

Integrated Series in Information Systems

Volume 30

Series Editors

Ramesh Sharda
Oklahoma State University, Stillwater, OK, USA

Stefan Voß
University of Hamburg, Hamburg, Germany

For further volumes:
<http://www.springer.com/series/6157>

Hsinchun Chen

Dark Web

Exploring and Data Mining
the Dark Side of the Web

Hsinchun Chen
Department of Management Information Systems
University of Arizona
Tucson, AZ, USA
hchen@eller.arizona.edu

ISSN 1571-0270
ISBN 978-1-4614-1556-5 e-ISBN 978-1-4614-1557-2
DOI 10.1007/978-1-4614-1557-2
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2011941611

© Springer Science+Business Media, LLC 2012

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.

The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Aims

The University of Arizona Artificial Intelligence Lab (AI Lab) 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 world, 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. The approaches and methods developed in this project contribute to advancing the field of Intelligence and Security Informatics (ISI). Such advances will help related stakeholders perform terrorism research and facilitate international security and peace.

Dark Web research has been featured in many national, international and local press and media, including: National Science Foundation press, Associated Press, BBC, Fox News, National Public Radio, Science News, Discover Magazine, Information Outlook, Wired Magazine, The Bulletin (Australian), Australian Broadcasting Corporation, Arizona Daily Star, East Valley Tribune, Phoenix ABC Channel 15, and Tucson Channels 4, 6, and 9. As an NSF-funded research project, our research team has generated significant findings and publications in major computer science and information systems journals and conferences. We hope our research will help educate the next generation of cyber/Internet-savvy analysts and agents in the intelligence, justice, and defense communities.

This monograph aims to provide an overview of the Dark Web landscape, suggest a systematic, computational approach to understanding the problems, and illustrate research progress with selected techniques, methods, and case studies developed by the University of Arizona AI Lab Dark Web team members.

Audience

This book aims to provide an interdisciplinary and understandable monograph about Dark Web research. We hope to bring useful knowledge to scientists, security professionals, counter-terrorism experts, and policy makers. The proposed work could also serve as a reference material or textbook in graduate level courses related to information security, information policy, information assurance, information systems, terrorism, and public policy.

The primary audience for the proposed monograph will include the following:

- **IT Academic Audience:** College professors, research scientists, graduate students, and select undergraduate juniors and seniors in computer science, information systems, information science, and other related IT disciplines who are interested in intelligence analysis and data mining and their security applications.
- **Security Academic Audience:** College professors, research scientists, graduate students, and select undergraduate juniors and seniors in political sciences, terrorism study, and criminology who are interested in exploring the impact of the Dark Web on society.
- **Security Industry Audience:** Executives, managers, analysts, and researchers in security and defense industry, think tanks, and research centers that are actively conducting IT-related security research and development, especially using open source web contents.
- **Government Audience:** Policy makers, managers, and analysts in federal, state, and local governments who are interested in understanding and assessing the impact of the Dark Web and their security concerns.

Scope and Organization

The book consists of three parts. In Part I, we provide an overview of the research framework and related resources relevant to intelligence and security informatics (ISI) and terrorism informatics. Part II presents ten chapters on computational approaches and techniques developed and validated in the Dark Web research. Part III presents nine chapters of case studies based on the Dark Web research approach. We provide a brief summary of each chapter below.

Part I. Research Framework: Overview and Introduction

• Chapter 1. Dark Web Research Overview

The AI Lab 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 world, 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.

- **Chapter 2. Intelligence and Security Informatics (ISI): Research Framework**

In this chapter we review the computational research framework that is adopted by the Dark Web research. We first present the security research context, followed by description of a data mining framework for intelligence and security informatics research. To address the data and technical challenges facing ISI, we present a research framework with a primary focus on KDD (Knowledge Discovery from Databases) technologies. The framework is discussed in the context of crime types and security implications.

- **Chapter 3. Terrorism Informatics**

In this chapter we provide an overview of selected resources of relevance to “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.

Part II. Dark Web Research: Computational Approach and Techniques

- **Chapter 4. Forum Spidering**

In this study we propose a novel crawling system designed to collect Dark Web forum content. The system uses a human-assisted accessibility approach to gain access to Dark Web forums. Several URL ordering features and techniques enable efficient extraction of forum postings. The system also includes an incremental crawler coupled with a recall improvement mechanism intended to facilitate enhanced retrieval and updating of collected content.

- **Chapter 5. Link and Content Analysis**

To improve understanding of terrorist activities, we have developed a novel methodology for collecting and analyzing Dark Web information. The methodology incorporates information collection, analysis, and visualization techniques, and exploits various web information sources. We applied it to collecting and analyzing information of selected jihad web sites and developed visualization of their site contents, relationships, and activity levels.

- **Chapter 6. Dark Network Analysis**

Dark networks such as terrorist networks and narcotics-trafficking networks are hidden from our view yet could have a devastating impact on our society and economy. Based on analysis of four real-world “dark” networks, we found that these covert networks share many common topological properties with other types of networks. Their efficiency in communication and flow of information, commands, and goods can be tied to their small-world structures characterized by small average path length and high clustering coefficient. In addition, we found that because of the small-world properties dark networks are more vulnerable to attacks on the bridges that connect different communities than to attacks on the hubs.

- **Chapter 7. Interactional Coherence Analysis**

Despite the rapid growth of text-based computer-mediated communication (CMC), its limitations have rendered the media highly incoherent. Interactional coherence analysis (ICA) attempts to accurately identify and construct interaction networks of CMC messages. In this study, we propose the Hybrid Interactional Coherence (HIC) algorithm for identification of web forum interaction. HIC utilizes both system features, such as header information and quotations, and linguistic features, such as direct address and lexical relation. Furthermore, several similarity-based methods, including a Lexical Match Algorithm (LMA) and a sliding window method, are utilized to account for interactional idiosyncrasies.

- **Chapter 8. Dark Web Attribute System**

In this study we propose a Dark Web Attribute System (DWAS) to enable quantitative Dark Web content analysis from three perspectives: technical sophistication, content richness, and web interactivity. Using the proposed methodology, we identified and examined the Internet usage of major Middle Eastern terrorist/extremist groups. In our comparison of terrorist/extremist web sites to U.S. government web sites, we found that terrorists/extremist groups exhibited levels of web knowledge similar to that of U.S. government agencies. Moreover, terrorists/extremists had a strong emphasis on multimedia usage and their web sites employed significantly more sophisticated multimedia technologies than government web sites.

- **Chapter 9. Authorship Analysis**

In this study we addressed the online anonymity problem by successfully applying authorship analysis to English and Arabic extremist group web forum messages. The performance impact of different feature categories and techniques was evaluated across both languages. In order to facilitate enhanced writing style identification, a comprehensive list of online authorship features was incorporated. Additionally, an Arabic language model was created by adopting specific features and techniques to deal with the challenging linguistic characteristics of Arabic, including an elongation filter and a root clustering algorithm.

- **Chapter 10. Sentiment Analysis**

In this study the use of sentiment analysis methodologies is proposed for classification of web forum opinions in multiple languages. The utility of stylistic and syntactic features is evaluated for sentiment classification of English and Arabic content. Specific feature extraction components are integrated to account for the linguistic characteristics of Arabic. The Entropy Weighted Genetic Algorithm (EWGA) is also developed, which is a hybridized genetic algorithm that incorporates the information gain heuristic for feature selection. The proposed features and techniques are evaluated on U.S. and Middle Eastern extremist web forum postings.

- **Chapter 11. Affect Analysis**

Analysis of affective intensities in computer-mediated communication is important in order to allow a better understanding of online users' emotions and preferences. In this study we compared several feature representations for affect analysis,

including learned n-grams and various automatically- and manually-crafted affect lexicons. We also proposed the support vector regression correlation ensemble (SVRCE) method for enhanced classification of affect intensities. Experiments were conducted on U.S. domestic and Middle Eastern extremist web forums.

- **Chapter 12. CyberGate Visualization**

Computer-mediated communication (CMC) analysis systems are important for improving participant accountability and researcher analysis capabilities. However, existing CMC systems focus on structural features, with little support for analysis of text content in web discourse. In this study we propose a framework for CMC text analysis grounded in Systemic Functional Linguistic Theory. Our framework addresses several ambiguous CMC text mining issues, including the relevant tasks, features, information types, feature selection methods, and visualization techniques. Based on it, we have developed a system called CyberGate, which includes the Writeprint and Ink Blot techniques. These techniques incorporate complementary feature selection and visualization methods in order to allow a breadth of analysis and categorization capabilities.

- **Chapter 13. Dark Web Forum Portal**

The Dark Web Forum Portal provides web-enabled access to critical international jihadist web forums. The focus of this chapter is on the significant extensions to previous work including: 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. A case study on identifying active jihadi participants in web forums is shown at the end.

Part III. Dark Web Research: Case Studies

- **Chapter 14. Jihadi Video Analysis**

This chapter presents an exploratory study of jihadi extremist groups' videos using content analysis and a multimedia coding tool to explore the types of video, groups' modus operandi, and production features that lend support to extremist groups. The videos convey messages powerful enough to mobilize members, sympathizers, and even new recruits to launch attacks that are captured (on video) and disseminated globally through the Internet. The videos are important for jihadi extremist groups' learning, training, and recruitment. In addition, the content collection and analysis of extremist groups' videos can help policy makers, intelligence analysts, and researchers better understand the extremist groups' terror campaigns and modus operandi, and help suggest counter-intelligence strategies and tactics for troop training.

- **Chapter 15. Extremist YouTube Videos**

In this study, we propose a text-based framework for video content classification of online video-sharing web sites. Different types of user-generated data (e.g., titles, descriptions, and comments) were used as proxies for online videos, and

three types of text features (lexical, syntactic, and content-specific features) were extracted. Three feature-based classification techniques (C4.5, Naïve Bayes, and SVM) were used to classify videos. To evaluate the proposed framework, we developed a testbed based on jihadi videos collected from the most popular video-sharing site, YouTube.

- **Chapter 16. Improvised Explosive Devices (IED) on Dark Web**

This chapter presents a cyber-archaeology approach to social movement research. Cultural cyber-artifacts of significance to the social movement are collected and classified using automated techniques, enabling analysis across multiple related virtual communities. Approaches to the analysis of cyber-artifacts are guided by perspectives of social movement theory. A Dark Web case study on a broad group of related IED virtual communities is presented to demonstrate the efficacy of the framework and provide a detailed instantiation of the proposed approach for evaluation.

- **Chapter 17. Weapons of Mass Destruction (WMD) on Dark Web**

In this chapter we propose a research framework that aims to investigate the capability, accessibility, and intent of critical high-risk countries, institutions, researchers, and extremist or terrorist groups. We propose to develop a knowledge base of the Nuclear Web that will collect, analyze, and pinpoint significant actors in the high-risk international nuclear physics and weapons communities. We also identify potential extremist or terrorist groups from our Dark Web testbed who might pose WMD threats to the U.S. and the international community. Selected knowledge mapping and focused web crawling techniques and findings from a preliminary study are presented.

- **Chapter 18. Bioterrorism Knowledge Mapping**

In this research we propose a framework to identify the researchers who have expertise in the bioterrorism agents/diseases research domain, the major institutions and countries where these researchers reside, and the emerging topics and trends in bioterrorism agents/diseases research. By utilizing knowledge mapping techniques, we analyzed the productivity status, collaboration status, and emerging topics in the bioterrorism domain. The analysis results provide insights into the research status of bioterrorism agents/diseases and thus allow a more comprehensive view of bioterrorism researchers and ongoing work.

- **Chapter 19. Women's Forums on the Dark Web**

In this study, we develop a feature-based text classification framework to examine the online gender differences between female and male posters on web forums by analyzing writing styles and topics of interests. We examine the performance of different feature sets in an experiment involving political opinions. The results of our experimental study on this Islamic women's political forum show that the feature sets containing both content-free and content-specific features perform significantly better than those consisting of only content-free features.

- **Chapter 20. US Domestic Extremist Groups**

U.S. domestic extremist groups have increased in number and are intensively utilizing the Internet as an effective tool to share resources and members with limited regard for geographic, legal, or other obstacles. In this study, we develop automated and semi-automated methodologies for capturing, classifying, and organizing domestic extremist web site data. We found that by analyzing the hyperlink structures and content of domestic extremist web sites and constructing social network maps, their inter-organizational structure and cluster affinities could be identified.

- **Chapter 21. International Falun Gong Movement on the Web**

In this study, we developed a cyber-archaeology approach and used the international Falun Gong (FLG) movement as a case study. The FLG is known as a peaceful international social movement, unlike the more violent jihadi movement. We employed Social Network Analysis and Writprint to analyze FLG's cyber-artifacts from the perspectives of links, web content, and forum content. In the link analysis, FLG's web sites linked closely to Chinese democracy and human rights social movement organizations (SMOs), reflecting FLG's historical conflicts with the Chinese government after the official ban in 1999.

- **Chapter 22. Botnets and Cyber Criminals**

In the last several years, the nature of computer hacking has completely changed. Cybercrime has risen to unprecedented sophistication with the evolution of botnet technology, and an underground community of cyber criminals has arisen, capable of inflicting serious socioeconomic and infrastructural damage in the information age. This chapter serves as an introduction to the world of modern cybercrime and discusses information systems to investigate it. We investigated the command and control (C&C) signatures of major botnet herders using data collected from the ShadowServer Foundation, a nonprofit research group for botnet research. We also performed exploratory population modeling of the bots and cluster analysis of selected cyber criminals.

Tuscon, Arizona, USA

Hsinchun Chen

About the Author



Dr. Hsinchun Chen is the McClelland Professor of Management Information Systems at the University of Arizona. He received a B.S. degree from the National Chiao-Tung University in Taiwan, an MBA degree from SUNY Buffalo, and his Ph.D. degree in Information Systems from New York University. Dr. Chen has served as a Scientific Counselor/Advisor of the National Library of Medicine (USA), Academia Sinica (Taiwan), and National Library of China (China). Dr. Chen is a Fellow of IEEE and AAAS. He received the IEEE Computer Society 2006 Technical Achievement Award and the INFORMS Design Science Award in 2008. He has an h-index score of 50. He is author/editor of 20 books, 25

book chapters, 210 SCI journal articles, and 140 refereed conference articles covering web computing, search engines, digital library, intelligence analysis, biomedical informatics, data/text/web mining, and knowledge management. His recent books include: *Infectious Disease Informatics* (2010); *Mapping Nanotechnology Knowledge and Innovation* (2008), *Digital Government: E-Government Research, Case Studies, and Implementation* (2007); *Intelligence and Security Informatics for International Security: Information Sharing and Data Mining* (2006); and *Medical Informatics: Knowledge Management and Data Mining in Biomedicine* (2005), all published by Springer. Dr. Chen was ranked #8 in publication productivity in Information Systems (CAIS 2005) and #1 in Digital Library research (IP&M 2005) in two bibliometric studies. He is Editor in Chief (EIC) of the new *ACM Transactions on Management Information Systems (ACM TMIS)* and Springer *Security Informatics (SI)* Journal, and the Associate EIC of *IEEE Intelligent Systems*. He serves on ten editorial boards including: *ACM Transactions on Information Systems*, *IEEE Transactions on Systems, Man, and Cybernetics*, *Journal of the American Society for Information Science and Technology*, *Decision Support Systems*,

and *International Journal on Digital Library*. He has been an advisor for major NSF, DOJ, NLM, DOD, DHS, and other international research programs in digital library, digital government, medical informatics, and national security research. Dr. Chen is the founding director of the Artificial Intelligence Lab and Hoffman E-Commerce Lab. The UA Artificial Intelligence Lab, which houses 20+ researchers, has received more than \$30M in research funding from NSF, NIH, NLM, DOD, DOJ, CIA, DHS, and other agencies. Dr. Chen has also produced 25 Ph.D. students who are placed in major academic institutions around the world. The Hoffman E-Commerce Lab, which has been funded mostly by major IT industry partners, features one of the most advanced e-commerce hardware and software environments in the College of Management. Dr. Chen was conference co-chair of ACM/IEEE Joint Conference on Digital Libraries (JCDL) 2004 and has served as the conference/program co-chair for the past eight International Conferences of Asian Digital Libraries (ICADL), the premiere digital library meeting in Asia that he helped develop. Dr. Chen is also (founding) conference co-chair of the IEEE International Conference on Intelligence and Security Informatics (ISI) 2003-present. The ISI conference, which has been sponsored by NSF, CIA, DHS, and NIJ, has become the premiere meeting for international and homeland security IT research. Dr. Chen's COPLINK system, which has been quoted as a national model for public safety information sharing and analysis, has been adopted in more than 3,500 law enforcement and intelligence agencies. The COPLINK research had been featured in the *New York Times*, *Newsweek*, *Los Angeles Times*, *Washington Post*, *Boston Globe*, and *ABC News*, among others. The COPLINK project was selected as a finalist by the prestigious International Association of Chiefs of Police (IACP)/Motorola 2003 Weaver Seavey Award for Quality in Law Enforcement in 2003. COPLINK research has recently been expanded to border protection (BorderSafe), disease and bioagent surveillance (BioPortal), and terrorism informatics research (Dark Web), funded by NSF, DOD, CIA, and DHS. In collaboration with selected international terrorism research centers and intelligence agencies, the Dark Web project has generated one of the largest databases in the world about extremist/terrorist-generated Internet contents (web sites, forums, blogs, and multimedia documents). Dark Web research supports link analysis, content analysis, web metrics analysis, multimedia analysis, sentiment analysis, and authorship analysis of international terrorism contents. The project has received significant international press coverage, including: *Associated Press*, *USA Today*, *The Economist*, *NSF Press*, *Washington Post*, *Fox News*, *BBC*, *PBS*, *Business Week*, *Discover magazine*, *WIRED magazine*, *Government Computing Week*, *Second German TV (ZDF)*, *Toronto Star*, and *Arizona Daily Star*, among others. Dr. Chen is also a successful entrepreneur. He is the founder of Knowledge Computing Corporation (KCC), a university spin-off IT company and a market leader in law enforcement and intelligence information sharing and data mining. KCC was acquired by a major private equity firm for \$40M in the summer of 2009 and merged with I2, the industry leader in crime analytics. The combined I2/KCC company was acquired by IBM for \$420M in 2011. Dr. Chen has also received numerous awards in information technology and knowledge management education and research including: AT&T Foundation Award,

SAP Award, the Andersen Consulting Professor of the Year Award, the University of Arizona Technology Innovation Award, and the National Chiao-Tung University Distinguished Alumnus Award. He was also named Distinguished Alumnus by SUNY Buffalo. Dr. Chen has served as a keynote or invited speaker in major international security informatics, medical informatics, information systems, knowledge management, and digital library conferences and major international government meetings (NATO, UN, EU, FBI, CIA, DOD, DHS). He is a Distinguished/Honorary Professor of several major universities in Taiwan and China and was named the Distinguished University Chair Professor of the National Taiwan University. Dr. Chen recently served as the Program Chair of the International Conference on Information Systems (ICIS) 2009, held in Phoenix, Arizona.

Contents

Part I Research Framework: Overview and Introduction

1 Dark Web Research Overview	3
1 Introduction.....	3
1.1 Web Sites.....	3
1.2 Forums	4
1.3 Blogs, Social Networking Sites, and Virtual Worlds	4
1.4 Videos and Multimedia Content	4
2 Computational Techniques (Data Mining, Text Mining, and Web Mining)	4
2.1 Dark Web Collection.....	5
2.2 Dark Web Analysis and Visualization	5
3 Dark Web Project Structure and Resources	7
3.1 Team Members (Selected)	8
3.2 Press Coverage and Interest	8
3.3 The IEEE Intelligence and Security Informatics Conference.....	9
3.4 Dark Web Publications.....	10
3.5 Dark Web Project Funding and Acknowledgments	16
4 Partnership Acknowledgments	17
Reference	18
2 Intelligence and Security Informatics (ISI): Research Framework.....	19
1 Information Technology and National Security.....	19
1.1 Problems and Challenges	21
1.2 Intelligence and Security Informatics Versus Biomedical Informatics: Emergence of a Discipline	22
1.3 Research Opportunities.....	24

2	ISI Research Framework.....	24
2.1	Caveats for Data Mining.....	27
2.2	Domestic Security, Civil Liberties, and Knowledge Discovery	28
2.3	Research Opportunities.....	29
	References.....	29
3	Terrorism Informatics	31
1	Introduction.....	31
2	Terrorism and the Internet.....	31
3	Terrorism Research Centers and Resources.....	33
3.1	Think Tanks and Intelligence Resources	33
3.2	Terrorism Databases and Online Resources.....	34
3.3	Higher Education Research Institutes	39
4	Conclusions.....	41
	References.....	41

Part II Dark Web Research: Computational Approach and Techniques

4	Forum Spidering	45
1	Introduction.....	45
2	Related Work: Focused and Hidden Web Crawlers	46
2.1	Accessibility.....	47
2.2	Collection Type.....	47
2.3	Content Richness	48
2.4	URL Ordering Features.....	48
2.5	URL Ordering Techniques	49
2.6	Collection Update Procedure	50
2.7	Summary of Previous Research	50
3	Research Gaps and Questions	52
3.1	Focused Crawling of the Hidden Web	52
3.2	Content Richness	52
3.3	Web Forum Collection Update Strategies.....	52
3.4	Research Questions.....	53
4	Research Design.....	53
4.1	Proposed Dark Web Forum Crawling System	53
4.2	Accessibility.....	53
4.3	Incremental Crawling for Collection Updating	54
5	System Design	54
5.1	Forum Identification.....	54
5.2	Forum Preprocessing	56
5.3	Forum Spidering	59
5.4	Forum Storage and Analysis.....	60
5.5	Dark Web Forum Crawling System Interface	61

6	Evaluation	62
6.1	Forum Accessibility Experiment	62
6.2	Forum Collection Update Experiment	63
6.3	Forum Collection Statistics	65
7	Conclusions and Future Directions	66
	References	67
5	Link and Content Analysis	71
1	Introduction	71
2	Literature Review	72
2.1	Terrorists' Use of the Web	72
2.2	Information Services for Studying Terrorism	73
2.3	Advanced Information Technologies for Combating Terrorism	74
3	A Methodology for Collecting and Analyzing Dark Web Information	75
3.1	The Methodology	75
3.2	Discussion of the Methodology	77
4	Jihad on the Web: A Case Study	77
4.1	Application of the Methodology	78
5	Results and Discussion	87
5.1	Expert Evaluation and Results	87
6	Conclusions and Future Directions	88
	References	89
6	Dark Network Analysis	91
1	Introduction	91
2	Topological Analysis of Networks	91
3	Methods and Data	93
4	Results and Discussion	94
4.1	Basic Properties	94
4.2	Small-World Properties	96
4.3	Scale-Free Properties	98
4.4	Caveats	99
4.5	Network Robustness	100
5	Conclusions	102
	References	103
7	Interactional Coherence Analysis	105
1	Introduction	105
2	Related Work	107
2.1	Obstacles to CMC Interactional Coherence	107
2.2	CMC Interactional Coherence Analysis	108
3	Research Gaps and Questions	113
4	System Design: Hybrid Interactional Coherence System	114
4.1	Data Preparation	115
4.2	HIC Algorithm: System Feature Match	116

4.3	HIC Algorithm: Linguistic Feature Match	117
4.4	HIC Algorithm: Residual Match.....	119
5	Evaluation	120
5.1	Test Bed	120
5.2	Comparison of Techniques	121
6	Conclusions.....	123
	References.....	124
8	Dark Web Attribute System.....	127
1	Introduction.....	127
2	Literature Review.....	128
2.1	Terrorism and the Internet.....	128
2.2	Existing Dark Web Studies	129
2.3	Dark Web Collection Building.....	129
2.4	Dark Web Content Analysis.....	132
3	Proposed Methodology: Dark Web Collection and Analysis	133
3.1	Dark Web Collection Building.....	134
3.2	The Dark Web Attribute System (DWAS)	136
4	Case Study: Understanding Middle Eastern Terrorist Groups.....	139
4.1	Building Dark Web Research Test Bed.....	140
4.2	Collection Analysis and Benchmark Comparison	142
5	Conclusions and Future Directions	149
	References.....	150
9	Authorship Analysis.....	153
1	Introduction.....	153
2	Literature Review: Authorship Analysis.....	154
2.1	Writing Style Features	154
2.2	Analysis Techniques	155
2.3	Online Messages	155
2.4	Multilingual Issues.....	156
3	Arabic Language Characteristics	156
3.1	Inflection	156
3.2	Diacritics	157
3.3	Word Length and Elongation	157
4	Research Questions and Research Design	158
4.1	Test Bed	158
4.2	Analysis Techniques	159
4.3	Addressing Arabic Characteristics.....	159
4.4	Feature Sets	160
5	Authorship Identification Procedure	162
5.1	Collection and Extraction	162
5.2	Experiment.....	163
6	Results and Discussion	163
6.1	Comparison of Feature Types	164
6.2	Comparison of Classification Techniques.....	164

7	Analysis of English and Arabic Group Models	165
7.1	Decision Tree Analysis	165
7.2	Feature Usage Analysis.....	166
8	Conclusions and Future Directions.....	168
	References.....	169
10	Sentiment Analysis.....	171
1	Introduction.....	171
2	Related Work.....	172
2.1	Sentiment Classification.....	172
2.2	Sentiment Analysis Tasks	173
2.3	Sentiment Analysis Features.....	175
2.4	Sentiment Classification Techniques	176
2.5	Sentiment Analysis Domains.....	177
3	Research Gaps and Questions	178
3.1	Web Forums in Multiple Languages	178
3.2	Stylistic Features.....	178
3.3	Feature Reduction for Sentiment Classification	179
3.4	Research Questions.....	179
4	Research Design.....	179
5	System Design	181
5.1	Feature Extraction.....	181
5.2	Feature Selection: Entropy Weighted Genetic Algorithm (EWGA)	184
5.3	Classification.....	188
6	System Evaluation.....	188
6.1	Test Bed	189
6.2	Experiment 1a: Evaluation of Features.....	190
6.3	Experiment 1b: Evaluation of Feature Selection Techniques	191
6.4	Results Discussion	192
7	Conclusions and Future Directions.....	197
	References.....	197
11	Affect Analysis.....	203
1	Introduction.....	203
2	Related Work.....	204
2.1	Features for Affect Analysis	206
2.2	Techniques for Assigning Affect Intensities	208
3	Research Design.....	209
3.1	Gaps and Questions.....	209
3.2	Research Framework.....	210
3.3	Research Hypotheses	214
4	Evaluation	214
4.1	Test Bed	214
4.2	Experimental Design.....	215

4.3	Experiment 1: Comparison of Feature Sets	216
4.4	Experiment 2: Comparison of Techniques	217
4.5	Experiment 3: Ablation Testing	218
4.6	Hypotheses Results	218
5	Case Study: Al-Firdaws vs. Montada	220
6	Conclusions	223
	References	223
12	CyberGate Visualization	227
1	Introduction	227
2	Background	229
2.1	CMC Content	229
2.2	CMC Systems	230
2.3	Need for CMC Systems Supporting Text Analysis	232
3	CMC Text Mining	233
3.1	Tasks	233
3.2	Information Types	234
3.3	Features	235
3.4	Feature Selection	236
3.5	Visualization	237
4	A Design Framework for CMC Text Analysis	238
4.1	Meta-requirements	240
4.2	Meta-design	240
5	System Design: The CyberGate System	241
5.1	Information Types and Features	241
5.2	Feature Selection	242
5.3	Visualization	243
5.4	Writeprints and Ink Blots	245
6	CyberGate Case Studies: Clearguidance.com and Ummah.com	248
6.1	Clearguidance.com	248
6.2	Ummah.com	250
7	Conclusions	252
	References	253
13	Dark Web Forum Portal	257
1	Introduction	257
2	Literature Review	258
2.1	Incremental Forum Spidering	258
2.2	Multilingual Translation	259
2.3	Social Network Analysis on Web Forums	259
3	Motivation and Research Questions	260
4	System Design	261
4.1	Data Acquisition	261
4.2	Data Preparation	263
4.3	System Functionality	263

5	Data Set: Dark Web Forums	263
6	System Functionality	264
6.1	Forum Browsing and Searching.....	265
6.2	Social Network Visualization	266
7	Case Study: Identifying Active Participants in Dark Web Forums Using Social Network Analysis.....	268
8	Conclusions and Future Directions.....	269
	References.....	269

Part III Dark Web Research: Case Studies

14	Jihadi Video Analysis.....	273
1	Introduction.....	273
2	Jihadi Groups' Videos.....	274
2.1	Dissemination of Extremist Groups' Videos	275
3	Collections of Extremist Groups' Videos	276
4	Content Analysis of Videos	278
4.1	Sample Collection.....	279
4.2	Coding the Videos.....	279
4.3	Inter-coder Reliability	280
5	Types of Videos Produced.....	280
5.1	Documentary Videos.....	280
5.2	Suicide Attack Videos.....	282
6	How the Groups Used the Videos.....	283
6.1	Groups Identified	284
7	Groups' Modus Operandi and Production Features.....	285
7.1	Production Features	286
8	Conclusions.....	287
	Appendix A List of Sample Videos	289
	Appendix B Coding Scheme.....	291
	References.....	292
15	Extremist YouTube Videos	295
1	Introduction.....	295
2	Literature Review.....	296
2.1	Video Domains.....	298
2.2	Feature Types	299
2.3	Classification Techniques.....	301
3	Research Gaps and Research Questions	302
4	System Design	304
4.1	Data Collection	304
4.2	Feature Generation.....	305
4.3	Classification and Evaluation.....	308
5	Test Bed and Hypotheses	309
5.1	Test Bed	309
5.2	Hypotheses.....	310

6	Experiment Results and Discussion	311
7	A Case Study: Domestic Extremist Videos	311
8	Conclusions and Future Directions	313
	References	314
16	Improvised Explosive Devices (IED) on Dark Web	319
1	Introduction	319
2	Cyber-Archaeology Framework for Social Movement Research	321
3	Case Study	323
3.1	Introduction	323
3.2	Cyber-Archaeology Framework Phase 1	324
3.3	Cyber-Archaeology Framework Phase 2	325
3.4	Cyber-Archaeology Framework Phase 3	332
4	Conclusion	337
	References	338
17	Weapons of Mass Destruction (WMD) on Dark Web	341
1	Introduction	341
2	Literature Review: Knowledge Mapping and Focused Web Crawling	342
2.1	Knowledge Mapping	342
2.2	Focused Web Crawling	343
3	The Capability-Accessibility-Intent Model for Nuclear Threat Detection: Nuclear Web and Dark Web	344
3.1	Capability	345
3.2	Accessibility	345
3.3	Intent	346
4	Case Study: Nuclear Web and Dark Web	346
4.1	Knowledge Mapping for Nuclear Web	346
4.2	Focused Web Crawling for Dark Web	347
5	Conclusions	352
	References	353
18	Bioterrorism Knowledge Mapping	355
1	Introduction	355
2	Literature Review	356
2.1	Bioterrorism Literature Analysis	356
2.2	Knowledge Mapping	356
3	Research Test Bed	358
4	Research Design	359
4.1	Data Acquisition	360
4.2	Data Parsing and Cleaning	360
4.3	Data Analysis	361

5	Analysis Results and Discussion	361
5.1	Productivity Status of Bioterrorism Research.....	361
5.2	Collaboration Status of Bioterrorism Researchers.....	362
5.3	Emerging Topics of Bioterrorism Research.....	364
6	Conclusions and Future Directions.....	366
	References.....	366
19	Women's Forums on the Dark Web	369
1	Introduction.....	369
2	Literature Review.....	370
2.1	Online Gender Differences	370
2.2	Online Text Classification.....	372
3	Research Gaps and Questions.....	376
4	Research Design.....	377
4.1	Web Forum Message Acquisition.....	377
4.2	Feature Generation.....	377
4.3	Classification and Evaluation.....	380
5	Experimental Study.....	381
5.1	Test Bed	381
5.2	Hypotheses.....	382
5.3	Experimental Results	382
5.4	Different Topics of Interest: Females and Males	382
6	Conclusions and Future Directions.....	385
	References.....	386
20	US Domestic Extremist Groups	391
1	Introduction.....	391
2	Previous Research.....	392
2.1	Social Movement Research on Extremists and the Internet	392
2.2	Web Harvesting Approaches.....	393
2.3	Web Link and Content Analysis	394
3	Proposed Approach.....	395
3.1	Collection Building.....	395
3.2	Link Analysis.....	397
3.3	Content Analysis.....	398
4	Test Bed: Collection of Domestic Extremist Web Sites	399
5	Analysis Results.....	400
5.1	Link Analysis Results	400
5.2	Content Analysis Results.....	402
6	Conclusions and Future Work.....	404
	References.....	404

21	International Falun Gong Movement on the Web	407
1	Introduction.....	407
2	Literature Review.....	408
2.1	Social Movement Theory.....	408
2.2	Social Movement Organizations and the Internet.....	409
2.3	Social Network Analysis.....	410
2.4	Writeprints	411
3	Research Design: A Case Study of the International FLG Movement.....	412
3.1	The Falun Gong Movement	412
3.2	Research Design.....	413
4	Research Results	417
4.1	Link Analysis	417
4.2	Web Content Analysis.....	419
4.3	Forum Content Analysis	419
5	Conclusions.....	423
	References.....	424
22	Botnets and Cyber Criminals	427
1	Introduction.....	427
1.1	The Underground Economy.....	429
1.2	Terrorism Online.....	430
2	Research Test Bed: The Shadowserver Foundation.....	431
2.1	Honeypots	431
2.2	Malware Analysis	431
2.3	Snooping	432
3	Investigating the Botnet World	432
3.1	Dataset Processing	433
3.2	Criminal Social Networks.....	435
4	Future Work	438
5	Conclusions.....	438
	References.....	439
	Index.....	441