Integrated Series in Information Systems 30
Series Editors: Ramesh Sharda · Stefan Voß

Hsinchun Chen

Dark Web

Exploring and Data Mining the Dark Side of the Web

Springer

Integrated Series in Information Systems

Volume 30

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

Hsinchun Chen

Dark Web

Exploring and Data Mining the Dark Side of the Web

Springer

Hsinchun Chen
Department of Management Information Systems
University of Arizona
Tuscon, 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

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 Writeprint 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,