization. CIIR Technical Report. (2005)
4. Brown, P.F, deSouza, P.V., Mercer, Pietra, V.J.D., Lai, J.C.: Class-based n-gram models of natural language. Comput. Linguist. **18**(4), 467–479 (1992)
5. Collection of Methods to Analyze the text. http://code.google.com/p/text-analysis/. Accessed on 1-12-2010
6. Dumais, S., Platt, J., Hackerman, D., Sahami, M.: Inductive Learning Algorithms and Representations for Text Categorization. CIKM'98. ACM. (1998)
7. Freund, Y., Schapire, R.E: A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. **55**, 119–139 (1997)
8. Hall, M., Frank, E., Holmes, G., Pfahringer,B., Reutemann, P., Ian H. Witten, I. H.: The WEKA Data Mining Software: An Update; SIGKDD Explorations, vol. 11(1). (2009)
9. Jing, L., Huang, J., Michael K. Ng., Rong, H.: A feature weighting approach to building classification models by interactive clustering. LNAI, pp. 284–294. Springer, Berlin (2004)
10. Joachims, T: A Statistical Learning Model of Text Classification for Support Vector Machines. In: 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. (2001)
11. Kyriakopoulou, A., Kalamboukis, T.: Using clustering to enhance text classification. In: 30th annual international ACM SIGIR 07, conference on Research and development in information retrieval. (2007)

12. Kyriakopoulou, A., Kalamboukis, T.: Combining Clustering with Classification for Spam Detection in Social Bookmarking Systems. RSDC, 2008
13. Kyriakopoulou, A.: Text Classification Aided by Clustering: A Literature Review. I-Tech Education and Publishing KG, Vienna, Austria (2008)
14. Lewis, D.D.: An evaluation of phrasal and clustered representations on a text categorization task. In: 15th International Annual Conference SIGR'92, pp. 37–50. (1992)
15. Li, Y., Hung, E.: Building a decision cluster forest model to classify high dimensional data with multi-classes. LNAI, pp. 263–277. Springer, Berlin (2009)
16. Li, Y., Hung, E., Chung, K., Huang, J.: Building a decision cluster classification model for high dimensional data by a variable weighting K-means method. In: AI '08 Proceedings of the 21st Australasian Joint Conference on Artificial Intelligence: Advances in Artificial Intelligence. (2008)
17. McCallum, A., Nigam, K.: A comparison of event models for naive bayes text classification. Technical Report . Workshop on Learning for Text Categorization, pp. 41–48 . (1998)
18. Moore, J., Hong, E., Han, S., Boley, D., Gini, M., Gross, R., Hastings, K., Karypis, G., Kumar, V., Mobasher, B.: Web Page Categorization and Feature Selection Using Association Rule and Principal Component Clustering. (1997)
19. National Commission on Terrorist Attacks Upon the United States. http://govinfo.library.unt.edu/911/report/911Report.pdf, (2004). Accessed on 25-08-2010
20. Nizamani, S., Memon, N., Wiil, U.K.: Detecting suspicious emails using improved features. In: IEEE International Conference on Modeling and Simulation Control, pp. 232–236. (2010)
21. Quinlan, J.R.: Induction of Decision Trees. J. Mach. Learn. **1**, 81–106 (1986)
22. Quinlan, J.R.: C4.5: Programs for machine learning. Machine Learning, vol. 16, pp. 235–240. Springer, Berlin (1993)
23. Renuka, D.K., Hamsapriya, T.: Email Classification for Spam Detection using Word Stemming. Int. J. Comput. Appl. **1**, 45–47 (2010)
24. Sebastiani, F.: Machine Learning in Automated Text Categorization. ACM Comput. surv. **34**(1), 1–47 (2002)
25. Schapire, R.E., Singer, Y.: Boostexter: A boosting based system for text categorization. Mach. Learn. **39**(2/3), 135–168 (2000)
26. Schlimmer, J.C., Fisher, D.: A case study of incremental concept induction. In: 5th National Conference on Artificial Intelligence, pp. 496–501. (1986)
27. Tan, P.N., Michael Steinbach, Vipin Kumar: Introduction to Data Mining. pp. 490–530. (2006)
28. Utgoff, P.E: ID5: An incremental ID3. In: 5th International Conference on Machine Learning, pp. 107–120. (1988)
29. Utgoff, P.E.: Incremental induction of decision trees. Mach. Learn. **4**, 161–186 (1989)
30. Utgoff, P.E., Berkman, N.C., Clouse, J.A.: Decision tree induction based on efficient tree restructuring. Mach. Learn. **29**, 5–44 (1997)
31. Vapnik, V.: The Nature of Statistical Theory. Springer, Berlin (1995)
32. Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S. , Zhou, Z.H., Steinbach, M., Hand, D.J., Steinberg, D.: Top 10 algorithms in data mining. Survey paper, Springer, Berlin (2007)
33. Yang, Y., Pederson, J.: Feature selection in statistical learning of text categorization. In: ZCML-97, pp. 412–420. (1997)
34. Yong, Z., Youwen, L., Shixiong, X.: An improved KNN text classification algorithm based on clustering. J. Comput. **4**(3), 230–237 (2009)
35. Zeng, H.J., Wang, X.H., Chen, Z., Ying, W.: CBC: Clustering based text classification. Requiring minimal labeled data. In: 3rd IEEE International Conference on Data Mining. (2003)

Effectiveness of Social Networks for Studying Biological Agents and Identifying Cancer Biomarkers

Ghada Naji, Mohamad Nagi, Abdallah M. ElSheikh, Shang Gao, Keivan Kianmehr, Tansel Özyer, Jon Rokne, Douglas Demetrick, Mick Ridley, and Reda Alhajj

Abstract Social networks form phenomena that exist and evolve; they are dynamic. These phenomena have been realized and studied by the anthropology and sociology research communities since 1930. However, the recent rapid development in information technology and the internet has increased the interest in social networks and as a model they have been adapted to more applications and domains. Though researchers first studied social networks of humans, for our study described in this chapter we argue that genes and proteins act collaboratively and exist in communities analogous to humans, animals, insects, etc. They complement each other and collectively achieve specific tasks where some would have major roles appearing upfront and others may play minor background roles. However, molecules turn into aggressive actors when their internal structure is augmented; consequently, they may deviate from their target, change camp, and disturb other molecules leading to disaster. Such mutations may be uncontrolled and unintentionally occur

G. Naji
Department of Biology, Faculty of Sciences III, Lebanese University, Tripoli, Lebanon

M. Nagi · M. Ridley
Department of Computing, School of Computing Informatics and Media, University of Bradford, Bradford, UK

A.M. ElSheikh · S. Gao · K. Kianmehr · T. Özyer · J. Rokne
Computer Science Department, University of Calgary, Calgary, AB, Canada

D. Demetrick
Departments of Pathology, Oncology and Biochemistry & Molecular Biology,
University of Calgary, Calgary, AB, Canada

R. Alhajj (✉)
Computer Science Department, University of Calgary, Calgary, AB, Canada
and
Department of Computer Science, Global University, Beirut, Lebanon
and
Department of Information Technology, Hellenic American University, NH, USA
e-mail: alhajj@ucalgary.ca

inside a body, or they may be intentional and controlled by humans to serve one of two purposes, treatment or bioterrorism. In other words, mutation in the molecules (genes) can lead to a change in behavior. This may lead to good or bad effect, e.g., recovery from illness or diseases that may severely affect the body causing disability or death. Once mutated outside the body, molecules may turn into harmful biological weapons of mass destruction. The latter process does not require sophisticated equipment and hence is extremely dangerous with the uprising global terrorism activities. Bioterrorism is therefore a serious concern for humanity. One could say that mutated biological agents outside the body once misused could be way more dangerous than mutated molecules within the body. In this chapter, we will elaborate on bioterrorism and its consequences; we will also propose a model to study social networks of genes within the body leading to the identification of disease biomarkers.

1 Introduction

We witness the establishment of social communities as part of our daily life. Since the dawn of humans on earth they have tended to come together and socialize leading to social communities that evolved into the currently known nations. The existing communities could dynamically change by having new members joining and some of the existing members leaving. For example, it is very common for people to move from one country to another and even from one location to another within the same country. People may change political parties. Employees may change employers and even departments within the same organization. This natural phenomenon has been studied by researchers in anthropology and sociology from early in the twentieth century. Social network research today is multidisciplinary requiring expertise from anthropology, sociology [22], behavioral science, psychology, statistics, mathematics, computer science, etc. Finding a balance between these domains of knowledge is by itself challenging and requires significant effort. We further argue that social network methodology is rich enough to successfully serve a variety of applications, including web mining [33, 51], web services, personalized search, patient monitoring, biological networks [38], disease biomarker detection, outlier detection, team work, wildlife monitoring, document and text analysis, database design and partitioning, traffic monitoring, homeland security, among others [25, 30, 46, 50, 52, 57, 58]. The analysis of social networks leads to valuable discoveries that may have essential social and economic impact. From the social perspective, the discoveries may highlight terrorism groups [25, 52, 57], common hobbies, family relationships, social functions, occupations, friendship, disease biomarkers, etc. From the economic perspective, the analysis may lead to certain target customer groups, the development of drugs, exceptional weather conditions, unusual trends in the stock market, etc.

For long time, the social network methodology was dependent on manual processes and hence concentrated on small networks involving mostly humans

[4, 9–11, 30, 46, 47, 58, 65]. The main theme was to analyze human interactions in specific environments in order to discover key persons, groups, etc. However, the recent development in information technology and the wide spread of the world wide web have highly influenced and shaped the research in social networks [33, 53, 64]. People are joining online social networks and socialize on the web. There is a clear shift from real to virtual life. Furthermore, researchers in different fields have started to realize the effectiveness of social network models in a variety of new applications.

1.1   The Social Network Model

The simplest model of a social network is a set of actors linked by certain type of relationship [11]. Actors (interchangeably called individuals) such as words, pieces of code, items in stock, routers in a network, employees, physicians, patients, trends in weather behavior, animals, humans, insects, genes, proteins, drugs, web pages, etc., can be grouped together into communities. Graphs form most attractive representation of social networks. The edges in a graph may represent social interactions, organizational structures, physical proximity, or even more abstract interactions such as hyperlinks or similarity, among others. The study of social networks has been extensively realized in the research community as we see more conferences emerging and more journals starting dedicated to the computing aspects of mining and analysis of social networks. This will then further interest in the field from more researchers who are willing to contribute to social network research.

If it is possible, though challenging to figure out social communities of abstract items like pieces of code or physical objects like flowers. Realizing the social communities of dynamic and moving objects like stars, humans and animals, e.g., [25] might require less effort. The actors in one community share something which distinguishes their community from the other existing communities. The overall structure of communities is in general hierarchical. For instance, humanity forms a single huge community sharing minimal characteristics of being human. This single huge community is composed of subcommunities that have further distinguishing characteristics which recursively split into subcommunities by considering further characteristics until we end up having each individual family as a community. In the same manner in a cell all the gene products form one large community which could be distributed into smaller communities by considering more features of the gene products. One gene product may belong to two or more communities and it would have a specific role within each community. Finally, social communities are mostly dynamic (whether modeled as stationary or dynamic actors) [9]; and hence the analysis of evolving communities is more demanding to cope with the dynamics of the changing network [9]. Data mining techniques have been effectively used to analyze and study social networks. The social network modeling framework described in this chapter is heavily based on data mining techniques.

1.2 Effectiveness of Data Mining

Data mining techniques are attractive for studying and analyzing huge amounts of data from different points of view and for summarizing the data into useful information through which knowledge hidden within the data can be extracted. In general, data mining techniques are used for social network analysis where people who have similar social profiles are grouped together in social communities [47]; for instance, people may be grouped into small communities based on their income and exchanged emails. Data mining also involves the process of finding correlations in the data. For instance, data mining is used by companies to target their promotions to certain individuals based on the individuals' purchase history; it can also be used in market prediction where it is possible to predict the market status based on the history. One of the data mining techniques successful for such applications is clustering, which has been extensively studied in the literature [41].

Clustering is simply the process of grouping a given set of instances into classes/clusters such that instances in each class are similar and instances in different classes are dissimilar. Clustering may lead to different grouping patterns based on the set of features considered in the process. For instance, people may be classified differently based on any combination of age, sex, weight, nationality, address, etc.

Clustering may be supervised when the classes are known in advance; it is unsupervised when the classes are yet to be determined. The former is called classification and involves two major steps: model construction by using training data to build a model which is tested using remaining unseen data instances in order to determine its accuracy. On the other hand, clustering as unsupervised learning process does not rely on predefined classes and training examples while classifying the data objects. It is the duty of the clustering approach to decide on the classes. Clustering has been successfully applied to different domains, including customer classification, web data, financial data, spatial data, explanatory pattern-analysis and classification, grouping, decision making, document retrieval, gene expression data analysis, image segmentation, among others. Objects are clustered based on some similarity measures, namely homogeneity and separateness. Determining the number of clusters is a problem common to most of the existing algorithms. Usually the number of clusters is user defined. The best number of clusters can be defined by running the algorithm for different values of the number of clusters, and then choose the most appropriate value based on majority voting of some validation indexes. We further argue that even clustering algorithms that claim to run without the number of clusters predefined (like DBScan) still require specifying some parameters that guide the clustering process; tuning and deciding on the latter parameters is equally challenging as deciding on the best number of clusters. In this study, we use k-means clustering algorithm, which requires the number of clusters as input. We decide on the number of clusters by considering the result from applying frequent pattern mining as described in the sequel.

1.3 Realizing Molecule Interactions as Social Network

In this chapter, we concentrate on social networks of molecules with particular emphasis on their mutations, which we define as a significant change in their natural function, since careful attention has to be given to the case of beneficial mutations and precautions have to be taken to stop/avoid/prevent the consequences of bad mutations. We first cover bioterrorism and then describe how the social network model fits well for identifying disease biomarkers. We mainly concentrate on social communities of gene products. A gene can be defined as a string of nucleotides from the alphabet $\{A, T, G, C\}$, where each letter represents a chemical compound. Further details of genetics are out of the scope of this chapter. These genes are the functional segments of DNA that encode mRNA which generates the proteins. Genes form the basic unit for heredity and genetic material transformation. Surprisingly, recent research has revealed that less that 5% of the whole DNA sequence in human involves functional genes. The number of genes differs among organisms; for example, a human genome encodes approximately 20,500 genes, and yeast has around 6,000 genes.

We argue that genes, RNA and proteins form communities. For instance, it is known that some proteins collectively function to maintain cellular protein conformation during stressful proteotoxic insults. Signaling pathways such as that regulating cell division can be thought of as a community. Likewise, investigating whole microbial communities instead of individual micro-organisms could guide scientists to answer fundamental questions such as how ecosystems respond to climate change or pollution. Actually, all the interactions between different organisms within an ecosystem must be taken into consideration in order to assess the environmental impact on microbial communities. This is not possible by following traditional approaches such as those which examine changes in gene expression of individual microbial cells. It is more realistic to investigate and analyze the gene expression of a whole community at once.

We describe a model for identifying disease biomarkers by mining and analysis of social communities of genes and/or gene products. The discoveries from the study described here will support and extend our previous findings related to biomarkers [6, 7]. The social network of gene expressions is watched over time in order to study the behavior of gene products: how they change camps to assume different functions, and how they affect other molecules within the cell. We employ different data snapshots related to the same set of gene products and collected from the same patients at different time points. A social network of the gene products is derived based on each snapshot. The different social networks are analyzed to identify and compare the expressed genes. The outcome will lead to highlight the benefit or thread caused by the analyzed genes.

A social network is derived by employing three perspectives. The first perspective is frequent pattern mining where gene expression and samples represent items and transactions, respectively. We utilize maximal-frequent sets of genes. In our discussion of the machine learning applications, the set of genes is frequent when it

satisfies a minimum frequency threshold by having the genes concurrently expressed in a certain number of samples. It is closed if none of its supersets has the same frequency in the analyzed snapshot; it is maximal-closed if in addition to being closed, none of its supersets is frequent. The second perspective is the k-means clustering algorithm where the number of clusters k is set to the number of derived maximal-closed frequent sets of genes. The coexistence of two genes in the frequent patterns is used to decide on whether to have them connected in the social network, where we use the term "gene" simply as a label for the "gene product". When the two genes exist in the same cluster, the weight of the connection is set to the number of frequent patterns hosting the two genes concurrently; otherwise, the weight is set to the reciprocal of the latter value. The weight is adjusted further based on a third perspective which considers the data snapshot directly as discussed later in Sect. 2. The constructed social networks from all the data snapshots are to be investigated further to derive social communities that lead to better identification of biomarkers, where we define "biomarker" to be a "biological property", and not in the conventional sense of a biological predictor of clinical behavior. Though our results reported in this chapter are related to cancer biomarkers, the approach is general enough to be applied to identify biomarkers of other diseases should the corresponding data be available for modeling and analysis. The latter study has been left as future work. Finally, support vector machines (SVMs) are used to classify the given samples based on the identified biomarkers.

The remaining parts of this paper are organized as follows. Section 2 is an overview of the social networks methodology. Bioterrorism is briefly discussed in Sect. 3. Section 4 summarizes the literature on disease biomarkers with emphasis on cancer biomarkers. Section 5 presents the proposed framework to identify social communities of genes and their analysis. Section 6 reports the experimental results. Section 7 is conclusions and future work.

2 Basic Methodology for Social Network Analysis

The simplest model of a social network involves a homogeneous set of actors which are connected based on certain criteria. For instance, in pharmacology the actors could be drugs, and two actors are connected if it is possible for them to appear together in the same prescription. In a university environment the actors could be students and two actors are connected if they are enrolled together in at least 3 courses during the current semester. Analyzing the first network will lead to communities of drugs that are used together and analysis of the second network will lead to communities of students who study together during the current semester. The identified students may then have the potential to be enrolled in the same courses in the future. The latter piece of information may be valuable in developing a recommendation system for students to select courses such that a set of courses is recommended to students who belong to the same community. The links may reflect either binary relationships (a missing link indicates no relationship) or weighted

connections to indicate the strength or degree of the relationship (which may be negative or positive). A graph that represents a social network is generally called a *sociogram*.

It is also possible to have more than one set of actors. The most common trend is to have two disjoined sets of actors like mRNA and proteins. Then, two actors can be connected if and only if they do not belong to the same set and they satisfy the criteria employed to derive the links, like a mRNA and a protein may be connected to show that the mRNA translates into the protein. This model is better represented using bipartite graph. For instance, drugs and diseases may be two sets of actors such that a link between a drug and a disease indicates the usage of the drug for treating the disease. Another example involves students and courses as two sets of actors such that a student is linked to every course he/she is enrolled in. The first network could be analyzed to discover the most important drugs used in treating most of the diseases. The second network leads to valuable information like identifying the most crowded courses.

The number of actors' groups in the model specifies the degree of the mode for the social network. The two versions described above are known as one-mode and two-mode social networks. These are the two most common settings. A two-mode social network is constructed for the third perspective that we have considered for enriching the social networks of genes tackled in this study. The two mode network is derived directly from the data snapshot where the two sets of actors are the genes and the sample. A link is present in the social network between gene g and sample s if and only if g is expressed in s.

It is possible to derive two one-mode networks from a two-mode network by applying a process known as folding, which operates directly on the adjacency matrix that corresponds to the social network. Assume the adjacency matrix for the two-mode network of genes and samples is constructed so that rows represent genes and columns represent samples. Multiplying the adjacency matrix by its transpose will produce a new adjacency matrix for a one-mode social network where the actors are the genes. The links between the actors in the latter network reflect the influence of the samples in the original two-mode network. On the other hand, multiplying the transpose by the original adjacency matrix will lead to a one-mode network where the actors are the samples. For our study, we are interested in the one-mode network of genes. The outcome from this perspective will influence the social network derived by considering the frequent pattern mining and clustering.

Once a model is constructed, social network analysis can be applied to knowledge discovery in the model. The analysis is classified into two categories, individual centric and group centric. The former analysis starts the from a single actor as a key player in the network and studies its neighborhood. The latter on the other hand considers the whole group at once, and studies the interactions within the group as a whole. The choice of which approach to follow depends on whether the interest is in studying the whole group at once or on identifying individual leaders and use them to influence the whole group. For instance, we may study the correlation between different terms in a book by analyzing how frequent the terms occur together in the same chapter of the book, and we may even fine-tune

the analysis to consider sections or paragraphs instead of chapters. We may also study the social communities of actors in a play by analyzing how often they come together in the same scene or sketch.

Formally, a social network is represented as a graph with weights, denoted by $G = (V, E, W)$, where V is the set of actors in the network, E is the set of edges connecting the actors to indicate relationship, and W is a $|V| \times |V|$ matrix of real values representing the weights of the different links between the $|V|$ actors. Normally, W is neglected in a social network with binary relationships between the actors. For totally connected graphs, $E = (V \times V)$, but in general $E \subseteq (V \times V)$. Once the graph is constructed, different metrics are employed in the analysis for knowledge discovery. The most commonly used metrics include density, centrality [19, 31], and cliques' identification (where every node is connected to every other node in a clique). Finding complete cliques may not be possible in real life; thus we try to find the maximally connected groups which are very close to form a clique.

Density is measured as the ratio of the number of edges in E over the total number of edges in a complete graph (which is $|V| \times (|V| - 1)$ for a complete directed graph). Density gives an indication about cohesion. Density may also be applied to subgroups by considering subgraphs instead of the whole graph. Within a subgraph, density measures in the subgraph the ratio of the actual connections to all possible connections. Measuring the density between two groups (subgraphs) is also possible.

Centrality generally refers to the importance of individual actors in a given group. Centrality may be also measured in terms of degree, betweenness [18, 40, 60], closeness and eigenvector. Degree is a simple measure realized for actor a as the number of actors connected to a divided by the total number of actors minus one (i.e., $\frac{degree(a)}{|V|-1}$); degree is distinguished as in-degree and out-degree of each node in directed graphs. In case graph G is weighted, any of the three values in-degree weight, out-degree weight, or gain in weight could be used to measure the degree of centrality. For actor a, and by considering each other actor i, the latter three weighted measures are computed as:

$$in\_degree(a) = \sum_i (W_{ia}). \tag{1}$$

$$out\_degree(a) = \sum_i (W_{ai}). \tag{2}$$

$$gain\_in\_weight(a) = \sum_i (W_{ia} - W_{ia}). \tag{3}$$

Anthonisse [8] and Freeman [35] independently introduced betweenness as a measure of centrality for the analysis of social networks which only considers the shortest paths in the graph. It refers to how a given actor could be considered as the hub of the network and this is determined by the number of shortest paths that pass via the given actor. In other words, other actors do not have direct link and must

communicate via the given actor. Formally, let $d_{i,j}$ be the shortest distance between two actors i and j, $\sigma_{i,j} = \sum_i^{|V|} \sum_j^{|V|}(d_{i,j})$ is the set of shortest paths between all pairs of actors i and j in the social network, and $\sigma_{i,a,j} = \sum_i^{|V|} \sum_j^{|V|}(d_{i,a} + d_{a,j})$ is the set of shortest paths that pass via actor a and connect i to j, i.e., $\sigma_{i,a,j} \subseteq \sigma_{i,j}$; the betweenness of actor a is computed as:

$$Betweenness(a) = \frac{\sigma_{i,a,j}}{\sigma_{i,j}}. \tag{4}$$

Betweenness may be also measured using Bavelas–Leavitt index [12], which is the reciprocal of (4). The Bavelas–Leavitt index of centrality for a given actor a, denoted $BL(a)$, is therefore computed as:

$$BL(a) = \frac{\sigma_{i,j}}{\sigma_{i,a,j}}. \tag{5}$$

By considering connected actors (whether directly or indirectly), closeness can be regarded as a measure of how long it will take information to spread from a given actor to other reachable actors in the network. It is the ratio of the number of links that an actor must follow to visit each of the other reachable actors in the network and the actor is considered more central when its ratio of closeness is closer to 1, i.e., it is directly connected to all reachable actors. Once individual groups are identified, it is possible to study their overlapping members in order to figure out the interactions between the different communities within the network. Having interrelated communities is preferred to isolated ones. This is equivalent to fuzzy clustering which is a more natural way of expressing membership for most real-world applications.

Eigenvector centrality measures the importance of an actor in a social network. It is computed based on the adjacency matrix A, where entry $A_{i,j} = 1$ if i and j represent adjacent (directly connected) actors in the social network and $A_{i,j} = 0$ otherwise. On the other hand, $A_{i,j} = W_{i,j}$ for weighted graphs. The eigenvector centrality measure for actor a, denoted e_a, is computed as:

$$e_a = \frac{1}{\lambda} \sum_{j=1, j \neq a}^{|V|} (e_j) = \frac{1}{\lambda} \sum_{j=1}^{|V|} (A_{i,j} \times e_j), \tag{6}$$

where λ is a constant called the eigenvalue when (6) is written in vector notation as $A\,e = \lambda e$. If there are n actors in the social network, then the adjacency matrix is of size $n \times n$. For such matrices, there are n eigenvalues and we are interested in the largest eigenvalue. The eigenvector e of the latter eigenvalue is used for the eigenvector centrality in the (6) measure for actor a.

For the study described in this chapter, we compute the number of times two genes occur together in the same maximal-closed frequent set of genes and the outcome is supported by the results from applying two other perspectives, namely

the outcome from k-means clustering and the outcome from folding the two-mode network of genes and samples. The framework is described next in Sect. 5.

3 Bioterrorism

Bioterrorism involves the aggressive and planned release of biological agents like viruses, bacteria, or other germs with the target to severely affect humans, animals, resources or the economy by causing illness or death in people, animals, or plants. The utilized biological agents are typically found in nature, but it is possible that they could be mutated or engineered to stimulate and/or increase their ability to spread into the environment and to cause damage as well as to make them resistant to current medicines.

Mutated biological agents can spread through the environment, including air, water, or food. Some agents used in bioterrorism, like the smallpox virus [1, 32, 42, 44, 49, 55, 56] can spread from person to person and other agents, like anthrax [24,34], can not. In his study Mayr [55] emphasized smallpox. He described many pox viruses and how they can be potential threats to human and animal life. The complications involved with post-vaccinal impairments are also discussed. Agosto [1] discussed the smallpox and the Variola viruses and he also elaborated on how these viruses may turn into epidemics causing panic and social unrest.

For a long time, humans have used in conflicts biological agents ranging from naive simple to recent sophisticated agents. For instance, it is claimed that back in the history the Assyrians poisoned the wells of their enemies with rye ergot. In the recent history, people used biological weapons to severely affect their enemies in a trial to go for quick victory. However, they never considered the consequences of having the effect of the biological weapons staying in the region after it is controlled.

Biological weapons were heavily used during World War I and World Ware II, e.g., [23]. However, there are tremendous efforts since 1970s to prevent the development of biological weapons. Unfortunately, the development of biological weapons is attractive to less developed countries as well as to terrorists [68] because biological agents are relatively easy and inexpensive to obtain or produce. They can be easily disseminated and they can cause widespread fear and panic beyond the actual physical damage they can cause.

The development in biological weapons has been much faster than the development of the medication to treat the affected casualities. This has led to epidemics in the targeted regions. However, over time, biological warfare became more complex and countries began to develop weapons which were much more effective on the targeted group, and much less likely to cause infection to the wrong party. One significant enhancement in development of biological weapons was the use of anthrax. Finally, for long time, bioterrorism has been used to target individuals and groups as well as the economy. For instance, there has been recently several incidences of delivering letters laced with infectious anthrax.

As far as the economy is concerned, there is no clear evidence whether the source of the infecting viruses is from the nature or manmade. For instance, the recent spread in viruses like the swine flu (influenza), bird flu and SARS has raised major concerns worldwide as people started speculations who could be the planner and what could be the goal. Major serious concerns could be raised after closer look and deeper investigation of the areas hit by the virus and this may lead to the belief that it is not a coincidence. One can imagine the SARS hitting in China where the economy is booming. The H1N1 hit in Mexico, one of the tourism attractions in North America. The foot-and-mouth disease virus affected the economy of UK in 2001 and 2007, in addition to a number of other countries since then though to a less extent. Finally, Blancou and Pearson [16] discussed the consequences of bioterrorism on the economy.

It is not possible at all to control the development and spread of biological weapons. Students studying genetics engineering are attractive targets for involvement in producing biological weapons. Accordingly, it is important to carefully select these students and to educate them how to use their knowledge and expertise solely to the benefit of the humanity. They should be armed with the knowledge that will guard them from being deceived by people and organizations which are mainly terrorists who may misuse their expertise and knowledge. As long as there are conflicts and competitions, people who feel themselves weaker and cornered like terrorists will preferentially choose easy to produce, transport and use harmful weapons, that is biological weapons. As long as there exist people who are willing to kill others, it will not be possible to eliminate biological weapons. Terrorists will always look for maximizing causalities and will not hesitate in using biological weapons once they get the ability to acquire them. In other words, ongoing efforts to use biological warfare has been more apparent in small radical organizations attempting to create fear in the eyes of large groups. Some efforts have only been partially effective in creating fear, due to the lack of visibility associated with modern biological weapon used by small organizations.

As long as we are not able to globally identify and get ride of all terrorists, it might be more feasible to educate people in two main directions. First, people should be warned not to be deceived by terrorists who try brainwash them and to turn them into dangerous individuals. Terrorists try to approach people emotionally by using different factors including poverty, political and ethnic discrimination, and religious speeches falsely manipulated to suit the propaganda of the terrorists. Thus, it is possible to avoid losing the fight against terrorism by concentrating on effective ways to handle the factors mostly misutilized by terrorists, by opening new job opportunities that will eliminate poverty, by appropriately dealing with all ethnic groups to address their concerns and meet their reasonable expectations and by spreading the right and true understanding of religion. Misunderstandings of religion has always created problems and conflicts. It is important to watch out for unknowledgeable scholars who play with the emotions of the youth and motivate them to turn into terrorist candidates. This issue is going out of control where it is easy for anyone to declare himself as a scholar and start spreading poison into the minds of the youth. Preventing this would be much easier than recovering

and cleaning up the mess created by false propaganda. However international coordination and collaboration will be needed to effectively tackle the problem, especially with the widespread of the internet based propaganda and TV channels. Equally important is to have people feel that they are equally treated and not discriminated against. Preferential treatment to certain groups is one of the main items in the propaganda of terrorists. Second, people should be educated how to protect themselves from biological weapons and how to be treated in case they are affected. The latter issue has been addressed to some extent by different groups. For instance, Bronze [21] describes potential vaccines and pharmaceutical strategies for either prevention or treatment of established infections. Blank et al. [17] discussed how to learn lessons from the anthrax attacks of 2001 in order to react better to future bioterrorism attacks. The authors also analyzed the effectiveness of the distribution of antibiotics before and after an attack. Binder et al. [15] commented on the importance of using medicine and science as well as public knowledge to defend against bioterrorism.

4 Related Work on Identifying Disease Biomarkers

Identifying disease biomarkers in general and cancer biomarkers in particular is an interesting research problem that has received the attention of a number of research groups who tackled the problem of determining the best biomarkers for different types of diseases like cancer, including leukemia, bladder, lung, prostate, liver, breast, etc. Devarajan [28] reported some biomarkers that could help in the early prediction of acute kidney injury. Sahab et al. [67] presented a good overview of biomarkers for different diseases including heart, rheumatoid arthritis, asthma and cystic fibrosis, in addition to several cancer biomarkers like prostate, breast, ovarian and lung.

Leukemia is one type of cancer which has been extensively studied in the literature. Golub et al. [39] may be considered as the first group who tried to distinguish between acute myelogenous leukemia (AML) and acute lymphocytic leukemia (ALL) based on gene expression data. They used a model of self-organizing maps in combination with a weighted voting scheme. They obtained a strong prediction of 29/34 samples in the test data using 50 genes. Furey et al. [36] applied SVMs to the AML/ALL data. Significant genes are derived based on a score calculated from the mean and standard deviation of each gene type. Tests are performed for the 25, 250, 500, and 1,000 top ranked genes. At least two test samples are misclassified in all SVM tests. Guyon et al. [75] also applied SVM, but with a different feature selection method called recursive feature elimination. For each iteration, the weights associated with the genes are recalculated by a SVM test, and genes with the smallest weights are removed. On 8 and 16 genes, the classification error on the test set is zero. Jinyan and Wong [48] used new feature selection method called emerging patterns. When they applied their method on the

AML/ALL data, they were able to identify one gene (zyxin), which was able to classify 31/34 of the samples.

Tseng et al. [72] studied collectively potential diagnostic biomarkers for four cancer types, namely liver, prostate, lung and bladder. They identified 99 distinct multi-cancer biomarkers in the comparison of all three tissues in liver and prostate and 44 in the comparison of normal versus tumor in liver, prostate and lung. They also found out that bladder samples have different list of biomarkers from the other three cancer types. Shariat et al. [69] studied whether the assessment of five previously characterized bladder cancer biomarkers (p53, pRB, p21, p27, and cyclin E1) could improve the ability to predict disease recurrence and cancer-specific survival after radical cystectomy in patients with pTa-3N0M0 urothelial carcinoma of the bladder.

Wagner et al. [74] discuss several classification-based approaches to finding protein biomarker candidates using protein profiles obtained via mass spectrometry; they assess the statistical significance of the discovered biomarkers; the target is to link the biomarkers to given disease states, and thus to narrow the search for biomarker candidates.

Sorensen and Orntoft [70] covered advances in biomarker discovery for prostate cancer by microarray profiling of mRNA and microRNA expression. The authors discussed limitations in the application of microarray-based expression profiling for identification of prostate cancer biomarkers and hence the strong need for promising biomarkers to enable more accurate detection of prostate cancer, improve prediction of tumor aggressiveness and facilitate discovery of new therapeutic targets for tailored medicine.

Toure and Basu [73] applied a neural network to cancer classification where 10 genes were used for classification purposes. The neural network was able to fully separate the two classes during the training phase. However, the classification of the test set samples did not achieve high accuracy since 15 samples were misclassified. In another work, Li et al. [54] applied GA/KNN method to the same data set where 50 genes were used for classification. GA/KNN correctly classified all training set samples, and all but one of the 34 test samples. Bijlani et al. [14] used independently consistent expression discriminator (ICED) for feature extraction. They could select 3 AML distinctors and 13 ALL distinctors, which were able to classify the training set without any errors; but one sample was misclassified in the test data. Zhang and Ke [79] used the 50 genes reported by Golub et al.[39] and applied SVM and central support vector machine (CSVM) for classification. Two misclassifications occurred using SVM, but no errors were reported when CSVM was used.

As the colon data set is concerned, Bicciato et al.[13] used Autoassociative Neural network model for classification. No sample was misclassified during the training session, while seven samples received wrong classification during the testing phase. In another study [76], Wang et al. used SVM for classification, 4 samples were misclassified. Tusher et al.[73] used two novel classification models. The first was combination of optimally selected self-organizing map, followed by fuzzy C-means clustering. And the second was the use of pair-wise fisher's linear discriminant. In the former model, 12% of the samples were misclassified and in the

latter model 18% of the samples were incorrectly classified. Several other studies investigated the colon data set, e.g., [14]. At least four misclassifications occurred in most of the studies carried out. Many other groups have demonstrated the power of fuzzy logic in microarray analysis, e.g., [66, 75]; but the main problem of how to choose the fuzziness parameter m has to be further investigated. Usually 2 is the preferred value for m. However, Dembele and Kastner [26] have shown that 2 is generally not appropriate value for all data sets. Our group has already reported interesting results regarding the better choice of the values of m [6, 7].

5 Identifying Social Communities of Genes by Frequent Pattern Mining, K-means and Network Folding

In this section, we describe our approach to analyzing gene expression data that has been employed to generate the social communities of genes, which are further analyzed to identify expression of key genes as disease biomarkers and we concentrate on cancer biomarkers for the data sets used in the testing. Given a set of actors which are the expressed genes in this particular study, our target is to find the links between them in order to establish the social network of the genes. This is possible by investigating the relationship between the different genes. Our approach is based on a unified framework that combines three perspectives, namely frequent pattern mining (maximal-closed sets of genes in particular), k-means clustering and folding of the two-mode social network that connects genes to samples. First, the frequent pattern mining process is employed to produce the initial adjacency matrix where two genes are linked to reflect the number of maximal frequent sets of genes that contain the two genes concurrently. Second, k-means clustering and the folding process are separately applied to adjust the weight from the result from the first perspective. Finally, we compute the average of all the values in the matrix. Based on the latter average we produce an adjacency matrix for all the genes by accepting two genes as adjacent if and only if their correlation value in the adjacency matrix is larger than the average. The adjacency matrix helps in finding communities of genes and the central gene within each community where the latter gene is considered as a disease biomarker.

5.1 Frequent Pattern Mining

Frequent pattern mining is one of the most attractive data mining techniques described in the literature [2, 3]. It was initially developed as the first of two steps for market basket analysis and later it has been adapted to different interesting applications. In general, frequent pattern mining investigates the correlation between items within the transactions in a given database. So, any problem that can be modeled in terms of transactions and items could be classified among the applications of the frequent pattern mining framework. Fortunately, the social network of genes

problem investigated in this work could benefit from frequent pattern mining by considering the samples as transactions and the genes correspond to the set of items.

To understand the frequent pattern mining model, we provide a brief coverage in the rest of this section. Consider a database of transactions, such that each transaction contains a set of items, the frequent pattern mining process determines sets of items that satisfy a minimum support threshold specified mostly by a domain expert where t support of a set X is the percentage of transactions that contain all items in X. It is also possible to derive the minimum support threshold in an automated way by considering characteristics of the data. However, this is outside the scope of the work described in this chapter.

Market basket analysis is one of the first applications of frequent pattern mining [3]. Organizations that deal with transactional data aim at using the analysis outcome to decide on better marketing strategies, to design better promotional activities, to make better product shelving decisions, and above all to use these as a tool to gain competitive advantage.

Agrawal et al. [2] first introduced frequent pattern mining in 1993, and since then it is one of the topics most frequently investigated by researchers in the data mining arena. During the past two decades, several research groups have provided solutions to this problem in many different ways, e.g., [37, 45, 62, 63]. The time and space scalability of the developed approaches greatly vary based on their techniques to mine the investigated databases. They mainly differ in the number of database scans, and hence the time consumed by the mining process, as well as in the data structures they use, which are mostly main memory resident. All the latter mentioned performance related issues are outside the scope of this chapter because the main target is to determine the maximal-closed frequent sets of items regardless of the performance of the approach to be utilized.

A very naive and brute force approach for finding all frequent patterns from a particular database is to generate all possible patterns from the database, and then check the corresponding frequency of each pattern against the database. The problem with this approach is that there can be 2^n (where n is the number of items, genes in our study) candidate patterns to be checked, and it is not computationally or space efficient to determine the frequency of such huge number of patterns.

Over the past two decades, researchers in the area have come up with numerous frequent pattern mining algorithms in an attempt to efficiently solve the problem, e.g., [37, 45, 62, 63]. Covering all frequent pattern mining approaches is outside the scope of this chapter; interested readers may refer to the literature for comprehensive coverage. In this chapter, the focus is on describing the general process of finding maximal-closed patterns which are the target of our study as one possible toos for producing the social network of genes.

5.2 Finding Frequent Sets of Expressed Genes

There are several algorithms for finding frequent patterns in a dataset. We will present the Apriori algorithm here which has been accepted as the first algorithm

developed for frequent pattern mining. It makes multiple passes over the data to find frequent sets of items of all lengths. In the first pass, it counts the support (frequency) of individual items and determines which ones of them are frequent, i.e., satisfy the minimum support constraint. In each subsequent pass, it uses the frequent sets of items generated from the previous pass to produce the new potentially candidate frequent sets of items, and counts the support of these candidate sets of items to find the ones that are indeed frequent.

The Apriori algorithm generates candidate sets of items in the current pass by only considering frequent sets of items from the previous pass. The intuition behind this is based on what is known as the Apriori-heuristic, which states that a set of items may be frequent if all its subset sets of items are frequent. This can be done by a self-join of the frequent sets of items of length k, say L_k with itself and then pruning from the result any set of items which has all of its subsets not included in L_k. This process results in generating a much smaller number of candidate sets of items. Therefore, candidate generation consists of two steps: the join step and the pruning step. After the pruning step, the remaining candidates are checked by scanning the database to determine their frequencies. This process is recursively repeated until it is not possible to construct more frequent sets of items.

The Apriori algorithm is level wise in nature. It requires multiple database scans to find the support count for a potentially large candidate set of items and this can be very time consuming. Moreover, the Apriori algorithm requires generating a large number of candidate sets of items at each level, especially for levels two and three. These can also be considered as CPU and memory intensive tasks.

There are several other Apriori-like algorithms, such as DHP [63], DCP [61], and DCI [62], which mainly focus on improving the performance of mining by reducing the candidate generation and/or by introducing special data structures that reduce the time for counting the support of candidates. On the other hand, algorithms like DIC [20] and CARMA [45] try to improve the performance by reducing the number of database scans.

5.3   Finding Maximal-closed Frequent Itemsets

Redundancy is the main problem with keeping all the frequent sets of genes as described in Sect. 5.2. The number of frequent sets of genes is directly related to how the genes are co-expressed across the samples. As the co-expression across the samples increases, the number of frequent sets of genes increases. However, many of the frequent sets of genes may share the same frequency and noticing this would help in minimizing the number of frequent sets of genes to maintain by keeping only maximal-closed frequent sets of genes. In other words, we reduce the number of frequent sets of genes by only concentrating on maximal-closed frequent sets of genes.

A frequent set of genes is said to be closed if and only if its support is different from the support of all its frequent supersets. A frequent set of genes is maximal-closed if and only if it is closed and none of its supersets is frequent. Based on

this, we keep frequent sets of genes of maximum size. To illustrate the process, consider five samples and six genes. For each sample, we list the expressed genes as follows: $s_1 = \{g_1, g_3, g_4\}$, $s_2 = \{g_2, g_3, g_5\}$, $s_3 = \{g_1, g_2, g_3, g_5\}$, $s_4 = \{g_2, g_5\}$, and $s_6 = \{g_3, g_5, g_6\}$. Assume the minimum support threshold is 2, i.e., a set of genes is said to be frequent if it is supported by at least two samples. Excluding the singleton frequent sets of genes, we could enumerate the following five frequent sets of genes: $\{g_1, g_3\}$, $\{g_2, g_3\}$, $\{g_2, g_5\}$, $\{g_3, g_5\}$, and $\{g_2, g_3, g_5\}$. Out of these sets of genes only four are closed frequent, namely $\{g_1, g_3\}$, $\{g_2, g_5\}$, $\{g_3, g_5\}$ and $\{g_2, g_3, g_5\}$. Finally, only $\{g_1, g_3\}$ and $\{g_2, g_3, g_5\}$ are maximal-closed frequent sets of genes. From this result, it is obvious that g_3 is the most important gene; it has the highest closeness, betweenness and degree centralities. Further, the three genes, g_2, g_3 and g_5 form a clique.

The identified maximal-closed frequent sets of genes will be sufficient for determining the frequency of each set of genes that are mostly co-expressed. This allows us to concentrate only on sets of genes that either have different support, or once have same support they do not totally overlap, i.e., none of them subsumes the other.

5.4 Constructing Social Network of Genes and Identifying Biomarkers

The outcome from the proposed three pronged approach is a rich source of information for constructing a social network. Given the n maximal-closed frequent sets of genes, say $MC_1, MC_2, \ldots, MC_n$ and the utilized m genes, say $g_1, g_2, \ldots, g_m$, we construct a matrix $M = m \times m$ to include one row and one column per gene. Entries in matrix M are computed by considering each frequent set of genes MC_k ($k = 1, n$) and increment $M(i, j)$ by 1 if the pair of genes g_i and g_j exist inside the maximal-closed frequent set of genes MC_k. In other words, $M(i, j) = r$, for all $1 \leq i \leq m$ and $1 \leq j \leq m$, where $0 \leq r \leq n$ is the number of maximal-closed frequent sets of genes in which the pair of genes (g_i, g_j) coexist. It is obvious that $M(i, i) = n$ for all $1 \leq i \leq m$.

To produce a more robust social network of the genes, a second perspective is applied. This perspective works directly on the original data which consists of the samples. A two-mode network of genes and samples is produced. Gene g_i is connected to sample s_j if and only if g_i is expressed in sample s_j. Then, we apply folding on the produced two-mode social network to derive a one-mode social network that covers only the genes. As mentioned above, the one-mode network is produced by multiplying the adjacency matrix of the two-mode network by its transpose. Two genes g_i and g_j are linked in the one-mode network to reflect the strength of having g_i and g_j co-expressed in the same samples. In other words, the one-mode social network of genes is a kind of weighted graph; the weight of each link reflects the degree of co-expressiveness of the two connected genes. The weight

of the link connecting genes g_i and g_j is added to the value in entry (i, j) in matrix M produced by the first perspective.

The third perspective is the k-means clustering algorithm which is used to produce n clusters of the m genes. The clustering result is reflected onto matrix M by considering the following strategy. For every two genes g_i and g_j, if g_i and g_j coexist in the same cluster then the entry $M(i, j)$ is maintained, otherwise (if g_i and g_j do not exist in the same cluster then $M(i, j)$ is replaced by $\frac{1}{M(i,j)}$ as a kind of punishment for the two genes. The basic idea behind this strategy is simple: if two genes are related then they should exist in the same cluster. Actually the test results confirm the validity of this strategy because we realized that genes that do exist together in the same cluster when they coexist in large number of maximal-closed frequent sets of genes.

After all entries in M are determined, we compute the average, say A_v of all the values in M as follows: $A_v = \frac{\sum_{i=1}^{m} \sum_{j=1}^{m} M(i,j)}{m^2}$. Based on the comparison of each entry in M with A_v, we normalize every entry $M(i, j)$ to $\frac{M(i,j)}{A_v}$ and then we set $M(i, j) = 0$ if and only if $M(i, j) < 1$. The revised matrix $\hat{M}$ represents the adjacency matrix of the actual social network where there exist an edge between genes g_i and g_j if and only if $M(i, j) > 0$. After M is transformed into adjacency matrix, genes are clustered into communities by considering the overlap of nonzero values in the corresponding rows. Each gene joins the community where it has more overlap. Then, we determine the central gene within each community by considering the degree of centrality which is determined by computing two values for each gene g in a given community:

1. The weighted degree centrality of gene g, denoted D_g is the sum of the values in the row of g in matrix M divided by the number of genes in the same community, say n_c, $D_g = \frac{\sum_{j=1}^{m} M(g,j)}{n_c}$.
2. The un-weighted degree of centrality is the number of non-zero entries in row g, denoted z_g.

Based on the values of D and z computed for each community, we find the most central gene g within each community as the gene that has high values for D_g and z_g. For this, we sort each of the two lists of values in descending order where the list of z values is given higher priority in the analysis because the values in list D are weighted and hence do not reflect the actual number of neighbor genes. The latter values are considered more seriously to differentiate genes that have closer z values. Central genes within the communities are analyzed further as the biomarkers.

6 Test Results

We have conducted two types of experiments. The first set of experiments is intended to illustrate the validity of our main argument that genes do socialize and form dynamic communities. In the second set of experiments, we directly applied the three-pronged approach to identify some potential diagnostic cancer biomarkers.

## 6.1 Illustrating the Dynamic Behavior of Genes

To illustrate the argument that gene products do socialize and change functional camps, we report the network of the yeast gene expression data is described in [27]. The authors carried out a comprehensive investigation of the temporal program of gene expression accompanying the metabolic shift from fermentation to respiration and they also studied genes whose expression was affected by deletion of the transcriptional co-repressor TUP1 or over-expression of the transcriptional activator YAP1. Using the same data described in [27] social networks are constructed at times 9.5, 13.5 and 20.5 h with 500 genes. The corresponding networks are shown in Figs. 1, 2, and 3, respectively. To explicitly shown the change in connections, we also plotted part of each of the three networks by zooming in to shown certain genes where the three zoomed parts labeled with names of genes are shown in Figs. 4–6, respectively. These figures explicitly demonstrate the changes in the network; and hence the change in the correlation between the expressed genes. This verifies our argument that gene products do socialize and change camps. However, more in depth biological analysis in the wet-lab may be needed to carefully study the communities of genes leading to a solid verification of our claim.
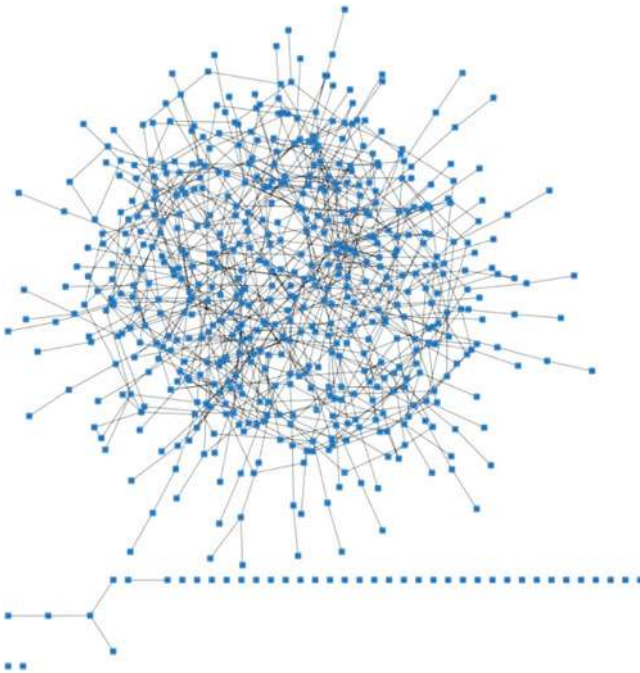
Fig. 1 The complete network of yeast at time 9.5 h
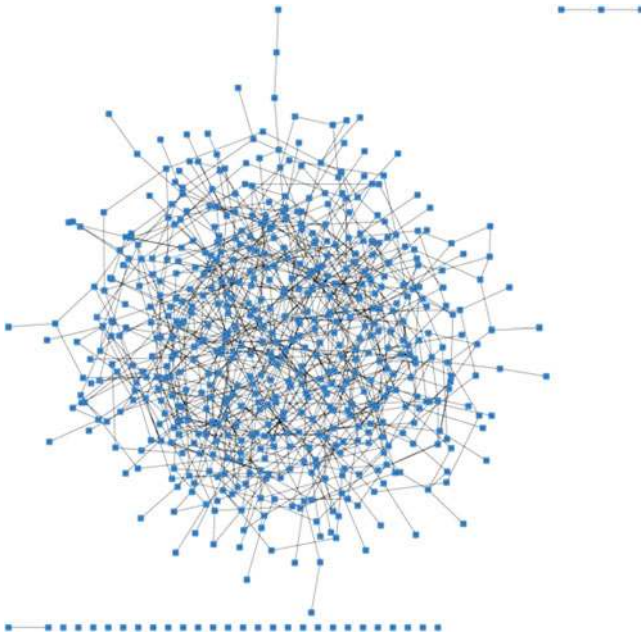
Fig. 2 The complete network of yeast at time 13.5 h
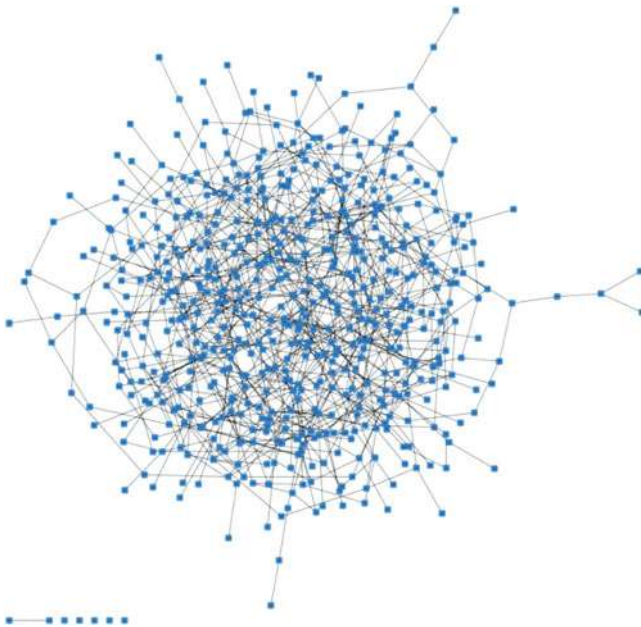
Fig. 3 The complete network of yeast at time 20.5 h
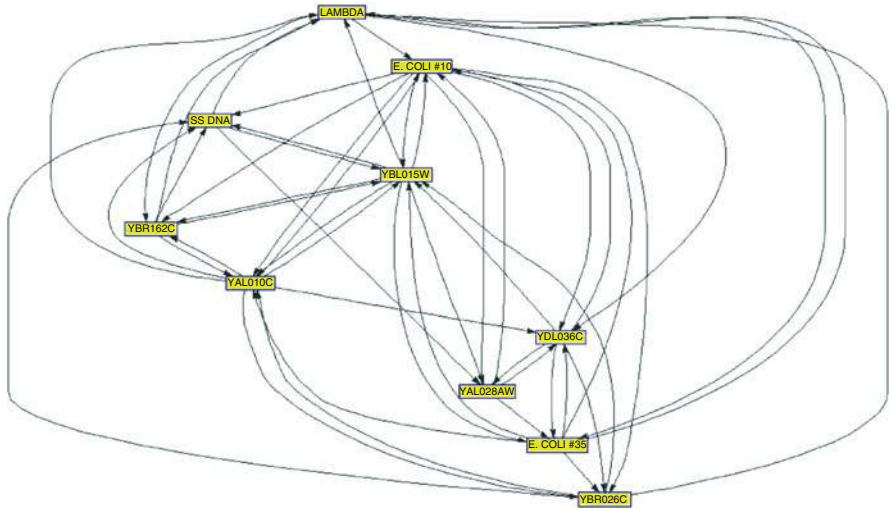
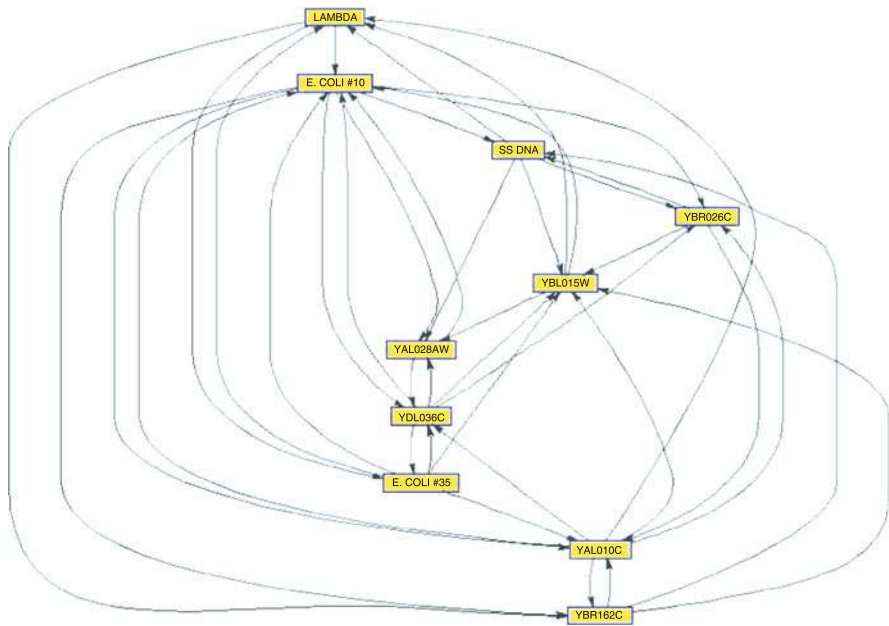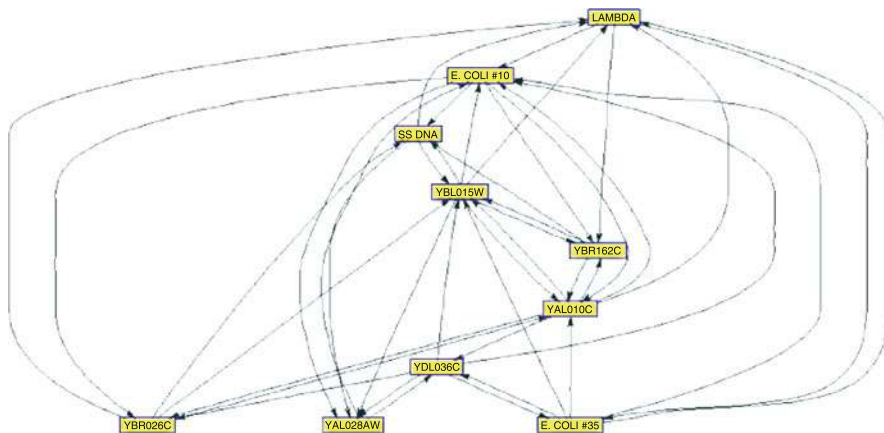**Fig. 4** Part of the network of yeast at time 9.5 h

Fig. 5 Part of the network of yeast at time 13.5 h

Fig. 6 Part of the network of yeast at time 20.5 h

## 6.2 Illustrating the Proposed Social Network Construction Framework

In this section, we discuss the conducted experiments. We highlight the results of our approach and evaluate its effectiveness and applicability. We report the results from the comparison of our approach with other existing methods which investigated the same cancer classification problem. We have also analyzed the results both in terms of accuracy and biological significance.

Data preprocessing has been conducted using matlab 7.0. To derive maximal-closed frequent sets of genes, we first used the CloSpan algorithm originally proposed by Yan et al. [78] and then applied on the result a postprocessing step to find the maximal-closed sets of genes. The clustering has been conducted using the k-means implementation in Matlab. Gene selection has been performed using t2test in matlab. For classification, we have used LIBSVM package implemented in matlab. LIBSVM is a free library for classification and regression available online at (http://www.csie.ntu.edu.tw/~cjlin/libsvm/).

In this work we have used two cancer data sets:

1. Acute myeloid leukemia (AML)/Acute lymphocytic leukemia (ALL) taken from [39]
2. Colon data set from [13]

The AML/ALL data contains 7,130 genes and 73 patient samples. We split this data as follows: 58 sample for training and 15 for testing. The colon cancer data studied has 2,000 genes and 62 patient samples with 40 tumor and 22 normal. Samples were split as follows, 48 samples were used for training and 14 samples were used for testing. As a normalization step, the intensity values have been normalized such
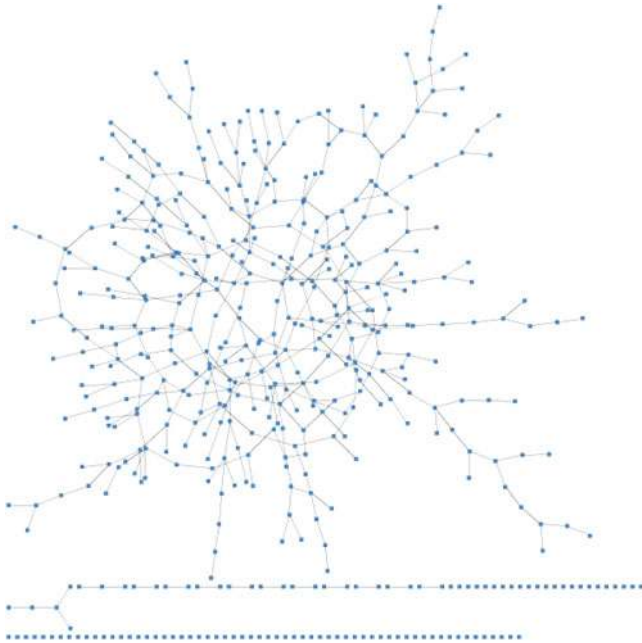
Fig. 7 The network of the lukemia data

that the overall intensities for each chip are equivalent. This is done by fitting a linear regression model using the intensities of all genes in both the first sample (baseline) and each of the other samples. The inverse of the "slope" of the linear regression line becomes the (multiplicative) re-scaling factor for the current sample. This is done for every chip (sample) in the data set, except the baseline which gets a re-scaling factor of one. A missing value means that the spot was not identified. Missing values were predicted according by to k-nearest neighbors strategy; we replace a missing value by the weighted average of the five nearest neighbors; we compute the weighted average by first dividing each of the five nearest neighbors by its distance from the missing value and the computed the average of the five produced values.

We applied the three-pronged approach using the training part from each of the two data sets. First, the frequent pattern mining perspective produced 150 and 62 maximal closed sets of genes for the AML/ALL and colon data, respectively. Second, we applied k-means clustering on each of the two data sets by setting the values of k to the reported number of maximal-closed frequent sets of genes. Third, we applied folding on the two mode social networks of the AML/ALL and colon data. Finally, we combined the results from the three perspectives as described in Sect. 5.4. The network produced from the lukemia data is shown in Fig. 7.

The analysis of the social network constructed for the Lukemia data set revealed six communities of genes. Then the most central gene within each community was

Table 1 Results from other works compared with our results where the errors in the training and testing set were provided and the number of features used for classification. **NA** stands for Not applied in the work

	[39]	[36]	[75]	[48]	[14]	[79]	[5]	Our method
Errors in training set	2	2	0	NA	0	0	0	0
Errors in test set	5	2–4	0	3	1	0	0	0
Number of features used for classification	50	25–1000	8–16	1	16	50	8	6

reported leading to six biomarker genes were a gene is accepted as central when it has high degree centrality and high closeness centrality within its community. We then used the discovered six genes (as data representatives) to build the SVM. The reported results of 100% accuracy (for the training set) and 100% cross validation (for the test set) illustrate the effectiveness of the proposed approach as robust framework for detecting biomarkers. As reported in Table 1, comparing our results to those mentioned in the literature it can be easily seen that the proposed framework is more effective and robust.

For the colon data set, the social network construction and analysis process reported 12 communities. Again we identified the most central gene in each community. The latter 12 genes were then used in building the SVM classifier and the reported results are 85% accuracy and 94% cross validation. Finally, the network for the colon data is shown in Fig. 8. At the end, comparing all the networks plotted in this chapter would lead to stronger support for the argument that genes do socialize and act collaboratively within the cell. In fact every cell contains the same genes and only certain genes are active in each cell type.

7   Summary and Conclusions

We argue that biological molecules within the cell form communities and act collaboratively within those communities to achieve certain goals. Unfortunately, optimal methods that could closely evaluate the interactions between the genes and lead to the extraction of the actual social networks are still lacking. In this work, we tried to construct the social network of gene products by employing a three-pronged approach leading to a robust framework. For visualization, the reader may think of the simple model of two communities of genes within each cell–one community containing expressed genes and the other community containing unexpressed genes for each particular cellular response to a stimulus. The community of expressed genes could be further divided based on the particular cellular goal of expressed genes (cell proliferation, metabolism, synthesis of a particular cellular product, etc.). The aim of our methods is to define the representative genes by analysis of the social communities. Our analysis allows us to find weighted links between
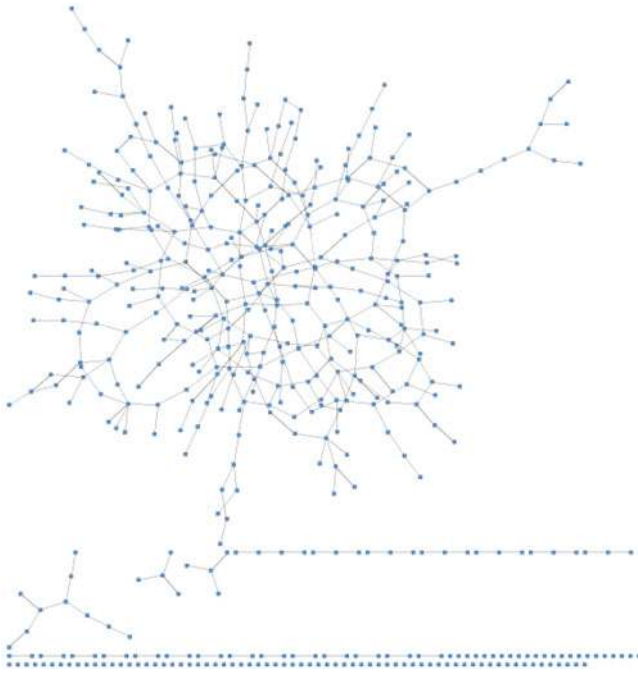
Fig. 8 The network of the colon data

genes based on their co-occurrence in the outcome from the three perspectives employed, namely frequent pattern mining, clustering and folding of the two-mode network. The outputs from the analysis of the social communities reported the most promising biomarker genes. After demonstrating its ability to identify potential cancer diagnostic biomarkers, the same methodology described in this paper can be applied to identify potential biomarkers of other diseases. Also, we are considering social communities of proteins as well as social communities that result from the interactions between genes and proteins. The latter networks are more challenging to study; these are at the center of our current research efforts.

References

1. Agosto, J.: Confronting bioterrorism: Epidemiologic, clinical, and preventive aspects of smallpox. Salud Publica de Mexico, pp. 298–309 (2003)
2. Agrawal, R., Imieliski, T., Swami, A.: Mining association rules between sets of items in large databases. In: Proceedings of ACM SIGMOD International Conference on Management of Data, pp. 207–216. Washington, D.C., May 1993
3. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proceedings of the International Conference on Very Large Data Bases, pp. 487–499. San Francisco, CA, (1994)

4. Albert, R., Barabosi, A.L.: Statistical mechanics of complex networks. Rev. Mod. Phys. **74**(1), 47–97 (2002)

5. Alshalalfa, M., Özyer, T., Alhajj, R., Rokne, J.: Discovering cancer biomarkers: From DNA to communities of genes. Int. J. NVO **8**(1/2), 158–172 (2011)

6. Alshalalfa, M., Alhajj, R.: Cancer class prediction: Two stage clustering approach to identify informative genes. Intell. Data Anal. **13**(4) (2009)

7. Alshalalfa, M., Alhajj, R., Rokne, J.: Identifying disease-related biomarkers by studying social networks of genes. In: Lim, C.P., Jain, L.C. (eds.) New Directions in Decision Support Systems: Methodologies and Applications. Springer, Berlin (2009)

8. Anthonisse, J.M.: The rush in a directed graph. Technical Report BN9/71, Stichting Mahtematisch Centrum, Amsterdam, Oct 1971

9. Backstrom, L., Huttenlocher, D., Kleinberg, J., Lan, X.: Group formation in large social networks: Membership, growth, and evolution. Proceedings of the ACM KDD (2006)

10. Barabosi, A.L., Albert, R.: Emergence of scaling in random networks. Science **286**(5439), 509–512 (1999)

11. Baumes, J., Goldberg, M., Magdon-Ismail, M., Wallace, W.: Discovering hidden groups in communication networks. In: Proceedings of NSF/NIJ Symposium on Intelligence and Security Informatics. (2004)

12. Bavelas, A.A.: A Mathematical model for group structures. Hum. Organ. **7**, 16–30 (1948)

13. Bicciato, S., Pandin, M., Didon, G., Di Bello, C.: Pattern identification and classification in gene expression data using an autoassociative neural network model. Biotechnol. Bioeng. **81**(5), 594–606 (2002)

14. Bijlani, R., Cheng, Y., Pearce, D.A., Brooks, A.I., Ogihara, M.: Prediction of biologically significant components from microarray data: Independently consisitent expression discriminator(ICED). Bioinformatics **19**, 62–70 (2003)

15. Binder, P., et al.: Medical management of biological warfare and bioterrorism: Place of the immunoprevention and the immunotherapy. Comp. Immunol. Microbiol. Infect. Dis. **26**(5–6), 401–421 (2003)

16. Blancou, J., Pearson, J.E.: Bioterrorism and infectious animal diseases. Comp. Immunol. Microbiol. Infect. Dis. **26**(5–6), 431–443 (2003)

17. Blank, S., Moskin, L.C., Zucker, J.R.: An Ounce of Prevention is a Ton of Work: Mass Antibiotic Prophylaxis for Anthrax, New York City, 2001. (Policy Review), Emerg. Infect. Dis., **9**(6), 615–612 (2003)

18. Brandes, U.: A faster algorithm for betweenness centrality. J. Math. Sociol. **25**(2), 163–177 (2001)

19. Brandes, U., Pich, C.: Centrality estimation in large networks. Int. J. Bifurcat. Chaos **17**(7), 2303–2318 (2007)

20. Brin, S., Motwani, R., Ullman, J.D., Tsur, S.: Dynamic itemset counting and implication rules for market basket data. In: Proceedings of ACM SIGMOD International Conference on Management of Data, pp. 255–264. Tucson, Arizona, May 1997

21. Bronze, M.S.: Preventive and therapeutic approaches to viral agents of bioterrorism. Drug Discov. Today 740–745 (2003)

22. Carley, K., Prietula, M. (eds.): Computational Organization Theory. Lawrence Erlbaum associates, Hillsdale, NJ (1994)

23. Christopher, G.W., et al., Biological warfare: A historical perspective. JAMA **278**(5), 412 (1997)

24. Cieslak TJ, Eitzen EM Jr. Clinical and Epidemiologic principles of anthrax, Emerg Infect Dis. **5**(4):552–5. Jul-Aug 1999

25. Croft, D.P., James, R., Thomas, P., Hathaway, C., Mawdsley, D., Laland, K., Krause, J.: Social structure and co-operative interactions in a wild population of guppies (poecilia reticulata). Behav. Ecol. Sociobiol. **59**(5), 644–650 (2006)

26. Dembele, D., Kastner, P.: Fuzzy c-means method for clustering microarray data. Bioinformatics **19**, 973–980 (2003)

27. DeRisi, J.L., Iyer, V.R., Brown, P.O.: Exploring the metabolic and genetic control of gene expression on a genomic scale. Science **278**(5338), 680–686 (1997)
28. Devarajan, P.: Novel biomarkers for the early prediction of acute kidney injury. Cancer Ther. **3**, 477–488 (2005)
29. Diestel, R.: Graph Theory, 2nd edn. Graduate Texts in Mathematics. Springer, Berlin (2000)
30. Domingos, P., Richardson, M.: Mining the network value of customers. In: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, San Francisco, CA (2001)
31. Everett, M.G., Borgatti, S.P.: The centrality of groups and classes. J. Math. Sociol. **23**(3), 181–201 (1999)
32. Ferguson, N.M., et al.: Planning for smallpox outbreaks. Nature **425** (2003)
33. Flake, G.W., Lawrence, S., Giles, C.L.: Efficient identification of web communities. Proceedings of ACM International Conference on Knowledge Discovery and Data Mining, pp.150–160 (2000)
34. Forrester, M., Stanley, S.: Calls about anthrax to the Texas Poison Center Network in relation to the anthrax bioterrorism attack in 2001. Vet. Hum. Toxicol. 247–248 (2003)
35. Freeman, L.C.: A set of measures of centrality based upon betweenness. Sociometry **40**, 35–41 (1977)
36. Furey T.S., Cristianini, N., Duffy, N., Bednarski, D.W., Schummer, M., Haussler, D.: Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics **16**(10), 906–14 (2000)
37. Ganti, V., Gehrke, J., Ramakrishnan, R.: Demon: Mining and monitoring evolving data. IEEE Trans. Knowl. Data Eng. **13**(1), 50–63 (2001)
38. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. **99**(12), 7821–7826 (2002)
39. Golub, T.R., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C., Lender, E.: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science **286**, 531–537 (1999)
40. Gould, R.V.: Measures of betweenness in non-symmetric networks. Soc. Networks **9**, 277–282 (1987)
41. Grabmeier, J., Rudolph, A.: Techniques of cluster algorithms in data mining. Data. Min. Knowl. Discov. **6**, pp.303–360 (2003)
42. Grais, R.F., Ellis, J.H., Glass, G.E.: Forecasting the geographical spread of smallpox case by air travel. Epidemiol. Infect. **131**, 849–857 (2003)
75. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: gene selection for cancer classification using support vector machines. Mach. Learn. **46**(1–3), 389–422 (2002)
44. Halloran, E., Longini Jr, I.M., Nizam, A., Yang, Y.: Containing Bioterrorist smallpox. Science **298** (2002)
45. Hidber, C.: Online association rule mining. In: Proceedings of ACM SIGMOD international conference on Management of data, pp. 145–156, Philadelphia, Pennsylvania (1999)
46. Janssen, M.A., Jager, W.: Simulating market dynamics: Interactions between consumer psychology and social networks. Artif. Life **9**, 343–356 (2003)
47. Jensen, D., Neville, J.: Data mining in social networks. In: Proceedings of the Symposium on Dynamic Social Network Modeling and Analysis (2002)
48. Jinyan, L., Wong, L.: identifying good diagnosis gene group from gene expression profile using the concept of emerging patterns. Bioinformatics **18**, 725–734 (2002)
49. Kaplan, E.H., Craft, D.L., Wein, L.M.: Emergency Response to a smallpox attack: The case for mass vaccination. PNAS **100**(7) (2003)
50. Kianmehr, K. and Alhajj, R.: Calling Communities Analysis and Identification Using Machine Learning Techniques. Expert. Syst. Appl. **36**(3), 6218–6226 (2009)
51. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. J. ACM **46**(5), 604–632 (1999)

52. Klerks, P.: The Network Paradigm Applied to Criminal Organisations: Theoretical nitpicking or a relevant doctrine for investigators? Recent developments in the Netherlands. CONNECTIONS **24**(3), 53–65 (2001)
53. Lawrence, S., Giles, C.L.: Accessibility of information on the web. Nature **400**, 107–109 (1999)
54. Li, L., Pedersen, L.G., Darden, T.A., Weinberg, C.R.: Class prediction and discovery based on gene expression data. Iostatistics Branch and Lab of Structural Biology, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina (2000)
55. Mayr, A.: Smallpox vaccination and bioterrorism with pox viruses. Comp. Immunol. Microbiol. Infect. Dis. **26**(5–6), 423–430 (2003)
56. Meltzer, M.L., et al.: Modeling potential responses to smallpox as a bioterrorist weapon. Emerg. Infect. Dis. **7**(6) (2001)
57. Memon, N., Larsen, H.L.: Structural Analysis and Mathematical Methods for Destabilizing Terrorist Networks. Proceedings of the International Conference on Advanced Data Mining Applications, Springer-Verlag Lecture Notes in Artificial Intelligence (LNAI 4093), pp. 1037–1048 (2006)
58. Menczer, F.: Evolution of document networks. Proc. Natl. Acad. Sci. USA **101**, 5261–5265 (2004)
59. Newman, M.E.J.: The structure of scientific collaboration networks. Proc. Natl. Acad. Sci. USA **98**, 404–409 (2001)
60. Newman, M.E.J.: A measure of betweenness centrality based on random Walks. Soc. Networks **27**, 39–54 (2005)
61. Orlando, S., Palmerini, P., Perego, R.: Enhancing the apriori algorithm for frequent set counting. In: Proceedings of ACM International Conference on Data Warehousing and Knowledge Discovery, pp. 71–82, London, UK (2001)
62. Orlando, S., Palmerini, P., Perego, R., Silvestri, F.: Adaptive and resource aware mining of frequent sets. In: Proceedings of IEEE International Conference on Data Mining, p. 338, Washington, DC (2002)
63. Park, J.S., Chen, M.S., Yu, P.S.: Using a hash-based method with transaction trimming for mining association rules. IEEE Trans. Knowl. Data Eng. **9**(5), 813–825 (1997)
64. Pennock, D.M., Flake, G.W., et al.: Winners don't take all: Characterizing the competition for links on the web. Proc. Natl. Acad. Sci. USA **99**(8), 5207–5211 (2002)
65. Powell, W.W., White, D.R., et al.: Network dynamics and field evolution: The growth of interorganizational collaboration in the life sciences. Am. J. Sociol. **110**(4), 1132–1205 (2005)
66. Ressom, H., Reynolds, R., Varghese, R.S.: Increasing the efficiency of fuzzy logic-based gene expression data analysis. Physiol. Genomics **13**, 107–117 (2003)
67. Sahab, Z.J., Semaan, S.M., Sang, Q.-X.A.: Methodology and Applications of Disease Biomarker Identification in Human Serum. Biomark Insights **2**, 21–43 (2007)
68. Stern, J.E.: Will Terrorists Turn to Poison? Orbis **37**(3), 393–410 (1993)
69. Shariat, S.F., et al.: Multiple biomarkers improve prediction of bladder cancer recurrence and mortality in patients undergoing cystectomy. Cancer **112**(2), 315–25 (2008)
70. Sorensen, K.D., Orntoft, T.F.: Discovery of Prostate Cancer Biomarkers by Microarray Gene Expression Profiling. Expert Rev Mol Diagn. **10**(1), 49–64 (2010)
73. Toure, A., Basu, M.: Application of neural network to gene expression data for cancer classification. In: Proceedings of IEEE International Joint Conference on Neural Networks, pp. 583–587 (2001)
72. Tseng, G.C., et al.: Investigating Multi-cancer Biomarkers and Their Cross-predictability in the Expression Profiles of Multiple Cancer Types. Biomarker Insights **4**, 57–79 (2009)
73. Tusher, V.G., Tibshirani, R., Chu, G.: Significant analysis of microarrays applied to the ionizing radiation response. PNAS **98**(9), 5116–5121 (2001)
74. Wagner, M., et al.: Computational protein biomarker prediction: a case study for prostate cancer. BMC Bioinformatics **5**(26) (2004). doi:10.11861471-2105-5-26
75. Woolf, P.J., Wang, Y.: A fuzzy logic approach to analyzing gene expression data. Physiol. Genomics **3**, 9–15 (2000)

76. Wang, J., Hellem, T., Jonassen, I., Myklebost, O., Hovig, E.: Tumor classification and marker gene prediction by feature selection and fuzzy c-means clustering using microarray data. BMC Bioinformatics **4**, 60–72 (2003)
77. Xu, J.J., Chen, H.: CrimeNet Explorer: A framework for criminal network knowledge discovery. ACM Trans. Inform. Syst. **23**(2), 201–226 (2005)
78. Yan, X., Han, J., Afshar, R.: CloSpan: Mining Closed Sequential Patterns in Large Datasets. Proc. of 2003 SIAM Int. Conf. Data Mining (SDM' 03) (2003)
79. Zhang, X., Ke, H.: ALL/AML cancer classification by gene expression data using SVM and CSVM approach. Genomics informatics **11**, 237–239 (2000)

Part III
Case Studies

From Terrorism Informatics to Dark Web Research

Hsinchun Chen

Abstract In this paper, we provide an overview of "Terrorism Informatics," a new discipline that aims to study the terrorism phenomena with a data-driven, quantitative, and computational approach. We first summarize several critical books that lay the foundation for studying terrorism in the new Internet era. We then review important terrorism research centers and resources that are of relevance to our Dark Web research. The University of Arizona Artificial Intelligence Lab's Dark Web project is a long-term scientific research program that aims to study and understand the international terrorism (Jihadist) phenomena via a computational, data-centric approach. We aim to collect "ALL" web content generated by international terrorist groups, including web sites, forums, chat rooms, blogs, social networking sites, videos, virtual worlds, etc. We have developed various multilingual data mining, text mining, and web mining techniques to perform link analysis, content analysis, web metrics (technical sophistication) analysis, sentiment analysis, authorship analysis, and video analysis in our research. We report our recent Dark Web Forum Portal research, which provides web enabled access to critical international jihadist web forums. The portal has several significant technical extensions from our previous work: increasing the scope of our data collection, adding an incremental spidering component for regular data updates; enhancing the searching and browsing functions; enhancing multilingual machine-translation for Arabic, French, German, and Russian; and advanced Social Network Analysis (SNA). A case study on identifying active participants is shown at the end.

H. Chen (✉)
Artificial Intelligence Lab, Management Information Systems Department,
The University of Arizona, Tucson, AZ, USA
e-mail: hchen@eller.arizona.edu

1   Introduction

Terrorism informatics is defined as the"application of advanced methodologies and information fusion and analysis techniques to acquire, integrate, process, analyze, and manage the diversity of terrorism-related information for national/international and homeland security-related applications" [1]. These techniques are derived from disciplines such as computer science, informatics, statistics, mathematics, linguistics, social sciences, and public policy. Because the study of terrorism involves copious amounts of information from multiple sources, data types, and languages, information fusion and analysis techniques such as data mining, text mining, web mining, data integration, language translation technologies, and image and video processing are playing key roles in the future prevention, detection, and remediation of terrorism. Although there has been substantial investment and research in the application of computer technology to terrorism research, much of the literature in this emerging area is fragmented and often narrowly focused within specific domains. There is a critical need to develop multi-disciplinary approach to answering important terrorism related research questions.

1.1   Terrorism and the Internet

Terrorism is the systematic use of terror especially as a means of coercion. At present, the international community has been unable to formulate a universally agreed, legally binding, criminal law definition of terrorism. Common definitions of terrorism refer only to those violent acts which are intended to create fear (terror), are perpetrated for an ideological goal, and deliberately target or disregard the safety of non-combatants (civilians). (http://en.wikipedia.org/wiki/Terrorism). Rooted in political science, terrorism study has attracted researchers from many social science disciplines, from international relations to communications, and from defense analysis to intelligence study. Bruce Hoffman of the Georgetown University School of Foreign Service and Brian Jenkins of the RAND Corporation are some of the prominent scholars in terrorism and counter-insurgency study.

More recently, several terrorism scholars have begun to look into modern terrorism with a more data and network centric perspective. Marc Sageman's two critically acclaimed books: "Understanding Terrors Networks" (2004) and "Leaderless Jihad" (2007) are good examples. As stated in the book cover:

"For decades, a new type of terrorism has been quietly gathering ranks in the world. American ability to remain oblivious to the new movements ended on September 11, 2001. The Islamic fanatics in the global Salafi jihad (the violent, revivalist social movement of which al Qaeda is a part) target the West, but their operations mercilessly slaughter thousands of people of all races and religions throughout the world. Marc Sageman challenges conventional wisdom about terrorism, observing that the key to mounting an effective defense against

future attacks is a thorough understanding of the networks that allow the new terrorist to proliferate." [2]

Based on intensive data collection and analysis of documents from international press and court hearings on 172 important jihadists, Sageman was able to look into the social bonds predated ideological commitment. Many important network-based observations about these groups, including: small-world network and cliques formation, network robustness, wide geographical distribution, fuzzy boundaries, the strength of weak bonds, etc. were identified with striking examples.

The Internet has also been found to affect the global jihad by making possible a new type of relationship between an individual and a virtual community [2]. In the 2007 "Leaderless Jihad" book, Sageman continued his rigorous and systematic analysis of detailed evidence-based terrorist (500+ members) database. He described that "the process of radicalization in a hostile habitat but linked through the Internet leads to a disconnected global network, the Leader Jihad." [3]. He urged terrorism study to go from anecdote to data, and from journalism to social sciences. Among his findings, he found that before 2004 face-to-face interactions were more common among the average 26-year jihadi members; while after 2004, most interactions were on the Internet, and for members of a much younger average age, 20 years old. In a matter of 3–4 years and consistent with the modern information communication and technology (ICT) evolution, the Internet and the social media are helping the radical Islamists create a global, virtual social movement.

In addition to Sageman's influential work, several scholars have also examined the impact of Internet on the proliferation and radicalization of the global jihadi movement. In his seminal book in 2006, "Terror on the Internet," Gabriel Weimann [4] of the Haifa University Department of Communications in Israel reported his 8-year study of the use of Internet by terrorist organizations and their supporters. Sophisticated web sites have been found to help these organizations raise funds, recruit members, plan and launch attacks, and publicize their chilling results. Weimann describes the Internet as the new media for promoting new terrorism to a new generation of audience and for winning the war over minds. In Johnny Ryan's 2007 book, "Countering Militant Islamist Radicalization on the Internet," [5] he presented the EU's perspective on such developments, especially for the Europe-based home-grown radical groups. Due to the ubiquity and scale of the Internet, he suggested pooling technical and linguistic resources to monitor extremism on the Internet for the EU member states. He also suggested disseminating moderate opinions of credible Muslim scholars and web opinion leaders and encouraged user-driven content in social media to counter radicalization and violence.

The Dark Web research program of the University of Arizona Artificial Intelligence Lab is a complementary effort in the emerging discipline of terrorism informatics [6]. Unlike the social sciences approach adopted by the abovementioned scholars, the Dark Web project adopts a data, system and computational approach to studying terror and terrorism on the Internet. By collecting and analyzing a large-scale, longitudinal and fluid collection of terrorist-generated content using computer programs, we offer our complementary perspective and approach in understanding the overwhelmingly complex international terrorism landscape. Hsinchun Chen's

2006 book on "Intelligence and Security Informatics" (2006) reports selected examples from the Dark Web project at its early stage. This book serves to report significant findings and observations from the recent Dark Web developments.

The edited volume of "Terrorism Informatics" (2008) by Chen, Reid, Sinai, Silke, and Ganor became the first manuscript dedicated to the terrorism informatics topic. The book is highly inter-disciplinary, with editors and contributors from both social sciences and computational science. The goal of the book is to present terrorism informatics along two highly intertwined dimensions: methodological issues in terrorism research, including information diffusion techniques to support terrorism prevention, detection, and response; and legal, social, privacy, and data confidentiality challenges and approaches.

1.2 Terrorism Research Centers and Resources

Terrorism informatics relies heavily on terrorism domain knowledge and databases. We summarize some critical terrorism research centers and resources below based on our Dark Web experience. They are grouped into three major categories: (1) think tanks and intelligence resources; (2) terrorism databases and online resources; and (3) higher education research institutes. Clearly, the summary is not intended to be exhaustive. We only hope to help the (terrorism) community outsiders (like us or other computer scientists) get a glimpse of possible terrorism-related resources for them to get familiar with this complex area. For each research unit, we also provide a web link for readers to find additional detailed information.

1.2.1 Think Tanks and Intelligence Resources

The RAND Corporation is a non-profit research and development outfit in USA. It has evolved from a think tank during World War II to an independent corporation. Two prominent terrorism scholars, Brian Jenkins and Bruce Hoffman, are affiliated with the RAND Corporation. The company has been influential with its many excellent timely and thorough reports of international relation, political violence, and terrorism studies (Fig. 1). For example, see: http://www.rand.org/pubs/online/terrorism-and-homeland-security.html

As part of the West Point military academy, the Combating Terrorism Center (CTC) provides counter-terrorism strategic analyses independent from academy curricula, Pentagon tactics, or U.S. government politics since 2003. Some of its notable topical reports include: Harmony project (making sense of DoD Al Qaeda document database) and Islamic imagery project (Fig. 2). For more detail, see: http://ctc.usma.edu/harmony/harmony_docs.asp

The International Institute for Counter-Terrorism (ICT), located in Herzliya, Israel provides coverage of Middle Eastern events from an Israeli perspective. The institute regularly produces reports, commentaries, and multimedia contents. It's

Fig. 1 Sample RAND report

Fig. 2 The CTC harmony reports

highly successful annual conference typically draws 400–800 participants from all over the world (Fig. 3). For more information, see: http://www.ict.org.il/

1.2.2 Terrorism Databases and Online Resources

Internet Haganah (Hebrew word for "defense") is heavily involved in monitoring and disabling terror web sites on the Web. It alerts counter-terrorism vigilantes and indirectly fosters information "warfare." Its Open Source Intelligence (OSINT) gathers and catalogues available intelligence documents of relevance to terrorism (Fig. 4). For more information, see: http://internet-haganah.com/haganah/

News » Multimedia » Hamas... Torturing Palestinians

Fig. 3 Sample ICT multimedia content

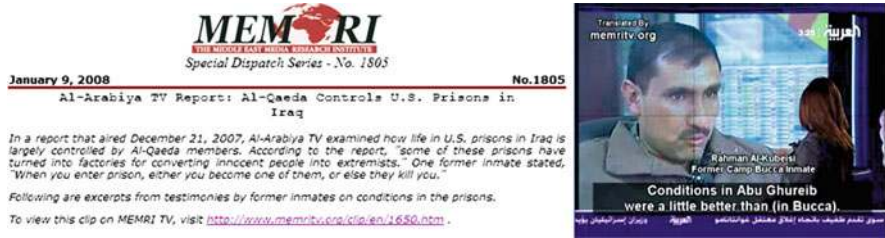**Fig. 4** Sample internet Haganah information on the web

Fig. 5 Sample MEMRI web content

Middle East Media Research Institute (MEMRI) is non-profit D.C.-based organization, with branches in Europe, Japan, and Israel. It regularly translates Arabic, Persian, and Turkish Media and annotates videos, news articles, and Web sites in the region. It also provides Islamic reformers a platform by translating their ideas and thoughts (Fig. 5). For more information, see: http://www.memri.org/

The Memorial Institute for the Prevention of Terrorism (MPIT) is a non-profit organization funded by the US Department of Homeland Security (DHS). Based in Oklahoma City, Oklahoma, the 1995 federal building bombing spurred interest in terrorism information repository, which resulted in creation of the MIPT Terrorism Knowledge Base (TKB). TKB contains two separate terrorist incident databases, the RAND Terrorism Chronology 1968–1997 and the RAND-MIPT Terrorism Incident database (1998–Present) (Fig. 6). The TKB ceased operations on March 31, 2008. For more information, see: http://www.mipt.org/

Also funded by the DHS, the University of Maryland's National Consortium for the Study of Terrorism and Responses to Terrorism (START) has extensive research about terrorist group formation and recruitment, persistence and dynamics, and societal responses to terrorist threats and attacks (Fig. 7).

It has provides a searchable open-source Global Terrorism Database (GTD), presenting information on terrorist events around the world since 1970 (currently updated through 2007), including data on where, when, and how each of over 80,000 terrorist events occurred. For more information, see: http://www.start.umd.edu/start/

Originally founded to document the crimes of the holocaust and finding the criminals responsible for genocide, the Simon Wiesenthal Center has targeted tolerance education mission through the Snider Social Action Institute and its Museum of Tolerance. It also produces the Digital Terrorism DVDs with Web extremist sites snapshots as part of larger educational mission (Fig. 8). For more information, see: http://www.wiesenthal.com/

Search for International Terrorist Entities (SITE) was founded in 2002 by undercover activist Rita Katz. The Site Intelligence Group, a for-profit organization, is now monitoring terrorist activities. It keeps translations of terrorist media and documents and makes them available as subscription to media, government, and corporations (Fig. 9). For more information, see: http://www.siteintelgroup.org/

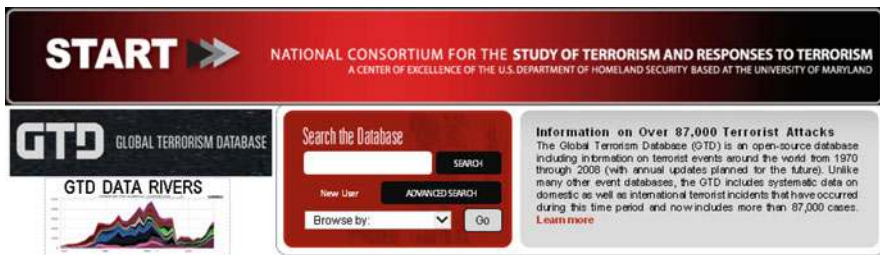**Fig. 6** MIPT and its terrorism knowledge base

Fig. 7 START and its global terrorism database

1.2.3 Higher Education Research Institutes

The University of St. Andrews Centre for the Study of Terrorism and Political Violence (CSTPV) provides a political science perspective to terrorism. The program offers subscription-based access to political analyses and awarding distance learning terrorism study certificates. Many influential political science grounded terrorism scholars were trained at St. Andrews (Fig. 10). For more information, see: http://www.st-andrews.ac.uk/~wwwir/research/cstpv/

The International Centre for Political Violence and Terrorism Research (ICPVTR) is a research and education centre within the S. Rajaratnam School

Product Detail

Name:	Digital Terrorism & Hate 2007
Product Code:	AL28
Description:	Item Number: AL28
Product Detail:	Digital Terrorism and Hate 2007 is the Simon Wiesenthal Center's newly released ninth annual, interactive report exposing terrorism and hate on the Internet. Compiled by Center researchers, the 2007 edition is culled from close to 7,000 problematic websites, blogs, newsgroups, youtubeTM and other on-demand video sites. Key sections of Digital Terrorism are available in English, French and Spanish.

Fig. 8 Simon Wiesenthal center and its digital terrorism DVDs

First Issue of "Echo of the Epics" – A Magazine from Al-Qaeda in Yemen
By SITE Intelligence Group

January 14, 2008

A translation of the index, editorial, and two articles are provided to our Monitoring Service subscribers.

Fig. 9 Sample SITE report
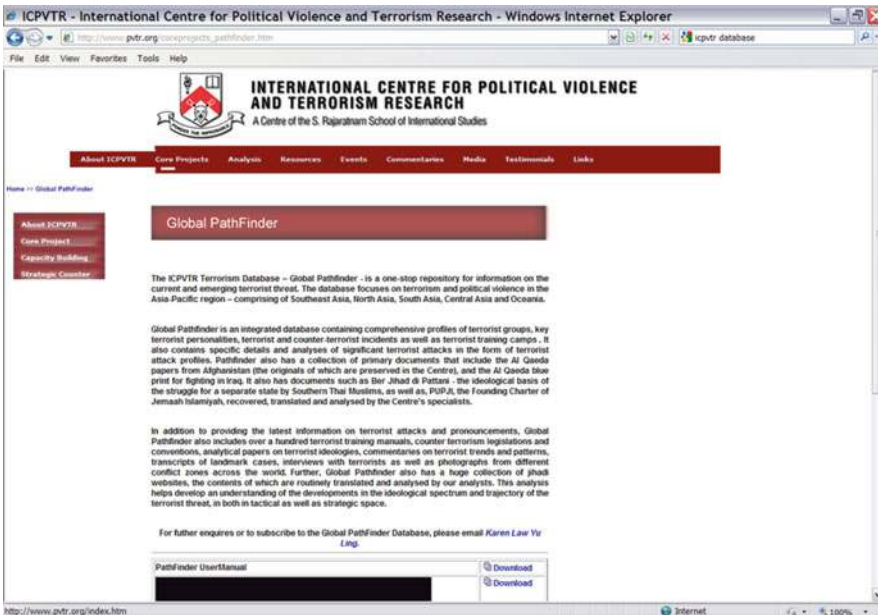
Fig. 10 Sample CSTPV reports

Fig. 11 ICPVTR terrorism database

of International Studies (RSIS) at Nanyang Technological University, Singapore. ICPVTR conducts research, training, and outreach programs aimed at reducing the threat of politically motivated violence and at mitigating its effects on the international system. Its Global Pathfinder system is a one-stop repository for information on the current and emerging terrorist threat. The database focuses on terrorism and political violence in the Asia-Pacific region – comprising of Southeast Asia, North Asia, South Asia, Central Asia and Oceania (Fig. 11). For more information, see: http://www.pvtr.org/

Fig. 12 The Dartmouth ISTS web site

Sponsored by many US government agencies, the Dartmouth Institute for Security Technology Studies (ISTS) focuses on cyber security, trust, and cyber terrorism. Its project topics include: hardening IT infrastructure against attack, image and video forensics, trusted digital certificate provision, information infrastructure risk assessment, etc. (Fig. 12). For more information, see: http://www.ists.dartmouth.edu/

As the field of terrorism informatics continues to grow and evolve, we anticipate broader collaboration between social scientists and computational researchers who are interested in counter-terrorism research. We also expect new methodologies, terrorism databases, and computational approaches to emerge and mature based on the rich content and complex interactions produced by the terrorists and extremists on the Internet.

2 Dark Web Research Overview

Gabriel Weimann of the Haifa University in Israel estimated that there are about 5,000 terrorist web sites as of 2006 [7]. Based on our actual spidering experience over the past 8 years, we believe there are about 100,000 sites of extremist and terrorist content as of 2010, including: web sites, forums, blogs, social networking sites, video sites, and virtual world sites (e.g., Second Life). The largest increase since 2006–2007 is in various new Web 2.0 sites (forums, videos, blogs, virtual

world, etc.) in different languages (i.e., for home-grown groups, particularly in Europe). We have found significant terrorism content in more than 15 languages.

We collect (using computer programs) various web contents every 2–3 months; we started spidering in 2002. Currently we only collect the complete contents of about 1,000 sites, in Arabic, Spanish, and English languages. We also have partial contents of about another 10,000 sites. In total, our collection is about 15 TBs in size, with close to 2,000,000,000 pages/files/postings from more than 10,000 sites. We believe our Dark Web collection is the largest open-source extremist and terrorist collection in the academic world. Researchers can have graded access to our collection by contacting our research center. We present below a summary of important Dark Web contents.

2.1 Web Sites

Our web site collection consists of the complete contents of about 1,000 sites, in various static (html, pdf, Word) and dynamic (PHP, JSP, CGI) formats. We collect every single page, link, and attachment within these sites. We also collect partial information from about 10,000 related (linked) sites. Some large well-known sites contain more than 10,000 pages/files in 10+ languages (in selected pages).

2.2 Forums

We collect the complete contents (authors, headings, postings, threads, time-tags, etc.) of about 300 terrorist forums. We also perform periodic updates. Some large radical sites include more than 30,000 members with close to 1,000,000 messages posted. We have also developed the Dark Web Forum Portal, which provides beta search access to several international jihadist "Dark Web" forums collected by the Artificial Intelligence Lab at the University of Arizona. Users may search, view, translate, and download messages (by forum member name, thread title, topic, keyword, etc.). Preliminary SNA visualization is also available.

2.2.1 Blogs, Social Networking Sites, and Virtual Worlds

We have identified and extracted many smaller, transient (meaning, the sites appear and disappear very quickly) blogs and social networking sites, mostly hosted by terrorist sympathizers and "wannabes." We have also identified more than 30 (self-proclaimed) terrorist or extremist groups in virtual world sites. (However, we are still unsure whether they are "real" terrorist/extremists or just playing the roles in virtual games.)

2.2.2 Videos and Multimedia Content

Terrorist sites are extremely rich in content, with heavy usage of multimedia formats. We have identified and extracted about 1,000,000 images and 100,000 videos from many terrorist sites and specialty multimedia file-hosting third-party servers. More than 50% of our videos are Improvised Explosive Devices (IED) related.

Our computational tools are grouped into two categories: (1) Collection and (2) Analysis and Visualization. Significant deep web spidering, computational linguistic analysis, sentiment analysis, SNA, and social media analysis and visualization research has been conducted by members of the AI Lab over the past 8 years.

2.3 Dark Web Collection

2.3.1 Web Site Spidering

We have developed various focused spiders/crawlers for deep web based on our previous digital library research. Our spiders can access password-protected sites and perform randomized (human-like) fetching. Our spiders are trained to fetch all html, pdf, and word files, links, PHP, CGI, and ASP files, images, audios, and videos in a web site. To ensure freshness, we spider selected web sites every 2–3 months.

2.3.2 Forum Spidering

Our forum spidering tool recognizes 15+ forum hosting software and their formats. We collect the complete forum including: authors, headings, postings, threads, time-tags, etc., which allow us to re-construct participant interactions. We perform periodic forum spidering and incremental updates based on research needs. We have collected and processed forum contents in Arabic, English, Spanish, French, and Chinese using selected computational linguistics techniques.

2.3.3 Multimedia (Image, Audio, and Video) Spidering

We have developed specialized techniques for spidering and collecting multimedia files and attachments from web sites and forums. We plan to perform stenography research to identify encrypted images in our collection and multimedia analysis (video segmentation, image recognition, voice/speech recognition) to identify unique terrorist-generated video contents and styles.

2.4   Dark Web Analysis and Visualization

2.4.1   Social Network Analysis

We have developed various SNA techniques to examine web site and forum posting relationships. We have used various topological metrics (betweeness, degree, etc.) and properties (preferential attachment, growth, etc.) to model terrorist and terrorist site interactions. We have developed several clustering (e.g., Blockmodeling) and projection (e.g., Multi-Dimensional Scaling, Spring Embedder) techniques to visualize their relationships. Our focus is on understanding "Dark Networks" (unlike traditional "bright" scholarship, email, or computer networks) and their unique properties (e.g., hiding, justice intervention, rival competition, etc.).

2.4.2   Content Analysis

We have developed several detailed (terrorism-specific) coding schemes to analyze the contents of terrorist and extremist web sites. Content categories include: recruiting, training, sharing ideology, communication, propaganda, etc. We have also developed computer programs to help automatically identify selected content categories (e.g., web master information, forum availability, etc.).

2.4.3   Web Metric Analysis

Web metrics analysis examines the technical sophistication, media richness, and web interactivity of extremist and terrorist web sites. We examine technical features and capabilities (e.g., their ability to use forms, tables, CGI programs, multimedia files, etc.) of such sites to determine their level of "web-savvy-ness." Web metrics provides a measure for terrorists/extremists' capability and resources. All terrorist site web metrics are extracted and computed using computer programs.

2.4.4   Sentiment and Affect Analysis

Not all sites are equally radical or violent. Sentiment (polarity: positive/negative) and affect (emotion: violence, racism, anger, etc.) analysis allows us to identify radical and violent sites that warrant further study. We also examine how radical ideas become "infectious" based on their contents, and senders and their interactions. We reply much on recent advances in Opinion Mining – analyzing opinions in short web-based texts. We have also developed selected visualization techniques to examine sentiment/affect changes in time and among people. Our research includes several probabilistic multilingual affect lexicons and selected dimension reduction and projection (e.g., Principal Component Analysis) techniques.

2.4.5 Authorship Analysis and Writeprint

Grounded in authorship analysis research, we have developed the (cyber) Writeprint technique to uniquely identify anonymous senders based on the signatures associated with their forum messages. We expand the lexical and syntactic features of traditional authorship analysis to include system (e.g., font size, color, web links) and semantic (e.g., violence. racism) features of relevance to online texts of extremists and terrorists. We have also developed advanced Inkblob and Writeprint visualizations to help visually identify web signatures. Our Writeprint technique has been developed for Arabic, English, and Chinese languages. The Arabic Writeprint consists of more than 400 features, all automatically extracted from online messages using computer programs. Writeprint can achieve an accuracy level of 95%.

2.4.6 Video Analysis

Based on previous terrorism ontology research, we have developed a unique coding scheme to analyze terrorist-generated videos based on the contents, production characteristics, and meta data associated with the videos. We have also developed a semi-automated tool to allow human analysts to quickly and accurately analyze and code these videos.

2.4.7 IEDs in the Dark Web Analysis

We have conducted several systematic studies to identify IED related content generated by terrorist and insurgency groups in the Dark Web. A smaller number of sites are responsible for distributing a large percentage of IED related web pages, forum postings, training materials, explosive videos, etc. We have developed unique signatures for those IED sites based on their contents, linkages, and multimedia file characteristics. Much of the content needs to be analyzed by military analysts. Training materials also need to be developed for troops before their deployment ("seeing the battlefield from your enemies' eyes").

3 Dark Web Forum Portal

In recent years, there have been numerous studies from a variety of perspectives analyzing the Internet presence of hate and extremist groups. The use of the Internet by such groups has provoked interest in terrorism researchers in various social sciences including psychology, sociology, criminology, and political science; computational scientists studying web mining and information extraction; and security analysts and others concerned with homeland and national policies and security.

Yet the websites and forums of extremist and terrorist groups have long remained an underutilized resource due to their ephemeral nature and persistent access and analysis problems. They emerge quickly, often just as quickly disappearing or, in many cases, seeming to disappear by changing their uniform resource locators (URLs) but retaining much of the same content [7]. Furthermore, some are hacked or closed by the ISPs. Thus, many researchers, students, analysts, and others face difficulties in identifying, collecting, and analyzing this content. Since terrorist and extremist groups are increasingly using the Internet to promulgate their agendas, it has become imperative that persistent access as well as user-friendly searching be provided to this content. Given the sheer volume of sites, their dynamic and fugitive nature, different languages, and noise, it has become clear that systematic and automated procedures for identifying, collecting, and searching these sites must be provided.

3.1 System Design

As shown in Fig. 13, the Dark Web Forum Portal system contains three components: Data Acquisition, Data Preparation, and System Functionality. The overall system design is similar to our previous paper [8]. But we have added an incremental spidering component to regularly update the collection. Each component is detailed in the following sections.

3.1.1 Data Acquisition

In this component, spidering programs are developed to collect the Web pages from online forums that contain Jihadist related content identified by domain experts. The spidering component is composed of complete spidering and incremental spidering (Fig. 14). Complete Spidering is applied to forums the first time they are added to our collection, while incremental spidering is adopted if the forums already exist in the collection. When a forum is first added to our collection, the complete spidering is applied to collect all available postings. Incremental spiders are designed to identify and collect postings posted after the last updating time of the forum, so that only a small portion of forum data is collected and therefore makes the spidering process much more efficient. To achieve this goal, an incremental spider is developed for each forum in the collection.

The incremental spidering consists of three main steps: Sub-Forum List Page Spidering, Thread List Page Spidering and Incremental Spidering. *Sub-Forum List Page Spidering:* Forums generally contain one or more sub-forums representing different discussion themes. In this step, incremental spiders first spider and parse sub-forum list pages of a forum and identify URLs of sub-forums. T*hread List Page*
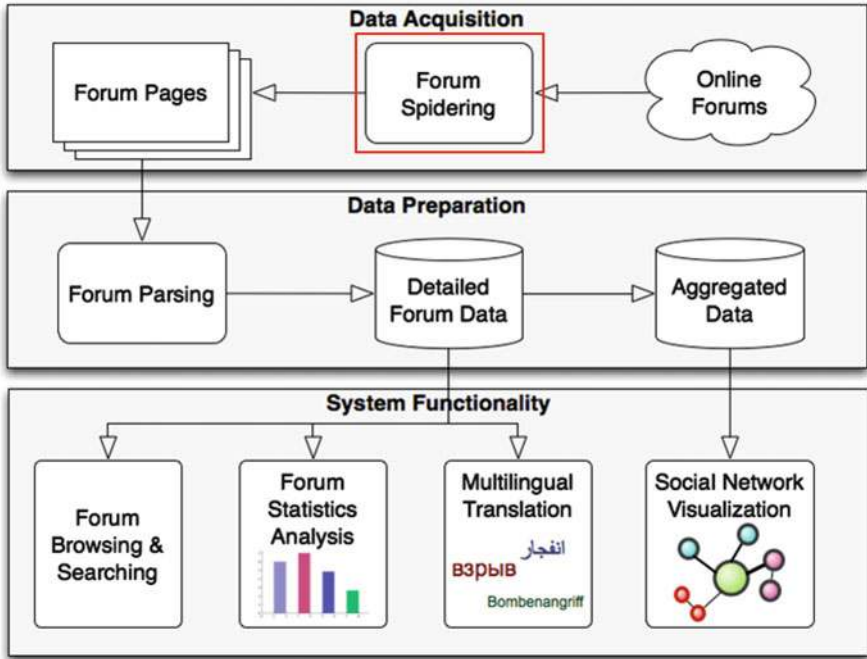
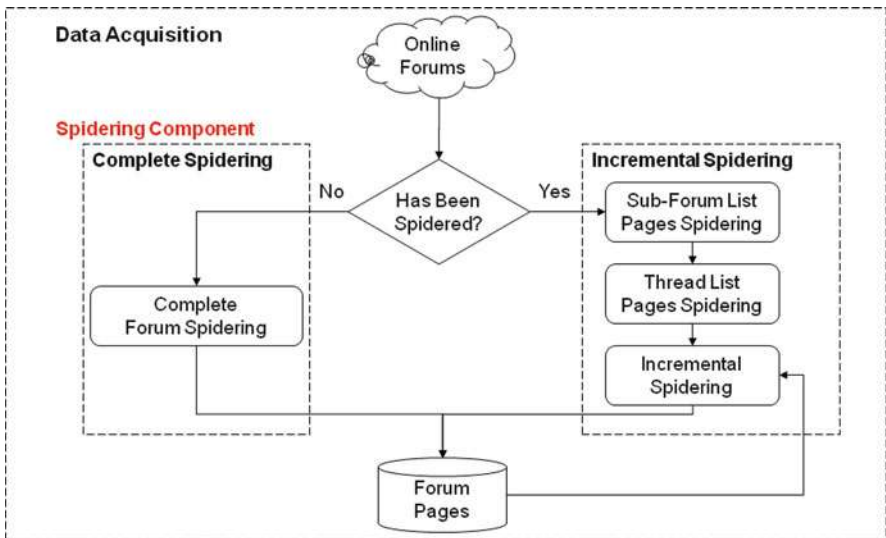**Fig. 13** System design of the Dark Web Forum Portal

Fig. 14 Framework of the spidering component

Spidering: Thread list pages contain the metadata of discussion threads (such as title, date of the last update, and author name) which are sorted by dates of the last update decreasingly. For each sub-forum, the incremental spider starts from downloading the first thread list page of the sub-forum; and dates of the last update of discussion thread are then extracted. Threads updated later than the date of the latest posting in the database are considered to be new threads and their URLs are collected. If every thread listed in the first thread list page is a new thread, the spidering will move to the next thread page. Otherwise, the spidering of this sub-forum is complete. *Incremental Spidering:* After collecting all the URLs of new threads, the incremental spider begins to download all of the postings within the new threads.

We conducted an experiment of incremental spidering using Hanin Net forum (http://www.hanein.info/vb/), which is the most active forum identified by our domain experts. Because this forum is the most active one in our forum list, the experiment result can be considered as the upper bound of the time required for the incremental spidering. In the experiment, we collected postings from 11/1/2009 through 12/10/2009 (about 6 weeks). The experiment results shows that during the time span, 3,504 threads and 29,016 postings were collected. The entire incremental spidering process was completed in 39 min. The result indicates that incremental spidering is a promising solution for keeping our forum data up-to-date.

3.1.2 Data Preparation

In this component, forum parsing programs are developed to extract the detailed forum data from the raw HTML Web pages and store it in a local database. For each forum, the structured, detailed forum data extracted include thread names, main message bodies, member names, and post dates.

3.1.3 System Functionality

Different functions are developed and incorporated into the system as real-time services, including single and multiple forums, browsing and searching, forum statistics analysis, multilingual translation, and social network visualization. The Dark Web Forum portal is implemented using Apache Tomcat and the database is implemented using Microsoft SQL Server 2008. For forum statistics analysis, Java applet-based charts are created to show the trends based on the numbers of messages produced over time. The multilingual translation function is implemented using Google Translation Service, which can automatically detect non-English texts and translate them into English. The social network visualization function provides dynamic, user-interactive networks implemented using JUNG (http://jung.sourceforge.net/) to visualize the interactions among forum members.

3.2   Data Set: Dark Web Forums

Table 1 lists the forums incorporated into our system. Currently, the portal contains 29 Jihadist forums, among which 17 are Arabic forums, 7 are English forums, 3 are French forums and the other 2 are in German and Russian, respectively. The forums have been carefully selected with significant input from terrorism researchers, security and military educators, and other experts. The Arabic-language forums selected include major jihadist websites, some of which include English language sections. The English-language forums represent both extremist and more moderate groups. The French, German, and Russian forums provide representative content for extremist groups communicating in these languages, and provide additional opportunity to evaluate multilingual translation. The total number of messages is about 13M; approximately 3M postings will be added annually through incremental spidering.

3.3   System Functionality

As we described before, the system has four types of functions: single and multiple forum browsing and searching, forum statistics analysis, multilingual translation, and social network visualization. In this section, we describe our enhanced browsing and searching and social network visualization. For the other two functions, see our previous paper [8].

3.3.1   Single Forum Browsing and Searching

The search function allows users to search message titles or bodies using multiple keywords. Users can choose the Boolean operations of the keywords to be either "AND" or "OR." Users can also express their search terms in English even when the forum is, for example, mainly Arabic. In that case, the search will return matches for both the English terms and the Arabic translations of those terms.

3.3.2   Multiple Forums Browsing and Searching

In addition to browsing and searching information in a particular forum, our portal also supports multiple forumsearching across all forums in the portal. For example, a total of 227 threads (Fig. 15) are retrieved across all forums that contain keywords "bomb," "Iraq," and "kill" (AND operation) in the thread titles or message bodies. Among them, 159 are from the forum "Gawaher," 56 are from forum "Ansar1," 5 are from forum "Ummah," etc., "Gawaher" has more discussions on this topicthan any of the other forums. Detailed searching results for each forum on these keywords

Table 1 Statistics of the dark web forums (Al-Firdaws and Islamic network are now inactive)

Name	Language	Time span	Number of members	Number of threads	Number of messages
Al-Boraq	Arabic	01/08/2006–01/02/2010	3,503	52,322	223,648
Al-Fallujah	Arabic	09/19/2006–01/02/2010	5,853	74,899	547,712
Al-Firdaws*	Arabic	01/02/2005–02/06/2007	2,187	9,359	39,715
Midad al-Suyuf	Arabic	03/18/2006–1/02/2010	1,597	11,232	38,382
Alokab	Arabic	04/08/2005–12/31/2009	1,547	8,096	55,947
Al-Qimmah	Arabic	11/23/2007–01/02/2010	287	12,097	23,709
Alsayra	Arabic	04/05/2001–12/31/2009	66,705	147,598	1,227,207
Ansar	Arabic	11/07/2008–01/02/2010	1,224	12,041	46,928
At-tahadi	Arabic	04/14/2008–01/02/2010	313	2,599	5,406
Hanin Net	Arabic	11/27/2006–01/12/2010	2,837	96,239	821,478
Hawaa World	Arabic	01/01/2001–01/02/2010	113,579	40,501	2,251,553
Hadramout	Arabic	11/25/2000–12/29/2009	29,491	151,694	1,552,227
Ma'arik	Arabic	07/29/2007–01/03/2010	1,880	15,288	57,047
Al-Mujahidin	Arabic	11/09/2007–01/02/2010	4,259	29,980	140,930
Montada	Arabic	09/25/2000–12/29/2009	40,291	120,181	1,412,028
Ana al-Muslim	Arabic	10/08/1985–11/26/2009	12,215	179,791	1,343,370
Shumukh	Arabic	03/21/2007–01/02/2010	3,938	46,666	289,201
Ansar	English	12/08/2008–01/02/2010	377	11,133	29,056
Gawaher	English	10/24/2004–01/01/2010	6,790	210,656	569,709
Islamic Awakening	English	04/28/2004–12/31/2009	2,361	25,112	116,009
Islamic Network*	English	06/09/2004–05/07/2008	1,573	11,974	87,314
Islamic Web-Community	English	11/14/2000–12/31/2009	745	6,262	24,850
Turn To Islam	English	06/02/2006–01/01/2010	9,926	38,702	308,970
Ummah	English	04/01/2002–12/31/2009	14,349	71,218	1,192,583
Al Minha Dj	French	06/01/2008–01/04/2010	313	2,007	6,421
Forums d'aslama	French	10/06/2004–01/03/2010	2,665	20,468	131,559
Al-Mourabitoune	French	05/05/2002–03/27/2009	3,198	7,905	72,140
Ansar	German	02/27/2009–01/02/2010	62	726	1,645
KavkazChat	Russian	03/21/2003–01/03/2010	5,634	6,144	558,042
Total			339,699	1,422,890	13,174,786

All Forums threads related to Topic: bomb, iraq, kill

This page shows all threads found which contain the search term.

Forum Name	Number of threads have been found
Forums in Arabic:	
Alboraq	0
AlFaloja	0
AlFirdaws	0
Almedad	0
Alokab	0
Alqimmah	0
Alsayra	0
AsAnsar	0
Atahadi	0
Hanein	0
Hawaa	0
Hdrmut	1
M3f	0
Majahden	0
Montada	0
Muslm	0
Shamikh	1
Forums in English:	
Ansar1	56
Gawaher	159
IslamicAwakening	3
IslamicNetwork	2
Myiwc	0
Ummah	5
TurnToIslam	0
Forums in French:	
Alminhadj	0
Aslama	0
Ribaat	0
Forums in German:	
DeAnsarnet	0
Forums in Russian:	
KavkazChat	0
IN ALL FORUMS	**227**

Fig. 15 Screenshot of the cross-forum search based on keywords "bomb," "iraq," and "kill" with AND operator

can be found by clicking the row corresponding to a particular forum. Figure 16 shows the screenshot of the detailed result for forum "Gawaher" based on the cross forum search in Fig. 15.

3.3.3 Social Network Visualization

The interface of the SNA function is shown in Fig. 17. It consists of three parts: the *search panel* (top box), *analysis panel* (middle box), and *visualization panel* (bottom box).

The search panel allows the user to choose three search criteria: forum, keyword and time period. The threads that meet these search criteria are identified as "related threads" and are used to construct the social network. Any of the forums listed in the portal can be selected to perform SNA. The keywords are selected by the user, in any language, separated with space of comma. Thread names, user names, and postings

Dark Web Forum Portal

Fig. 16 Screenshot of the detailed result for forum "Gawaher" based on the cross forum search shown in Fig. 15

Fig. 17 SNA function interface

are searched using these keywords, and a thread is identified as a related thread if the thread name, or at least one posting, or at least one user name, contain any of the keywords. The start date and the end date are used to constrain the postings in the search result. When related threads are returned, the social network will be constructed based on the structure of these threads.

The analysis panel allows the user to select different metrics for SNA, and to set the parameters for graph visualization. Every node in the social network has a set of attributes, including the screen name, the number of postings, and various social network metrics. After the social network is constructed, all nodes are ranked in descending order based on the number of postings. Since the resulting social network usually contains a large number of message authors, which makes the graph too crowded for analysis, the slide bar can be used to display only a portion of the top authors based on the ranking in order to make the graph easier to read. The label as well as the value of the selected metrics can be displayed beside each node by checking the corresponding box. An isolated node is defined as a node that has no connections to any other node. Removing isolated nodes is a useful function when too many cause noise in the graph. Checking the "Link to User Post" box will change the color of every node from green to red, and a click on any node will pop up a new window that shows all postings by this user during the selected time period.

The visualization panel displays the graph based on the settings in the analysis panel, with the thickness of the link proportionate to the intensity of interactions between two nodes. Any node can be dragged to any position in the panel, and all connected nodes and corresponding links will be highlighted when holding the mouse button pressed during the move of the node. Different layouts are also provided for graph visualization. Four types of layout algorithms are integrated into the component, including static layout, circle layout, 3 force-based layouts (Fruchterman–Reingold, Kamada–Kawai, and Spring), and a self-organizing layout (ISOM). If users want to perform advanced analysis on the graph using other SNA tools such as UCINET, Pajek and so on, clicking the "Export Graph" button allows the graph to be exported to a ".net" format file, which is the Pajek graph file format recognized by most SNA tools.

3.4   Case Study: Islamic Awakening Forum Search and SNA

In this case study, we demonstrated a scenario on how to use the SNA component to identify active users on a particular topic of interest (in this case, "religious beliefs"). We chose Islamic Awakening for this case study. We searched "Muslim, Islam, Sharia, Sunni and Shia" as keywords and generated the topic-based social network. The time period selected was 01/01/2009–12/31/2009 (1 year). As shown in Fig. 18, "Bint ul Islam," "Iloveislam," and "Abuhannah" were the three most active participants on these topics based on the "Number of Messages." We can show
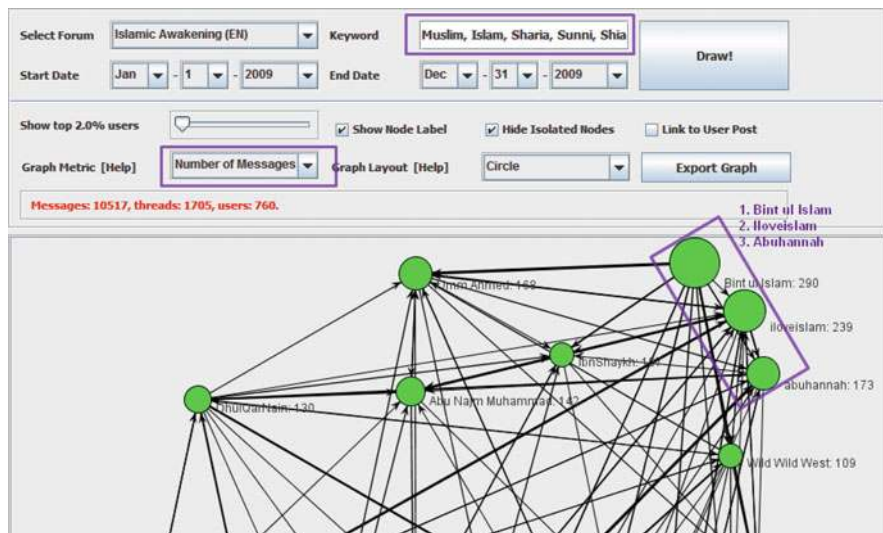
Fig. 18 Active participants in Islamic Awakening

the SNA based on other different graph metrics such as "Betweeness," "PageRank," "In-degree," and "Out-degree."

By checking the "Link to User Post" box, the user can see the detailed messages. Example messages on this religious topic include the following: "Sister, since you are a new Muslim, Allah will test you as He Says in the Qur'an: "Do men think that they will be left alone on saying, "We believe", and that they will not be tested? We did test those before them, and Allah will certainly know those who are true from those who are false" [Al-Qur'an 29:2–3] You have to remain strong. . . . " by "iloveislam."

"Between The Past And The Future Imam Ibn ul Qayyim al Jawziyyah al-Fawaa'id, pp 151–152 Al-Istiqaamah, No. 2 Your life in the present moment is in between the past and the future. So what has preceded can be rectified by tawbah (repentance), nadam (regret) and istighfar (seeking Allaah's forgiveness). ..." by "Bint ul Islam."

4 Conclusions and Future Directions

In this paper, we present a review of "terrorism informatics" and our work on developing the Dark Web research program. The Dark Web Forum Portal is an infrastructure which integrates heterogeneous Jihadi forum data, and will serve as a strong complement to the current terrorism informatics databases, news reports and other sources available to the research community. Please contact the

University of Arizona Artificial Intelligence for access to selected data and for future collaboration.

References

1. Chen, H., Reid, E., Sinai, J., Silke, A., Ganor, B. (eds.): Terrorism Informatics: Knowledge Management and Data Mining for Homeland Security. Springer, Heidelberg (2008)
2. Sageman, M.: Understanding Terror Networks. University of Pennsylvania Press, Philadelphia, Pennsylvania (2004)
3. Sageman, M.: Leaderless Jihad. University of Pennsylvania Press, Philadelphia, Pennsylvania (2008)
4. Weimann, G.: www.terror.net: How modern terrorism uses the Internet. Special Report, United States Institute of Peace. Retrieved Oct 31, 2004. http://www.usip.org/pubs/specialreports/sr116.pdf
5. Ryan, J.: Countering Militant Islamist Radicalization on the Internet: A User Driven Strategy to Recover the Web. Institute of European Affairs (2007)
6. Chen, H.: Intelligence and Security Informatics for International Security: Information Sharing and Data Mining. Springer, New York (2006)
7. Weimann, G.: Terror on the Internet: The New Arena, the New Challenges. US Institute of Peace Press, Washington, DC (2006)
8. Zhang, Y., Zeng, S., Fan, L., Dang, Y., Larson, C., Chen, H.: Dark Web Forums Portal: Searching and Analyzing Jihadist Forums. In: Proceedings of the IEEE International Intelligence and Security Informatics Conference, Dallas, Texas, 8–11 June 2009

Investigating Terrorist Attacks Using CDR Data: A Case Study

Fatih Ozgul, Ahmet Celik, Claus Atzenbeck, and Nadir Gergin

Abstract Call Detail Records (CDR) are commonly used by police and intelligence services all over the world. In many countries, GSM operators are obligated to keep CDR data for all of their subscribers. For prosecutors, courts, and judges investigating criminal cases it is beneficial to use CDR data. This case study shows how mining CDR data helped in the investigation of a terrorist attack that happened in Istanbul, Turkey. A truck was put on fire by a terrorist organization to protest against the conditions of a terrorist leader in prison. Arsonists were identified and arrested after the interrogation of suspects. The judge asked GSM operators to provide the suspects' CDR data for verifying their testimonies. Five different attributes retrieved from the CDR were merged. Date, time, and location of suspects were compared with the log of signals received by base stations. In order to find out whether subjects are acquainted with each other, their phone calls on the day of the attack were matched using GSM line numbers. Furthermore, suspects' cell phone handsets were matched using International Mobile Equipment Identity (IMEI) numbers in order to find out whether same handsets or SIM cards have been used in the past. Friendship, spatiotemporal, GSM line number, and IMEI number analysis using CDR data revealed that some testimonies were wrong and thus helped in identifying the suspects who carried out the attack.

F. Ozgul (✉)
Faculty of Computing, Engineering and Technology, University of Sunderland, Sunderland, UK
e-mail: fatih.ozgul@istanbul.com

A. Celik
Diyarbakir A. Gaffar Okkan Vocational School, Turkish National Police, Diyarbakir, Turkey
e-mail: acelik@rutgers.edu

C. Atzenbeck
Institute of Information Systems, Hof University, Hof, Germany
e-mail: claus.atzenbeck@iisys.de

N. Gergin
Diyarbakir Police Department, Turkish National Police, Diyarbakir, Turkey
e-mail: nadirgergin@yahoo.com

1 Introduction

Terrorist attacks happen regularly in different parts of the world. They are generally solved in many ways, including the use of human oriented information (e.g., testimonies of suspects and witnesses, human intelligence during police interrogations), investigating physical objects (e.g., bomb producing material and its mechanism), use of forensic methods (e.g., fingerprint), signal based technical intelligence (e.g., Call Detail Records, phone tapping, and eaves dropping), or open source intelligence (e.g., personal web sites, web forums, data gathered from social networking sites).

Literature shows that it is current practice to use personal web sites, social networking sites, web forums, e-mails, police arrest data, or scenario based information for detecting terrorist groups via open source intelligence. TMODS and COPLINK are examples of systems that are specialized in criminal and terrorist network detection. TMODS [3, 4, 7, 12–14], developed by twenty-first century technologies, automates the tasks of searching and analyzing instances of particular threatening activity patterns. With TMODS, the analyst can define an attributed relational graph representing the pattern of threatening activity he or she is looking for. TMODS then automates the search for that threat pattern through an input graph representing the observed data. TMODS pinpoints the subset of data that matches the defined threat pattern and thus transforms a manual search into an efficient automated graph matching tool. The user defined threatening activity or pattern graph can be produced with terrorist network ontologies and matched against the observed activity graph. At the end, humans analyze the match that is highlighted against the input graph. TMODS requires a pattern graph and an analyst. Like a supervised learning algorithm, TMODS tries to tailor the results according to predefined threatening activities.

Similar to TMODS, the CrimeNet Explorer framework [1, 10, 18, 19] is used for automated network analysis and visualization using police arrest reports and texts. It is based on a concept space algorithm, COPLINK connect, and COPLINK detect structures and obtains link data from arrest text reports among other offender data. The algorithm used by the Tucson Police Department for collecting documents, co-occurrence analysis, and associative retrieval. CrimeNet Explorer uses a Reciprocal Nearest Neighbour (RNN) based clustering algorithm to find out links between offenders, as well as discovery of previously unknown groups.

Call Detail Records (CDR) are another source on which data mining techniques can be applied in order to retrieve valuable information. It has been proven valuable in helping to solve certain criminal cases. In fact, CDR of terrorists' mobile phones can be often considered as one of main sources of evidences.

For example, on June 2, 2008, seven suspected members of the "Scarecrow Bandits" got caught after two police chases and a standoff at a Costco store in Plano. They were called as Scarecrow Bandits due to their habit of wearing plaid shirts and floppy-brim straw hats during some of their suspected 20 bank robberies in North Texas [6]. Finally, the criminals were identified via a novel method of using CDR data, rather than traditional law enforcement methods. FBI agents obtained a huge

amount of data from mobile companies that hold CDR related to the time of the bank robberies. They revealed that two phones belonging to two of the suspects were used for calls at around the same times when 12 bank robberies happened [2].

This example shows the benefits of using transactional information of telephone calls (i.e., calling and called numbers, time, duration) for crime investigations. However, storing all call traffic in order to decide later which messages are worthy for further investigation, is usually not feasible. In the US, police agencies have previously been able to obtain call details of individual phones, collected via specialized devices (pen register or trap-and-trace devices) in response to court orders. In comparison, CDR databases are maintained by telephone companies due to several reasons (such as billing) and hold calling information about all subscribers over long time periods. They are rich sources of information about customer activities, revealing both the structure of organizations and the behavior of individuals. Several telephone companies accept the demand of CDR data by police and intelligence organizations in response to governmental pressure [5].

2   Using CDR Data and Crime Data Mining

Data mining is the search for valuable information in large volumes of data, involving humans and computers [17]. It helps us to extract patterns from data. There is a history of data mining used in business intelligence to cluster, classify, regress, and associate rule learning of information.

Data mining is in its infancy era in the field of crime prevention and counterterrorism. Naturally, there are concerns about data mining violating peoples' privacy or data security. In order to support privacy, organizations have developed regulations. For example, the Organization for Economic Co-operation and Development (OECD) established privacy guidelines in 1980: The purpose for data collection and its limitations should be specified, openness should be a general policy, reasonable security safeguards should be established, and individuals should have right to learn what kind of information is/was being collected [8]. Furthermore, a new method of data mining – called *privacy preserving data mining* – emerged, dealing with privacy issues of certain information.

CDR may contain valuable crime-related information that can be exploited with data mining techniques. They include all details pertaining to a call, such as time, duration, origin, and destination of the call. CDR data is collected at base stations. Not surprisingly, a billion calls are made every month, leading to terabytes of data stored [15]. As Table 1 indicates, mining CDR data and its implications, such as privacy considerations, vary from country to country.

Although it is very common among police and intelligence practitioners to use CDR data for solving crime cases, we could not find any example case in the current computer science literature. There are, however, examples of using CDR data for business intelligence. For example, retrieving information from CDR can provide major insights to GSM operators for designing effective strategies [16]. Beside that,

Table 1 Current legal requirements and conditions for using CDR data. (Partly obtained from [11])

	Obtaining phone record	Data mining the call records	Privacy international ranking
Australia	According to the Telecommunication Act, phone records are considered private unless an exception is applied. One agency could obtain records and share them with others	According to National Law and Ratified International Law, data mining is not permissible, even if there is a national security matter	2 – Systematic Failure to Uphold Safeguards
Canada	PIPEDA (Personal Information Protection and Electronic Documents Act) protects individual privacy against secret information gathering practices. It prevents phone records from being disclosed by service provider companies without consent, subpoena, warrant or court order unless case recognized as matter of national security	Reasonable expectation of privacy: Privacy is not automatically lost when information gets disclosed to a third party custodian. Third party records have not the same constitutional protection as searching a person's home. Data mining call records is possible if it is related to national security and effective. Intrusiveness should be under an investigation and should not be shared with other agencies	1 – Significant Protection and Safe Guards
Turkey	According to Turkish Penal Code article 135, phone records are considered private unless a judge verdict is given to disclose. The police, gendarmerie, intelligence services may obtain these records from Turkish Communications Centre, which gets data from service providers for demanded numbers in a given period of time	Mining CDR data (not sound records) is permitted after a judge verdict is obtained. Experts are asked to mine specific problems based on the judge's, prosecutor's, or lawyer's demands to prove alibi, relationship, ownership of handsets, etc	N/A
UK	ATCSA (Anti-Terrorism, Crime and Security Act) and RIPA (Regulation of Investigatory Powers Act) permit government requested data in the scope of national security	The issue has not been addressed by any law i in the UK, but it is already widely used	3 – Endemic Surveillance Society
USA	FISA (Foreign Intelligence Surveillance Act) and PATRIOT Act gives authority to government agencies to collect phone record for national security purposes	Reasonable expectation of privacy: Person does not have legitimate privacy expectancy in records that were voluntarily conveyed to a third party. Data mining call records is inefficient and many false positive occur	4 – Extensive Surveillance Society

only little work exists that utilizes CDR for discovering user patterns with data mining techniques. One example is the approach developed in [9], which aims at capturing moving patterns of mobile users from CDR rather than from redundant moving logs.

3 Used CDR Data Set for Case Study

The CDR data we used for this cases study is provided by those GSM operators to which the suspected terrorists of our case were subscribed. CDR comes either in digital form or printed on paper. The latter must be digitalized using OCR software. The final data format is relational. Figure 1 depicts the features of a CDR document. It includes the following fields:

1. Automated row number of records (removed)
2. Caller number (kept)
3. Date and time of call (kept)
4. Type of call, such as "called", "received call", "sent SMS", "received SMS", "directed to number" (kept)
5. Called person's number (kept)
6. Duration of the call (removed)
7. Address of the person (removed)
8. Base station name and number of phone call signal delivered via (kept)
9. The columns 9–11 hold various demographics data about the called person (removed)

We removed fields that were not required for our analysis. They are indicated in the list above. The type of call was converted to two categorical values "called" and "SMS sent".

For spatiotemporal and friendship analysis of CDR data, only records of calls between mobile users are retained, while calls from or to fixed line numbers are removed. This is because fixed line numbers usually cannot be mapped to a single user. Note that CDR data does not only include voice calls, but also data calls (e.g., delivery of short messages).

Fig. 1 Used CDR data (blurred due to privacy)

4   Case Study

Our case study is based on what happened on the last day of 2007 [20], in the Aksaray district of Istanbul, Turkey. Arsonists attacked a truck in order to protest against the government. Five terrorists were detained and taken into custody. Their mobile phones and SIM cards were taken and they confirmed their numbers verbally. With a judge's verdict, the phone call signal data information for the past 10 days was retrieved from their GSM operator's database. During this time span the preparation for the attack and the attack itself took place. The retrieved signal data, received in separate tables – one for each suspect, was merged into a single transactions table. Unnecessary features got removed and some others converted, as described in the previous section.

4.1   Criminal Network Creation and Friendship Analysis

The second field (number of caller) and fifth field (called person's number) were used for the network creation. The number of calls was calculated by counting distinct transactions. After the network was built, we could easily identify someone's acquaintance.

  According to the created social network (Fig. 2) we can understand that P.3 has the highest centrality and betweenness value. Since P.2 received only calls from P.1, we can assume that P.2 takes orders from P.1 and gives orders to P.3. Higher hierarchical position holders generally give call to a subordinate but they are only
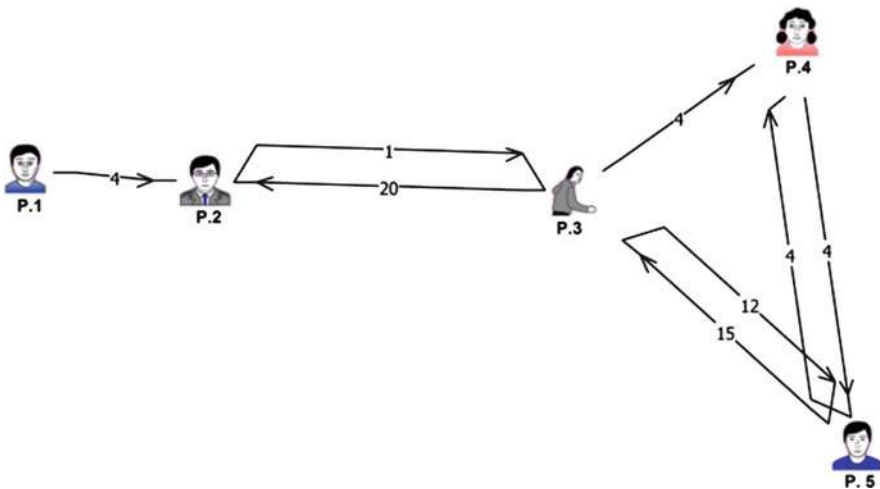
Fig. 2 Obtained terrorist network information from CDR data

called when they ask for feedback. In the observed ten days, P.1 and P.2 used their mobile phones fewer times then other members. According to eyewitnesses' statements, P.3 performed the terrorist attack, assisted by P.4 and P.5. However, in order to conceal their relationships, P.3 came from a different direction than P.4 and P.5, who entered the crime scene together. During the interrogation and trial phase, P.4 denied that she was there. She claimed that she was very far from the vehicle when it was burned by two men.

4.2   Spatiotemporal Analysis of Movements

By knowing which base station received signals from a cell phone reveals the position of its user. During the attack, however, P.1's cell phone was switched off, thus we did not get information about his location, while P.2 was in the south (Fig. 3). P.3 was approaching the crime location (Fig. 4) from the north.

   Around the same time, P.4 was approaching from the west, getting closer to the crime location (Fig. 5). According to some statements, P.5 was approaching towards P.3 (Fig. 6), while P.4 was behind him. We know from the location data analysis that
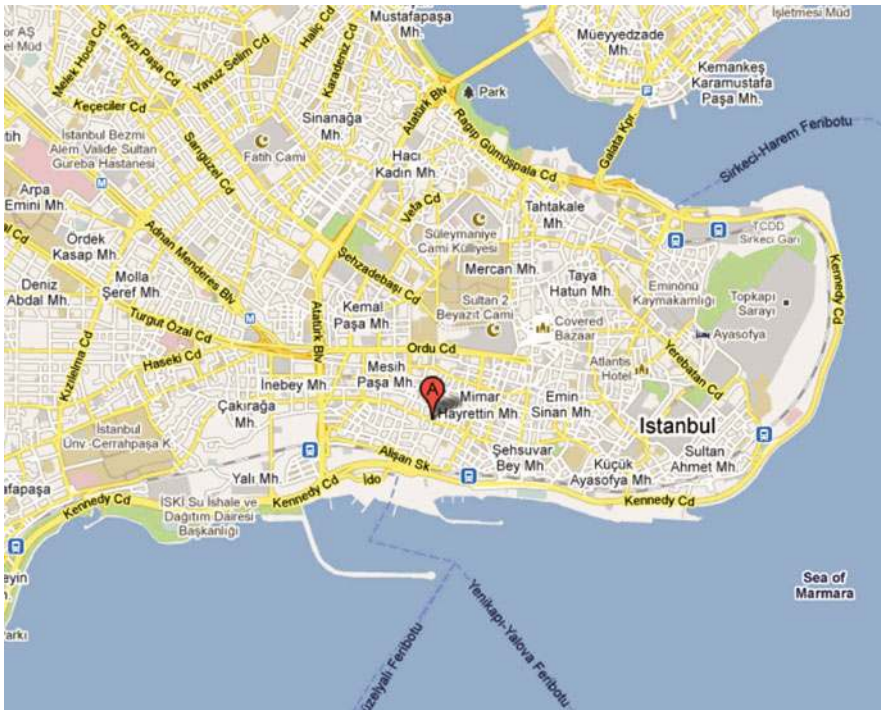
Fig. 3   Location of P.2 during attack (*red pin*)

Fig. 4 Location of P.3 during attack (*red pin*)

Fig. 5 Location of P.4 during attack (*red pin*)

Fig. 6 Location of P.5 during attack (*red pin*)

all members were close to the crime scene. The only exception is P.1, as we do not have his cell phone signal and thus do not know whether he was around. The spatial and temporal analysis reveals that P.4 was moving together with P.5 towards P.3. In the range of 100–500 m accuracy in urban areas, GSM base station signals may be deceiving. Then, signals are mainly coming from the same base stations.

4.3   IMEI Number and GSM Line Number Analysis

Another analysis method over mobile phone call data is matching International Mobile Equipment Identity (IMEI) numbers of mobile phone headsets' IMEI number with SIM cards' line number. By this we can find out if there are any multi-user SIM cards within the same cell at a given time period. In our case, we see that P.4 used two SIM cards with two mobile phones; one of them was also used by another, unknown person (Fig. 7). This shared handset was used 73 times by P.4 and 8 times by the unknown person. Analyzing which SIM card P.4 used while the
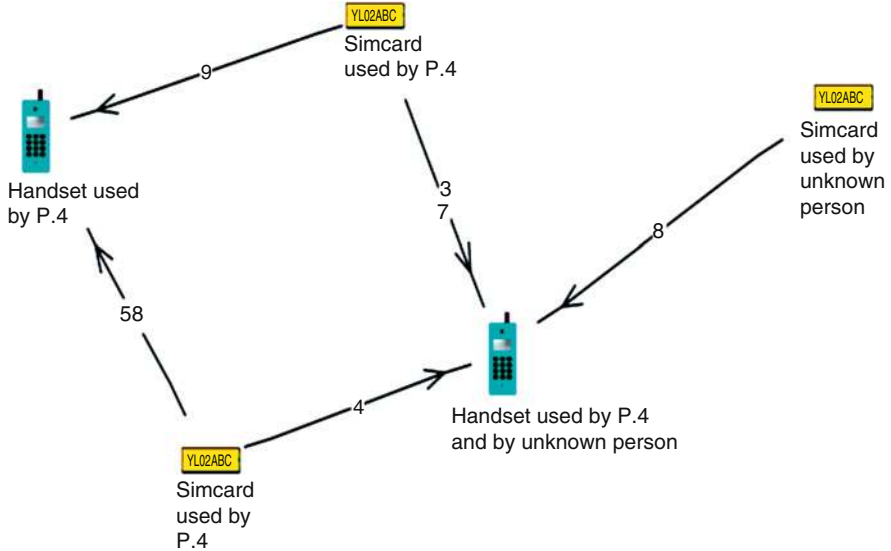
Fig. 7 Handset IMEI numbers and line number of SIM cards are matched. *Yellow nodes* are simcards and *green images* are mobile phones

crime happened gives us the desired answer. In this case, P.4 can be using three SIM cards with two handsets or two persons used one handset with two different SIM cards. One of the problems in identifying the actual user of a mobile phone is that its SIM card may have been purchased by another person. Thus, the analysis of IMEI numbers of handsets and dialing numbers of lines is also valuable for this purpose.

5 Conclusion

After an arson attack performed by a terrorist group, CDR data was legally obtained from GSM operators providing information from cell phone base stations close to the crime scene. The CDR data was used to identify movement and cell phone communication on the crime day and several days before. The following CDR attributes have been considered:

1. Number of lines from a person
2. Number of lines to a person
3. Type of call
4. Date/time of call
5. GSM mast station location
6. IMEI number of used mobile phone

First, we used three attributes of the CDR data for creating a one-mode friendship network including all suspects. This revealed some information about the relationships between the suspects. Then, the spatiotemporal analysis of the suspects' movements showed the suspects' location while the attack happened. Finally, we checked whether a mobile handset was used by more than one suspect by creating a two-mode ownership network based of IMEI numbers and GSM number ownership data stored in SIM cards. We showed handset and GSM line relationships by plotting them as two-mode social network graphs.

With this case study we presented the benefits of using CDR data for solving crime and terrorist events. In the future, more research should be done for spatiotemporal mining of organizational crime. Signals coming from mast stations can be used for more accurate spatiotemporal analysis. Possibly patterns of movements can be identified by showing timely movements of suspects on dynamic maps.

References

1. Chen, H., Schroeder, J., et al.: COPLINK connect: Information and knowledge management for law Enforcement. Decis. Support Syst. **34**, 271–285 (2002)
2. CNET NEWS, Feb.11, 2010: Declan McCullagh: Feds Push for Tracking Cell Phones. http://news.cnet.com/8301-13578_3-10451518-38.html. Accessed 27 Oct 2010
3. Coffman, T., Greenblatt, S., Marcus, S.: Graph-based Technologies for Intelligence Analysis. Commun. ACM **47**(3), 45–47 (2004)
4. Coffman, T.R., Marcus, S.E.: Pattern Classification in Social Network Analysis: A case study. In: Proceedings of IEEE Aerospace Conference, IEEE. 6–13 Mar 2004
5. Diffie, W., Landau, S.: Communications Surveillance: Privacy and Security Risk. Commun. ACM **52**(11), 42–47 (2009)
6. Dallas/Fort Worth Local News From CBS 11 & TXA 21 "Police Arrest 7 Scarecrow Bandit Suspects". http://cbs11tv.com/local/scarecrow.bandits.bank.2.738338.html. Accessed 27 October 2010
7. Greenblatt, S., Coffman, T., Marcus, S.: Emerging Information Technologies and Enabling Policies for Counter Terrorism. In: Behavioral Network Analysis for Terrorist Detection, Wiley-IEEE Press, Hoboken (2005)
8. Han, J., Kamber, M.: Data mining: concepts and techniques. Morgan Kaufmann, California (2006)
9. Hung, C.-C., Peng, W.-C., Huang, J.-L.: Exploring Regression for Mining User Moving Patterns in a Mobile Computing System. In: Proceedings of the 1st International Conference on High Performance Computing and Communications, pp. 878–887. Sept 2005
10. Kaza, S., et al.: Predicting Criminal Relationships Using Multivariate Survival Analysis. In: Proceedings of the 8th Annual International Conference on Digital Government Research: bridging Disciplines & Domains, pp. 290–291. Digital Government Society of North America, Philadelphia, Pennsylvania, USA (2007)
11. MacArthur, A.P.: The NSA phone call database: The problematic acquisition and mining of call records in the United States, Canada, the United Kingdom, and Australia. Duke J. Comput. Int. Law **17**, 441 (2007)
12. Marcus, S., Coffman, T.: Terrorist Modus Operandi Discovery System 1.0: Functionality, Examples, and Value. 21st Century Technologies, Austin, TX, USA (2002)
13. Marcus, S.E., Moy, M., Coffman, T.: Social Network Analysis. In: Cook, D.J., Holder, L.B. (eds.) Mining Graph Data. Wiley, Hoboken, New Jersey, USA (2007)

14. Moy, M.: Using TMODS to Run Best Friends Group Detection Algorithm. In: Proceeding of 21st Century Technologies, Austin, TX, USA (2005)
15. Nanavati, A.A., Gurumurthy, S., Das, G., Charraborty, D., Dasgupta, K., Mukherjea, S., Joshi, A.: On the Structural Properties of Massive Telecom Call Graphs: Findings and Implications. In: Proceedings of CIKM '06, Arlington,Victoria, US, pp. 435–444. ACM, Nov 5–11 2006
16. Wu, B., Ye, Q., Yang, S., Wang, B.: Group CRM: A New Telecom CRM Framework from Social Network Perspective. In: Proceedings of CNIKM '09, pp. 3–10. ACM, 6 Nov 2009
17. Weiss, S.M., Indurkhya, N.: Predictive data mining: a practical guide. Morgan Kaufmann, California (1998)
18. Xu, J., Chen, H.C.: Fighting Organised Crimes: Using Shortest-path Algorithms to Identify Associations in Criminal Networks. Decis. Support Syst. **38**(3), 473–487 (2003)
19. Xu, J., Chen, H.C.: CrimeNet Explorer: A Framework for Criminal Network Knowledge Discovery. ACM Transactions on Information Systems **23**(2), 201–226 (2005)
20. Zaman, Turkish Daily Newspaper (news in Turkish about terrorist attack in Istanbul, TR). http://www.zaman.com.tr/haber.do?haberno=631223. Accessed 7 May 2009

Multilingual Real-time Event Extraction for Border Security Intelligence Gathering

Martin Atkinson, Jakub Piskorski, Erik Van der Goot, and Roman Yangarber

Abstract This chapter gives an overview of tools developed for Frontex, the European Agency for the Management of Operational Cooperation at the External Borders of the Member States of the European Union, to facilitate the process of extracting structured information on events related to border security from on-line news articles, with a particular focus on incidents and developments in the context of illegal migration, cross-border crime, and related crisis situations at the EU external borders and in third countries. A hybrid event extraction system has been constructed, which consists of two core event extraction engines, namely, *NEXUS*, developed at the Joint Research Centre (JRC) of the European Commission and *PULS*, developed at the University of Helsinki. These systems are applied to the stream of news articles continuously gathered and pre-processed by the Europe Media Monitor (EMM) – a large-scale multilingual news aggregation engine, developed at the JRC. In order to bridge the automated analysis phase with in-depth human analysis phase an event moderation tool has been developed, which allows the user to access the database of automatically extracted event descriptions and to clean, validate, group, enhance, and export them into other knowledge repositories.

M. Atkinson · E. Van der Goot
Joint Research Centre of the European Commission, Ispra, Italy
e-mail: martin.atkinson@jrc.europa.eu; erik.van-der-goot@jrc.europa.eu

J. Piskorski (✉)
Research and Development Unit, Frontex, Warsaw, Poland
e-mail: jakub.piskorski@frontex.europa.eu

R. Yangarber
Department of Computer Science, University of Helsinki, Helsinki, Finland
e-mail: roman.yangarber@cs.helsinki.fi

1 Introduction

In the last decade, various security authorities and organisations around the globe have acknowledged that a significant amount of information relevant for early detection of certain threats and for situation monitoring is publicly available. In particular, an ever-growing amount of information published on the Internet led to an emergence of advanced software tools that combine techniques from text mining, machine learning, statistical analysis and computational linguistics to help analysts and intelligence experts to manage the overflow of information, filter out the relevant from the irrelevant, and to extract valuable and actionable knowledge from on-line sources.

This chapter gives an overview of tools developed for Frontex, the European Agency for the Management of Operational Cooperation at the External Borders of the Member States of the European Union, to facilitate the process of extracting structured information on border-security related events from on-line news articles, with a particular focus on incidents and developments relevant in the context of illegal migration, cross-border crime, and related crisis situations at the EU external borders and in third countries. The rationale behind exploiting on-line media sources for this purpose is threefold. First, information on certain border security-related events might not be available from any other sources. This applies in particular to developments in third countries. Second, such information might be available from other sources (e.g., from Member States or through Intelligence Liaison Officer networks), but there might be a significant delay before it becomes available via official channels. Third, open-source information, in particular from on-line media, can be used for cross-checking with information available in-house or obtained from other sources.

The need for strengthening the capabilities for tracking the security situation in the source, transit and target countries for illegal immigration into the EU has been identified and acknowledged by the European Commission (EC). Specifically, the Commission Communication COM (2008) 68[1] proposes the creation of an Integrated European Border Surveillance System (EUROSUR), where step six of Policy Option one suggests development and deployment of new tools for strategic information to be gathered by Frontex from various sources (e.g., from open sources) in order to recognize patterns and analyze trends, supporting the detection of migration routes and the prediction of risks for Common Pre-frontier Intelligence Picture (CPIP).

In order to support the Commission in the process of the development of EUROSUR, Frontex has been tasked to carry out a EUROSUR Pilot Project, whose main objective is to set up an information exchange network between six Member States (Poland, Finland, Slovakia–land borders group, and France, Italy

---

[1] '*Examining the creation of EUROSUR*', http://eur-lex.europa.eu/LexUriServ/-LexUriServ.do?uri=COM:2008:0068:FIN:EN:PDF.

and Spain–maritime borders group) and for Frontex to share information of common interest. The information of common interest is divided into three layers: incident information, analysis information, and operational information. The OSINT (Open Source Intelligence) tools for event extraction presented in this chapter will be used by Frontex specifically to populate the incident information layer of the EUROSUR pilot information exchange network and to provide input (information on the developments in third countries) for the creation of CPIP in the analysis layer. In this manner, the event extraction tools will facilitate and support Frontex in performing its two core tasks: (a) situation monitoring, i.e., providing a constant and short-term picture of the situation, at the EU-external borders and beyond, and (b) intelligence-based risk analysis, which drives the operational work of the agency.

The particularities of the EUROSUR pilot information exchange network and specific tasks of Frontex outlined above impose certain requirements on the tools for on-line news event extraction, in particular, they should:

- Extract information in real or near-real time (situational awareness)
- Extract as fine-grained event descriptions as possible (risk analysis)
- Process news articles in many different languages, since a significant fraction of relevant events are only reported in non-English, local news, with Italian, Spanish, Greek, French, Turkish, Russian, Ukrainian, Portuguese, and Arabic being the most important at the moment

To meet these requirements a hybrid event extraction system has been constructed, which consists of two core event extraction engines, namely, *NEXUS* [1,2], developed at the JRC of the EC and *PULS* [3, 4] developed at the University of Helsinki. These systems are applied to the stream of news articles continuously gathered by the EMM[2]–a large-scale platform for electronic media monitoring, developed at the JRC of the EC. EMM retrieves thousands of news articles per day from circa 2,500 news sources in all EU and major world languages, clusters these articles and classifies them according to a large number of pre-specified categories. In order to bridge the gap between the automated event extraction and in-depth analysis the event extraction system has been equipped with tools for storing and moderation of the event descriptions, and provides functionalities for the visualization of the automatically extracted and moderated events on a map.

The rest of this chapter is organized as follows. In Sect. 2, related work on event extraction is presented. In Sect. 3, our event extraction task is described in more detail. An overview of the complete event extraction framework, including the description of EMM and the two core event extraction engines deployed, *NEXUS* and *PULS*, is given in Sect. 4. Section 5 presents some results of the evaluation of the integrated systems. The aspects of event visualization and event moderation are described comprehensively in Sects. 6 and 7, respectively. Section 8 presents conclusions and directions for future work.

---

[2]http://press.jrc.it.

> '...Illegal network providing forged documents to obtain regular work permits dismantled on 20 October 2010 in Almeria by Guardia Civil. 2 Egyptians were arrested...'

Fig. 1 An example of structured information extracted from a news article describing interception of an illegal network providing forged documents in Spain

2 Event Extraction and Related Work

Formally, the task of event extraction is to automatically identify events in free text and to derive detailed information about them, ideally identifying *who did what to whom, through what methods (instruments), when, where and why*. Information about an event is usually represented in a so-called *template* which can be seen as an attribute-value matrix. An example illustrating an event template is presented in Fig. 1. Automatic event extraction is a complex task [5] due to the complexity of natural language and due to the fact that, in news, a full event description is usually scattered over several sentences and articles. Event extraction is a higher-level task in the problem domain known as Information Extraction (IE); event extraction relies on lower-level IE tasks, which include identifying named entities (e.g., persons, organizations) and relations among them.

Research on automated event extraction [5] was initially pushed by the DARPA-initiated Message Understanding Conferences (MUCs),[3] and later by the Automatic Content Extraction (ACE) programme.[4] Early event extraction systems in the 1980s were handcrafted, monolingual and usually able to process documents only in one particular domain. An example of such early systems is given in [6]. In the 1990s one could observe an emergence of automatically trainable event extraction systems, which relied on large amounts of annotated data, and which still required a lot of effort to be customised to a new domain. The last decade was characterised by moving towards deployment of weakly-supervised machine learning techniques for

---

[3]http://www.itl.nist.gov/iaui/894.02/related/projects/muc.

[4]http://projects.ldc.upenn.edu/ace.

training the systems [7,8], in order to further reduce the time needed to adapt them to new domains and scenarios. The focus also shifted towards tackling multilinguality and greater utilisation of cross-document information fusion techniques to improve the coverage and accuracy of the event extraction results.

Although, a considerable amount of work on automatic extraction of events has been reported so far, it appears to be a less-studied area in comparison to the lower-level IE tasks such as named-entity and relation extraction. Precision/recall figures around 60% are considered to be a good result. Examples of current capabilities of event extraction technology deployed in the security domain, with a focus on identification of disease outbreaks and conflict incidents, are given in [9, 10] and [11]. The most recent trends and developments in the area of event extraction are reported in [12].

The deployment of event extraction technology in the domain of border security appears to be new, although it bears some similarity to previous work. Extracting events from news reports related to terrorist attacks is a topic studied since MUC-3 [13] and MUC-4 [14] in the 1990s. More recently, related work may be found in the domain of epidemic surveillance, e.g., *MedISys* and *PULS* [15], *HealthMap* [16], and *BioCaster* [17]. In fact, epidemic surveillance is in part (though not entirely) subsumed by the security domain, since the spread of epidemics impacts border security as well. The event schema used in epidemic surveillance is similar to that used in the border security domain, however, in the latter domain, the schema is considerably more complex, and requires covering many similar and partly overlapping event types, which complicates the text analysis significantly. Other work similar to ours on event extraction from online news have been reported in [18] and [19].

3   Event Extraction Task for Frontex

This section outlines the event extraction task. We aim at the detection and extraction of structured information on events related to: (a) illegal migration (e.g., illegal border crossings), (b) human trafficking (e.g., human organs trafficking), (c) smuggling (e.g., smuggling in drugs), and (d) crisis situations (e.g., natural and man-made disasters, armed conflicts). Each of the four main event types are subdivided into subtypes. Some subtypes are further divided into fine-grained classes, e.g., earthquakes, storms, wild fires, and outbreaks of infectious diseases are subtypes of NATURAL DISASTER type. The event type hierarchy,[5] including only main and subtypes is depicted in Fig. 2. The subtype OTHER of CRISIS branch in the type hierarchy is used to tag crisis events, which do not fall under

---

[5]As of today, not all of the event subtypes in the event type hierarchy are within the scope of EUROSUR.

```
* ILLEGAL-MIGRATION
    - ILLEGAL-ENTRY            - ILLEGAL-EXIT
    - ILLEGAL-STAY            - FACILITATOR-RELATED-INCIDENT
* HUMAN-TRAFFICKING
    - PROSTITUTION            - FORCED-LABOUR
    - BEGGING                - HUMAN-ORGANS
* SMUGGLE
    - SMUGGLE-DRUGS          - SMUGGLE-GOODS
    - SMUGGLE-WASTE          - SMUGGLE-CBRN
    - SMUGGLE-ARMS
* CRISIS
    - NATURAL-DISASTER       - MAN-MADE-DISASTER
    - VIOLENCE               - ARMED CONFLICT
    - HUMANITARIAN CRISIS    - AREST
    - KIDNAPPING             - HOSTAGE-RELEASE
    - DEPORTATION            - INTERCEPTION
    - SENTENCE/TRIAL         - OTHER
    - NON-DEFINED
```

Fig. 2 The event type hierarchy

any other subtype of CRISIS. The NON-DEFINED class is used to tag events, which potentially might be considered as 'crisis' events, but the system was unable to make an unambiguous decision about the event type.

For all event types of interest there is a harmonised event template structure, which is depicted in Fig. 3. We briefly clarify some of the slots in this structure. The slot LOCATION refers to the most specific place name, which can be extracted. The ZONE slot refers to the area in which a particular event occurred (e.g., *EU territorial waters*, *land border*, *third country*, etc.). The CONFIDENCE slot indicates the automatically computed confidence with which the system has extracted the event, whereas the value of the RELEVANCE slot gives the relevance of the event, i.e., system's guess on how important given event might be for the end user (these may be used as criteria for filtering). Finally, the SNIPPET slot is filled with the text fragment which triggered the extraction of a particular event, i.e., the context in which extraction pattern(s) were matched. The rationale behind introducing this slot is to facilitate tracking the context in the article, from which the information was extracted.

The set of obligatory slots, i.e., the slots which are always filled by the system for any event type, includes: TYPE, EVENT_DESCRIPTOR, SOURCE, and PUBLICATION_DATE. Any other slot depicted in Fig. 3 is filled whenever the system is able to extract some information (usually a text fragment) from the news article(s) and in case it is appropriate to fill such a slot–some slots are event-specific. For instance, the slots ITEM and MEANS might be filled with the information on "what items were smuggled" and "what means of transportation were used to cross the border", respectively.

Fig. 3 The event template structure

4 Event Extraction Framework

This section describes the core event extraction framework. We start with the description of the overall system architecture in Sect. 4.1, which is followed in Sects. 4.2 and 4.3 by an overview of EMM on top of which the system has been implemented. In Sects. 4.4 and 4.5, we briefly present the two core event extraction systems deployed, namely, *NEXUS*, and *PULS*.

4.1 System Architecture

The event extraction system architecture is depicted in Fig. 4. First, news articles are gathered by a large-scale news aggregation engine, the EMM developed at the JRC of the EC [20]. EMM retrieves more than 100,000 news articles per day from more than 2,500 news feeds in 42 languages. These news articles are geo-located,
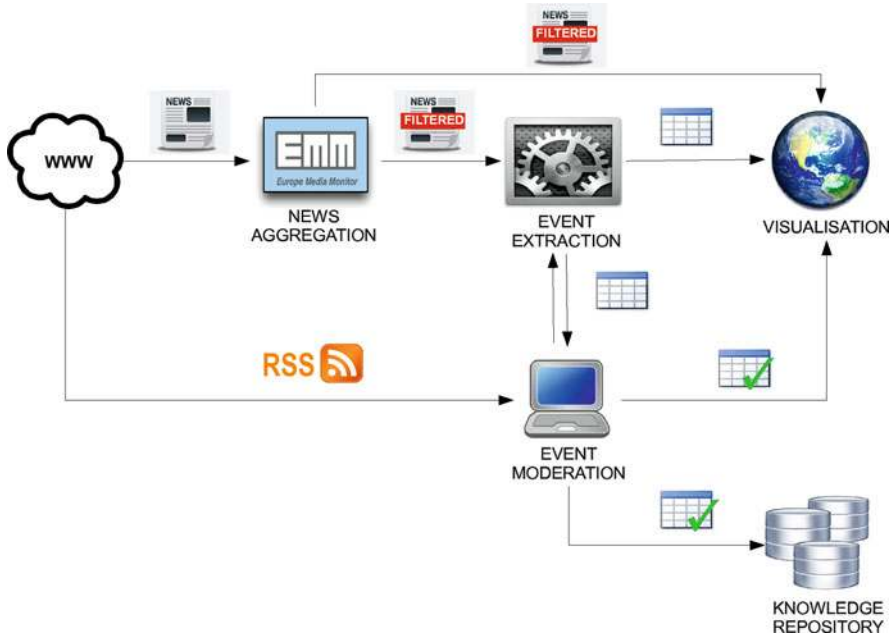
Fig. 4 The system architecture

tagged with meta-data and further filtered using standard keyword-based techniques (see Sect. 4.2) in order to select those articles, which potentially refer to incidents and events relevant for the border security domain (see Fig. 2). In addition, the news articles harvested within a 4-h window are grouped into clusters according to content similarity. The filtering and clustering process is performed every 10 min.

Next, the stream of filtered news articles and clusters is passed to the event extraction engine, which consists of two core event extraction systems, namely, *NEXUS*, developed at JRC, and *PULS* developed by the University of Helsinki. *NEXUS* follows a shallow cluster-centric approach, which makes it more suitable for extracting information from the entire cluster of topically-related articles, whereas *PULS* performs a more thorough analysis of the full text of each news article, which allows us to handle events which may be scattered more widely throughout the article. The deployment of the two event extraction systems was not only designed to obtain richer coverage, but also motivated by the fact that at present they are tuned to detect sets of event types that are not entirely overlapping, i.e., they cover different event types and languages. Furthermore, the application of two different event extraction systems also may provide a means to investigate and study these two approaches in a real-world scenario.

The output produced by the core event extraction engine, i.e., a stream of instantiated event templates, is made accessible to the geographical browser for visualisation of the extracted events on a map (see Sect. 6). Complementing the

structured information returned by the event extraction systems, news articles that were found using keyword-based heuristics as potentially relevant to border security domain are also accessible in the geographical browser, in a separate layer.

In order to bridge the gap between the automated event extraction phase and an in-depth analysis phase, an event moderation system (see Sect. 7) provides functionality to access the database of automatically extracted event descriptions and to clean, validate, compare, group, filter, enhance, and export them. In particular, cleaned and validated event descriptions can be made accessible in a separate 'moderated events' layer in the geographical browser and they can be exported into other knowledge repositories. The event moderation tool also makes it possible to apply the core event extraction engine on any RSS feed[6] on demand, e.g., an RSS feed related to an information source not covered by EMM.

4.2 EMM/FMM

The EMM and its instantiation in the form of the Frontex Media Monitor (FMM)[7] use a combination of techniques from text mining, automated machine learning and statistical analysis to determine *who*, *what*, *where*, and *when* from news articles published on the World Wide Web in near-real time. EMM is a web based multi-lingual news aggregation system that collects hundreds of thousands news articles per day from thousands of sources in many languages. The system exploits information retrieval techniques to automatically classify, locate and identify who and what is mentioned in each article. Every 10 min EMM automatically groups articles talking about the same subject in each language, called News Clusters, and displays the ten largest News Clusters per language. Additionally, a lot of related research work has been conducted using the flow of articles or News Clusters for example to derive reported social networks [21], to provide automatic multi-document summaries [22] and to analyse tonality/bias of articles towards people [23]. The public website (http://emm.newsbrief.eu) provides a user interface to all this information. It is visited on a regular basis by some 30,000 human users, and gets some 1.2M hits per day. The system runs 24/7 only on a few servers without the use of any database technology.

4.3 EMM Processing and Information Retrieval

At the core of the EMM system there is a processing chain of lightweight extensible processes each independently running and connected together using a basic but

---

[6]Really Simple Syndication: http://www.rssboard.org/rss-specification.

[7]FMM is a customised instance of EMM dedicated to monitoring media for border security-related events and topics, i.e., using specialized sources and filters.

robust and reliable web-service architecture developed in-house and based on HTTP post. Articles flow through the processing chain as RSS items. As each item passes through a node in the chain, it gets enriched with meta-data that the node produces, before going onto the next node in the chain. At the end of the chain the generated meta-data is exploited as needed, for example, in the public web site as aggregated statistics on charts.

The first process in this chain is the Scaper System. This system has a list of predefined sites together with the frequency that each site should be visited according to how often the site gets updated. It uses this list to visit each site periodically, then gets the list of articles published on it. It generates a simple RSS feed containing this list. Sites that publish multiple RSS feeds are aggregated into a single feed. Sites that don't publish RSS feeds are processed using a custom HTML parser that converts the web pages into XHTML which is then transformed into RSS.

The next module in the chain, called the Grabber, detects the new articles published, and, using a patent-pending text extraction process, determines the main content of the new articles. Grabber produces a new RSS feed for each site, containing title, link, description and text for all new articles detected, which is then passed on to the next process in the chain, the automated language detection, based on word frequency tables. The Information Clustering and Story tracking processes populate these continuously updated frequency tables using a technique similar to an infinite input response filter [24].

Next, an Entity Recognition process detects people and organisation names in the article from a home-grown multi-lingual information base. This information base of entities and organisations is populated by an automated (offline) entity recognition system.

Geo-Tagging is the next process in the chain that detects potential place names in the text. Again, the place names come from a multi-lingual gazetteer like Geonames.[8] In addition to the place name and location, all place names are classified based on administrative size and within their administrative location, like province, region and country. All this information is used for disambiguation: common place names like *Paris, France* from *Paris, Texas*; common words like *in Beaucourt-sur-l'Hallue in France*; and removes previously identified entities like *Lewis Hamilton*, to avoid selecting the capital of Barbados. A final round of disambiguation takes place after the clustering phase of the system (see [2] for more details).

The RSS items then flow into a Classification System. This system uses a mix of multi-lingual weighted regular expression based patterns together with combination of boolean logic to match text in the article in order to tag it as belonging to a given category. Articles are assigned multiple category tags according to their content together with a rank value indicating the relative position in the text. An example of a category is given in Fig. 5. At a basic level this system can be used to identify the occurrence of a wide range of general to specific events. In order to avoid noise, for

---

[8]The multilingual gazetteer mixes open resources from http://www.geonames.org and an in-house developed geo-spatial gazetteer.

With word proximity of 50 look for any one of:

sin+papeles, sans+papiers, undocumented+immigrant%, i_migr%+ilegal, inmigrante%+irregular%, illegal+migra%, illegal+immigra%, illegal+resident%, illegale+immigrant%, illegale+immigratie, illegale+migranten, illegale+migration, imigrant%+clandestin%, imigrante%+ilega%, imigração+clandestina, clandestin%, imigração+ilegal, immigr%+clandestin%, immigrant%+illega%, migrazion%+irregolar%,immigration%+irr_guli_r%, irregular+migr%, overstayer%

AND any one of:

barca, barche, mer, barcos,bateau, passera, passeras, bateaux, ships, seas, carrett%+del+mare, pater%, harrag%, waters,mare boat%, navio%, cayuco%, ship, barco, vessel%, naufrag%, guarda+coste, Guarda+Coste, guardia+costera, Guardia+Costera, Coast+Guar%, Guardia+Costiera, Garde+côtière, mediterranean, atlantic, Lampedusa, Ténérife, Canaries, Canarie, sbarc%, acque+libiche, Canarias, aguas, Canarias

But NONE of:

EEUU, somalia, estados+unidos, USA, US, mexico, america, amerika, chile, united+states, australia

Fig. 5 An example of the FrontexSeaBorder category definition. Key: A '+' represents any amount of *white space*, a '%' represents zero or more characters, a '_' represents exactly one character

the Frontex event extraction task a filter is applied that only allows articles that have been tagged with any number of border-security related categories. It is important to mention that the Classification System is also capable of dealing with Arabic and similar languages with noun prefixing, and Chinese and other ideograph languages, where no whitespace separates words.

Next, an Article Filtering System allows articles to be selected based on meta-data content. The selection can be boolean combinations of meta-data values allowing expressions like "articles occurring in Spain and talking about sea borders".

Subsequently, the articles flow into the Clustering and Story Tracking Cache, where every 10 min the last 4 h of articles are hierarchically clustered in every language individually. Initially each news article is considered to be a cluster, the process is agglomerative and employs average group linkage to determine the distance between the clusters using a simple cosine measure. The clustering process continues until the maximum cosine distance falls below a certain set threshold which is a function of the theoretical density of the feature vectors, where a higher density leads to a higher threshold value. The article feature vectors are simple word count vectors, constructed using a simple bag of words approach, with some additional ad-hoc rules like: ignore top 100 frequent words, ignore words of two characters. A cluster only remains if it has at least two articles that are not duplicates from at least two different news stories. This algorithm has also been modified and tested on ideogram-based texts in order to cluster Chinese language articles. The clustering system forms the last phase of the IR processing.

Event extraction and other more sophisticated IE processes can be plugged in after the Classification component in order to process individual articles, or after the Clustering System in order to extract events from topically-related articles.

4.4 NEXUS

NEXUS, follows a shallow cluster-centric approach [1, 2]. Each cluster of topically related articles undergoes a shallow linguistic analysis (fine-grained tokenization, morphological analysis, gazetteer look-up, sentence boundary detection, etc.) and a cascade of simple finite-state extraction grammars is applied to each article in the cluster. While the lower-level grammars are used to extract person names (*Osama bin Laden*), person groups (*Algerian immigrants*), numerical expressions (*two hundred*), quantifiers (*More than*), and other small-scale structures (*small boats*), the top-level grammar consists of 1–2 slot extraction patterns, like the ones shown below, where the labels in angle brackets refer to the semantic roles which will be assigned to the matched objects.

```
PER-GROUP <DEAD> "was shot by" PER-GROUP <PERPETRATOR>
PER-GROUP <PERPETRATOR> "have released"
"have released" PER-GROUP <RELEASED>
PER-GROUP <DISPLACED> "fled their homes"
PER-GROUP <IMMIGRANT> "sbarcati clandestinamente"
```

For instance, the application of the last pattern above to the text *15 palestinesi sbarcati clandestinamente in Sicilia* would result in assigning the group of 15 Palestinians (*15 palestinesi*) the semantic role IMMIGRANT, i.e., a corresponding slot in the event template structure is filled with *15 palestinesi*.

The grammars are encoded and processed with *ExPRESS*, a highly efficient finite-state based extraction pattern matching engine [25]. Patterns in *ExPRESS* are regular expressions over flat feature structures, i.e., non-recursive typed feature structures, whose attributes are strings.

The system processes only the top sentence and the title of each article for the following reasons:

- News articles are written in the "inverted-pyramid" style, i.e., the most important parts of the story are placed in the beginning of the article and the least important facts are left toward the end
- Processing the entire text might involve handling more complex language phenomena (e.g., anaphora, ellipsis and complex syntax), which is hard and requires knowledge-intensive processing
- If some crucial information has not been captured from one article in the cluster, it might be extracted from other articles in the same cluster

We demonstrate the handling complex language phenomena by considering the following two sentences:

*The United Nations says **Somali gunmen** who hijacked a U.N.-chartered vessel carrying food aid for tsunami victims **have released the ship** after holding it for more than 2 months.*

***Somali gunmen have released the ship** after holding it for more than 2 months.*

In the first sentence, proper extraction of the fact that Somali gunmen have released the ship requires deeper syntactic parsing in order to detect that *who hijacked a U.N.-chartered vessel carrying food aid for tsunami victims* is a relative clause that describes the Somali gunmen. Otherwise, the application of a linear extraction pattern (e.g., `PER-GROUP <PERPETRATOR> "have released"`) would result in tagging *tsunami victims* as a perpetrator of the `HOSTAGE-RELEASE` event, i.e., as those, who released the ship, which is clearly wrong. In the case of the second sentence, the application of a linear pattern would yield a correct extraction of *Somali gunmen* as the actor of the event. Noteworthy, the title and initial sentences of most of the news articles on crisis-related events exhibit relatively simple syntactical structure, like the one in the second sentence in the above example or in the example in Fig. 6.

Since the information about events is scattered over different articles, the last step consists of cross-article cluster-level information fusion in order to produce full-fledged event descriptions, i.e., information extracted locally from each single article in the same cluster is aggregated and validated. This process encompasses mainly three tasks: entity role disambiguation (as a result of extraction pattern application the same entity might be assigned different roles), victim counting and event type classification, all accomplished through heuristics. Consider as an example the titles and corresponding first sentences from articles on an event related to killing 44 people at a wedding in Turkey, shown in Fig. 6. '*44*' is selected as an estimate of the number of dead (victims) since it appears most frequently ('*45*' appears only once in the last article, therefore it is discarded). Furthermore, article three is the only one mentioning 22 women and children being involved in the event and Kurdish region as the place of the event. Through cross-article information fusion this information becomes present in the final event description. If the same entity has two roles assigned in the same cluster, preference is given to the role assigned by the most reliable group of patterns, e.g., two-slot extraction patterns are considered most reliable, patterns for the extraction of the `KILLED` slot is considered more reliable than the ones for extraction of the `ACTOR` (perpetrator) slot. In our example, *44 people* will be assigned the role of the victim since patterns for detecting `KILLED` (victims) has priority over the patterns for extracting the actors (e.g., in *44 people killed* the application of some extraction patterns, e.g. `PER-GROUP <PERPETRATOR> "killed"`, might assign the role of the perpetrator to *44 people*). It is important to note that *NEXUS* detects and extracts only the main event for each cluster. The details of information fusion are discussed in detail in [1, 26]

NEXUS is capable of processing news in several languages, including, i.a., English, Italian, Spanish, French, Portuguese, Russian and Arabic. Due to a linguistically light-weight approach, *NEXUS* can be adapted to process texts in a new language in a relatively short time [27, 28]. In particular, weakly supervised methods (bootstrapping techniques) are deployed for facilitating the process of creating language-specific resources, i.e., domain-specific lexica [28] and extraction patterns [29].

44 shot dead during Turkey ceremony
Eight gunmen suspected of fatally shooting 44 people at an engagement ceremony in south-east Turkey have been arrested . . .
44 killed in attack in Turkey (AP)
AP – Turkish security forces on Tuesday detained eight gunmen suspected of fatally shooting 44 people, many of whom were praying, at an engagement ceremony in the rural southeast of the country . . .
Turkey – Forty-four people killed at wedding party in Kurdish region
*Forty-four people, including **22 women and children**, were killed in an attack on a wedding party late on Monday in Turkey's **Kurdish region** . . .*
Masked gunmen kill 45 people at engagement party
ANKARA, Turkey – Masked assailants with grenades and automatic weapons attacked an engagement ceremony in southeast Turkey on Monday, killing 45 people . . .

Fig. 6 Titles and corresponding first sentences from articles (in the same cluster) reporting on a same event related to killing of forty-four people at wedding in Turkey

Although the *NEXUS* architecture is language-independent, the system is quite flexible regarding the grammar design approach that it can implement. In fact, while surface pattern-based grammars proved to be largely effective for English language, they turned out to perform significantly worse for Romance languages, due to a number of phenomena within this language family. For instance, verb phrases that describe events appear to be structurally more complex, with crucial event information conveyed by productive prepositional phrases rather than by the lexical head of the verb phrase–as in English–as is evident in the following example from an Italian article:

*Saturday, February 02, 2007 08:19:00 AM CEST Due donne italiane **sono state barbaramente uccise a colpi di pietra** sull'isola di Sal, . . .*
   [Two Italian women **were ruthlessly stoned to death** on Sal island . . .]

Moreover, in the compressed style of news, phrases are often nested rather than linearly segregated, as in the following excerpt from a Spanish article, where the locative Prepositional Phrase (PP), which is part of the verb phrase "localizada", is linearly nested within and splits the subject NP phrase:

Localizada[*una patera* [*en Cadiz*$_{PP}$] *con cuatro sin papeles muertos*$_{NP}$]
   [Detected a boat in Cadiz with four dead clandestines]

All these phenomena in these languages, which are potentially hard to tackle for surface form-based approaches, are partially coped with by modeling extraction patterns and noun phrases via rules referencing more abstract morphological information and by factorizing them into their component parts.

The cluster-centric approach to event extraction described above appeared to work satisfactorily in case of crisis-related events (e.g., man-made and natural disasters, violent events, etc.) [2]. However, extracting information on illegal

migration incidents and related cross-border crimes poses additional challenges. In particular, our empirical observations revealed that:

1. Information about the target incidents is often only implicit in text–in contrast with more traditional IE domains (environmental crisis), i.e., their type is not explicitly encoded in language
2. Such incidents are often reported locally, i.e., in local news, in one or two languages only, so the cluster-centric approach might not be applicable (at least the information fusion part of this approach)
3. In most cases (due to low number of articles), relevant information (e.g., description of the illegal immigrants) is neither reported in the title nor first sentence of any article available
4. Geo-locating is more difficult since there are usually several geo-references in an article (significantly more than in the case of crisis events), e.g., to event place, country of departure, destination country/place, places visited on the illegal migration route, references to countries of origins of the victims and perpetrators, the origin of the means used by some authority involved (e.g., boat of the Coast Guards in certain region/city, etc.)

In order to adapt *NEXUS* to the extraction of illegal migration incidents and related cross-border crimes we are currently exploring several new strategies to tackle the problems listed above.

For instance, the extraction of violent events in principle boils down to modelling the mapping from some predicates (e.g. ARRESTED, INJURED, RELEASED) to linguistic constructions (verb groups) that express them, whereas in illegal migration domain, the target event types have high-level definitions, and frequently, they have to be inferred from: (a) other event types typically co-occurring with them, (b) various semantic features of these events' participants. Event type information is rarely carried by the verb groups only. In the following example, the qualification as '*immigrants*' of the Agent role filler of a generic motion event and an humanitarian '*rescue*' event seem to be sufficient to trigger an 'Illegal Entry Attempt' and 'Interception' event, resp.:

> *Siete pateras con 100 inmigrantes a bordo llegan a Almeria durante el fin de semana*
> [Seven boats with 100 immigrants on board arrived in Almeria during the weekend]
>   *Rescatan una patera con 16 inmigrantes en Almeria, la sexta en 24 horas. . . .*
> [A boat with 16 immigrants is rescued in Almeria, the sixth in 24 h]

This example demonstrates that a knowledge-based inference mechanism [30] has to be modeled based on empirical analysis of appropriate news corpora for this particular domain. Moreover, richer frame structures are required, allowing at least for secondary event information.

The problems two and three listed above can only be solved by going beyond the title and the first sentence, as the *PULS* system does, or by merging automatically generated event descriptions on the same event, which did not appear in the same cluster (were reported on different days). Our initial experiments on processing entire articles (see Sect. 5.1) revealed that the recall can be significantly improved, but the precision drops, as was expected.

The problem with the geo-locating, i.e., improving the accuracy thereof could be solved by some kind of decomposition of location prepositional phrases, in order to filter out certain geo references, e.g., `going via <GEO-NAME>` pattern could be used to filter out certain country name, which is on the route. Clearly, the countries on the route are relevant information to be extracted, but the most important geo information is the location where a concrete incident took place (e.g. illegal border crossing attempt, interception, detection, deportation, etc.).

For further details on *NEXUS* see [1, 2, 26, 27].

4.5   PULS

PULS, the *Pattern Understanding and Learning System*, is developed at the University of Helsinki to extract factual information from natural-language text. The *PULS* IE system has been adapted to analyse documents that have triggered EMM/FMM alerts, as well as alerts from the epidemiological domain.[9] In the Frontex application domain, *PULS* analyzes the entire text of the incoming articles for facts about outbreaks of communicable diseases, as well as illegal migration incidents and related cross-border crime, such as human trafficking and smuggling. In this section, we survey the following aspects of the system: (a) high-level overview, (b) highlights about tailoring the system to the security domain, (c) information aggregation, and (d) *PULS* in the domain of outbreaks of epidemics.

IE technology aims to deliver detailed information about specific incidents of interest that are reported in the retrieved articles. The purpose of the detailed analysis is to boost the precision of the results, since, as mentioned previously, keyword-based queries may trigger on documents which may be off-topic but happen to contain the alerts in unrelated contexts. In this way, IE aims to make the retrieved information more useful for the end-user. The pattern matching component in IE is one of the main mechanisms that assure that the keywords appear only in relevant contexts. This is of value to users who are interested in on-going surveillance of *specific, focused scenarios*, as opposed to users who wish to monitor documents that mention the alerts in any context.

The approach to IE taken in *PULS* is similar in essence to the approach described above in Sect. 4.4. *PULS* employs patterns that are written in a regular-expression formalism. *PULS* also has preliminary processing phases, where it identifies proper names, expressions of several fixed types–dates, monetary expressions, numeric expressions, etc.–and the basic components of the sentence, such as noun and verb phrases.

PULS differs from the above approach in that *PULS* analyses the entire text article. This means that *PULS* needs to address the problems mentioned in the preceding section. To deal with the problem that information may be scattered

---

[9]http://puls.cs.helsinki.fi.

over several sentences, *PULS* applies anaphora resolution; this is a technique for linking referring expressions–such as pronouns ('*they*') and proper noun phrases ('*the immigrants*')–with the expressions to which they refer elsewhere in the text.

Covering the complete text means that *PULS* may be able to discover more details about interesting events, but it also means that information found later in the article may be less relevant. To handle this problem, *PULS* computes a *relevance score* for each extracted fact. Relevance depends on many *features* of the news article. A useful feature is the relative position of the sentence that contains the event within the document. Later sentences typically have lower relevance–but not always, since a fact that is important for the user may not be the focus of the article, and may be mentioned as secondary information (and therefore, not at the very beginning). Conversely, a fact mentioned in the first sentence may have low relevance, e.g., if the article is describing events that occurred some time ago. Another feature that helps identify the relevance is, therefore, the difference between the publication date and the date of the event. The length of the article is yet another indicator of relevance–longer documents are more likely not to be about specific current event, but rather surveys with broader scope.

Another set of features is the specific words mentioned in the text. For example, if the text contains the words '*conference*' or '*scientific study*' then the article is likely not describing current events. These types of word-dependent features are called *lexical* features.

These and other features we employ are not deterministic–they do not indicate definitively whether the fact is relevant. We do not speak about precise indicators, but rather in terms of tendencies, or *probabilities*. Therefore the features are combined in a probabilistic fashion, through standard machine-learning techniques. *PULS* uses Support Vector Machines (SVM) and Naive Bayes classifiers to predict the relevance of each event and each article, based on the pre-defined set of features.

Probabilistic modeling also spares the system designer from having to hand-craft the particular lexical features to be used. Doing this manually would be both labor-intensive and error-prone. Probabilistic methods also allow the system use to a very large number of lexical features, and to learn directly from the data which are useful and which are not.

For each document, the IE system extracts a set of events reported in the text. The event templates–the structure of the extracted information–are identical to the ones described above. Each event, as above, consists of a set of attributes.

In the epidemic domain, the event will contain (if found) the location and country of the incident, the name of the disease/condition, the date of the incident, and descriptive information about the victims–the type of victim (people, animals), number, whether they survived, etc. The incident may cover a single occurrence *'80 chickens were reported dead of bird flu on the farm'* or larger time interval, as in *'Two people in the region have contracted the disease since the start of the month.'* Text may also contain 'periodic' incidents: *'according to the Health Ministry, 330 people die of malaria in Uganda each month'*. The system also identifies events in which the disease is *unknown*, or undiagnosed, which are especially important for surveillance.

The system relies on several domain-independent and domain-specific *knowledge bases*. The hierarchical gazetteer containing names of countries, states or provinces, cities, etc., is an example of a domain-independent knowledge base. An example of a domain-specific knowledge base is the medical ontology, containing names of diseases, viruses, drugs, etc., organised in a conceptual hierarchy. The *PULS* ontology currently contains 2,700 disease terms; 400 vectors (organisms that transmit disease, like rats, mosquitoes, etc.); 1,500 political entities–countries, their top-level divisions and name variants; and over 70,000 location names (towns, cities, provinces).

For illegal migration incidents and related cross-border crime, as well as for epidemic events, *PULS* analyses the DATE of the event, and the LOCALE, i.e., the country and a more specific location within the country, when possible; in case of illegal migration, it tries to extract the second country/locale as well. The other attributes are domain-specific, and follow the schema defined in Fig. 3. In the Frontex domain, *PULS* looks for attributes relating to the target event types: the ITEM slot is filled with names of drugs, arms, and goods that are typically smuggled between countries, VICTIM_DESCRIPTOR, and PERPETRATOR_DESCRIPTOR for the human trafficking scenario and for several kinds of suspects (people who enter or stay in countries without appropriate permits) for the illegal migration scenario. An example of news article text on smuggling arms is given below.

> *On Nov. 3, authorities arrested three men, including an active duty U.S. Navy SEAL, for conspiring to smuggle and sell weapons to an undercover federal agent in Nevada and Colorado. According to a criminal complaint, SEAL Nicholas Bickle of San Diego smuggled 80 AK-47 weapons from Iraq...*

Since the types and number of items, suspects, victims, and authorities are typically not mentioned in a compact form in the same sentence, we use a wider context to try to find all the attributes for a given event. *PULS* tracks concepts in the vicinity of the event pattern that are suitable to fill the missing slots.

Populating the knowledge bases requires a significant investment of time and manual labour. *PULS* employs weakly-supervised learning to reduce the amount of manual labour as far as possible, by bootstrapping the knowledge bases from large, un-annotated document collections [31, 32].

Information Aggregation: News stories considered important by the news media may be repeated multiple times across different sources. This may create a situation where the redundancy will potentially cause the user to consume extra time on sifting through stories that are identical or nearly identical.

Developing stories that evolve over time also tend to generate multiple reports in the news. Epidemic outbreaks are a common example of this, where the outbreak may have an initial onset, a peak, and a subsequent decay; multiple stories related to the outbreak will appear in the media, which may complement, correct, update, or even contradict one another. To handle these types of relationships among reported stories, *PULS* attempts to aggregate the extracted facts into groups of related events.

Linguistic coverage: At the present time, *PULS* analyzes articles in English for the illegal migration domain, and in English and French for the medical domain. Extension to further languages is on-going work.

5 Evaluation

This section gives and overview of some results of the evaluation of both *NEXUS* and *PULS* in terms of their coverage and precision. Figures regarding accuracy of geo-locating events are also given. Finally, some numbers reflecting the quantity of processed data and extracted events over time are presented.

5.1 NEXUS *Evaluation*

Before *NEXUS* was adapted to detect events related to illegal migration and cross-border crime it has been mainly deployed for detecting violent events and man-made/natural disasters. Some evaluation of its performance in this particular domain has been carried out. For instance, in an experiment on processing 368 English-language news clusters from 24 January 2008, *NEXUS* achieved 93% coverage and 81% precision regarding the task of detecting violent and man-made/natural disaster events. With respect to DEAD and INJURED slots, an accuracy of 80 and 93% could be achieved, resp. Similar evaluations for other language-specific instances of *NEXUS* have been carried out. For example, the evaluation for Italian on 213 clusters of news articles gathered during four consecutive days in July 2008 yielded the precision/recall figures presented in Fig. 7. Here, two instances of the event extraction system were compared, a baseline version using analogous resources as for the English system (low linguistic abstraction and phrase decomposition), and an advanced one relying on more abstract rules along the lines sketched in Sect. 4.4, and described in more detail in [27]. As can be observed, deploying more abstract patterns improves recall for Italian, and similar behaviour was found in some experiments on Spanish and Portuguese.

We have carried out some experiments on evaluation of recall and precision for detecting illegal migration incidents and related cross-border crimes. For the evaluation of the precision of *NEXUS* for Italian and Spanish we have selected for each language 50 automatically extracted event descriptions and measured weighted precision as follows. Let p_i denote precision of the extraction of attribute (slot) i, and let w_i denote the weight of attribute i (the higher the weight the more important the

|	Dead		Injured		Arrested	
	BL	LR	BL	LR	BL	LR
P	0.82	**0.91**	0.66	**0.77**	**0.83**	0.70
R	0.48	**0.84**	0.15	**0.53**	0.55	**0.66**

Fig. 7 Comparison of the extraction accuracy for Italian for DEAD, INJURED, and ARRESTED roles with the *BL* baseline system and *LR* linguistically rich grammar. *P*, *R* stand for precision and recall

attribute). The weighted precision is then calculated as $wp = \sum p_i \cdot w_i / \sum w_i$. The following attributes and corresponding weights have been considered for the evaluation of the weighted precision: TYPE (weight five), SUBTYPE (1), COUNTRY_ID (5), DATE (5), PERPETRATOR_DESCRIPTION (5), NUMBER_AFFECTED (1), NUMBER_ARRESTED (1), WOMEN/MINORS_INVOLVED (5), MEANS (1), ITEMS (1). For Italian we obtained weighted precision of 95.91%, whereas for Spanish it amounted to 85.72%. The slightly worse result for Spanish is due to the incompleteness of the extraction grammars for Spanish. The SUBTYPE and COUNTRY_ID slots turned to be the most 'difficult' ones, although it has to be mentioned that for some slots, e.g., MEANS and ITEMS, the values were rarely extracted, whereas the value for SUBTYPE and COUNTRY_ID was extracted in nearly 100% cases.

Although, precision of extracting slots for illegal migration and related cross-border crime events is comparable with the results obtained for the crisis-related events (e.g., man-made and natural disasters, violent events, etc.), the performance figures for the event detection (event filtering) task appear to be significantly worse since the cluster-centric approach to event detection in the illegal migration domain is not applicable due to the problems mentioned in Sect. 4.4. In particular, we have computed the recall and precision for the *event filtering* task, which measures how many of the events that are reported in the articles are correctly picked-up – without regard for accuracy of filling the individual slots. For this purpose we have selected and annotated 300 news articles in Italian related to illegal migration and cross-border crime, which were pre-filtered with EMM alerts. We run two versions of *NEXUS* on this news article collection. The first 'classical' version, analyses only the first sentence of each news article, whereas the second 'enhanced' analyses the entire text of each article and the title as well. The 'classical' *NEXUS* obtained 75% precision and 20% recall, whereas the 'enhanced' version scored 60% precision and 79% recall. In particular, the results obtained by the latter version of *NEXUS* indicate that the analysis of the entire text of news articles appears to be essential in the illgal migration domain, which confirmed our earlier empirical observation that significant amount of relevant information is not mentioned in the first sentence/title. Nevertheless, these results were achieved with a preliminary version of the system adapted to illegal migration domain for Italian, so they just give a coarse-grained picture.

An interesting aspect of evaluating an event extraction system is the number of event descriptions it is capable to extract from a given news article stream over time. The diagram in Fig. 8 shows a comparison of the number of automatically extracted events related to smuggling and illegal migration from the news in Italian and Spanish in the last quarter of 2010. A change in the trend can be observed, i.e., the number of illegal migration incidents decreased, while the number of smuggling events rose, which reflects the real situation. In particular, Italian and Spanish authorities reported ever-decreasing numbers of illegal migration attempts in this particular period, while the number of incidents in Greece (rarely reported in Italian and Spanish news) significantly rose in the second half of 2010. In other words, the main entry point of illegal migrants to the EU shifted from Italy's and Spain's maritime borders to Greek-Turkish border.
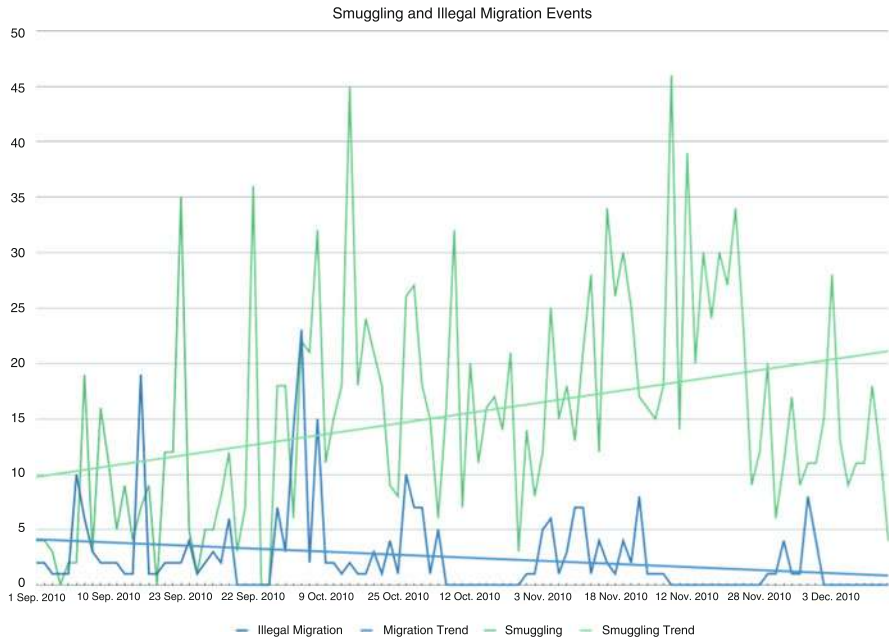
Fig. 8 Smuggling versus illegal migration incidents extracted from the news in Italian and Spanish in the period of September 2010–December 2010

5.2 *PULS Evaluation*

PULS continually undergoes several types of evaluation, ranging from the formal MUC/ACE-style evaluations, which measure performance on every slot in the template, to more coarse-grained evaluations, which focus on extracting a certain subset of "key" slots. The performance, measured on a training corpus of 100 documents, ranges from 65 to 70% recall and 75 to 80% precision. The false-negative rate, measured on a corpus of 200 documents selected as candidates by MedISys using keyword-based matching, was approximately 15%. Results from *PULS's* evaluation on other domains can be found in [15].

The first quantitative evaluation of *PULS* on the domain of illegal migration and related cross-border crime detection (for English) yielded 61% recall and 80% precision for the event detection (event filtering) task, which is higher than the currently deployed 'classical' *NEXUS* version, but exhibits less recall and is more precise than the 'enhanced' *NEXUS* version. A collection of circa 30 news articles in English was used for the aforementioned evaluation. The main conclusion of performing the evaluation of the event filtering task revealed that significantly better results can be obtained in case the entire text of a news article is analysed. The evaluation of the precision of extraction of the various attributes by *PULS* has not been accomplished yet, but first figures indicate slightly lower scores than those obtained by *NEXUS*.
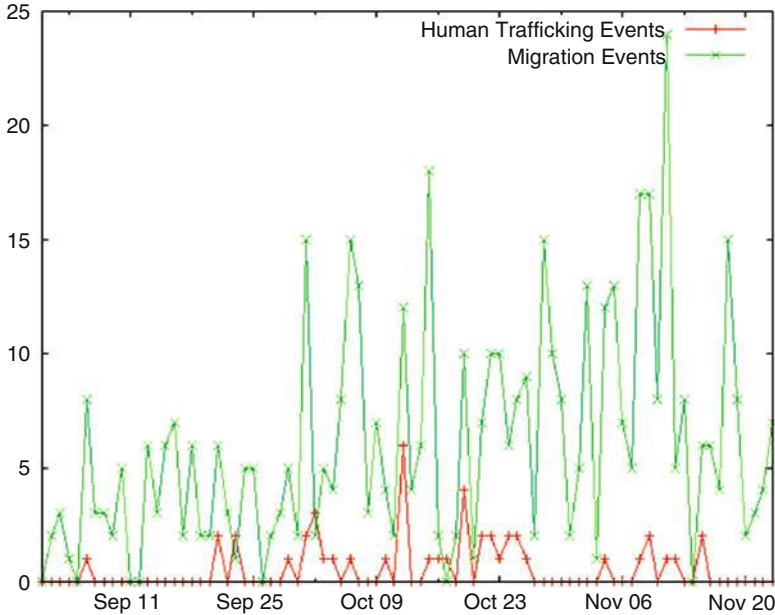
Fig. 9 Illegal migration versus human trafficking events extracted from the news in English in the period of September 2010–December 2010

Analogous to the evaluation of *NEXUS*, we have computed the numbers of events automatically extracted by *PULS* from news articles in English over time. We have observed that human trafficking events are less frequently reported in the news than illegal migration incidents are. The Fig. 9 shows a comparison of the numbers of human trafficking and illegal migration related incidents extracted from the news in English in the period of September 2010–December 2010.

The average number of news related to health-threats arriving from EMM is about three times higher than the number of border-security related documents, whereas the amount of medical events extracted is 3–4 times higher than the border-security events, as depicted in Fig. 10. The proportionally higher number of medical events extracted might be the result of the fact that news on health threats tend to report on more events than the news on typical border-security events, e.g., illegal migration attempts, trafficking, etc. The peak in medical events in Fig. 10 reflects the outbreak of cholera in Haiti, which started at the end of October 2010.

5.3 Geo Tagging Evaluation

One of the important aspects of the event extraction system in the domain of border security is to automatically geo-locate events. Actually, it was expected that the new domain will present new challenges to the accuracy and recall of event locations, especially where complex location information is encoded in the event
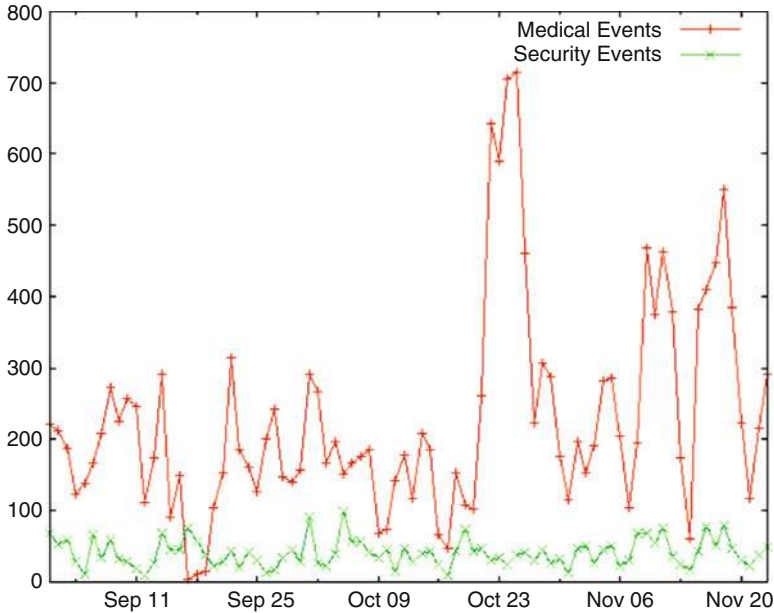
Fig. 10 Medical (health-threat) events versus border-security related events extracted from the news in English in the period of September–December 2010

like: '*Migrants left port X, pass via point Y and entered at location Z*'. In order to verify this hypothesis an evaluation of the performance of the geo-tagging system was undertaken and the results are reported here. The main premise of the geo-coding system was reported in [2] which relies on the use of clustered articles for the final geo-disambiguation stage. Typical border-security related events, i.e., illegal migration and related cross-border crime, cannot profit from this information since the events are taken from individual articles and not clustered articles.[10] This evaluation looked at events related to smuggling, generated by the 'Italian' system. Out of 207 events, 118 were geo-located correctly, 62 were wrongly geo-located and 27 events were not geo-located at all, which yields 65% precision and 57% recall. In a simmilar experiment for the crisis domain (natural/man-made disasters and violent events) we obtained for Italian 99% precision and 89% recall. For the aforementioned experiment a corpus of 84 news articles was used. These results reflect the increased complexity of the geo-locating in the domain of illegal migration and related cross-border crime. As already mentioned at the end of Sect. 4.4 an application of IE-oriented patterns to filter out certain geographical names from being considered a location of an event could potentially alleviate the problem. For instance, the geographical names matched by a pattern like

---

[10]Border security related events are extracted from single articles since the appearance of such events is often singularly reported, clustered articles only report large impact events.

passing through `<GEO_NAME>` could be scored lower than geographical names matched by patterns like `were intercepted in <GEO-NAME>`.

6 Event Visualisation

A fundamental property of events is where they occur. Therefore, already once we got the initial version of the event extraction system, a KML (Keyhole Markup Language)[11] services were set up to visualise the events on a map (Google Map). Since then the number of languages that we are able to extract events from has increased, as have the event types. Also, integrating results from different extraction systems that apply different algorithms has become possible. Clearly, the number of events displayed on one layer has now got to a level were a bit of layered rationalisation is needed. However, before talking about layered rationalisation it is important to consider an increasingly important aspect of this visualisation namely Icon Policy.

6.1 Icon Policy

The first icon policy that we introduced was a simple one where we displayed the same flag icon located where the event was reported and its colour depended on the severity of the event: black if the event reported people killed, red if it reported injuries, and green for all others This was in line with the typical representation on Google (and on other mapping systems) of push pins with different colours. We quickly came to realise that this approach was not particularly useful in the context of events. At this point in time there were only violent and natural disaster events where the severity was strongly linked to the event type. Hence, we came up with a set of icons based on traffic symbol principles, where the general shape of the border of the icon has some Event Type class significance. Then, the actual event type is depicted as an icon in the centre. On a map view that already has a lot of complex patterns in terms of features, this approach calls the fast and intuitive identification of the event typology and to a certain extent consequence at any given location.

The diagram in Fig. 11 shows the icon policy aligned with the EUROSUR Pilot Project which is one of the latest evolutions of our icon policy to be implemented. The specific policy used here follows the following logic:

- Events related to border security appear with a circular form exploiting a popular road 'no entry' sign for events related to illegal migration and a circular restricted sign for related cross-border crime. The exceptions to this policy are facilitator arrests and deportations which appear in an inverted triangle since they are also pertinent to crisis events (e.g., arrest, trial, . . .).

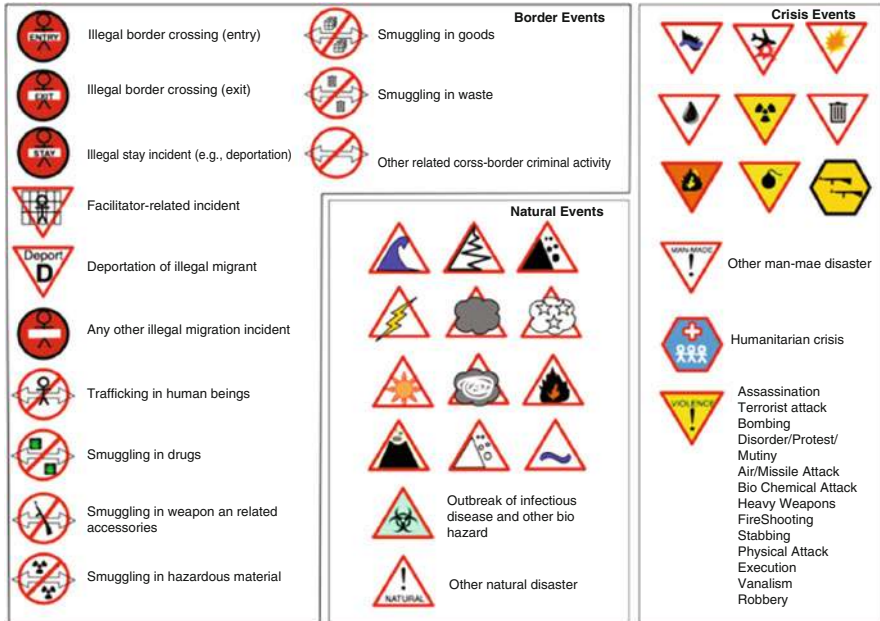[11]KML – http://www.opengeospatial.org/standards/kml.

Fig. 11 Event icon policy inline with EUROSUR requirements

- Natural disasters appear inside a normal triangle, with a pertinent pictogram of the natural phenomenon.
- Other crisis events that encapsulate the remaining event types (e.g., man-made disasters and violent events) appear in an inverted triangles. The exception to this are armed conflicts and humanitarian crisis that appear in a hexagon, to separate these particular subtypes (they are consequences of the other smaller-scale events). Like road traffic signs that have subtle differences in different countries the general notion of the pictogram in the centre has a universal language and possibly cultural neutrality.

6.2 Layer Rationalization

As stated earlier, layer rationalisation is necessary, especially in a system extracting events in multiple languages from news, where large stories proliferate across languages and hence so do the events that get extracted. The main remit of layer rationalisation was therefore to provide layers that have the potential to provide the user with the possibility to see quickly events of certain classes together based on language. Another factor influencing layer rationalisation is the frequency of event identification. Some types of events are generated rather infrequently over a 24 h

time span, whereas others are more prolific. Therefore, it was decided to adopt the following layered breakdown:

- The top layers would consist of the following general event types: border events, crisis events, health threats (medical events, e.g., outbreaks of infectious diseases), and moderated events.
- Each top layer would be decomposed into temporal layers depending on the event frequency of reporting like: last 24 h, last 7 days and last month.
- Each temporal layer would be divided into source language of the news article from which the event was extracted.
- Finally, layers would be created to provide background information, e.g., layers showing news articles filtered using EMM classification system that were on a particular topic (in case events were missed by the event extraction system).

With the constrains of the layered approach the services that produce the information were modified in order to accomplish the required decomposition. Clearly, these services were designed to accommodate any rationalisation approach requiring time, event type and language as parameters.

The screenshot in Fig. 12 shows the end result of this approach, where the layer organisation can be seen on the left hand side of the image, decomposed by hierarchy and switched (ticked) on or off. Then, on the right hand side of the pane it's possible to see the events from layers that are switched on. Clicking on the icon

Fig. 12 Layer rationalisation depicted in Google Earth with some illegal migration, smuggling and crisis events

shows the event detail in terms of title and description of the article that triggered the event, under this the values of the extracted information are shown.

7   Event Moderation

Event moderation provides the fundamental link from structured information generated by an automated extraction system to further in-depth situation analysis. In order to accurately spot trends in event proliferation, geo spatially and temporally, trends in event typology and impact it is necessary to have an information store with events that are validated. This validation can only be completed by a human. The main objective, therefore, of the Event Moderation System is to empower a human operator to achieve information validation of resulting information flowing out of the automated event extraction systems. In order to achieve this objective the event moderation system must be a physical as well as logical bridge between the automated extraction systems and the information store used for analysis.

Our choice in implementation of this part of the system was to develop a server based moderation service system and a Rich Internet Client for end user interaction. This allows for Internet access to the event moderation interface without the need to install dedicated software. The server provides common REST-ful services over HTTP. An example of the client user interface is given in Fig. 13.
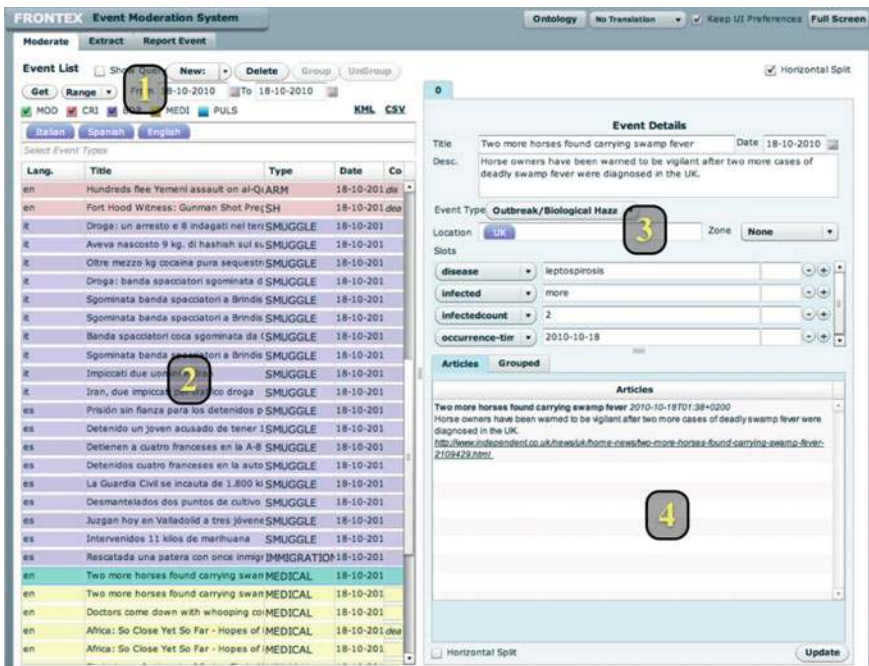

Fig. 13 Main event moderation user interface

Fig. 14 Zoom in of access and filtering part of the event moderation system's user interface

The following sections outline a number of essential services deemed necessary for completing the event moderation-related tasks together with corresponding examples of user interface components that were deployed.

7.1 Fundamental Moderation Tasks

In this section, we consider the fundamental tasks of moderation decomposed into event retrieval, evaluation, moderation and then storage. These tasks can be supported by a number of discrete services like filtering, moderating and event grouping.

Access and event filtering is necessary to reduce the number of events being processed and to focus on a particular event type or time period. The filtering service also has a secondary role of providing dedicated layers to other visualisation tools. In our instance of this system the filtering mechanism allows the selection of events based on automatically extracted meta-data like: date of event occurrence, type of event, system that generated the event, language of article from which the event was extracted. An example zoom of a user interface that provides an intuitive access to the filtering service is shown in Fig. 14. An important feature to avoid the generation of duplicate events is to allow access to previously moderated events, both in the retrieval process and in the view filtering. In our user interface access and filtering is achieved through a unique service and user interface.

The moderation service is very simple, it allows new events previously unmoderated to be inserted as moderated events, update of previously moderated events and delete of a moderated event. An example of a moderation interface is shown in Fig. 13. The numbers in the figure represent: (1) The access to event filtering and selection, (2) A tabular overview of the events that have been selected (each row corresponds to an event) – pre-selected meta-data values are shown, (3) The selection of row allows the detail of the event to be displayed the right hand side of the window, and (4) Displays links to the article(s) from which the event description was extracted. Links back to the original articles are maintained. All the information

of the event can be moderated by the user including title, description, event type and subtype, location (see Sect. 7.3), and more general event descriptors (see Sect. 7.4).

The event grouping service allows similar events or sub events of a main event to be grouped together logically. Grouping typically is needed when events or the resulting sub events have a relatively long time duration like earthquakes and armed conflicts, where after the main event there is substantial reporting of sub events.

7.2 Client-side Translation

It is not assumed that an operator of the moderation system is able to understand all relevant languages. Thanks to the rise in performance of statistical machine translation it is now possible to provide inline dynamic translation of the main title and description of event information. The Event Moderation System uses the Google AJAX API[12] and uses the user interface client to call directly the Google Translation Systems, without passing through our services. Given the language information that EMM associates with articles it is possible to translate all event descriptions into a selected language irrespective of the source language.

For clients that are sensitive to the use of Google services, there has been good progress towards building independent capacity of translation services [33], the main obstacles in this area are the lack of parallel corpora for certain languages. None, the less good open source systems are available [34].

7.3 Gazetteer and Event Mapping

Events currently get assigned the geo-tag of the article assigned by the geo-tagging process of EMM. To support the use in validating this value a user interface component to the gazetteer service in the form of a type ahead field is deployed. An example of this component is shown in Fig. 15.

A two stage gazetteer service is deployed by our system in order to support the precise validation of event location. In stage one, the interface is waiting for user input as soon as the user has input some letters and the component detects a pause, it sends a query to the gazetteer service. The names supplied in return for this initial query are taken from different administrative levels in the Gazetteer like country names, region names, province names and then place names that match part of the typed input. This query returns from the server and the results are now stored in the interface component that now enters stage two. Now, in stage two, as the user continues to type the component shows matches to the stored names and displays
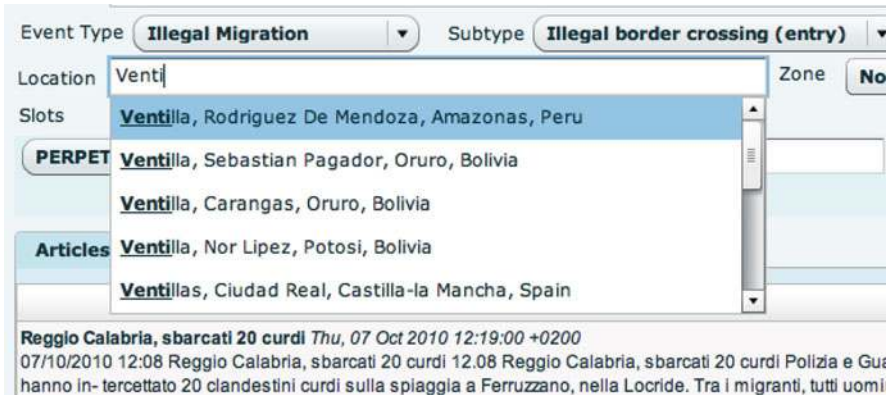
---

[12]http://code.google.com/apis/ajaxlanguage/.

Fig. 15 User interface component linked to multi-lingual gazetteer

possible candidates, the user can select a candidate or continue typing where if no names are matched then the component returns to stage one and makes a new request to get matching names from the server. Note that once a location is selected then the administrative path of the selection is shown in the drop down list in Fig. 15.

7.4 Dynamic Ontology

The purpose of the ontology in the Event Moderation System is to allow the users to define how they wish events generated by the event extraction to be interpreted and displayed. Interpretation here is currently implemented at the basic level of inheritance for the purpose of event classification.

Additionally, the ontology is used as a means to improve the user experience of the system by driving a number of aspects of the user interface in order to fit better the context of event generation to the expected use of the events from the moderations store. It acts as a gap shortener in this sense, for instance, by providing mapping of native event types to event type hierarchy of the required moderation system (target application). Another aspect of the configuration is that it maps the native description of the slot values to a better fitting nomenclature of the moderation system.

A side effect of the use of the ontology here is also to configure, for example, which languages the moderation tool should provide support for.

Finally, the ontology is used by services that deliver events in terms of KML descriptions in order to enforce a common icon policy for event visualization.

The ontology is described using a basic XML format, also a dedicated super-user interface has been developed in order to facilitate the editing in real time of the Ontology (such that modifications in it's structure and content are taken into account in real time).

Fig. 16 User interface supporting manual event reporting

7.5 Manual Event Entry

The recall of any automated event extraction system can't be expected to be 100 percent accurate. There will be some cases where events might not be reported immediately in the printed/online news but by other news sources or means. Hence, it's important to provide a service that allows manual event entry, for which the user interface is shown in Fig. 16. This is a very simple service that allows users via a simple web interface to enter new events that are not related to any particular article or RSS Feed. The web interface follows the same format and template layout as the other event detail entry screens in the Event Moderation System.

Once entered into the interface the event can be accessed via the main system moderation interface, it can be recalled by queries like any other moderated event and can be visualised through the visualisation services.

7.6 Assisted Event Entry Using Event Extraction

Again, in order to facilitate user entry of events perhaps existing in open sources not covered by EMM/FMM, a further set of services have been implemented to assist event entry using the automated extraction system. The main purpose of this service is to allow a user to select an open source, push it through an automated event extraction system, view and moderate the events that are extracted, and then finally store them in the moderation store.

Fig. 17 Assisted event entry user interface showing extraction service results

An example of the visualisation of this service is given in Fig. 17. This particular instance provides the opportunity to select any RSS feed (1) and to download this RSS into a visualisation pane (2). Here, the RSS is checked for ISO language codes (REF). If the language is incorrect the translation service can be called to translate the articles into a language on which the event extraction system can operate or if the language is valid the operator can just select to have the codes added. Next, the RSS is then sent to a dedicated event extraction service, which in turn, returns the event descriptions that it extracts from the RSS. The results can be seen in the right side of the window (3). Finally, a selected event can be moderated and saved to the moderation store (4).

8 Conclusions and Future Work

In this chapter, we have reported on the development of a real-time multilingual event extraction framework for automatically capturing structured knowledge on events related to border security from online news. This framework was created for Frontex to reduce the overload of online news information, and to particularly focus on detection of illegal migration incidents, cross-border crimes, and related

crisis situations at the EU external borders and in third countries. It is fully operational 24/7 and consists of a core event extraction system, a geo-browser for visualisation of the events on a map, and an event moderation component, which bridges the gap between automatically extracted event descriptions and in-depth analysis. The core event extraction system combines two event extraction engines, namely, *NEXUS*, developed at the JRC of the EC and *PULS*, developed at the University of Helsinki. The rationale behind deployment of two different event extraction engines was not only to ensure better coverage, but also to provide a framework to explore and compare two different underlying approaches to event extraction. The core integrated event extraction system is applied to the stream of news articles gathered by the EMM – a large-scale multilingual news aggregation engine, and FMM, a Frontex-tailored version thereof, both developed at the JRC, and is capable of efficiently extracting events from news in English, Italian, French, Spanish, Portuguese, Russian, and Arabic, with a 10-min delay. The range of events the system already detects is relatively wide, however, for some of the aforementioned languages the process of adapting the system to cover all events related to border security is still work in progress. Once information is captured we have shown possibilities for its organisation and visualisation as well as the necessary tools and standards that are required in order to support the needs for a fully operational application for analysts and intelligence officers. An evaluation of the system performance in terms of its' coverage, precision and usability for the end user is currently being carried out. We have presented some of the results of this evaluation, which shows satisfactory performance and strong application potential of the automated event extraction for real-time situation monitoring in the domain of border security. In particular, in 2011 the event extraction framework will be integrated with the EUROSUR pilot information exchange network to facilitate the task of populating by Frontex of the incident information layer.

The current work focuses mainly on customizing the systems to cover all relevant event types in all relevant languages, with particular focus on French, Arabic, Russian, and Ukrainian. Customization for Turkish and Greek is envisaged too. Although, the system performance in case of detecting classical crisis-related events (e.g., man-made and natural disasters, violent events, etc.) is satisfactory, extracting information on the more specific border-security events, including illegal migration incidents and related cross-border crimes poses additional challenges. Firstly, information about the target incidents is often only implicit in text. Secondly, such incidents are often reported in local news only, in one language, and relevant information is scattered over the whole article. Finally, automated geo-locating of such events is more difficult since there are usually several geo-references in an article, e.g., in case of illegal migration incidents, there are geo-references to: event place, country of departure, destination country/place, places visited on the illegal migration route, countries of origins of the victims and perpetrators, origin of the means used by some authority involved, etc. All the aforementioned challenges require deployment of slightly more sophisticated linguistic processing of the news articles in this domain. We are currently exploring the ways how to tackle these problems.

There are other challenges, not particularly related to the border security domain, to be addressed in the future, which include: automated detection of duplicate events, linking of event sequences, sub-event recognition, automatic ranking of events in terms of their relevance for the end users (partially provided). Furthermore, in order to provide more accurate slot values of automatically generated event descriptions, we are evaluating the usefulness of cross-lingual information fusion. Since it has been acknowledged that security-relevant information can also be found and/or discovered from electronic social media (e.g., as reported in the work on the Dark Web archive [35], which consists of a huge collection of postings gathered from deep web fora used by extremist and terrorist groups), we are also exploring the possibility to extend the coverage of the system through extracting information from blogs, not only to detect events of certain types, but also to detect sentiment on certain border-security related topics in third countries. Finally, we plan to carry out a more thorough evaluation of the usability of the event extraction tools for the end-user community in Frontex and Member States, with a particular focus on the evaluation of the event moderation tool, and identification of future extensions of its' functionalities, e.g., analytical ones.

References

1. Tanev, H., Piskorski, J., Atkinson, M.: Real-Time News Event Extraction for Global Crisis Monitoring. In: Proceedings of the 13*th* International Conference on Applications of Natural Language to Information Systems (NLDB 2008). Lecture Notes in Computer Science, vol. 5039, pp. 207–218. Springer, Berlin (2008)
2. Piskorski, J., Tanev, H., Atkinson, M., Van der Goot, E.: Cluster-centric approach to news event extraction. In: Proceedings of the International Conference on Multimedia & Network Information Systems, Wroclaw, Poland, IOS Press (2009)
3. Grishman, R., Huttunen, S., Yangarber, R.:
   Information extraction for enhanced access to disease outbreak reports. J. Biomed. Inform. **35**(4) (2003)
4. Yangarber, R., Jokipii, L., Rauramo, A., Huttunen, S.: Extracting information about outbreaks of infectious epidemics. In: Proceedings of the HLT-EMNLP 2005, Vancouver, Canada (2005)
5. Appelt, D.: Introduction to Information Extraction Technology, Tutorial held at IJCAI-99, Stockholm, Sweden (1999)
6. Llytinen, S., Gershman., A.: ATRANS: Automatic processing of money transfer messages. In: Proceedings of the 5th National Conference of the American Association for Artificial Intelligence, IEEE Computer Society Press (1986)
7. Kehler, A.: Learning embedded discourse mechanisms for information extraction. In: Proceedings of AAAI Spring Symposium on Applying Machine Learning to Discourse Processing (1998)

8. Yangarber, R., Grishman, R., Tapanainen, P., Huttunen, S.: Unsupervised discovery of scenario-level patterns for information extraction. In: Proceedings of ANLP-NAACL 2000, Seattle, USA, (2000)

9. Grishman, R., Huttunen, S., Yangarber, R.: Real-time event extraction for infectious disease outbreaks. In: Proceedings of Human Language Technology Conference 2002, San Diego, USA (2002)

10. Yangarber, R., Etter, P.V., Steinberger, R.: Content collection and analysis in the domain of epidemiology. In: Proceedings of DrMED 2008: International Workshop on Describing Medical Web Resources at MIE 2008: the 21st International Congress of the European Federation for Medical Informatics 2008, Goeteborg, Sweden (2008)

11. King, G., Lowe, W.: An automated information extraction tool for international conflict data with performance as good as human coders: A rare events evaluation design. Int. Organ. **57**, 617–642 (2003)

12. Ashish, N., Appelt, D., Freitag, D., Zelenko, D.: In: Proceedings of the Workshop on Event Extraction and Synthesis, Held in conjunction with the AAAI 2006 conference, Menlo Park, California, USA (2006)

13. Proceedings of the 3rd Conference on Message Understanding, MUC 1991, San Diego, California, USA, May 21-23, 1991. ACL 1991, ISBN 1-55860-236-4

14. Proceedings of the 4th Conference on Message Understanding, MUC 1992, McLean, Virginia, USA, June 16-18, 1992. 1992, ISBN 1-55860-273-9

15. Steinberger, R., Fuart, F., van der Goot, E., Best, C., Etter, P.V., Yangarber, R.: Text Mining from the Web for Medical Intelligence. In: Fogelman-Soulie, F., Perrotta, D., Piskorski, J., Steinberger, R., (eds.) Mining Massive Data Sets for Security. IOS Press, Amsterdam, the Netherlands (2008)

16. Freifeld, C., Mandl, K., Reis, B., Brownstein, J.: HealthMap: Global infectious disease monitoring through automated classification and visualization of internet media reports. J. Am. Med. Inform. Assoc. **15**(1) (2008)

17. Doan, S., Hung-Ngo, Q., Kawazoe, A., Collier, N.: Global health monitor–a web-based system for detecting and mapping infectious diseases. In: Proceedings of the International Joint Conference on NLP (IJCNLP). (2008)

18. Naughton, M., Kushmerick, N., Carthy, J.: Event extraction from heterogeneous news sources. In: Proceedings of the AAAI 2006 workshop on Event Extraction and Synthesis, Menlo Park, California, USA (2006)

19. Wagner, E., Liu, J., Birnbaum, L., Forbus, K., Baker, J.: Using explicit semantic models to track situations across news articles. In: Proceedings of the AAAI 2006 workshop on Event Extraction and Synthesis, Menlo Park, California, USA (2006)

20. Atkinson, M., van der Goot, E.: Near real time information mining in multilingual news. In: Proceedings of the 18th World Wide Web Conference. (2009)

21. Pouliquen, B., Tanev, H., Atkinson, M.: Extracting and learning social networks out of multilingual news. In: Learning Social Networks out of Multilingual News. Proceedings of the social networks and application tools workshop (SocNet-08), Skalica, Slovakia, 19–21 Sept 2008

22. Kabadjov, M., Atkinson, M., Steinberger, J., Steinberger, R., Van der Goot, E.: NewsGist: A Multilingual Statistical News Summarizer. In: Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD'2010). (2010)

23. Balahur, A., Steinberger, R., Van der Goot, E., Pouliquen, B., Kabadjov, M.: Opinion Mining on Newspaper Quotations. In: Proceedings of the workshop 'Intelligent Analysis and Processing of Web News Content' (IAPWNC), held at the 2009 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, Milano, Italy. (2009)

24. Lyons, R.: Understanding Digital Signal Processing, Second Edition. ISBN-13: 978-0-13-108989-1, Prentice Hall (2004)

25. Piskorski, J.: ExPRESS – Extraction pattern recognition engine and specification suite. In: Proceedings of the 6th International Workshop Finite-State Methods and Natural language Processing 2007 (FSMNLP 2007), Potsdam, Germany (2007)

26. Piskorski, J., Tanev, H., Atkinson, M., van der Goot, E.: Online News Event Extraction for Global Crisis Surveillance. Transactions on Computational Collective Intelligence, Springer.
27. Zavarella, V., Tanev, H., Piskorski, J.: Event extraction for Italian using a cascade of finite-state grammars. In: Proceedings of the $7th$ International Workshop on Finite-State Machines and Natural Language Processing. (2008)
28. Tanev, H., Zavarella, V., Linge, J., Kabadjov, M., Piskorski, J., Atkinson, M., Steinberger, R.: Exploiting Machine Learning Techniques to Build an Event Extraction System for Portuguese and Spanish. Linguamática: Revista para o Processamento Automático das Línguas Ibéricas **2**, 550–566 (2009)
29. Tanev, H., Oezden, P.: Learning to populate an ontology of violent events. In: Fogelman-Soulie, F., Perrotta, D., Piskorski, J., Steinberger, R., (eds.) Mining Massive Data Sets for Security. IOS Press, Amsterdam, the Netherlands (2008)
30. Pinheiro, V., Furtado, V., Pequeno, T., Nogueira, D.: Natrual language processing based on semantic inferentialism for extracting crime information from text. In: Proceedings of 2010 IEEE Conference on Intelligence and Security Informatics, Vancouver, Canada (2010)
31. Yangarber, R.: Counter-training in discovery of semantic patterns. In: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, Sapporo, Japan, July 2003
32. Lin, W., Yangarber, R., Grishman, R.: Bootstrapped learning of semantic classes from positive and negative examples. In: Proceedings of the ICML Workshop on The Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining, Washington, DC, Aug 2003
33. Turchi, M., Flaounas, I., Ali, O., De Bie, T., Snowsill, T., Cristianini, N.: Found in translation. In: Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases, Springer (ECML/PKDD 2009) (2000)
34. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Zens, R., Constantin, A., Federico, M., Bertoldi, N., Dyer, C., Cowan, B., Shen, W., Moran, C., Bojar, O., Herbst, E.: Moses: open source toolkit for statistical machine translation. In: Proceedings of ACL 2007, Demonstration Session, pp. 177–180. (2007)
35. Zhang, Y., Zeng, S., Huang, C.N., Fan, L., Yu, X., Dang, Y., Larson, C., Denning, D., Roberts, N., Chen, H.: Developing a dark web collection and infrastructure for computational and social sciences. In: Proceedings of ISI 2010, Vancouver, Canada (2010)

Mining the Web to Monitor the Political Consensus

Federico Neri, Carlo Aliprandi, and Furio Camillo

Abstract Communication is becoming more and more crucial in the competitive political arena: politicians can monitor electors' suggestions or claims, or the perception they might have about leaders' statements, by analyzing blogs, newsgroups and newspapers. They try to take account of the complexity of public views in order to design populist measures and increase dramatically their consensus. The Web sources are more accessible, ubiquitous, and valuable than ever before. But the most valuable information is often hidden and encoded in blog posts or pages, which are often neither structured, nor classified, being free textual. The process of accessing all these raw data, heterogeneous in terms of source and lexicon, and transforming them into information is therefore strongly linked to automatic textual analysis and conceptual synthesis. This paper describes a Sentiment Mining study performed on over 1,000 news articles or forum/blog posts, concerning the Italian Prime Minister Silvio Berlusconi, involved in the escorts' scandal.

1 Introduction

The Web is a huge virtual space where to express and share individual opinions, influencing any aspect of life, with implications for companies and political parties alike. Nowadays communication is becoming more and more crucial in the competitive political arena: politicians and their consultants can monitor electors' suggestions or claims, or the perception they might have about leaders' statements and decisions, by analyzing blogs, newsgroups, and newspapers in real time. Spin

F. Neri (✉) · C. Aliprandi
Synthema Language and Semantic Intelligence, Pisa, Italy
e-mail: federico.neri@synthema.it; carlo.aliprandi@synthema.it

F. Camillo
Department of Statistical Sciences, Faculty of Economics, University of Bologna, Bologna, Italy
e-mail: furio.camillo@unibo.it

doctors take account of the complexity of public views in order to design populist measures and increase dramatically their leaders' consensus.

The revolution in information technology is making open sources more accessible, ubiquitous, and valuable, making Open Source Intelligence and Web Sentiment Analysis available at less cost than ever before. The world today is really in the midst of an information explosion. Anyway, the availability of a huge amount of data in Internet and in all the open sources information channels has lead to the well-identified modern paradox: an overload of information means, most of the time, a no usable knowledge. In fact, all the electronic texts are – and will be – written in various native languages, but these documents are relevant even for non-native speakers. The most valuable information is often hidden and encoded in pages which for their nature are neither structured, nor classified. Nowadays everyone experiences a mounting frustration in the attempt of finding the information of interest, wading through thousands of pieces of data. The process of accessing all these raw data, heterogeneous both for source, type, protocol, and language used, transforming them into information, is therefore inextricably linked to the concepts of automatic textual analysis and synthesis, hinging greatly on the ability to master the problems of multi-linguality.

Despite much progress in Natural Language Processing (NLP), the field is still a long way from a full Natural Language Understanding (NLU). In fact, understanding requires processing and knowledge that goes beyond parsing and lexical lookup and that is not explicitly conveyed by linguistic elements. Contextual understanding is needed to deal with the omissions. Ambiguities are a common aspect of human communication. Speakers are cooperative in filling gaps and correcting errors, but automatic systems generally are not. A mixed qualitative and quantitative approach can bridge this gap. This paper describes a Sentiment Mining study performed on over 1,000 blog posts or news articles about the Italian Prime Minister Silvio Berlusconi, by using a content enabling system – iSyn Semantic Center – that provides deep semantic information access and dynamic classification features for large quantities of distributed multimedia data. This Knowledge Mining platform, which has been adopted by some security sector-related government institutions and agencies in Italy to limit information overload in OSINT, has been designed and developed by Synthema together with some operative officers of the Information & Security Department of Italian State Defense. Synthema, in fact, has been supporting Intelligence operative structures in Italy both on technological and on substantive content matter issues, in order to provide hands-on expertise on Open Source Intelligence operations at strategic and tactical levels. This system uses a two steps process. First, the lexical analysis identifies the relevant knowledge from the posts, by detecting semantic relations and facts in texts. The automatic linguistic analysis is based on Morphological, Syntactic, Semantic Role, Semantic, Semiometric, and Statistical criteria. Second, the classification of documents uses Supervised and Unsupervised Clustering criteria.

1.1 State-of-the-Art of Semantic Information Systems

Current-generation information retrieval (IR) systems excel with respect to scale and robustness. However, if it comes to deep analysis and precision, they lack power. Users are limited by keywords search, which is not sufficient if answers to complex problems are sought. This becomes more acute when knowledge and information are needed from diverse linguistic and cultural backgrounds, so that both problems and answers are necessarily more complex. Developments in the IR have mostly been restricted to improvements in link and click analysis or smart query expansion or profiling, rather than focused on a deeper analysis of text and the building of smarter indexes. Traditionally, Text and Data mining systems can be seen as specialized systems able to convert more complex information into a structured database, allowing people to find knowledge rather than information. For some domains, Text Mining applications are well-advanced; for example, in the domains of medicine, military and intelligence, and aeronautics [1]. In addition to domain-specific miners, general technology has been developed to detect Named Entities [2], co-reference relations, geographical data [3] and time points [4].

The field of knowledge acquisition is growing rapidly with many enabling technologies being developed that eventually will approach NLU. However NLU and semantic interpretation in particular require the identification of every individual conceptual component and the semantic roles it play. In addition, understanding requires processing and knowledge that goes beyond parsing and lexical lookup and that is not explicitly conveyed by linguistic elements. First, contextual understanding is needed to deal with the omissions. Ambiguities are a common aspect of human communication. Speakers are cooperative in filling gaps and correcting errors, but automatic systems are not. Second, lexical knowledge does not provide background or world knowledge, which is often required for non-trivial inferences. Any automatic system trying to understand a simple sentence will require – among others – accurate capabilities for Named Entity Recognition and Classification (NERC), full Syntactic Parsing, Word Sense Disambiguation (WSD), and Semantic Role Labelling (SRL) [5].

Current baseline information systems are either large-scale, robust but shallow (standard IR systems), or they are small-scale, deep but ad hoc (Semantic-Web ontology-based systems). Furthermore, these systems are maintained by experts in IR, ontologies or language-technology and not by subject matter expertise.

Finally, hardly any of the systems is multilingual, yet alone cross-lingual and definitely not cross-cultural.

In Table 1, we summarize some key features of different state-of-the-art semantic information systems, as recently surveyed in the framework of the EU – FP7 ICT co-funded KYOTO project [6], thus comparing ad-hoc Semantic web solutions, Wordnet-based information systems and tradition IR with iSyn Semantic Center.

Table 1 Comparison of semantic information systems

Key features	Semantic web	Wordnet based	IR	iSyn semantic center
Large scale and multiple domains	No	Yes	Yes	Yes
Deep semantics	Yes	No	No	Yes
Multi-lingual	No	Yes	Yes	Yes
Cross-lingual	No	Yes	No	Yes
Data and fact mining	Yes	No	No	Yes

1.2   State-of-the-Art of Automatic Translation Systems

Automatic Translation (AT) includes a wide range of technologies, that take the analysis of human language as focus. Human language plays a fundamental role in sentiment analysis, since it is the medium through which communication takes place. In the business field, English has become the dominant language within global organizations and it is the de-facto answer to the language barrier problem. Anyway, in general, relevant documents can be written in various native languages, resulting of interest even to non-native speakers.

Although relatively old, this discipline is very lively and constantly evolving. AT was one of the first arenas for computational linguistics in the mid-1950s. The dream of the universal, global machine translation able to overcome all problems in human communication clashed soon with reality: language revealed itself to be an object far too complex to be described with a finite set of rules and heuristics, a real bottleneck for any real-world application. As a result, large-scale funding of AT research came to a virtual standstill for the next 20 years. Nevertheless, now there's a revival of interest for AT: the need for translation continues to increase significantly due to the Web advent. In literature AT systems are classified according to the level of automation of translation as opposed to the need for post-revision and correction. These approaches are usually classified as follows [7–10]:

- Machine translation (MT)
- Computer aided translation (CAT)

1.2.1   Machine Translation

In pure MT systems, the machine output produced is accepted as it is, without the need of any revision. Despite recent improvements, translation technology is not very accurate when used as a standalone solution. In MT systems, three main approaches are usually followed: Direct, Transfer and Interlingua. In the Direct model, a "primitive form of transfer" is adopted, since simply word-for-word replacement is foreseen. In the actual Transfer model, the system has to provide syntactically correct target language text by transforming source-language representations into suitable target-language representations. Both the Transfer and the

Interlingua approaches require linking rules to map the surface text to some form of internal representations [11]. In case of Transfer based MT systems, however, this internal representation is assumed to vary widely from language to language: this determines that for each source-language/target-language pair specific sets of rules have to be created. The Interlingua approach is based on the idea that the analysis of the source-language text is somehow independent from the source-language. The target-language text is then generated from a neutral, inter-lingual representation, which works as a sort of pivot among different languages. Moreover, different paradigms of MT research may be identified [11–14], those that:

- Rely on linguistic techniques: for these approaches, MT is grounded on principles of linguistic theory [15]. Constraints at syntax, lexical and semantic level are used to identify the target language equivalent.
- Rely only on statistics without recurring to linguistic knowledge: these approaches are enabled by the availability of great computational power and huge corpora which are used for training and for storing examples of translation.
- Use a combination of the two approaches, adopting hybrid methods.

1.2.2 Computer Aided Translation

When not only the level of text comprehension, but also the grammar and stylist levels must be accurate, CAT tools are used. They ensure the same level of consistency, accuracy, and control for repetitive tasks. CAT tools add a new dimension to the translation process by providing useful Terminology Management (TM), Dictionary enhancements and Memory Exploring features for all the phases, from pre-production – needed to tune the translation environment and determine the terminology to use – to the final linguistic review. TM supports the translator in defining the proper terminology by searching existing dictionaries and performing contextual search in source material and bilingual reference Memories. The final result of TM can also be integrated in MT systems. Memory Explorer tools are designed to help translators re-use previously translated bilingual material.

2 iSyn Semantic Center, the Knowledge Mining Platform

The system used in this study is built on the following components:

1. A Crawler, an adaptive and selective component that gathers documents from Internet/Intranet or Database sources.
2. A Semantic Engine, which identifies relevant knowledge in the texts, by detecting semantic relations and facts.
3. A Search Engine that enables Natural Language, Semantic and Semantic-Role queries.
4. A Machine Translation Engine, which enables AT of search results.

5. A Georeferentiation Engine, which enables an interactive geographical representation of documents.
6. A Classification Engine which classifies search results into clusters and subclusters recursively, highlighting meaningful relationships among them, or assigns documents to predefined thematic groups.

2.1 The Crawler

The crawler is a multimedia content gathering and indexing system, whose main goal is managing huge collections of data coming from different and geographically distributed information sources. It provides a very flexible and high performance dynamic indexing for contents retrieval. Its gathering activities are not limited to the standard Web, but also operate with other type of sources like remote databases by ODBC protocol, other Web sources (HTTPS-FTP-Gopher), Usenet news (NNTP), WebDav and SMB shares, mailboxes (POP3-POP3/S-IMAP-IMAP/S), file systems and other proprietary source protocols. The crawler provides default plug-ins to extract text from most common types of documents. Even more complex sources, such as audio and/or video files, might be suitably processed so as to extract a textual-based labeling, based on both the recognition of speeches, videos and images [7, 16].

2.1.1 Focused Crawling

Focused crawling [17] aims to crawl only the subset of the Web pages related to a specific category of interest. The major problem in focused crawling is performing the appropriate credit assignment to different documents along the crawling path, such that short-term gains are not pursued at the expense of less-obvious crawling paths, that ultimately yield larger sets of valuable pages. To address this problem the focused crawling algorithm builds a context model within which topically relevant pages occur on the Web. This algorithm shows significant performance improvements in crawling efficiency over standard focused crawling. In fact, the credit assignment can be significantly improved by equipping the crawler with the capability of modelling the context within which the topical materials is usually found on the Web. Such a context model has to capture typical link hierarchies within which valuable pages occur, as well as describe off-topic content that co-occurs in documents that are frequently closely associated with relevant pages. The general framework and the specific implementation of such a context model are called Context Graph. The Context Focused Crawler (CFC) uses the limited capability of search engines – like Google – to allow users to query for pages linking to a specified document. This data can be used to construct a representation of pages that occur within a certain link distance (defined as the minimum number of link necessary to move from one page to another) of the target documents. This

representation is used to train a set of classifiers, which are optimized to detect and assign documents to different categories based on the expected link distance from the document to the target document. During the crawling stage the classifiers are used to predict how many steps far from a target document the current retrieved document is likely to be. This information is then used to optimize the search. There are two distinct stages to using the algorithm when performing a focused crawl session:

- An initialization phase when a set of context graphs and associated classifiers are constructed for each of the seed documents.
- A crawling phase that uses the classifiers to guide the search, and performs online updating of the context graphs.

Counting on a rapid and efficient initialization phase, Focused Crawling is suitable for real-time services.

2.2   The Semantic Engine

This component identifies the relevant knowledge from the whole raw text, by detecting semantic relations and facts in texts. Concept extraction is applied through a pipeline of linguistic and semantic processors that share a common knowledge. The shared knowledge base guarantees a uniform interpretation layer for the diverse information from different sources and languages. The extracted knowledge and information will be indexed by the crawler, that can handle fast semantic search.

2.2.1   Lexical and Semantic Resources: Specific Domain Dictionaries and Knowledge Bases Creation and Updating

Generally speaking, the manual construction and maintenance of multilingual language resources is undoubtedly expensive, requiring remarkable efforts. The growing availability of comparable and parallel corpora have pushed Synthema to develop specific methods for semi-automatic updating of lexical resources. They are based on NLU and Machine Learning. These techniques detect multilingual lexicons from such corpora, by extracting all the meaningful term or phrases that express the same meaning in comparable documents. These objects enrich existing multilingual dictionaries and may constitute the basic lexical units for any Knowledge Base, helping on overcoming linguistic barrier. As an example, let's consider a corpus made of parallel documents written in English and in Italian, used as training set for the topic of interest. This case is quite straightforward, due to the fact that everyone normally use English language as reference. The major problem consists in the different syntactic structure and words definition these two languages may have. So a direct phrasal alignment is often needed. The following bilingual morphological analysis recognises as relevant terminology

only those terms or phrases, that exceed a threshold of significance. A specific algorithm associates an Information Quotient to each detected term and ranks it on its importance. The Information Quotient is calculated taking in account the term, it's Part Of Speech tag, its relative and absolute frequency, its distribution on documents. This morphological analysis detects significant Simple Word Terms (SWT) and Multi Word Terms (MWT), annotating their headwords, their relative and absolute positions. Synthema strategy on multilingual dictionary construction consists in the assumption that, having taken in account a specific term S and its phrasal occurrences, its translation T can be automatically detected by analysing the correspondent translated sentences. Thus, semi-automatic lexicon extraction and storage of multilingual relevant descriptors become possible. Each multilingual dictionary, specifically suited for the cross-lingual mapping, is bi-directional and contains multiple coupled terms f(S,T), stored as Translation Memories. Each lemma is referenced to syntax or domain dependent translated terms, so that each entry can represent multiple senses. Besides, the multilingual dictionaries contain lemmas together with simple binary features, as well as sophisticated tree-to-tree translation models, which map – node by node – whole sub-trees [18].

2.2.2 Lexical and Semantic Analysis

The automatic linguistic analysis of the textual documents is based on Morpho-Syntactic, Semantic, Semantic Role and Statistical criteria. At the heart of the lexical system is the McCord's theory of Slot Grammar [19]. The system analyzes each sentence, cycling through all its possible constructions. It tries to assign the context-appropriate meaning – the sense – to each word by establishing its context. Each slot structure can be partially or fully instantiated and it can be filled with representations from one or more statements to incrementally build the meaning of a statement. This includes most of the treatment of coordination, which uses a method of 'factoring out' unfilled slots from elliptical coordinated phrases. The parser – a bottom-up chart parser – employs a parse evaluation scheme used for pruning away unlikely analyses during parsing, as well as for ranking final analyses. It builds the syntactical tree incrementally. By including the semantic information directly in the dependency grammar structures, the system relies on the lexical semantic information combined with Semantic Role relations. The WSD algorithm considers also possible super-subordinate related concepts in order to find common senses in lemmas being analyzed.

The anaphora resolution is based on the detection of pronominal references, on the interpretation of specific named entities in the text, as well as assertions about them. Antecedent words for which co-reference is morphologically or syntactically excluded are automatically filtered out [20]. A salience weight is associated with each anaphora candidate, so that empirically and linguistically motivated heuristics can select the most appropriate referent in the list of possible interpretations.

Beside Named Entities, locations, time-points, etc, it detects relevant information like noun phrases which comply with a set of pre-defined morpho-syntactic patterns

and whose information exceeds a threshold of salience [21]. The detected terms
are then extracted, reduced to their Part Of Speech (Noun, Verb, Adjective, Adverb,
etc.) and Semantic Role (Agent, Object, Where, Cause, etc.) tagged base form [4,5].
The 96% of the words in a sentence is normally classified without any ambiguity,
while the complete syntactic tree for the sentence is extracted in the 77% of the
cases. The lemmatization speed is about 300 words per second. Once referred to
their lemma inside the multilingual Knowledge Base, they are used as documents
metadata [22,23].

The Semantic Engine supports English, Italian, German, French, Spanish,
Brazilian-Portuguese, and Arabic.

As an example, when analyzing the sentence

```
''The approval rating of Italian Prime Minister Silvio
Berlusconi, who has been dogged by a messy divorce
and accusations of cavorting with teenage girls
and escorts, has fallen below 50 percent for the
first time''
```

its Morpho-Syntactic Analysis returns as result (Table 2):

```
1:the{d(sg,def),0,3}
2:approval{n(cn,sg),4,8}
3:rating{n(cn,nom,sg),13,6}
4:of{p,20,2}
5:Italian{g,nationality,23,7}
6:[prime|'=minister']{n(cn,sg),h&a&title,31,14}
7:Silvio Berlusconi{n(prop,acc,sg),46,17}
9:who{n(pron(wh),sg),h&a,65,3}
10:have{v(aux),69,3}
11:be{v(pass),73,4}
12:dog{v(pastpart),78,6}
13:by{p,85,2}
14:a{d(sg,indef),88,1}
15:messy{g,90,5}
16:divorce{n(cn,acc,sg),96,7}
17:and{o(coord),104,3}
18:accusation{n(cn,acc,pl),108,11}
19:of{p,120,2}
20:cavort{v(prespart),123,9}
21:with{p,133,4}
22:teenage{g,138,7}
23:girl{n(cn,acc,pl),f&a&h,146,5}
24:and{o(coord),152,3}
25:escort{n(cn,acc,pl),156,7}
27:have{v(aux),165,3}
28:fall{v(pastparta),169,6}
```

Table 2 Part of speech labels

Label	Description
d	Determiner
p	Preposition
n	Noun (cn: common noun; prop: proper noun; pron: pronoun)
v	Verb (aux: auxiliary; pastpart: past participle; pass: passive)
o	z Coordination

Table 3 Semantic role labels

Label	Description
AGENT[x, y]	x is the subject for the action y
QUAL[x, y]	y qualifies or describes x
OBJ[x, y]	y is the object of the action x
COMP[x, y]	y is related to x by a complement
HOW[x, y]	y describes how action x occurs
WHEN[x, y]	y describes when action x occurs

```
29:below{p,176,5}
30:['50'|'=percent']{n(num,acc),percent,182,10}
31:for{p,193,3}
32:the{d(sg,def),197,3}
33:first{g,detadj&adjnoun,201,5}
34:time{n(cn,acc,sg),tm&tmdetr,207,4}
0:approval rating{n(@Nn),,4,15}
0:italian prime minister{n(@Gn),,23,22}
0:messy divorce{n(@Gn),,90,13}
0:teenage girl{n(@Gn),,138,13}
0:first time{n(@Gn),,201,10}
```

Instead, its Semantic Role Analysis returns as result (Table 3):

```
AGENT [3:rating,28:fall]
QUAL [3:rating,2:approval]
COMP [3:rating,7:Silvio Berlusconi] prep(of)
AGENT [16:divorce,12:dog]
QUAL [16:divorce,15:messy]
AGENT [18:accusation,12:dog]
OBJ [12:dog,7:Silvio Berlusconi]
HOW [28:fall,30:ST_number] prep(below)
COMP [28:fall,0:first time] prep(for)
```

In Fig. 1, the final result of the semantic annotation of the example, in the form of a functional graph, is presented. Each node represents a term and each arch represents a Semantic Role connecting two nodes. Different colors are used to represent the Part Of Speech values in nodes and Semantic Roles values in arches: for example, the purple color is used to represent nouns, the yellow to identify verbs
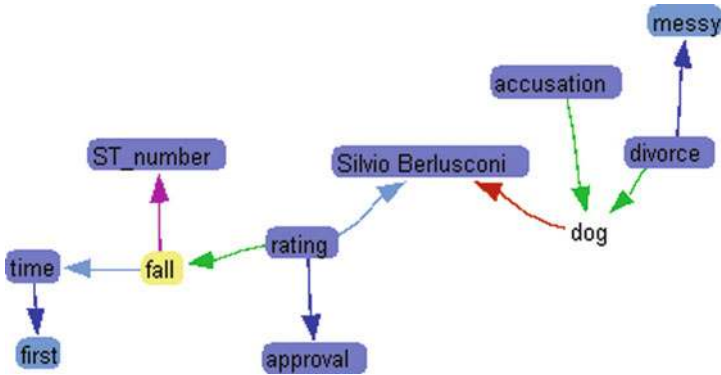
Fig. 1 Visual representation of lexical, semantic and semantic role analysis

and light blue is used for adjectives, whilst green color is used to describe subject-action relationships, the red for action-object relationships.

2.2.3 Sentiment Analysis

The Sentiment Analysis is based not only on the negative or positive polarity of words, but also on the syntactical tree of the sentence being analyzed. The system tries to read between lines, identifying idiomatic or colloquial expressions, giving interpretation to negations, modifying polarity of words basing on the related adverbs, adjectives or conjunctions. For example, when analyzing the sentence

```
''Gianfranco Fini has enraged the prime minister in
recent months by calling for greater morality in
government, complaining of Berlusconi's dictatorial
style of government.''
```

the system automatically gives it a negative polarity (Sentiment Score = negative), by identifying as sentiment factors

```
morality [positive]
enrage [negative]
complain [negative]
dictatorial [highly negative]
COMP(call for, morality) [highly negative]
```

Note that even if "*morality*" has a positive polarity, "c*all for morality*" turns it to an highly negative one. The system assigns a sentiment score to sentences and documents, taking account the polarity of all the words and logic-functional relationships among them.

```
1:Gianfranco Fini{n(prop,nom,sg),politician&h&a,0,15}
2:have{v(aux),16,3}
```

Fig. 2 Sentiment polarity
gauge: visual representation
of sentiment score

Sentiment Score: -5
complain [-1]
QUAL(morality, great) [1]
OBJ(enrage, prime minister) [-1]
QUAL(style, dictatorial) [-2]
COMP(call, morality) [-2]

```
3:enrage{v(pastparta),20,7}
4:the{d(sg,def),28,3}
5:[prime|'=minister']{n(cn,acc,sg),h&a&title,32,14}
6:in{p,47,2}
7:recent{g,50,6}
8:month{n(cn,acc,pl),tm,57,6}
9:by{p,64,2}
10:call{v(prespart),67,7}
11:for{p,75,3}
12:great{g,79,7}
13:morality{n(cn,acc,sg),87,8}
14:in{p,96,2}
15:government{n(cn,acc,sg),99,10}
17:complain{v(prespart),111,11}
18:of{p,123,2}
19:Berlusconi{n(prop,gen,sg),sn&politician&h&a,126,10}
20:apos{o(punc),136,2}
21:dictatorial{g,139,11}
22:style{n(cn,acc,sg),151,5}
23:of{p,157,2}
24:government{n(cn,acc,sg),160,10}
```

In Fig. 2, the final result of sentiment score is presented, for the example, in the
form of gauge. The Sentiment Polarity Gauge displays the polarity score of the
overall sentence in an easy and compact format: when the indicator is in the red
section of the bar, the polarity score is negative, whilst when the indicator is in
the green part, the polarity score is instead positive. It displays also the terms and
expressions which contribute to the Sentiment Polarity identification: even though
"*morality*" contributes to a positive score, other expressions with high negative
impact like "*COMP[call, morality]*", or "*dictatorial*", or "*enrage*" or "*complain*"
make it negative.

2.3 The Search Engine

2.3.1 Natural Language Search

Users can search documents by Natural Language queries, expressed using normal conversational syntax [24, 25]. In general, users can express their interest by using normal conversational syntax. Reasoning over facts and ontological structures makes it possible to handle diverse and more complex types of questions. Traditional Boolean queries in fact, while precise, require strict interpretation that can often exclude information that is relevant to user interests. So this is the reason why the system maps the query into concepts, having identified the most relevant terms contained and their semantic and functional interpretation, and use them as semantic keywords.

2.3.2 Semantic Search

Users can search by conceptual keywords combined by Boolean operators [24, 25].

2.3.3 Semantic Role Search

Users can search and navigate by semantic roles, exploring sentences and documents by the functional role played by each concept [24, 25]. Users can search by combining concepts into SAO (Subject–Action–Object) triples. By mapping a query to concepts and relations very precise matches can be generated, without the loss of scalability and robustness found in regular search engines that rely on string matching and context windows. The search engine returns as result all the documents which contain the query concepts/lemmas in the same functional role as in the query, trying to retrieve all the texts which constitute a real answer to the query. A chart displaying all the concepts and the relations among them is provided as a visual investigative component specifically designed to bring clarity to complex investigations. This chart automatically enables investigative information to be represented as visual elements that can be easily analyzed and interpreted: concepts are represented by nodes, relations are displayed as arches; users can explode them and have access to the set of sentences/documents characterized by the selected search criteria. Functional relationships – *Agent, Action, Object, Qualifier, When, Where, How* – among human beings and organizations can be searched for and highlighted; pattern and hidden connections can be instantly revealed to help investigations, promoting efficiency into investigative teams. Should human beings be cited, their photos or factsheets can be shown by simple clicking on the related icon.

2.4   The Machine Translation Engine

The system uses a combination of MT and CAT approaches, enabling the automatic translation of all the pages of interest. It uses a dual feed recursive process, where the reuse of translated text or the re-translation of text continuously improve the quality level.

2.5   The Georeferentiation Engine

The system allows users to search for information in a geographic map, timely projected. Geo-referencing can be based either on sources, or documents contents: for example, users can navigate into a geographical interactive map basing on the newspapers headquarters, or on the towns mentioned in a news article. Then users have the opportunity to understand how news spreads throughout the world, or understand what are the primary sources of information; to perceive easily and intuitively which is the original source, or if there's a coordinated press campaign.

2.6   The Classification Engine

The automatic classification of documents is made fulfilling both the Supervised and Unsupervised Classification schemas. The application assigns texts to predefined categories and dynamically discovers the groups of documents which share some common traits.

2.6.1   Supervised Clustering

The categorization model is created during the learning phase, on representative sets of training documents focused on the topics of interest. The Bayesian method is used as the learning method: the probabilistic classification model is normally built on around 300 documents for each thematic category.

2.6.2   Unsupervised Clustering

Documents are represented by a sparse matrix, where lines and columns are normalized in order to give more weight to rare terms. Each document is turned to a vector comparable to others. Similarity is measured by a simple cosines calculation between document vectors, whilst clustering is based on the K-Means algorithm. The application provides a visual summary of the clustering analysis. A map shows

the different groups of documents as differently sized bubbles and the meaningful correlation among them as lines drawn with different thickness.

3 Monitoring the Italian Prime Minister's Web Sentiment

3.1 Introduction

Communication is crucial in the competitive political arena. Taking account of the complexity of public views makes it possible to dramatically increase the political leaders' consensus, and promptly react to negative opinions expressed by journalists and opinion leaders in blogs, forums and newspapers. Synthema has been asked to investigate if there has been a coordinated foreign press campaign against the Italian prime minister Silvio Berlusconi, to represent his consensus and reputation in newspapers and magazines on behalf of some influential Italian politicians.

3.2 Collecting the Data

Around one thousands of Web pages has been collected by having focused crawled some of the most influential newspapers and media sources in the world, such as BBC, Wall Street Journal, Times, Reuters, ABC News, beside social networks, blogs, and forums. Only the textual contributions related to Silvio Berlusconi's sexual scandal have been taken in account, conceptually analyzed, and indexed.

3.3 Navigating the Data

3.3.1 Search and Cluster

Users can search documents by expressing their own interest by Natural Language queries, using normal conversational syntax. For example, when searching for *"prostituzione"* (namely *"prostitution"*) (see Fig. 3), the system returns as result all the documents talking about prostitution (type, persons involved, addresses, phone numbers, etc.), as displayed in Fig. 4. The Fig. 4 shows Patrizia D'Addario, the girl involved in the sexual scandal, as a specific escort resident in the South of Italy.

When clustering these results, the system identifies a group of documents dealing with the European parliament seat, which – basing on the phone callings transcriptions – was offered to the escort; another cluster of documents talks about the complaints expressed by the electors of People of Freedom, the Berlusconi's political party, or the decline of his rating of approval. Other documents concern the judicial acts or the scandal audio recordings. A map shows the different clusters as
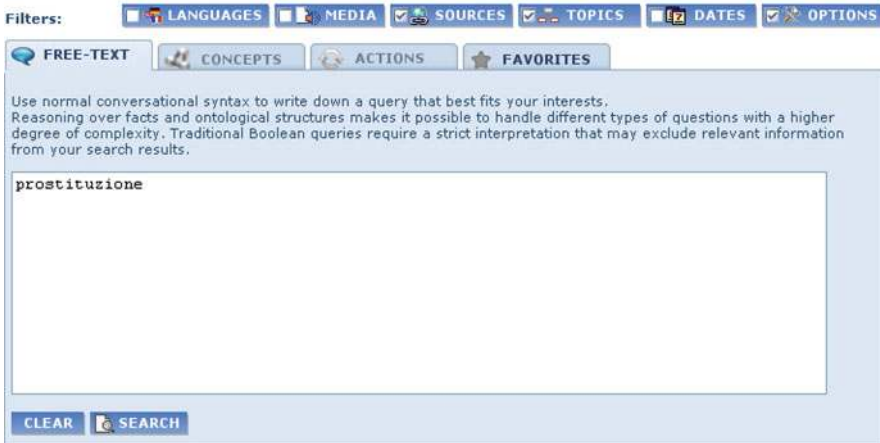
Fig. 3 Natural language search

Fig. 4 Results for natural language search: documents, named entities

differently sized bubbles, the meaningful correlation among them as lines drawn with different thickness (see Fig. 5). The difference in thickness indicates the strength of links: the stronger the link, the thicker the line. Analysts can search for relationships in clusters, explore links among topics. For example, when exploring the link between the *"complains"* and the decline of *"approval rating"*, the system suggests that it's due to *"escort says"*, *"recording"*, *"to pay sex"*, giving analyst a new automatic perspective on cited facts (see Fig. 5).

3.3.2 Search and Explore in Depth

In example, when searching for "dog" as action and "Silvio Berlusconi" as object (see Fig. 6), the system returns as result three sentences, as displayed in Table 4.
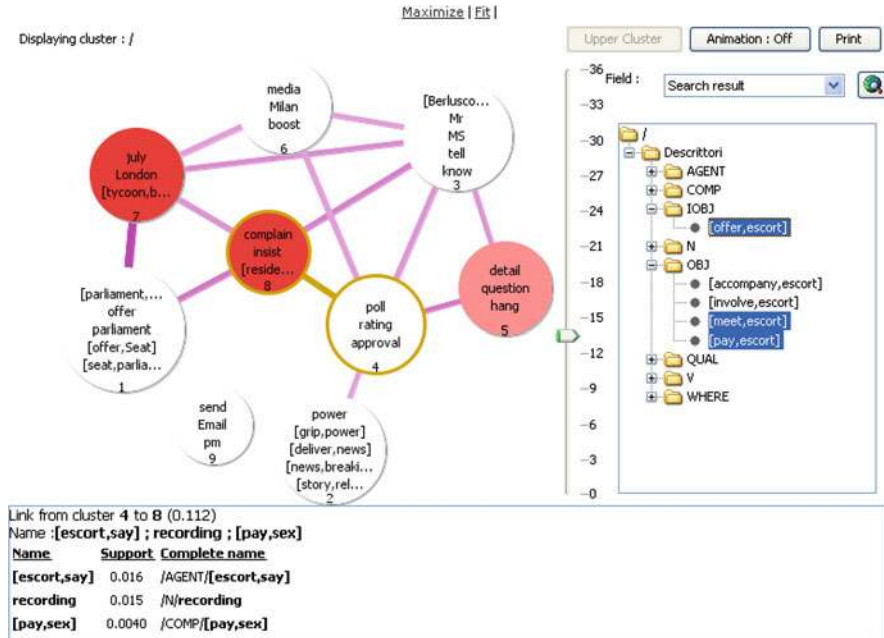
Fig. 5 Thematic map and semantic role projection on clusters

Table 4 Results of SAO search

Search	Results
to dog Silvio Berlusconi	...scandals dogging prime minister Silvio Berlusconi
	Silvio Berlusconi, who has been dogged by a messy divorce and accusation of ....
	Silvio Berlusconi, who has been dogged by a messy divorce and accusation of ....
	... scandals dog Berlusconi

Note that the system has automatically turned any passive sentence into an active one, mapping its passive subject into the object or turning the agent complement into the subject of the equivalent normalized active sentence. The system has also resolved any pronominal reference, mapping "who" with "Silvio Berlusconi" (see Fig. 7). The relations chart displayed in Fig. 8 can be considered a visual investigative component specifically designed to bring clarity to complex investigations. In fact, it automatically enables investigative information to be represented as visual elements, that can be easily analyzed and interpreted [25]. Then, navigating on the relations chart by simply clicking on conceptual nodes or relation arches, expanding them and having access to set of sentences/documents characterized by the selected criterion, the system suggests that sexual scandal weakens the Italian prime minister's credibility (see Fig. 8).
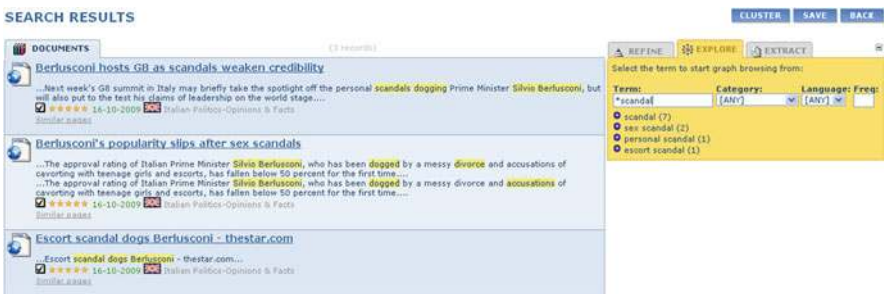
Fig. 6 SAO search
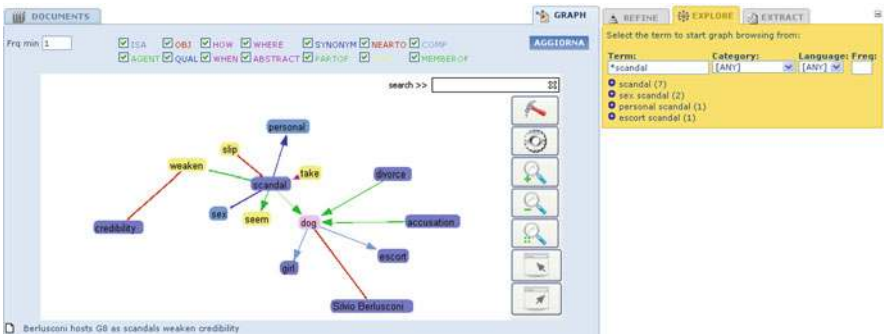

Fig. 7 Results for SAO search

Fig. 8 Semantic role chart

3.3.3   Explore Statistical Data

The system allows users to explore statistical – time and frequency – distributions for concepts, taken into account on the basis of their *Part Of Speech (Noun, Verb, Adjective, Adverb,* etc.) and *Semantic Role (Agent, Object, Where, Cause,* etc.). The system automatically labels adjectives, adverbs, nouns, or even sentences with a subjective positive, negative, or objective score, giving indications about perception people may have about persons or facts [26]. As an example, Table 5 summarizes the

Table 5 Adjectives frequency

Adjective	Frequency	Adjective	Frequency
Attractive	5	Contrary	2
Defamatory	4	Concerned	2
Damaging	4	Immoral	2
Defiant	3	Inappropriate	2
Illegal	3	Abusive	2
Sober	3	Juicy	2
Nice	3	Pornographic	2
Obscene	3	Precious	2
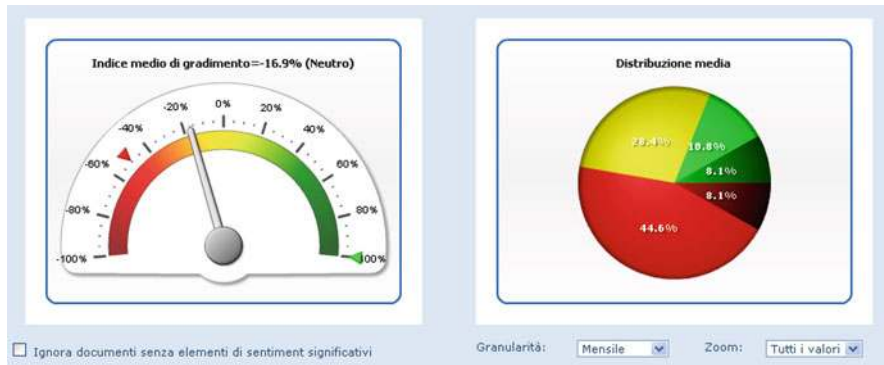
Fig. 9 Average sentiment polarity gauge and distribution of consensus chart

frequency distribution for all the adjectives extracted from the documents matching the search query; in red the negative concepts, in yellow the neutral ones, in green the positive ones. In this case, given the prevalence of negative scores, it is easy to infer a negative perception for Mr. Berlusconi and his sexual scandal:

Several visualization options and dashboards are available for showing results and distribution of polarity for the whole set of documents matching the search criteria. For example, an Average Sentiment Polarity Gauge (see Fig. 9) shows a global average polarity score; a Distribution of Consensus Chart (see Fig. 9) shows the distribution of documents across polarity classes. When analyzing the time projection for consensus, analysts can note its free-fall in coincidence of the publication of sexual scandal audio recordings. Thereafter the prime minister's consensus starts increasing in coincidence with the emerging financial crisis. In the fall-winter 2009, in fact, the crisis tests EU foundations, emerging a concrete risk for a sovereign default in Dubai and in some East Europe countries. The approval rate increases, although the journalists and bloggers start to write negative comments about the prime minister (see Fig. 10).
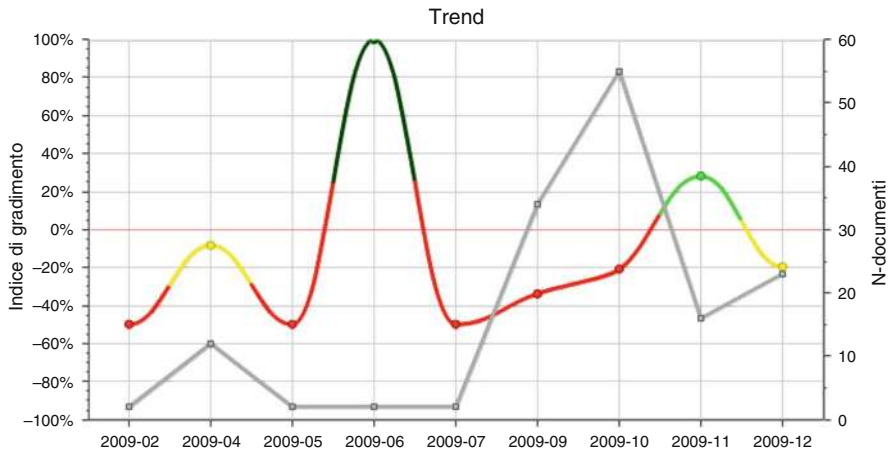
Fig. 10 Time distribution of consensus

Fig. 11 Georeferentiation analysis

3.3.4 Explore Geographical Data

Having had the opportunity to understand how the scandal news have spread throughout the world, or understand what were the primary sources of information, analysts can perceive easily and intuitively that there was not a coordinated press campaign against the Italian prime minister. In fact, by timely projecting the news and their sources, analysts can discover that primary sources of information were Italian (see Fig. 11).

4 Conclusions

This paper describes a Sentiment Mining study on the Italian prime minister Silvio Berlusconi, performed by a Knowledge Mining system used by some security sector-related government institutions and agencies in Italy to limit information overload in OSINT and Web Mining. The linguistic and semantic approaches implemented in this system enable the research, the analysis, the classification of great volumes of heterogeneous documents. The sentiment analysis is simply an additional layer of information extraction. This system helps documental analysts to cut through the information labyrinth, political spin doctors to take account of complexity of public views, assigning automatically a sentiment polarity to posts and documents, allowing to rapidly access all the potential texts of interest, or monitor the consensus. So, this system tries to limit the huge virtual space – the Web – where people express and share opinions, influencing any aspect of life, giving analysts just a synthetic and intuitive grid able to limit the information overload.

References

1. Grishman, R., Sundheim, B.: Message Understanding Conference – 6: A Brief History. In: Proceedings of the 16th International Conference on Computational Linguistics (COLING), I, Kopenhagen, 1996, pp. 466–471 (1996)
2. Hearst, M.: Untangling text data mining. In: ACL'99, pp. 20–26. University of Maryland, June 1999
3. Miller, H.J., Han, J.: Geographic Data Mining and Knowledge Discovery, CRC Press (2001)
4. Wei, L., Keogh, E.: Semi-Supervised Time Series Classification. SIGKDD 2006 (2006)
5. Carreras, X., Marquez, L.: Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling. In: CoNLL-2005. Ann Arbor, MI USA (2005)
6. Vossen, P., Agirre, E., Bond, F., Bosma, W., Fellbaum, C., Hicks, A., Hsieh, S., Isahara, H., Huang, Ch., Kanzaki, K., Marchetti, A., Rigau, G., Ronzano, F., Segers, R., Tesconi, M. (fc.): KYOTO: a Wiki for establishing semantic interoperability for knowledge sharing across languages and cultures. In: Handbook of Research on Culturally-Aware Information Technology: Perspectives and Models. Ed. Dr. E. Blanchard (Mc Gill University (Canada) and Dr. D. Allard (Dalhousie University), IGI Global USA, p. 265–294, ISBN 978161520883 (2011)
7. Dorr, B.J., Jordan, P.W., Benoit, J.W.: Surveys of current Paradigms in Machine Translation. Technical Report: LAMP-TR-027, UMIACS-TR-98-72, CS-TR-3961, University of Maryland, College Park (1998)
8. Bruderer, H.E.: The Present State of Machine and Machine-Assisted Translation. Commission of the European Communities, Third European Congress on Information Systems and Networks: Overcoming the Language Barrier. vol. 1. Verlag Dokumentation, Munich: 529–556 (1977)
9. Hutchins, W.J.: Progress in Documentation: Machine Translation and Machine-Aided Translation. J. Chem. Doc., **34**(2), 119–159 (1978)
10. Hutchins, W.J.: The Evolution of Machine Translation Systems. In Lawson 1982, pp. 21–37 (1982)

11. King, M.: EUROTRA: An Attempt to Achieve Multilingual MT. In Lawson 1982: 139–147 (1982)
12. Johnson, R.L.: Parsing – an MT Perspective. In: Jones, K.S., Wilks, Y. (eds.) Automatic Natural Language Parsing. Ellis Horwood, Ltd., Chichester, Great Britain (1983)
13. King, M.: Design Characteristics of a Machine Translation System. In: Proceedings of the Seventh IJCAL Vol. 1, Vancouver, B.C., Canada: 43–46 (1981)
14. Loh, S.-C.: Machine Translation: Past, Present, and Future. ALLC Bulletin. **4**(2), 105–114 (1976)
15. Mori, R.: Advances in Machine Translation Research in IBM. In: MT Summit III, Washington, D.C., 1–4 July (1991)
16. Baldini, N., Bini, M.: Focuseek searchbox for digital content gathering. In: AXMEDIS 2005 – 1st International Conference on Automated Production of Cross Media Content for Multi-channel Distribution. Proceedings Workshop and Industrial, pp. 24–28. (2005)
17. Diligenti, M., Coetzee, F. M., Lawrence, S., Giles, C. L., Gori, M.: Focused Crawling Using Context Graphs, In: Proceedings of 26th International Conference on Very Large Databases, VLDB, pp. 527–534, 10/9 – 12/9 (2000)
18. Neri, F., Raffaelli, R.: Text Mining applied to Multilingual Corpora, Knowledge Mining: In: Proceedings of the NEMIS 2004 Final Conference, Springer Verlag Pub., Spiros Sirmakessis Ed., ISBN-13: 978-3540250708 (2004)
19. McCord, M.C.: Slot Grammar: A System for Simpler Construction of Practical Natural Language Grammars. Natural Language and Logic 1989: 118–145 (1989)
20. McCord, M.C.: Slot grammars. Am. J. Comput. Linguist. **6**(1), 31–43 (1980)
21. Cascini, G., Neri, F.: Natural language processing for patents analysis and classification. In: ETRIA World Conference, TRIZ Future 2004, Florence, Italy, Proceedings, Cascini Ed., pp. 199–212, Firenze (IT), 03-05/11/2004
22. Baldini, N., Neri, F., Pettoni, M.: A multilanguage platform for open source intelligence, data mining and information engineering 2007. In: 8th International Conference on Data, Text and Web Mining and their Business Applications, The New Forest (UK), Proceedings, ISBN: 978-184564-081-1, WIT Transactions on Information and Communication Technologies, vol. 38, 18–20 June 2007
23. Neri, F., Pettoni, M.: Stalker, a multilanguage platform for open source intelligence, open source intelligence and web mining symposium. In: 12th International Conference on Information Visualization, Proceedings, ISBN:978-0-7695-3268-4, pp. 314–320, IEEE Computer Society, LSBU, London (UK), 8–11 July 2008
24. Neri, F., Geraci, P.: Mining Textual Data to boost Information Access in OSINT, Open Source Intelligence and Web Mining Symposium, 13th International Conference on Information Visualization, IV09, Proceedings, ISBN: 978-0-7695-3733-7, pp. 427–432, IEEE Computer Society, Barcelona (ES), 16-17/07/2009
25. Neri, F., Geraci, P.: Overcoming Linguistic Barriers in Open Source Intelligence, Governmental Sidebar, Semantic Conference, San Jose (CA), USA, 17/06/2009
26. Camillo, F., Neri, F.: Sentiment Mining: Automatic linguistic analysis and natural language understanding in an e-CRM systems. In: Mola, F., et al. (eds.) EURISBIS 2009, Proceedings, TILAPIA, p. 136, Cagliari (IT), 01/06/2009

Exploring the Evolution of Terrorist Networks

Nasrullah Memon, Uffe Kock Wiil, Pir Abdul Rasool Qureshi, and Panagiotis Karampelas

Abstract This paper discusses advancements and new trends in terrorist networks. We investigate a case regarding a recent terror plan that took place in Denmark and we present the analysis of the thwarted plot. Analyzing covert networks after an incident is practically easy for trial purposes. Mapping clandestine networks to thwarted terrorist activities is much more complicated. The network involved in the recent Danish terror plan is studied through publicly available information. Based on that information we mapped a part of the network centered on David Headley, who recently confessed to have planned a terrorist attack to take place on Danish soil. Despite its deficiencies, the map gives us an insight into new trends in terrorist organizations and people involved in terrorist plots.

N. Memon (✉)
Counterterrorism Research Lab, The Maersk Mc-Kinney Moller Institute, University of Southern Denmark, Odense, Denmark
and
Hellenic American University, Manchester, NH, USA
e-mail: memon@mmmi.sdu.dk

U.K. Wiil · P.A.R. Qureshi
Counterterrorism Research Lab, The Maersk Mc-Kinney Moller Institute, University of Southern Denmark, Odense, Denmark
e-mail: ukwiil@mmmi.sdu.dk; parq@mmmi.sdu.dk

P. Karampelas
Hellenic Air Force Academy, Athens, Greece
and
Hellenic American University, Manchester, NH, USA
e-mail: pkarampelas@gmail.com

1 Introduction

The events of 9/11 instantly changed the perception of the words "terrorist" and "network", and the United States and other countries rapidly started to gear up to fight a new kind of war against a new kind of enemy. In conventional warfare, conducted in specific locations, it is important to understand the terrain in which the battles will be fought. In the war against terror, there is no specific location and time. After 9/11, we know that the battleground can be anywhere and at any time. It is now clear that the terrorists' power base is not geographic; rather, they operate in networks, with members distributed across the globe. To fight such an enemy, we need to understand the new "terrain": networks – how they are constructed and how they operate.

Advanced and emerging information technologies like investigative data mining (IDM) offer key assets in confronting a secretive, asymmetric networked enemy. IDM is a powerful tool for intelligence and law enforcement agencies fighting terrorism [12]. IDM is a combination of data mining and subject-based automated data analysis techniques. Data mining is an approach which uses algorithms to discover predictive patterns in datasets. Subject-based automated data analysis applies models to data to predict behavior, assess risk, determine associations, or perform other types of analysis [13].

How can we mine terrorist networks? Traditional methods of machine learning and data mining, taking a random sample of homogeneous objects from a single relation as input, may not be appropriate. The data comprising terrorist networks tend to be heterogeneous, multi-relational, and semi-structured. IDM embodies descriptive and predictive modeling. By considering links (relationships between the entities), more information is made available to the mining process. Mathematical methods used in the research on IDM [12–15] are clearly relevant to intelligence analysis and may provide tools and techniques to discover terrorist networks in their planning phase and thereby prevent terrorist acts from being carried out. Relevant patterns to investigate include connections between actors (meetings, messages), activities of the involved actors (specialized training, purchasing of equipment), and information gathering (time tables, visiting sites).

IDM offers the ability to firstly map a covert cell, and to measure the specific structural and interactional criteria of such a cell. IDM aims to connect the dots between individuals and map and measure complex, covert, human groups, and organizations. The methods focus on uncovering the patterns of interaction, and correctly interpreting these networks to predict behaviors and decision-making within the network. IDM borrows techniques from social network analysis (SNA) and graph theory for connecting the dots.

IDM enables us to understand the information flow within a network and the nodes which play a central role in communication between remote nodes in the same network. Some of the most important techniques to model and understand information flow used in IDM are discussed in [6, 24]. The modeling of the information flow within a network can enhance the understanding of how the

network operates. In addition, this modeling improves prediction of how the remote members convey their messages to the rest of network and how this helps the network to overcome geographic constraints. This identifies some points that are important in order to determine how the terrain networks are constructed and operated.

Modeling and predicting the information flow within a network is also helpful in identifying the important members of a network, which are necessary in a network to maintain the stability and continuation of information flow [18]. Terrorism nowadays has evolved to an international rather than a national matter. The multi-culture security sensitive information from different geographic locations and international boundaries is collected to a central authority, no matter even if it lies in a cave in Afghanistan. The collected information is processed; the most feasible proposals are selected and then realized. However, the described approach has a weakness, i.e. the central authorities must adhere to a limited number of dependable nodes, and that group of nodes is common in, if not all, most of the networks operations. The study [18] attempts to model those important characteristics of terrain networks and attempts to identify the most dependable nodes, which are the ones with high dependence values [18] for the whole network. These nodes are considered very important, if the disruption of a terrorist network is the desired goal rather than full elimination, which is normally, the only feasible option available.

In IDM, a number of variations exist in the literature. One is known as link analysis (see for example [4, 21]). Link analysis research uses search and probabilistic approaches to find structural characteristics in the network such as hubs, gatekeepers, pulse-takers [22], or identifies potential relationships from relational data mining. Link analysis alone is insufficient as it looks at one side of the coin and ignores complex nonlinear relationships that may exist between the attributes. Another approach depends purely on visualization, such as NetMap [7]. Unfortunately, these tools that depend on visualization alone – despite being useful to provide some insight – are insufficient and rely on the user to carry out many tedious and time consuming tasks, many of which could be automated.

Uncovering a relationship among or within attributes (connecting the dots) is an important step, but in many domains it is more important to understand how this relationship has been evolved. Hence, understanding network dynamics and evolution is needed in order to complete the picture. Once we understand the dynamics and evolution of these relationships, we can search for ways to disconnect the dots if and when needed. This brings about several new tasks [16]: (a) subgroup detection; (b) object classification; (c) community detection; (d) object dependence; (e) hidden hierarchy detection; and (f) topological characteristics understanding.

In this paper, we use IDM techniques to study new trends regarding the recent Denmark terror plan, in which David Headley recently confessed to conspiring between October 2008 and October 2009 with his associates to plan and carry out terrorist attacks, including murder and maiming, against the facilities of Jyllands-Posten, a Danish newspaper, and two of its employees, Editor A and Cartoonist A [23].

In Sect. 2, we briefly present the case study of the recent Denmark terror plan. Section 3 introduces IDM techniques to detect key players and the role of dependence centrality [18] to determine the informational flow. In Sect. 4, we report and discuss our analysis results from the case study. Section 5 concludes the paper and presents future research directions.

2 Case Study

David Coleman Headley [1], formerly known as Daood Sayed Gilani, (born June 30, 1960) is a Pakistani-American businessman based in Chicago. He recently confessed his involvement in terrorist plots against India and Denmark. David Coleman Headley and Tahawwur Hussain Rana were accused by U.S. federal authorities in Chicago, in complaints unsealed on October 27, 2009, of plotting against the employees of a newspaper in Copenhagen. Headley is accused of traveling to Denmark to scout the building of the Jyllands–Posten newspaper, and a nearby Synagogue, for facilitating the organization of an attack by terrorists [23]. On December 8, 2009, the FBI also accused Headley of conspiring to bomb targets in Mumbai, India; providing material support to Lashkar-i-Taiba, a militant Pakistani extremist group; and aiding and abetting the murder of U.S. citizens [23]. There are some online data sources containing structured terrorist information like http://www.trackingthethreat.com/, http://www.globalsecurity.org/, etc. As we did not find the information about the entities present in the David Headley case from these manually updated online sources, we harvested the information [17] about the David Headley network from publicly available news and information sources. The harvested information was combined with the information present at "http://www.trackingthethreat.com/" and the result is shown in Fig. 1.

The red rectangle in Fig. 1 shows the main entities under our investigation and connected directly with the David Headley case. We have applied IDM techniques over the network and the application of these techniques is discussed in the following sections.

3 IDM Techniques for Detecting Communities and Key Players

In this section, we discuss various techniques to detect the core members of a terrorist network related to the specific case study.

3.1 Subgroup/Community Detection

One of the most common interests in analyzing terrorist networks is the identification of the substructures that may be present in the network. Subgroups are
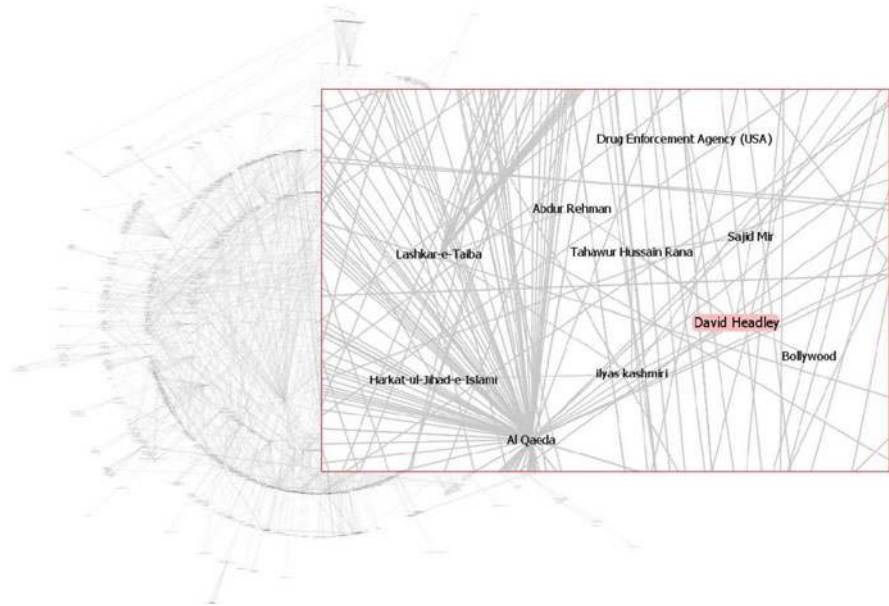
Fig. 1 The David Headley terrorist network

subsets of actors among whom there are relatively strong, direct, intense, frequent, or positive ties. We use a bottom-up approach for the detection of subgroups [16]. This approach begins with basic groups, and seeks to determine how far this kind of close relationship can be extended. The notion is to build outward from single ties to construct the network. The substructures that can be identified by bottom-up approaches include cliques, n-cliques, s-cliques, and k-plexes. We discuss each concept briefly [8]:

- A *clique* is defined as a maximal sub-graph in which every member of the graph is connected to every other member of the graph. Every member is connected to n-1 others and the distance between every pair is one. In practice, complete cliques are not very useful. They tend to overlap heavily and are limited in their size.
- An *n-clique* is a sub-graph in which every person is connected by a path of length n or less.
- A group is an *s-clique*, if it has local maximal SMI (Segregation Matrix Index). That a group G has local maximal SMI means that no other group has a higher SMI value. In addition, no other group has the same SMI value with one more element or one less element than G.
- A *k-plex* is a sub-graph in which every person is connected to at least n-k other people in the graph (recall in a clique everyone is connected to n-1, so this relaxes that condition) [8].

In addition, we have used the most popular CNM algorithm (which discovers clear communities in the network) introduced by Clauset, Newman, and Moore which maximizes modularity with greedy approach [5].

3.2 Object Classification

In traditional classification methods, objects are classified on the attributes that describe them. A particular important challenge is to classify in a large network those individuals who play key roles – such as leaders, facilitators, communications "go betweens", and so on. To understand the calculations used to single out the core members in a network, we need to discuss some measures of object classification [11, 23]:

- *Degree.* A basic measure of SNA that turns out to be important in IDM is the degree of a node – that is, the number of other nodes directly connected to it by edges. In a graph (network) describing a terrorist network, nodes of high degree represent "well connected" people, often the leaders.
- *Closeness.* This measure indicates for each node how close it is to other nodes in a graph. Analysts consider this measure a good indication of how rapidly information can spread through a network from one node to others. This measure relates to the closeness or the distance between nodes. A core member (central actor) can reach other actors through a minimum number of intermediary positions and is therefore dependent on fewer intermediary positions than a peripheral actor.
- *Betweenness.* The measure gives each node a score that reflects its role as a stepping-stone along geodesic (shortest) paths between other pairs of nodes. The idea is that if a geodesic path from node A to node B (there may be more than one) goes through node C, then node C gains potential importance. Such nodes – or the people that they represent a terrorist network – can have important roles in providing connections (for example, facilitating communications) between sets of nodes that otherwise have few other connections, or perhaps no other connections. This measure explores an actor's ability (say for example, node C) to be "irreplaceable" in the communication of two random actors (say for example, nodes A and B). It is of particular interest in the study of destabilizing terrorists by network attacks, because at any given time the removal of maximum betweenness actor seems to cause maximum damage in terms of connectivity and average distance in a network.

3.3 Node Dependence and Information Flow

The information flow enables the significant understanding of how the network operates, but it is also difficult to model. The traditional social network information

flow models further classify the links between nodes into multiple classes in a quest to identify the most probable communication link. In the social network literature, researchers have examined a broad range of types of ties [19]. These include communication ties (such as who talks to whom or who gives information or advice to whom), formal ties (such as who reports to whom), affective ties (such as who likes whom, or who trusts whom), material or work flow ties (such as who gives bomb making material or other resources to whom), and proximity ties (who is spatially or electronically close to whom). Networks are typically multiplex, that is, actors share more than one type of tie. For example, two terrorists might have a formal tie (one is a foot-soldier or a newly recruited person in the terrorist cell and reports to the other, who is the cell leader) and an affective tie (they are friends); and may also have a proximity tie (they are residing in the same apartment and their flats are two doors away on the same floor).

Network researchers have distinguished between strong ties (such as family and friends) and weak ties such as acquaintances [9, 10]. This distinction will involve a multitude of facets, including affect, mutual obligations, reciprocity, and intensity.

In information flow, the strong ties are particularly valuable when an individual seeks socio-emotional support and often entail a high level of trust. Weak ties are more valuable when individuals are seeking diverse or unique information from someone outside their regular frequent contacts. This point towards the fact that even weak ties are significantly important in terrorist networks, where information flows from the geographically distant parts of world through a lot of actors who does not even have met with each other [9].

The dependence centrality [18] is proposed to identify the most feasible link for information to flow and reach the farthest part of the graph. These links are said to be the most important links over which the network depends and the nodes connected with these links are detected by corresponding high dependence values. Consider a network representing a symmetrical relation, "communicates with" for a set of nodes. When a pair of nodes (say, u and v) is linked by an edge so that they can communicate directly without intermediaries, they are said to be adjacent. A set of edges linking two or more nodes (u, v, w) in such a way that node u would like to communicate with w, using node v. The dependence centrality can discover how many times node u uses node v to reach node w and how many shortest paths node u uses to reach node w. There can, of course, be more than one geodesic linking between any pair of nodes.

Let $\zeta(u, v)(w)$ = dependence factor of the node u on node v to reach any other node (i.e., node w) in the graph of communication as shown below:

$$\zeta_{u,v}(w) = \frac{\text{occurence}(u, v)}{d(u, v) * \text{path}(u, w)}$$

where,
occurence(u, v) = thenumberoftimes (shortest paths), the node u uses node v in the communication with one another,
path(u, w) = thenumbershortestpathsbetweennode u and node w, and
$d(u, v)$ = thegeodesicdistancebetweennode u and node v.

The dependence centrality and the corresponding theory has produced good results known in Al Qaeda case studies such as in [20] in which it has been found that the theory is in agreement with the reality.

4   Analysis Results

We have collected information about the network using the iMiner harvesting facility [16] and developed a network as shown in Fig. 1. The identified network has some interesting characteristics, if weighted against different SNA measures. By identifying the cohesive parts of the network with different algorithms, it has been found that David Headley is the most important node (Fig. 1) followed by Kashmiri and Rana. The degree, closeness, and betweenness of the different nodes shown in Fig. 1 are also in agreement with the importance of these nodes as shown in Fig. 2.

The different clique based and k-plex sub-graph detection algorithms also reveal the same results regarding the importance of different nodes. The number of sub graphs containing David Headley is the highest as shown Fig. 3.

David Headley is the most irreplaceable node of the network (as shown in Fig. 3), since its absence makes the maximum of sub-graphs incomplete. This may be due to the fact that Headley is the most socialized member of the network and has many links with common people outside the network. In contrary to 9/11, where the terrorists rarely interacted with the outside world and hiding themselves from media, this time the trend has changed. The people involved in the conspiracy are even connected with the superstars of Bollywood (film industry of India). A U.S. Department of Defense official who is following the case closely and spoke on condition of anonymity because of the criminal investigation in progress has stated:
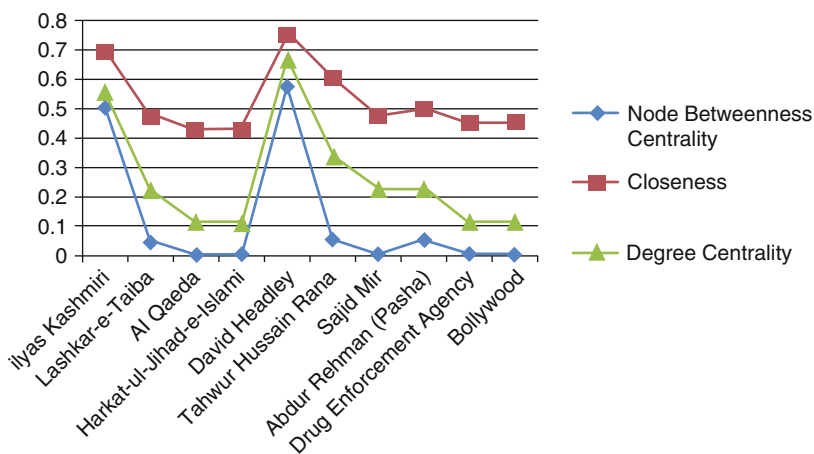

Fig. 2 Centrality measures of the David Headley network
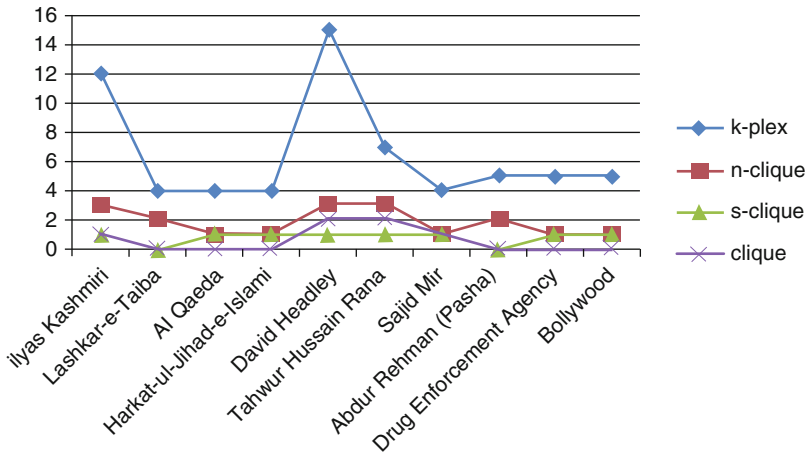
Fig. 3 Sub-graph detection in the David Headley network

The present and future is less about individual groups conducting attacks, and more about combinations of individuals, and groups and facilitators that come together.... Together they have the resources, the means, and the insights to execute those attacks [2].

The people involved in conspiracies are common people or at least they behave like common people and make it difficult for the investigators to distinguish and identify them. This fact is emphasized in the following statement:

Foreigners on reaching their destination in India and checking in at a hotel or hostel are required to fill in and sign a form for scrutiny by the immigration authority. It is the responsibility of the management of the establishment to get the forms filled in and to ensure their submission to the town/city immigration authorities along with the passport of the individual. Headley's passport should have been scrutinized in over 12 towns during his nine visits and passed through the hands of at least three dozen officers but not one of them found anything unusual in it [3].

The ties with the outside world and film industry can also be proved essential in strategies for fund raising or even for using the film industry to money laundering, which is of course not a new thing. Although, it has not been disclosed whether Headley had a talk with Bollywood stars about financing; it is surely a relevant suspicion.

As seen in Fig. 3, the sub-graph detection mechanisms using cliques do not clearly identify the key nodes. However, the k-plex algorithm clearly identifies the key nodes (Kashmiri and Headley).

Therefore, we have further investigated the network using the k-plex algorithm to find out additional features of interest. Finding sub-graphs with minimum size of three using the k-plex algorithm yielded 21 groups or k-plexes. All groups detected along with their members, density, size, and cohesion index are listed in Table 1. The statistical graph shown in Fig. 4 by plotting the density, cohesion index, and size against each k-plex shows that k-plex5 and k-plex10 are the ones with highest

Table 1 Subgroups (k-plexes) of the network along with the details of the members, size, density, and cohesion index

| K-Plex	Size	Density	Cohesion index	Members
k-Plex1	3	0.667	3.5	Ilyas Kashmiri, Lashkar-e-Taiba, Al Qaeda
k-Plex2	3	0.667	3.5	Ilyas Kashmiri, Lashkar-e-Taiba, Harkat-ul-Jihad-e-Islami
				Ilyas Kashmiri, Lashkar-e-Taiba, David Headley,
k-Plex3	4	0.667	2.286	Abdur Rehman (Pasha)
k-Plex4	3	0.667	2.333	Ilyas Kashmiri, Lashkar-e-Taiba, Tahwur Hussain Rana
k-Plex5	3	0.667	4.667	Ilyas Kashmiri, Al Qaeda, Harkat-ul-Jihad-e-Islami
k-Plex6	3	0.667	1.75	Ilyas Kashmiri, Al Qaeda, David Headley
k-Plex7	3	0.667	2.8	Ilyas Kashmiri, Al Qaeda, Tahwur Hussain Rana
k-Plex8	3	0.667	1.75	Ilyas Kashmiri,Harkat-ul-Jihad-e-Islami, David Headley
				Ilyas Kashmiri, Harkat-ul-Jihad-e-Islami,
k-Plex9	3	0.667	2.8	Tahwur Hussain Rana
				Ilyas Kashmiri, David Headley, Tahwur Hussain Rana,
k-Plex10	4	0.833	3.333	Sajid Mir
k-Plex11	3	0.667	1.75	Ilyas Kashmiri, David Headley, Drug Enforcement Agency
k-Plex12	3	0.667	1.75	Ilyas Kashmiri, David Headley, Bollywood
				David Headley, Tahwur Hussain Rana,
k-Plex13	3	0.667	2	Abdur Rehman (Pasha)
				David Headley, Tahwur Hussain Rana,
k-Plex14	3	0.667	2.333	Drug Enforcement Agency
k-Plex15	3	0.667	2.333	David Headley, Tahwur Hussain Rana, Bollywood
k-Plex16	3	0.667	2.333	David Headley, Sajid Mir, Abdur Rehman (Pasha)
k-Plex17	3	0.667	2.8	David Headley, Sajid Mir, Drug Enforcement Agency
k-Plex18	3	0.667	2.8	David Headley, Sajid Mir, Bollywood
				David Headley, Abdur Rehman (Pasha),
k-Plex19	3	0.667	2.8	Drug Enforcement Agency
k-Plex20	3	0.667	2.8	David Headley, Abdur Rehman (Pasha), Bollywood
k-Plex21	3	0.667	3.5	David Headley, Drug Enforcement Agency, Bollywood
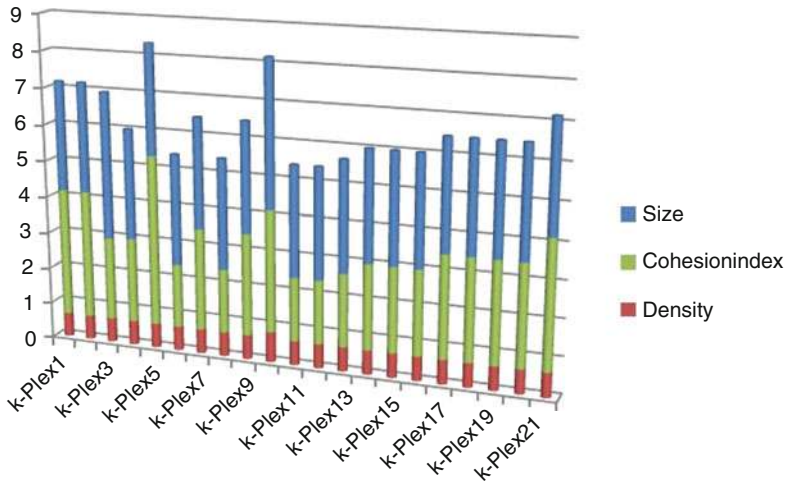
Fig. 4 Subgroups (k-plexes) in the David Headley network

bars representing that these are the most connected and important parts of the graph. The other valuable information that these k-plexes identify is that Kashmiri is the channel of communication between the terrorist organization and the other members of the graph as he is the only common member in both k-plexes. If we analyze k-plex5 alone it contains the members like Al-Qaeda and Huji (Harkat-ul-Jihad-e-Islami) along with Kashmiri, thus the members of this cell have the expertise to carry out a terrorist attack.

In case of k-plex10, it contains people who can be associated with strategic analysis and planning. They do not have expertise and experience to carry out terrorist attacks directly. Thus, analysts (planners) are separate from attackers as k-plex10 has people who can carry out analysis, initial investigation, and surveillance to prepare a master plan of the attack and k-plex5 can really implement such a plan. As in any project, the project manager is the person who coordinates the analysis and implementation teams of the project. This analysis shows that a similar role was played by Kashmiri in this conspiracy.

Our analysis points out a new dimension in criminal investigation and counter-terrorism research which indicates that terrorist organizations are using third party surveillance teams for the investigation and analysis part of their attack. These third party teams consist of people that are not linked with any sort of terrorism directly – possibly white collar people that have passports of countries like Canada or even the United States. Therefore, law enforcement agencies have some sort of trust in them and they can freely travel to any destination to scout. It is possible that multiple terrorist organizations share such teams each for their own cause. In the David Headley network, terrorist organizations such as Al Qaeda, HUJI and LeT (Lashkar-e-Taiba) may be sharing the same group of people for surveillance.

Fig. 5 Community structure in the David Headley network

This theory is also supported when we weigh the nodes of the graph with the CNM algorithm [5] for detecting the communities. The application of the CNM algorithm on the Headley network resulted in the graph containing three communities as shown in Fig. 5 (the communities are named G1, G2, and G3). G3 is the community of analysts and surveyors and this group consists of people that would not be suspected to be terrorists – people that are connected with famous people, that have worked for the law enforcement agencies, and that are influential enough to getting clearance (visa) to visit any country. G1 and G2 are the communities containing the terrorist organizations that are using the G3 community for the best of their interest – in this case terrorism.

Figure 6 shows the information flow hierarchy constructed by iMiner [18] with the help of dependence centrality [18]. The figure depicts that information flow from David Headley to Ilyas Kashmiri and similarly, Tahwur Hussain Rana also coordinated with Sajid Mir regarding the operation. In reality, David Headley has visited the targeted locations to collect information by surveying the potential
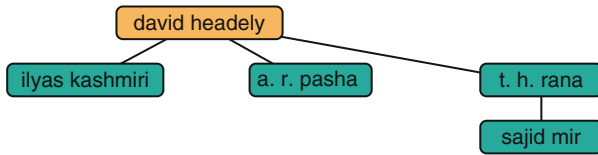
Fig. 6 Information flow hierarchy estimated by iMiner [16]

targets and surroundings, and is said to have visited Pakistan in order to meet Ilyas Kashmiri and other Lashkar-e-Taiba (LeT) members [23]. Thawur Hussain Rana was the owner of the company, for which David Headley was working. He was not directly involved in communication with Lashkar-e-Taiba, but was obviously informed about the proceedings by David Headley transmitting the information to other sources. Thus, the information flow depicted by this figure is very close to reality.

5 Conclusions and Future Work

The paper has presented the analysis of a recent terrorist plan against a target in Denmark. Investigative data mining techniques were used for analyzing the terrorist network which was centered on David Headley. The data used to construct the terrorist network were collected from publicly available information. Based on the findings discussed, the contribution of the paper is twofold:

1. We have shown that an investigative data mining system (iMiner [16]) can be used to harvest relevant information about a particular terrorist case from open sources and that a terrorist network can be constructed based on open source information. The constructed terrorist network can also be analyzed to reveal relevant information about the people participating in the network and their connections – such as detecting key persons and subgroups of the network.
2. Our analysis moreover indicates a new dimension of how terrorist organizations are using third party surveillance teams for the investigation and analysis part of their attacks. This is in contrast to, for instance, the 9/11 and other past incidents. Currently, there is much focus on suspected terrorists, which makes it difficult for them to travel to locations and plan attacks. Therefore, involvement of third party (white collar) people with no prior record of terrorism related activities may be a new trend that will be generalized in the future as terrorist organizations try to deal with the increased levels of surveillance and monitoring by many countries. The information flow theories and techniques such as dependence centrality can be used to identify such third party surveyors as they are not involved in terrorism directly, but are major stakeholders in information sharing processes and preplanning of terrorist attacks.

We are working on some new mathematical models which may help in detecting white collar terrorism think-tanks. As the most of the work in terrorist network analysis is borrowed from SNA, it is of crucial importance to have the researchers work with the experts of the intelligence world in order to design and develop new models to predict terrorist threats.

References

1. Sami, A.: Story of David Headley-suspect of planning attacks on Danish newspaper. International The News, November 22, 2009 available online at http://www.thenews.com.pk/updates.asp?id=91981
2. An article published in a comprehensive Punjabi journal "Wichar". http://www.wichaar.com/news/286/ARTICLE/17345/2009-11-20.html
3. An article published in Deccan Herald. http://www.deccanherald.com/content/40626/lashkar-e-taibas-us-connection.html
4. Barlow, M., Galloway, J., Abbass. H.: Mining evolution through visualization. In Proceedings of Workshop on Beyond Fitness: Visualization Evolution at the 8th International Conference on the Simulation and Synthesis of Living System. (2002) http://www.alife.org/alife8/workshops/15.pdf
5. Clauset, M.E.J. Newman, Moore, C.: Finding community structure in very large networks. Physical Review E, 70:066111, 2004. http://pre.aps.org/pdf/PRE/v70/i6/e066111
6. Chen, H., Chung, W., Xu, J.J., Wang, G., Qin, Y., Chau, M.: "Crime Data Mining: A General Framework and Some Examples," Computer, vol. 37, no. 4, pp. 50–56, Apr. 2004, doi:10.1109/MC.2004.1297301
7. DeRosa M.: Data Mining and Data Analysis for Counterterrorism, CSIS Report. (2004)
8. Fershtman, M.: Cohesive group detection in a social network by the segregation, Social Networks 19, 193–207 (1997)
9. Granovetter, M.: The Strength of Weak Ties. Am. J. Sociol. **81**, 1287–1303 (1973)
10. Granovetter, M.: The Strength of Weak Ties: A Network Theory Revisited. In: Collins, R.(ed.) Sociological Theory, pp. 105–130 (1982)
11. Krebs, V.E.: Mapping network of terrorist cells. Connections 24(3): 43–52 (2202)
12. Memon, N., Larsen H.L.: Practical approaches for analysis, visualization and destabilizing terrorist networks. In: The proceedings of ARES 2006: The First International Conference on Availability, Reliability and Security, Vienna, Austria, IEEE Computer Society, pp. 906–913 (2006)
13. Memon, N., Larsen, H.L.: Practical algorithms of destabilizing terrorist networks. In the proceedings of IEEE Intelligence Security Conference, San Diego, Lecture Notes in Computer Science, Vol. 3976: pp. 398–411. Springer, Berlin (2006)
14. Memon, N., Larsen, H.L.: Detecting Terrorist Activity Patterns using Investigative Data Mining Tool. International Journal of Knowledge and System Sciences, **3**(1), 43–52 (2006)
15. Memon, N., Qureshi, A.R.: Destabilizing terrorist networks. In WSEAS Transactions on Computers. **11**(4), 1649–1656 (2005)
16. Memon, N.: Investigative data mining: Mathematical models for analyzing, visualizing and destabilizing Terrorist Networks. PhD Dissertation, Aalborg University, Denmark. (2007)
17. Memon, N., Hicks D., Larsen H.L.: Harvesting Terrorist Information from Web. In proc. International Conference on Information Visualization (IV 2007), Zurich, Switzerland, July 4–6, 2007, pp. 664–671 (2007)
18. Memon, N., Hicks D., Larsen H.L.: How investigative data mining can help intelligence agencies to discover dependence of nodes in terrorist networks. Advanced Data Mining and Applications. Lecture Notes in Computer Science, 2007, vol. 4632/2007, pp. 430–441 (2007). DOI: 10.1007/978-3-540-73871-8-40

19. Monge, P.R., Contractor, N.: Theories of Communication Networks. Oxford University Press, New York (2003)
20. Memon, N., Kristoffersen, K.C., Hicks, D.L., Larsen, H.L.: Detecting Critical Regions in Covert Networks: A Case Study of 9/11 Terrorists Network. In: Proceedings of International Conference on Availability, Reliability, and Security 2007, Vienna, Austria, March 10–13, 2007 (2007)
21. Taskar B., Abbeel, P., Wong, M.-F., Koller, D.: Label and link prediction. In relational data. In: IJCAI Workshop on Learning Statistical Models from Relational Data. (2003) http://kdl.cs.umass.edu/srl2003_upload/files/taskar-paper.pdf
22. Q&A with Professor Karen Stephenson, April 18, 2006 http://www.elearningpost.com/articles/archives/qa_with_professor_karen_stephenson/
23. U.S. Department of Justice Press Release, January 14, 2010 available at http://www.justice.gov/opa/pr/2010/January/10-nsd-038.html
24. Xu, J., Chen, H.: Criminal network analysis and visualization. Commun. ACM 48, 6 (June 2005), 100–107 2005 DOI=10.1145/1064830.1064834 http://doi.acm.org/10.1145/1064830.1064834

Part IV
Alternative Perspective

The Ultimate Hack: Re-inventing Intelligence to Re-engineer Earth

Robert David Steele

Abstract It is for me a huge honor to be invited to deliver the counter-balancing view within this distinguished gathering of counter-terrorism experts. Let me begin by emphasizing what I consider to be the "three strikes" or in international terms, the "side-out" of government counter-terrorism thinking today.

Strike One: Terrorism is a symptom, not a root cause or threat. Going after Al Qaeda has nothing to do with the more threatening fact that a corrupt Saudi monarchy embraces infidels at the same time that it is destroying its own commonwealth and sponsoring Wahhabism world-wide.

Strike Two: No government is trained, equipped, nor organized to understand, much less defeat, terrorism in detail. Although the United Nations (UN) has recently discovered Coherence, with the implementation meme of "Deliver as One," and the US Government claims to be interested in Whole of Government operations, neither is actually real (yet).

Strike Three: Earth is at a tipping point, with multiple catastrophic possibilities, all of them interacting in complex unpredictable ways. In this context, terrorism is the canary in the coal mine, not the lethal atmosphere itself.

I will review these, and then outline a strategy for all of us that addresses all ten high-level threats to humanity within which terrorism is but number nine, and then only because of a potential mass destruction event.

R.D. Steele (✉)
Earth Intelligence Network, Oakton, VA, USA
e-mail: robert.david.steele.vivas@gmail.com
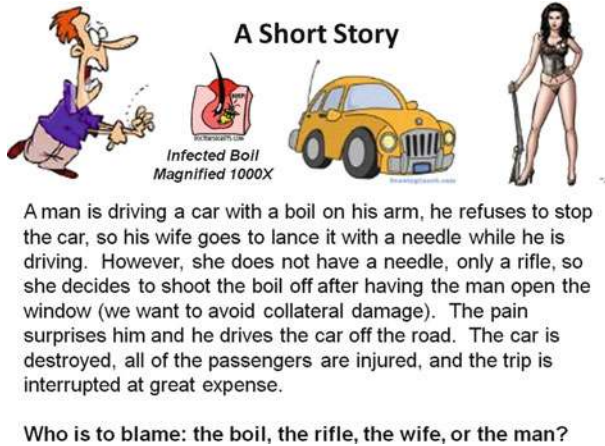
Fig. 1 Terrorism is a boil – a symptom

1 Strike One: Symptom Not Root Cause

Consider, please, the illustration in Fig. 1 and the short story that accompanies it-terrorism is a boil, nothing more.[1]

I was inspired to develop this example after interviewing Steven Carmel, a senior executive with Maersk Line Limited, who pointed out to me in an interview for my latest book [39], that most terrorist or piracy events were in effect a minor traffic accident, and that it was the terrible decisions of the U.S. Government that were costing the shipping industry billions of dollars. What we do in the way of counter-terrorism is vastly more costly and more harmful to global and local societies than terrorism.

2 Strike Two: Politicized Ignorance, Institutionalized Incompetence

Figure 2 is the first of two graphics on this shortfall. We simply do not have access to, nor the ability to understand, most of the available relevant information necessary to achieve good governance at all levels.[2]

Figure 2 is adapted with permission. I have added a depiction of how Google's programmable search engines can show you not what you need to know, but what

---

[1]The original briefing as delivered with words in Notes format, can be found at http://www.tinyurl.com/SteeleDK5. A longer earlier version for engineers as delivered at the University of British Columbia can be found at http://www.oss.net/HACK.

[2]Stephen Arnold, "Search panacea or ploy: Can collective intelligence improve findability?," in Tovey [43], pp. 375–388. Google portion added to this depiction.
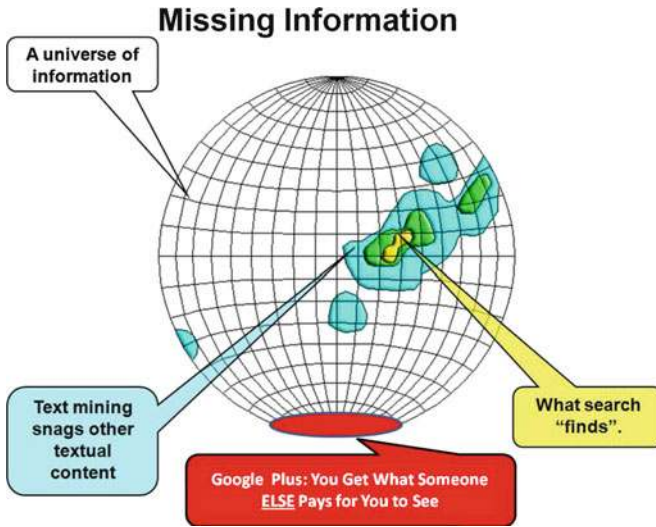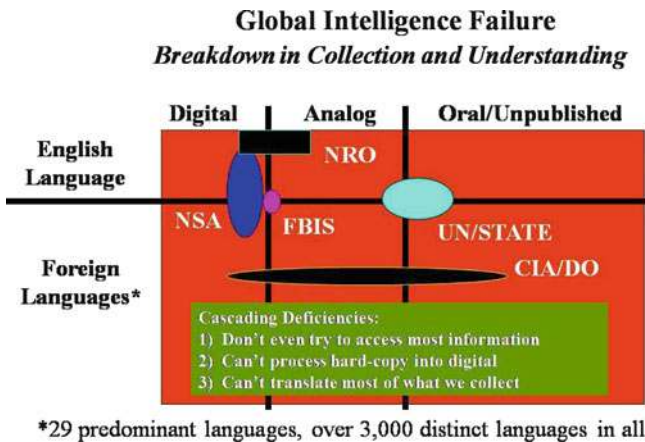
Fig. 2 Missing information

Fig. 3 Global intelligence collection failure

someone else is paying for you to see. Still and all, even with the best of intentions and the application of all available technologies, we see a fraction of the relevant information-most of it unpublished or analog and in languages we do not read or understand.

The second graphic (Fig. 3), created in the 1990s, is still valid today, and shows the degree to which secret sources and methods fail to capture all of the relevant information in many foreign languages. Back then we focused on 29 languages and then on 33 languages, as well as twelve distinct versions of Arabic. Today the essential number is 183, with growing awareness of the urgency of capturing historic

indigenous knowledge of sustainable agricultural and water conservation practices we have ignored for centuries.

My second graduate thesis[3] captured the essence: the average Embassy accesses roughly 20% of the relevant information, and in the process of sending it back to their varied home agencies, spill 80% of that – Washington – and other capitals – are operating on 2% of the relevant information.

This politicized ignorance and institutionalized incompetence is tolerated because political ideology, not reality-based policy, is the dominant governance paradigm.

Before proceeding to discuss Strike Three, we must first elaborate on Strike Two, as political ignorance and institutionalized incompetence are condemning nation-states to bankruptcy, separatism, and internal strife.

Our political pathology consists of several related pathologies:

- Political segmentation within a two-party tyranny
- Disconnect between revenue (means) and ways and ends
- Stovepiped perspectives enabled by both of the above
- Fragmentation of knowledge across all disciplines and domains
- Persistent data pathologies and information asymmetries

All of the above can be viewed as a meta-architecture within which counter-terrorism as a focus of effort is doomed to failure. Every element of counter-terrorism, from the secret and open source intelligence needed to understand it to the political and financial arrangements needed to construct programs to deter, prevent, and recover from counter-terrorism, to the dissemination of knowledge to help citizens and private sector organizations prosper in a world where terrorism is now a "given," these are all severely distorted, divorced from reality, and therefore lacking integrity in the "whole system" sense of the word. We cannot "do" counter-terrorism absent major changes in our system.

2.1   Political Segmentation Within a Two-party Tyranny

The United States of America (USA) is in my view no longer a full democracy, but rather a hybrid regime in which a corrupt Federal Reserve serves as the front for a corporate state that has captured both Congress and the legislature. The USA is a two-party tyranny that no longer abides by the principles of the Constitution. Figure 4 is one depiction.[4]

---

[3]Robert Steele, National Security C3I3H3-Command, Communications, and Computing, Inter-Agency, Inter-Disciplinary, Inter-Operability, Heuristics of the Community Intelligence Cycle (Norman, OK: University of Oklahoma, unpublished MPA thesis, 1987); online at http://tinyurl.com/Steele1987.

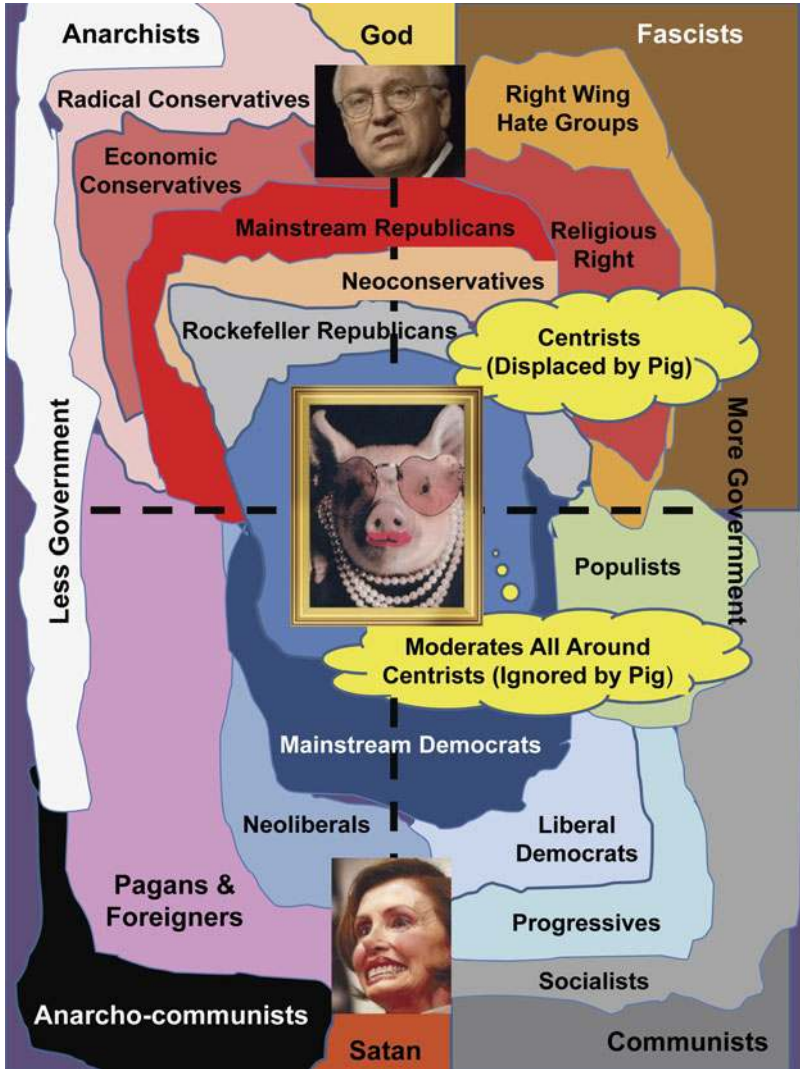[4]Adapted with permission from Crane [12]. See also Amato [2].

Fig. 4 Political segmentation in a two-party tyranny

2.2 Disconnect Among Revenue (Means), Ways, and Ends

The failure of Congress, when combined with the failure of the Executive, has created a government in which the taxes that citizens and small businesses pay are earmarked by a corrupt Congress toward equally corrupt corporations – how the U.S. Government (USG) spends money has little to do with the public interest and all to do with perpetuating a two-party tyranny that feeds a corporate regime

Fig. 5 Global terrorism in perspective

practicing unilateral militarism, virtual colonialism, and predatory capitalism – in effect, a global reign of terror in multiple forms.

Figure 5 integrates a few book covers [7, 9, 19, 22, 24, 26, 28, 30] representative of the pathologies within our system that make terrorism inevitable, and a rough depiction of the disconnect between citizens paying taxes, and how the tax revenue is spent.

2.3 Perspective Stovepiping Enabled by Both of the Above

This global reign of state-based and corporate-driven terrorism is made possible by the concentration of wealth after it leaves the public's pockets, and by the fragmentation of perspective and knowledge. We have lost all perspective including the all-important moral perspective. Here and on the next page are effect and cause.

The point in Fig. 6 is that all of these distinct communities are talking past each other – failing to communicate, they cannot make sense or counter terrorism by agreeing to eradicate poverty and other roots of violent rebellion.
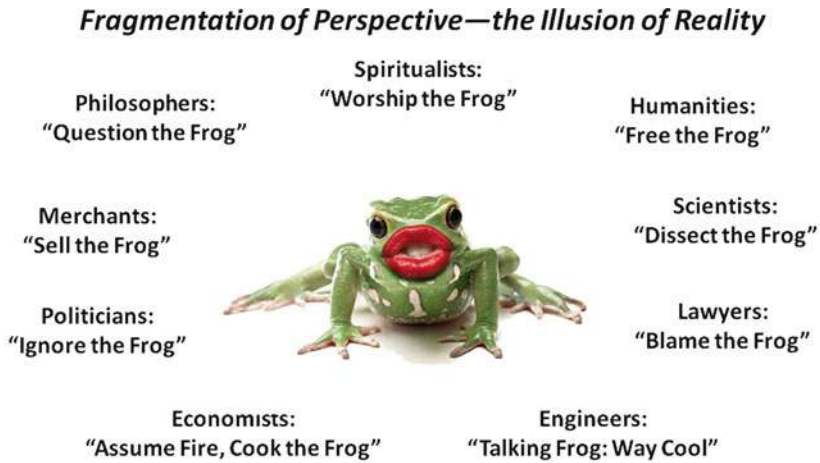
Fig. 6 Empire of illusion (Cf. [15, 35])

2.4 Fragmentation of Knowledge Across All Disciplines and Domains

Here are illustrated the cause of our impotence as citizens, not just in the USA, but around the world: the poor are illiterate while the rich are ignorant (Fig. 7).

Figure 7 was created by Kevin Boyack in partnership with Dick Klavans and Brad Ashcroft, themselves pioneers in citation analysis, scientific and technical forecasting, and commercial intelligence.[5]

2.5 Persistent Data Pathologies and Information Asymmetries

The fragmentation illustrated in Fig. 7, in combination with corrupt governments that allow ideology to displace intelligence and technology to displace thinking, leads to multiple data pathologies and information asymmetries (Fig. 8), all of them very costly to the public and the Earth.

---

[5]Cf. http://www.sandia.gov/news/features/mapping_science.html# and also see http://scimaps. org/maps/map/maps_of_science_fore_50/.
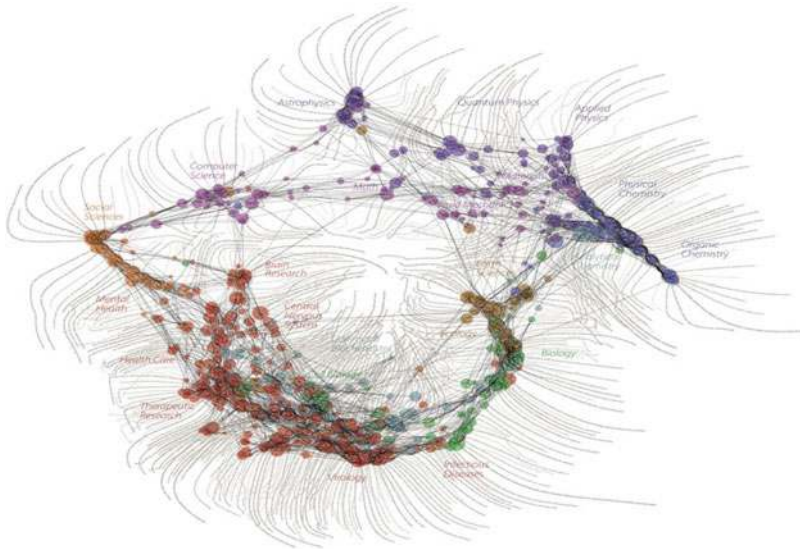
Fig. 7 Web of fragmented knowledge

Fig. 8 Data pathologies and information asymmetries

I devised this modest typology based on the illustrated works by others [5, 11, 16, 24, 25, 29, 31, 34], and have recently added Forgotten Knowledge.

All of the above prevent clarity, diversity, and integrity from achieving sustainability - they create lose–lose rather than non-zero outcomes for most.

Now, finally, we come to Strike Three. It is in the context of the multiple potential catastrophes that face the Earth that it is easiest to see that terrorism – and therefore counter-terrorism – are minor isolated and largely cosmetic symptoms of a system failure so large, so pervasive, and so imminent as to render completely irrelevant

Fig. 9 Map of world conflict and human rights violations

all efforts to "do" counter-terrorism. In this context, trying to do counter-terrorism is the equivalent of trying to design a better race car by focusing only on oil leaks. Figure 9 is a general depiction.

Created by Berto Jongman (NL), this map [20] is no longer updated annually but remains one of the best single depictions of the complexity and fragility of the human condition around the globe. It is a useful context for thinking about high-level threats to humanity.

3 Strike Three: Earth at Tipping Point, High-level Threats to Humanity

Ever since Lakhadar Brahimi (AG) led the Panel on UN Peace Operations (2000), the UN has been evolving. Two other panels have been especially important: the Panel on High-Level Threats, Challenges, and Change (2004), and the High-Level Panel on System-Wide Coherence (2006). Prior to the 2004 report [44] from the UN High-Level Panel on Threats, Challenges, and Change, no one had ever put together a credible list of the top threats to humanity, nor placed them in any sort of priority order. This panel did both, and the fact that LtGen Dr. Brent Scowcroft, USAF (Ret),

Poverty
Infectious Disease
Environmental Degradation
Inter-State Conflict
Civil War
Genocide
Other Atrocities
Proliferation
Terrorism
Transnational Crime

Fig. 10 Ten high-level threats to humanity

was the participant representing the USA, may help explain both the coherence of the report, and the lack of US ideological interference with the findings.

I list the ten high-level threats to humanity here in their original priority order (Fig. 10).

This is not the place for a detailed discussion of these threats, other than to point out that "terrorism" is on this list solely because of its potential to cause mass casualties with nuclear, biological, chemical, or radiological weapons – something the USA corporate state does on a daily basis.

It also bears emphasis that poverty is the root condition accelerating and extending all of the other threats to humanity.

Taken together, the three endeavors represent the evolution of the UN away from a Member-dependent body for consultations, and toward an intelligence-driven body for harmonizing spending and actions in the field by both the UN and by others. On the next page is a coherent view of the ten threats combined with twelve core policies and eight demographic challengers (Fig. 11).[6]

In this context terrorism is no more than one-tenth of the threat. Coherence demands that the UN, Members states, and all other parties not only understand the "true cost" of every course of action (and spending) under consideration, but that they understand this in a systemic context.

As Russell Ackoff [1] has taught us, and Buckminster Fuller [18] would heartily agree – what is good for one part of the system may be very bad for another part of the system. It is all about balance, and balance cannot be achieved without both understanding by all parties, and harmonization of spending and behavior by all parties. Figure 12 is one means of seeing the Whole Earth.

As with Fig. 11, a red circle shows where the counter-terrorism investment is in terms of both relevance and scope. It is both excessive in cost within the USA – and negligible to counter-productive in relation to all else. It is helpful to appreciate the time-energy-cost of each alternative option for policy-spending emphasis.

---

[6]As devised by the 26 co-founders of the Earth Intelligence Network (EIN), at http://www.earth-intelligence.net. Its front end is http://www.phibetaiota.net.
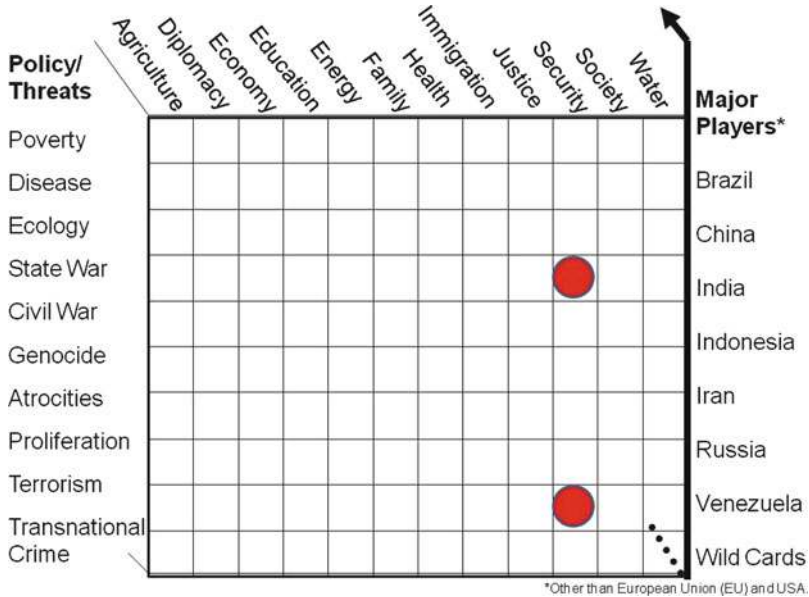
Fig. 11 Strategic analytic matrix for achieving coherence
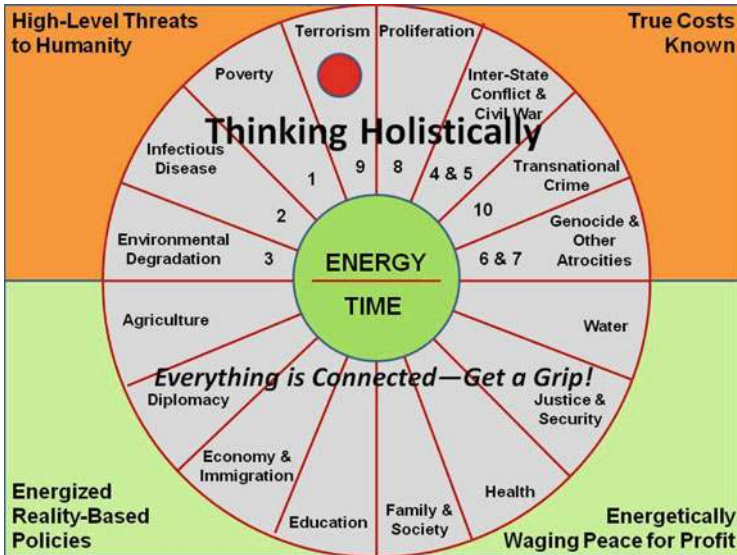

Fig. 12 Coherence circle for planning and operations

There are just a few points that need to be made as we flesh out the context within which most counter-terrorism plans, programs, budgets, and operations are next to worthless in terms of achieving sustainable outcomes.

Keeping this at a very high level:

- Changes to the Earth that used to take 10,000 years now take three. Education, Intelligence, and Research (EIR) must move as quickly as possible toward real-time science and real-time policy.[7]
- Catastrophic accelerators are everywhere – from paving over the wetlands to storing chlorine in massive rusted tanks above super-highways; from cow shit into our spinach to cow methane into our atmosphere; from toxins and vaccines to new forms of resistant bacteria – we have created a petri dish for human extinction.[8]
- Nothing the USA does in the next 20 years is going to make a difference unless it decides to be serious about EIR, and foster a global environment in which every person has a cell phone, call centers educate the poor one cell call at a time, and the emphasis is on helping the eight demographic powers create infinite sustainable wealth, achieving the non-zero solution.[9]

3.1 What Is to Be Done?

The UN, despite its being an unhinged bureaucracy,[10] is on the right track. Intelligence is still a dirty word and most UN "elements" (none of them under the effective "command and control" of the Secretary General) would prefer to avoid any semblance of coherence. One third of the UN staff is comprised of individuals hired under a nepotism – patronage system, and one third of the UN staff appears to be professional intelligence officers whose invitations to lunch are always welcome, but whose incompetence at their cover jobs places a further burden on the remaining one third of the UN staff that is the "real deal."

HOWEVER, significant progress has been made in three areas.

First, the value of intelligence as decision-support is now understood by the more serious and committed among the UN leadership. The Department of Peacekeeping

---

[7]I cannot find the specific book, and since George Mason University accepted my entire library as a gift when I joined the UN, I am at a loss. I was influenced by Brand [8].

[8]Cf. Perrow [30]. His book, especially, relegates terrorism to the minor league in comparison with what our own nuclear, chemical, biological, and radiological industries do every day.

[9]I am indebted to Tom Atlee [3, 4] – his mentoring changed my life. This specific point is made by a book he guided me to, Wright [47]. See also Benckler [6], Carter [10], Stewart [40], and Toffler [42].

[10]The Secretary General has little meaningful authority over anything, and no authority at all over the Specialized Agencies (SA) that comprise the bulk of the UN. The UN "management system" is mostly about moving internal information (inputs and outputs), not about managing and certainly not about forecasting and adapting.

Operations (DPKO) is the most serious of all, moving forward in the aftermath of the many constructive initiatives undertaken while MajGen Patrick Cammaert, RN NL (Ret) was Military Advisor to the Secretary General and then Force Commander, Eastern Congo. The Joint Operations Centres (JOC) and the Joint Military Analysis Centres (JMAC) are now standard. The Department of Political Affairs (DPA) and the Department of Safety & Security (DSS) are beginning to hire analysts, but they do not fully understand the intelligence process[11] and they especially do not understand how to maintain the integrity of the intelligence human, technical, and analytic processes.

Second, the Department of Field Support (DFS) has been split off and enhanced to the point that every element of the UN System that is in the field can rely in the future on world-class communications, computing, and logistics support. Eventually, this could mean that the United Nations Open-Source Decision-Support Information Network (UNODIN) would be integral to every UN office at every location.[12]

Third, and for the first time, the UN has created a hybrid organization, the International Commission Against Impunity in Guatemala (CICIG). It is a hybrid organization because it was chartered to be an arm of the government it is supporting, with extraordinary intelligence and counterintelligence powers to include full access to human and technical sources, and also the unique right to polygraph all personnel, both international and local, both assigned to CICIG and within the government, so as to assure the integrity of the mission.[13]

It is my view that the Swedish military, long in the forefront of peacekeeping and peacekeeping intelligence, has put forth the correct solution: the future of international peace and prosperity is a hybrid future, in which the Swedish concept, as modified by EIN, provides the coherence: Multinational, Multiagency, Multidisciplinary, Multidomain Information-Sharing and Sense-Making (M4IS2).

---

[11]Requirements definition; collection management; source discovery and validation; multi-source automated and human fusion; application of human expert judgment; visualization; and compelling timely actionable presentation. Cf. the various handbooks available at http://www.phibetaiota.net/category/handbooks.

[12]As discussed in [39].

[13]As the US prepares to abandon Afghanistan, one book Ossman [27] posits a UN hybrid International Reconciliation and Reconstruction Agency (IRRA) staffed almost exclusively by Muslim experts on detail from Indonesia, Iran, Malaysia, Nigeria, and Turkey. Such an organization, as an arm of the Afghan government but empowered with the polygraph and able to control corruption while avoiding the inevitable baggage of "infidels," would be the new model for addressing failed states, and readily adapted to the needs of Haiti, Somalia, Yemen, and others. What has become clear is that neither governments nor non-governmental organizations such as the Red Cross can be relied upon in isolation. Coherence demands a hybrid approach that enables intelligence-driven harmonization of effort by all parties, while also providing the integrity protection of counter-intelligence.
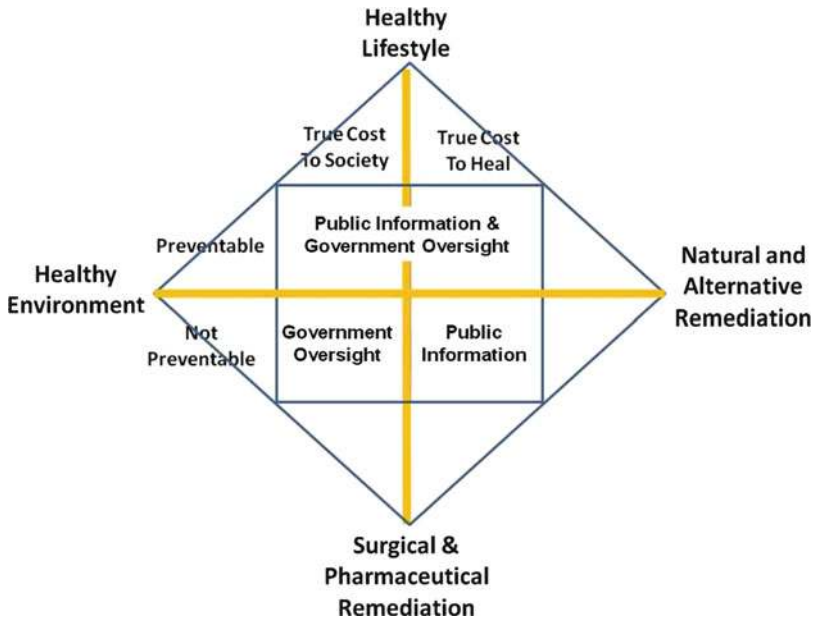
Fig. 13 Analytic strategy for health and public information

3.2   Creating the World Brain and Global Game[14]

Terrorism, to use another analogy, is a cancer within the body of humanity. Pursuing that analogy, Fig. 13 is an illustration of the four distinct aspects of health.

Our approach to counter-terrorism, as with our approach to health or any other topic, is not just fragmented, it is focused on only one of the four critical aspects of achieving a healthy body politic. To defeat terrorism, we must eradicate the underlying conditions, including corruption, dictatorships, and poverty, that alienate the public from governments and corporations that are predatory and therefore lacking in legitimacy, concepts examined in Manwaring [23].

3.3   Whole Systems and M4IS2 Information Exploitation

The balance of this contribution will focus on how to eradicate terrorism, along with the other nine high-level threats to humanity, by using a Whole Systems approach, and the new M4IS2[15] concept for harmonizing how all relevant actors spend and

---

[14]Robert David Steele, "World Brain as EarthGame**TM**," in Tovey [43], pp. 389–398.

[15]M4IS2: Multinational, Multiagency, Multidisciplinary, Multidomain Information-Sharing and Sense-Making.

Agriculture
Diplomacy
Economy
Education
Energy
Family
Health
Immigration
Justice
Security
Society
Water

Fig. 14 Core policies for information-enabled harmonization

behave across the twelve core policies. I list them in Fig. 14 for emphasis. This is a minimalist approach, a simple beginning.

Here are the elements of the World Brain and Global Game that the collective minds of the 26 co-founders of EIN have developed[16]:

- Universal strategy
- Information operations cube
- Four quadrants from knowledge to intelligence
- Fifteen slices of human intelligence (HUMINT)
- Six bubbles for digital information exploitation
- Global to local range of needs and gifts table
- Intelligence maturity scale

3.4 Universal Strategy

Ethics – the heart of what Will Durant addressed in his 1916 doctoral thesis and seminal work, *Philosophy and the Social Problem* [13] is how civilization can evolve with each generation learning from the past at less cost of blood, treasure, and spirit. Morality, he and Ariel Durant tell us in *Lessons of History* [14] is a priceless strategic asset. The "Golden Rule," common to all religions, can be universally implemented in a non-zero manner (Fig. 15).

3.5 Information Operations Cube

Earlier we addressed the pathologies that have to this day concentrated wealth while fragmenting knowledge. The catastrophic separation of the sciences from

---

[16]Earlier summaries include the EIN brochure at http://www.earth-intelligence.net, and the chapter "World Brain and EarthGame™" [43]. EarthGame™ is trademarked to EIN co-founder Medard Gabel, Global Game is a more generic term.
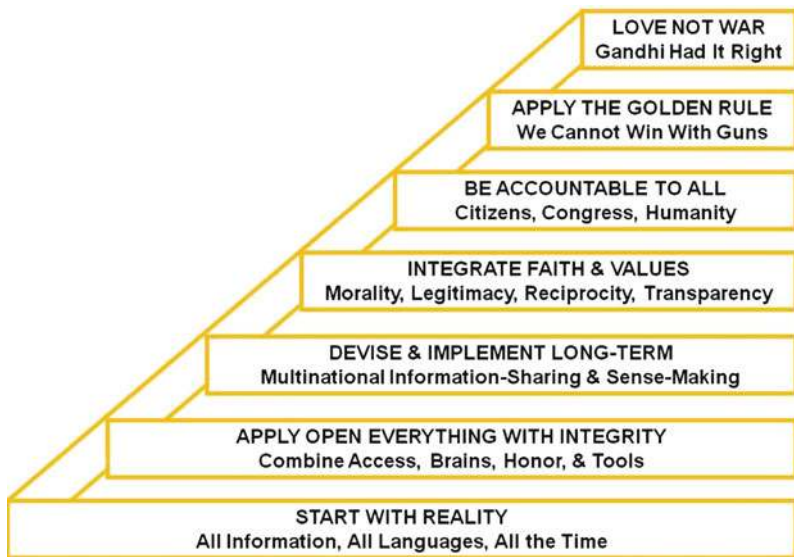
Fig. 15 Strategy to create a prosperous world at peace

the humanities [46], the retardation of the social sciences, the vanishing of history, and the proliferation of data pathologies and information asymmetries must all be reversed if we are to implement the strategy. Figure 16 is one way of looking at how we must respect all information in all languages all the time.

Figure 16 is intended to show both the diversity of points of view that exist, and that must be reconciled in order to arrive at a "best truth."

3.6 Four Quadrants from Knowledge to Intelligence

Appreciating the importance of accessing and exploiting all information in all languages all the time [37, 38], we now turn to the four quadrants, each representative of a stage of modern civilized information management (Fig. 17).

Most organizations are in Quadrant I, *avant guarde* individuals in Quadrant II. Very few are deeply into Quadrant III, and virtually no one is in Quadrant IV.

Here again, Fig. 17 strives to illustrate the diversity of opportunities to communicate and make sense that must be integrated to achieve full capacity.

3.7 Fifteen Slices of HUMINT

The on-going dialog over creating Whole of Government capabilities that can apply "soft power" is at root about integrating all human capabilities and enabling the sharing of all information across all human slices. From an expeditionary or
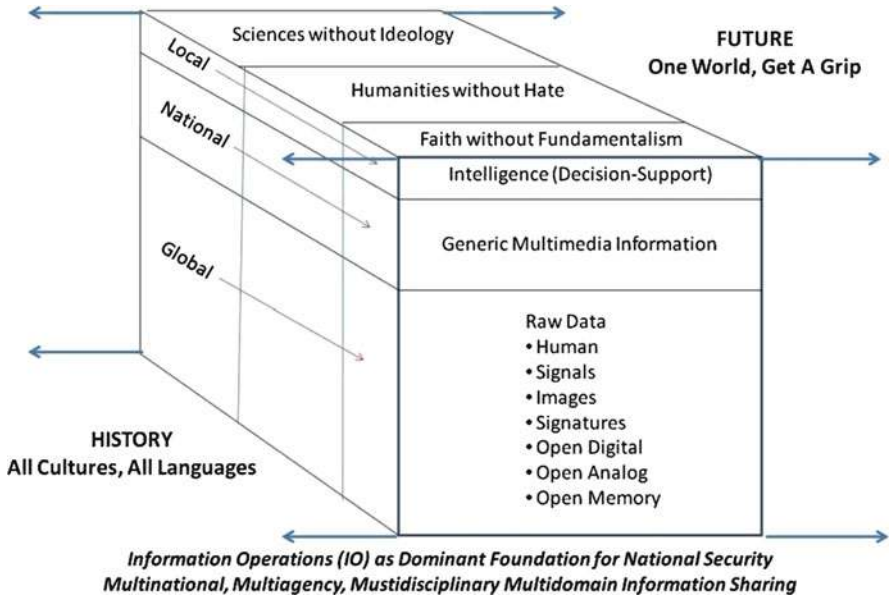
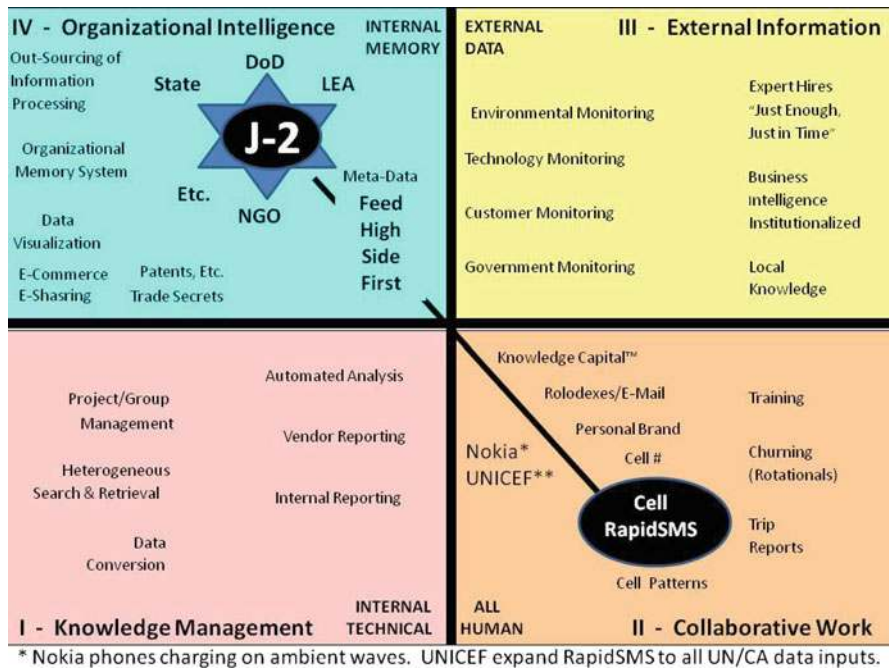**Fig. 16** Information operations for world prosperity and peace

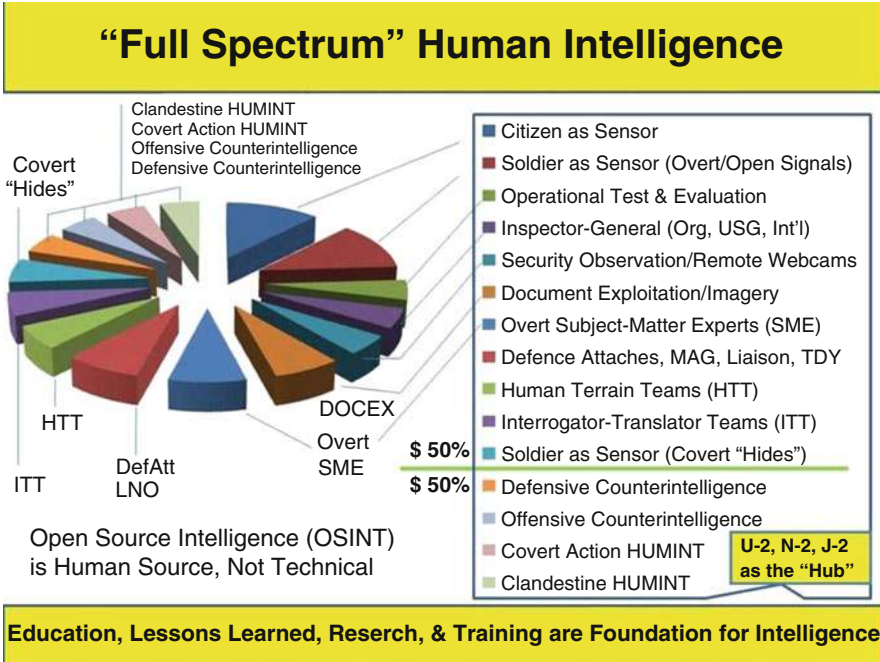**Fig. 17** Knowledge management to organizational intelligence

Fig. 18 Achieving "full spectrum" human intelligence

"terrorism-prevention" point of view, Fig. 18 shows fifteen slices of HUMINT that are not being harnessed in tandem today because we have governments focused on turf and things rather than on "making sense" [36].

90% or more of what can be known and shared in not federal, not expensive, and not secret. I have written and spoken extensively on this reality.

3.8 Six Bubbles for Digital Information Exploitation

Inspired by the UN High-Level Panel on Threats, Challenges, and Change, I funded and organized EIN to reflect on how to change the information and intelligence paradigm so as to achieve full-spectrum understanding. We concluded that the cell phone, not the laptop, was the device of the future, and that RapidSMS and Twitter were promising signs of the future of information sharing and sense-making. We can build this today (Fig. 19).

Both oligarchs and governments are fearful of this, not realizing that their current wealth is insignificant and not worth redistributing – only the creation of infinite wealth sufficient for all will do – this does that – and non-violently.
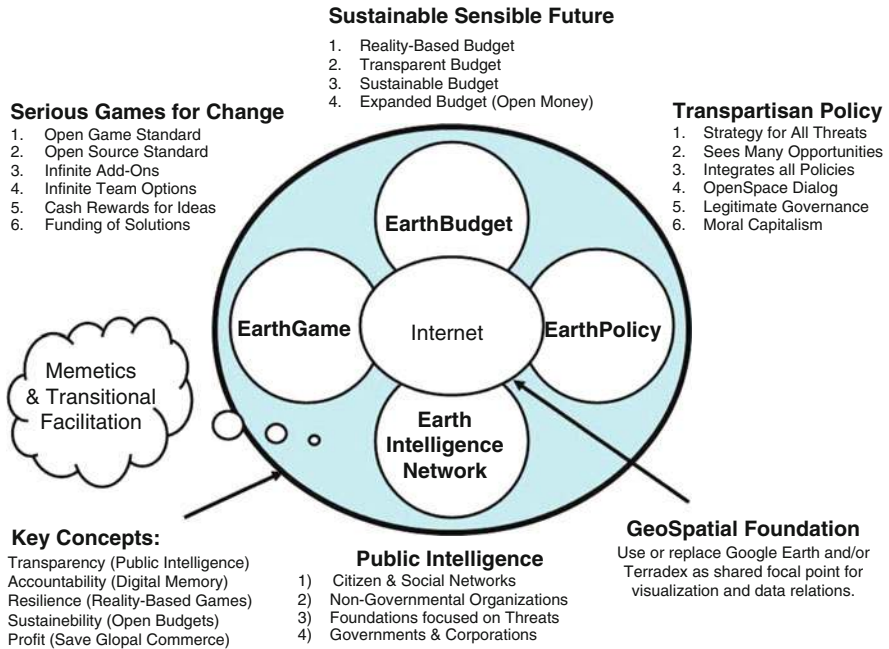
Fig. 19 Connecting all humans to all information

3.9 Global to Local Range of Needs and Gifts Table

One insight that we achieved in the process of studying data pathologies and information asymmetries as facilitated by fragmented knowledge and pathological political arrangements, was that corruption and intermediaries are virtually synonymous – corruption here refers to both witting and unwitting misallocation of resources. We concluded that enabling intelligence-driven micro-giving and *ad hoc* social networking focused on "one gift at a time," was the way to create a prosperous world at peace (Fig. 20).

At the same time, we realized that for lack of transparency, major investments by the eight tribes of intelligence suffered for lack of harmonization.

3.10 Intelligence Maturity Scale

Our final insight was based on the realization that secret intelligence is very immature as well as very wasteful – in the USA $75 billion a year produces "at best" 4% of what the President and a tiny handful of others require in the way of decision-support. Here, we capture the needed change of perspective (Fig. 21).
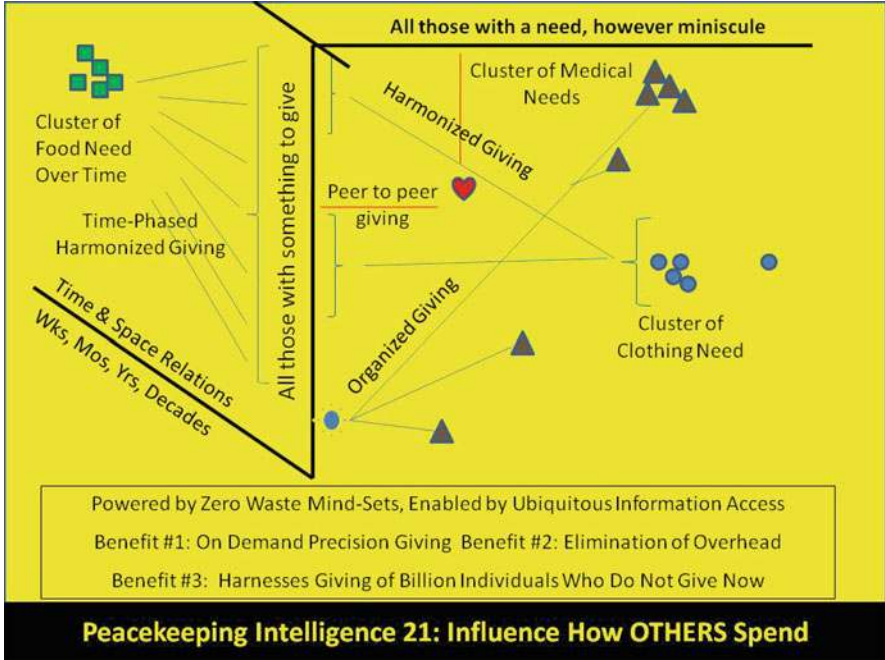
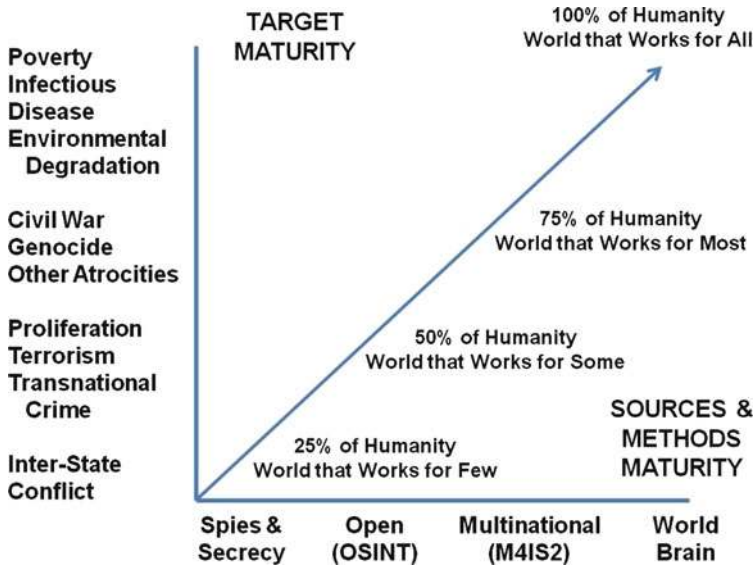**Fig. 20** Intelligence-driven harmonization of gifts and spending

Fig. 21 Intelligence maturity scale

OLD: GIVE THEM FISH

CURRENT: TEACH THEM

Information Arbitrage: The process of converting information into intelligence, and intelligence into profit. *Information costs money, intelligence makes money.*

FUTURE: ANCIENT WAYS, MODERN MEANS

Concluding Observations & Recommendations

1. Five billion poor are the center of gravity
2. Indigenous knowledge & energy are the way
3. Information arbitrage is the means
4. Denmark can influence other people's money
5. Denmark can make UN & NATO "Smart"
6. Denmark can save the North & restore the Earth

Fig. 22 Five billion poor, ancient ways, new means

Using the ten high-level threats to humanity, all associated with terrorism as a symptom, and our own knowledge of the possibilities for multinational, multifunctional information-sharing and sense-making, we conclude that the US Government is "out of control" in how it spends and that the public desperately needs to understand existing waste and imminent opportunity.

4 Conclusion Part I: Information Arbitrage

Long ago I coined the term "Information Arbitrage," along with "Information Peacekeeping," and I have focused for over two decades on how to use information to create a prosperous world at peace. Although we have a good ways to go, the plane of change is no longer flat – we are at the beginning of an almost vertical trajectory of positive change (Fig. 22).

Denmark specifically, and the other Nordic nations combined with Belgium, The Netherlands, and Luxembourg can make a difference. I recommend this new counter-terrorism center within the University of Southern Denmark seek to focus on counter-terrorism as a symptom in context, not as a threat in isolation.

5 Conclusion Part II: Arcs of Crisis and Collaboration

Although I have not spoken as much as I would have liked about Buckminster Fuller [18] and Russell Ackoff [1], they are present in my every breath. It is not possible to address terrorism without addressing the whole – isolated endeavors against terrorism alone are unaffordable and thus unsustainable. We must achieve economies of both scale and holistic understanding. I will end with this view of the Earth, one that I observe is essential to planning for the delivery of clean water, free electricity, and a sufficiency of food and shelter to every human being on the Earth (Fig. 23).

With "tough love," I must conclude that this center has promise, but only if you expand your thinking to include all ten threats across all twelve policies.

5.1 Next Steps

Responding now to questions and critical commentary on the presentation, I add some observations on stakeholders and next steps.

5.2 Stakeholders

There is only one stakeholder of consequence, the public. The public is *A Power Governments Cannot Suppress* [48] and it is the public that comprises an *Unconquerable World* [33]). It is a mistake to assume, expect, or seek stakeholders among governments, corporations, or even the UN – especially the UN. Figure 24 illustrates the shift in power and method from the past to the present and future.
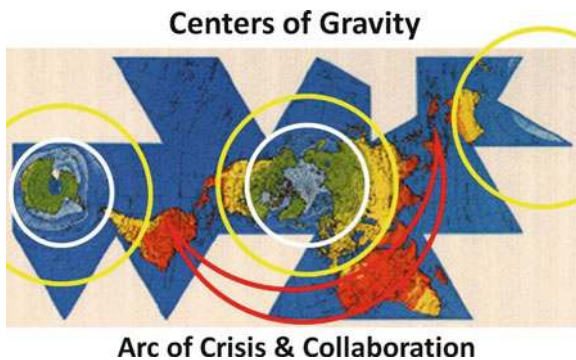

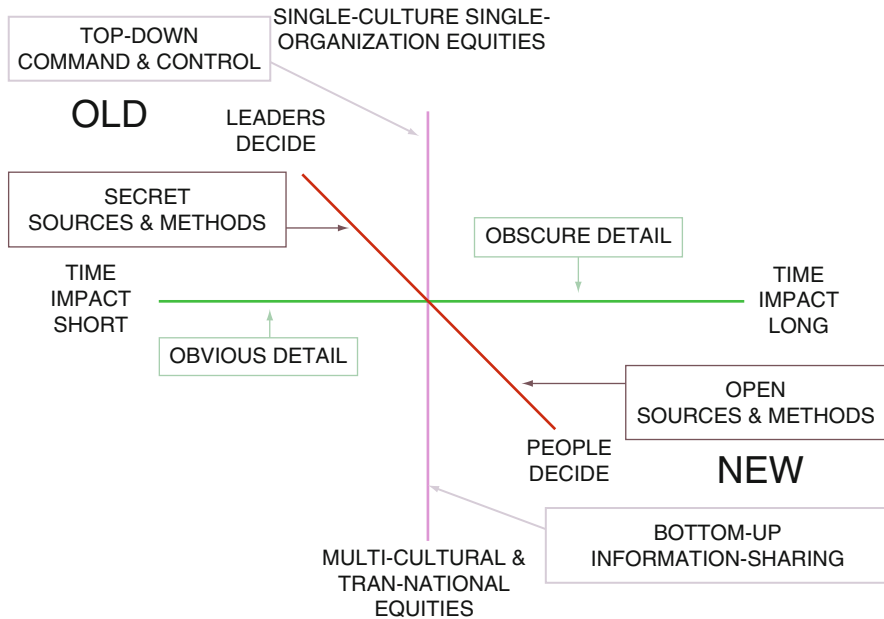**Fig. 23** Rescuing the arc of crisis with information-collaboration

Fig. 24 Epoch B bottom-up multinational public leadership

It is now known that Central Banks, the International Monetary Fund (IMF), the World Trade Organization (WTO) and all other "international" organizations have in fact been participating predators in a Global Class War [17]. It is now known that capitalism lost sight of the circular connection between profiting from consumer purchases and recycling profits into communities of consumers. It is in this context that any strategy for the future of humanity must reject fixed structures as solutions, and especially reject any organization that aspires to be secretive, avoiding transparency and accountability. Instead, the stakeholders of the future will be the eight tribes of intelligence, networks of humans who share information and make sense together. This is illustrated in Fig. 25.

The Information Era is the anti-thesis of the Industrial Era. Hierarchies are rigid and do not adapt. Hierarchies are part of the problem. Bureaucracy and the hoarding of information are part of the problem. As David Weinberger has pointed out, in this new era everything is "miscellaneous" and must be allowed to float freely precisely because any datum acquires more value the more people can access it [45].

5.3 Public Process Over Private Privilege

Transparency is the foundation for accountability as well as coherence and integrity of purpose and process. Corruption, so characteristic of the Industrial Era with its
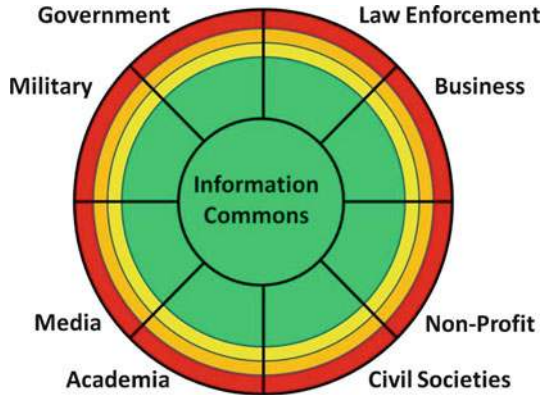
Fig. 25 The global information commons

stovepipes and cubbies of privilege, is pushed back by transparency. Just as plane spotters exposed the Central Intelligence Agency (CIA) rendition flights in support of transfers for torture, so also is the public now delving deeply into the precise manner that Goldman Sachs, Citi-Bank, Morgan Guaranty, and others deliberately and with malice aforethought "exploded" the US economy first, and then the global economy.

It has become clear to the public that all of this started in 1981 with the Presidency of Ronald Reagan.[17] Worse, it was carried on by Democratic Administrations – Wall Street bought Congress and the White House, with the result that wages for 95% of all US citizens dropped in relative terms, while "earnings," generally from financial manipulations rather than physical productions, increased by five times a year for the top 5%.

In combination with the new awareness of the importance of "true cost" information – costs to the social and ecological fabric that are externalized to the public and not included in the "price" of a good or service, it is clear that the Information Era demands – and enables – public intelligence in the public interest.

5.4 Hybrid Multinational Networks

Jean-Francois Rischard, at the time Vice President for Europe of the World Bank, is the first person I noticed who understood that the future would be governed by hybrid multinational networks sharing information and sharing the challenge of sense-making [32].

---

[17]An entire literature is emerging on this point. Two examples are David K. Johnston, "Scary New Wage Data," in *Tax.com*, Oct. 25, 2010 04:35 AM EDT and Paul Craig Roberts, "America's Jobs Losses are Permanent," *Counterpunch*, October 28, 2010.

Having worked for the CICIG as the Threat and Risk Analyst, I can testify to the fact that hybrids do work, but I must also testify to the fact that hybrids corrupt faster and more deeply than traditional institutions if they are not forced to be completely transparent to the varied stakeholders and particularly the public.

I am persuaded that henceforth governments, corporations, and international organizations will be the beneficiaries of hybrid multinational networks that emerge to share information and to create public intelligence in the public interest. They will not be the benefactors, and I see many of the wasteful initiatives funded by government gradually being eradicated as the public learns how to spot and then kill initiatives that are proposed to benefit the few rather than the many.

5.5 Intermediate Goals

Recently in the USA, our two leading comedian commentators, Jon Stewart and Stephen Colbert, sponsored a Rally for Sanity and/or Fear. Close to 250,000 people showed up, completely filling the Mall between the US Capitol and the Washington Monument; tens of millions more watched the rally on television.

The overwhelming message of this Rally was that the majority of Americans are in the middle of the political spectrum, desire civil discourse and sensible solutions, and are now "fed up" with the extremists on both sides of the political divide who substitute ideology for intelligence, and fear for common sense. The public is ready to demand and provide public intelligence.

The first and most important intermediate goal is to promulgate across the USA and then into other countries, a Strategic Analytic Model such as the EIN has developed and as has now been put forth at the *Huffington Post*.

The second intermediate goal is to offer the below graphic, created by Medard Gabel, founder of BigPictureSmallWorld and the life-long assistant to Buckminster Fuller in creating and managing the analog World Game (Fig. 26).[18]
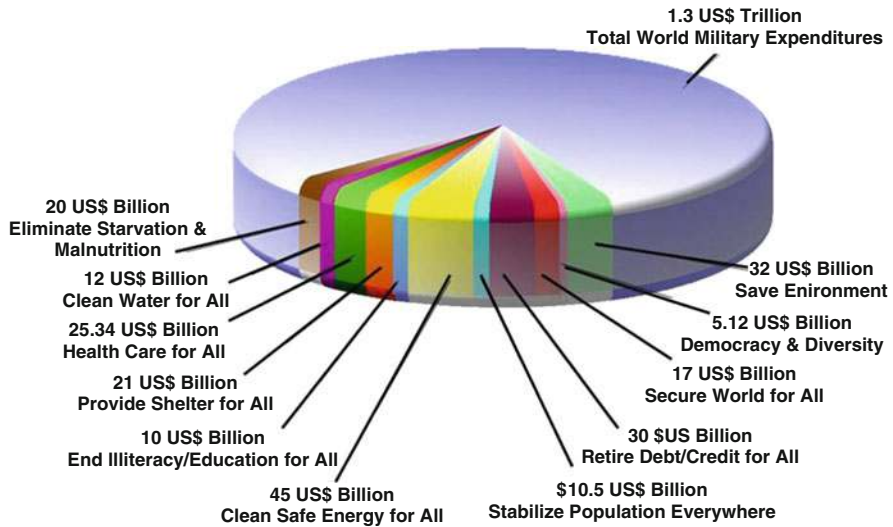
The third and final intermediate goal is to encourage all citizens to be students for life, and to focus on the "true cost" of every good and service (Fig. 27).

It took Jason "JZ" Liszkiewicz a full year of detailed research including extensive direct telephone and other personal conversation, to arrive at the "true cost" calculation depicted on the next page with very complete information. This must be done for every good and every service. As that information becomes public, there will be a shocking transformation of the public mind, and I anticipate a substantial turn away from petroleum and chlorine-based products [41], and a substantial turn away from meat, which consumes 5,000 gallons per pound of meat serves, as opposed to twenty gallons per pound of lettuce served [21].

E Veritate Potens-From Truth, We the People Are Made Powerful.

---

[18]For EIN, Professor Gabel has also calculated how to build the digital EarthGame, at a cost of less than $3 million a year. This would allow everyone to play themselves, permitting self-governance on all issues across all boundaries.

1.3 US$ Trillion
Total World Military Expenditures

20 US$ Billion
Eliminate Starvation &
Malnutrition

12 US$ Billion
Clean Water for All

25.34 US$ Billion
Health Care for All

21 US$ Billion
Provide Shelter for All

10 US$ Billion
End Illiteracy/Education for All

45 US$ Billion
Clean Safe Energy for All

$10.5 US$ Billion
Stabilize Population Everywhere

30 $US Billion
Retire Debt/Credit for All

17 US$ Billion
Secure World for All

5.12 US$ Billion
Democracy & Diversity

32 US$ Billion
Save Enironment

1.3 $ Trillion for War, When $227.96 Billion Could Buy BOTH Peace & Prosperity

Fig. 26 Cost of war versus cost of peace and prosperity for all

True Cost of One White CottonT-Shirt
Non-Organic, Foreign-Made, 200g/7oz

- Water : 570 gallons (45% irrigation)
- Energy: 8 kWh (machines), 11 to 29 gallons fuel
- Travel: 5,500 to 9,400+ miles
- Emissions: No_x, SO_2, CO, CO_2, N_2O, volatile compounds
- Toxins: 1-3g pesticides, diesel exhaust, heavy metals (dyes)
- Child Labor: 17 countries, 50 cents/day
- Buy Online: http://true-cost.re-configure.org

Fig. 27 True cost of one cotton t-shirt

Note

The author is the #1 Amazon reviewer for non-fiction, across ninety-eight categories. See especially:

Worth a Look: Book Review Lists (Positive)
http://tinyurl.com/SteeleBooksPositive
and Worth a Look: Book Review Lists (Negative)
http://tinyurl.com/SteeleBooksNegative.

See the *Journal of Public Intelligence* and all past production from 800 international contributors on intelligence reform and open intelligence at
http://www.phibetaiota.net.

References

1. Ackoff, R.: Redesigning Society. Stanford University, Stanford, CA (2003)
2. Amato, T.: Grand Illusion: The Myth of Voter Choice in a Two-Party Tyranny. New Press, New York, NY (2009)
3. Atlee, T.: Reflections on Evolutionary Activism. CreateSpace, Seattle, WA (2009)
4. Atlee, T.: The Tao of Democracy. The Writer's Collective, Cranston, RI (2003)
5. Beinhart, L.: Fog Facts: Searching for Truth in the Land of Spin. Nation Books, New York, NY (2006)
6. Benckler, Y.: Wealth of Networks. Yale University Press, New Haven, CT (2007)
7. Bogle, J.: The Battle for the Soul of Capitalism. Yale, New Haven, CT (2006)
8. Brand, S.: Clock of the Long Now. Basic Books, New York, NY (2000)
9. Butler, S.: War is a Racket: The Antiwar Classic by America's Most Decorated Soldier. Feral House, Port Townsend, WA (2003)
10. Carter, B.: Infinite Wealth. Butterworth-Heinemann, Burlington, MA (1999)
11. Chomsky, N., Herman, E.: Manufacturing Consent: The Political Economy of the Mass Media. Pantheon, New York, NY (2003)
12. Crane, M.: The Political Junkie Handbook. S.P.I. Books, New York, NY (2004)
13. Durant, W.: Philosophy and the Social Problem. Promethean Press, Frisco, TN (2008)
14. Durant, W., Durant, A.: Lessons of History. Simon & Schuster, New York, NY (2010)
15. Edges, C.: Empire of Illusion: The End of Literacy and the Triumph of Spectacle. Nation Books, New York, NY (2009)
16. Ellul, J.: Propaganda: The Formation of Men's Attitudes. Vintage, New York, NY (1973)
17. Faux, J.: The Global Class War: How America's Bipartisan Elite Lost Our Future – and What It Will Take to Win It Back. Wiley, New York, NY (2006)
18. Fuller, B.: Critical Path. St. Martin's Griffin, New York, NY (1982)
19. Johnson, C.: The Sorrows of Empire: Militarism, Secrecy, and the End of the Republic. Metropolitan Books, New York, NY (2004)
20. Jongman, B.: World Conflict and Human Rights Map (Project on Interdisciplinary Research on the Root Causes of Gross Human Rights Violations (PIOOM)). Leiden University, Leiden (1997–2002)
21. Lappe, F.M.: Diet for a Small Planet. Ballentine, New York, NY (1991)
22. Leebaert, D.: The Fifty-Year Wound: How America's Cold War Victory Has Shaped Our World. Back Bay Books, New York, NY (2003)
23. Manwaring, M., et al.: The Search for Security: A U.S. Grand Strategy for the Twenty-First Century. Praeger, New York, NY (2003)
24. Marrs, J.: Rule by Secrecy: The Hidden History That Connects the Trilateral Commission, the Freemasons, and the Great Pyramids. Harper Paperbacks, New York, NY (2001)
25. McKibben, B.: The Age of Missing Information. Random House, New York, NY (2006)
26. Naim, M.: Illicit: How Smugglers, Traffickers, and Copycats are Hijacking the Global Economy. Anchor, New York, NY (2006)
27. Ossman, J.D.: Surrender to Kindness: One Man's Epic Journey for Love and Peace. Adagio Press, Sarasota, FL (2010)
28. Palmer, M.: Breaking the Real Axis of Evil: How to Oust the World's Last Dictators by 2025. Rowman & Littlefield Publishers, New York, NY (2005)
29. Parry, R.: Lost History: Contras, Cocaine, the Press & 'Project Truth'. Media Consortium, San Francisco, CA (1999)
30. Perrow, C.: The Next Catastrophe: Reducing Our Vulnerabilities to Natural, Industrial, and Terrorist Disasters. Princeton University Press, Princeton, NJ (2007)
31. Rampton, S., Stauber, J.: Weapons of Mass Deception: The Uses of Propaganda in Bush's War on Iraq. Tarcher, New York, NY (2003)
32. Rischard, J.-F.: HIGH NOON: 20 Global Problems, 20 Years to Solve Them. Basic Books, New York, NY (2003)

33. Schell, J.: Unconquerable World: Power, Nonviolence, and the Will of the People. Holt Books, New York, NY (2004)
34. Shattuck, R.: Forbidden Knowledge: From Prometheus to Pornography. Mariner Books, New York, NY (1997)
35. Steele, R.D.: Election 2008: Lipstick on the Pig. Earth Intelligence Network, Oakton, VA (2008)
36. Steele, R.D.: Human Intelligence (HUMINT): All Humans, All Minds, All the Time. Strategic Studies Institute, Carlisle, PA (2010)
37. Steele, R.D.: Information Operations: All Information, All Languages, All the Time. Open Source Solutions Network, Oakton, VA (2006)
38. Steele, R.D.: Information Operations: Putting the "I" Back Into DIME. Strategic Studies Institute, Carlisle, PA (2006)
39. Steele, R.D.: Intelligence for Earth: Clarity, Diversity, Integrity, & Sustainability. Earth Intelligence Network, Oakton, VA (2010)
40. Stewart, T.: Wealth of Knowledge. Crown Business, New York, NY (2003)
41. Thornton, J.: Pandora's Poison: Chlorine, Health, and a New Environmental Strategy. MIT, Cambridge, MA (2001)
42. Toffler, A., Toffler, H.: Revolutionary Wealth. Knoph, New York, NY (2006)
43. Tovey, M. (ed.): Collective Intelligence: Creating a Prosperous World at Peace. Earth Intelligence Network, Oakton, VA (2008)
44. United Nations: Report of the High-Level Panel on Threats, Challenges, and Change, A more secure world: Our shared responsibility. United Nations, New York, NY (2004)
45. Weinberger, D.: Everything Is Miscellaneous: The Power of the New Digital Disorder. Holt, New York, NY (2008)
46. Wilson, E.O.: Consilience: The Unity of Knowledge. Vintage, New York, NY (1999)
47. Wright, R.: Non-Zero: The Logic of Human Destiny. Vintage, New York, NY (2001)
48. Zinn, H.: A Power Governments Cannot Suppress. City Lights Books, San Francisco, CA (2006)