SECURING CYBERSPACE: ANALYZING CYBERCRIMINAL COMMUNITIES THROUGH WEB AND TEXT MINING PERSPECTIVES

by

Victor Benjamin

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF MANAGEMENT INFORMATION SYSTEMS

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2016

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, I certify that I have read the dissertation prepared by Victor Benjamin, titled Securing Cyberspace: Analyzing Cybercriminal Communities through Web and Text Mining Perspectives and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

_____ Date: 04/26/2016
Hsinchun Chen

_____ Date: 04/26/2016
Jay Nunamaker

_____ Date: 04/26/2016
Joseph Valacich

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

_____ Date: 04/26/2016
Dissertation Director: Hsinchun Chen

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of the requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that an accurate acknowledgement of the source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, However, permission must be obtained from the author.

SIGNED: Victor Benjamin

ACKNOWLEDGEMENTS

I hold great appreciation for my dissertation committee members, Drs. Hsinchun Chen, Jay Nunamaker, and Joseph Valacich, for your continuous guidance and encouragement. I am especially grateful for my advisor, Dr. Hsinchun Chen, for supporting me and acting as my mentor since my time as an undergraduate at the University of Arizona. I have learned many life lessons and have had many great experiences over the decade I have worked together. Thank you for the many opportunities you provided for me to work on meaningful projects and make an impact.

I sincerely thank Cathy Larson and the rest of my colleagues at the Artificial Intelligence Lab for your generous assistance and insight in various projects. I also thank you for your friendship. I hold many fond memories of the success I achieved together, as well as the more personal moments that made working at the Artificial Intelligence Lab so special.

I owe deep gratitude to all members of the Management Information Systems (MIS) department. There are several staff and faculty members I have interacted with over the past decade. Thank you all for your continuous encouragement and assistance.

Also thanks to the following for your kindness and support over the years: Dr. Ronald Breiger (Sociology, University of Arizona), Dr. Salim Hariri (Electrical & Computer Engineering, University of Arizona), Dr. Mihai Surdeanu (Computer Science, University of Arizona), and Dr. Thomas Holt (Criminology, Michigan State University). You have all enriched my studies and I am extremely grateful.

I also owe my deepest gratitude to my loving and supportive family, friends, and Xiao for being there with me. My dissertation has been partly supported by grants from the National Science Foundation (SES-1314631, DUE-1303362, CBET-0730908) and the Defense Threat Reduction Agency (HDTRA1-09-1-0058).

DEDICATION

This dissertation is dedicated to my parents, Awiyai and Evalyn;

To my brothers, Karl and Martin;

And to Lucky, my four-pawed friend.

TABLE OF CONTENTS

LIST OF FIGURES 8
LIST OF TABLES 9
ABSTRACT 10
1. INTRODUCTION 11
2. ESSAY I: DIVING INTO THE CYBERCRIMINAL WEB: A GUIDE FOR CONDUCTING LARGE-SCALE CYBERCRIMINAL FORUM RESEARCH 14
2.1. Introduction 14
2.2. Background 16
2.3. Conducting Cybercriminal Forum Research 20
2.3.1. Cybercriminal Forum Identification 20
2.3.2. Keyword Searches 21
2.3.3. Snowball Collection 22
2.3.4. Deep web Hidden Services 23
2.4. Forum Collection 25
2.4.1. Creating a Secure Forum Collection Environment 27
2.4.2. Web Crawler Setup 28
2.4.3. Parsing Collected Webpages 30
2.5. Analytical Directions 30
2.5.1. Cybercriminal Content Analysis 31
2.5.2. Cybercriminal Language Modeling 31
2.5.3. Cybercriminal Network Analysis 32
2.5.4. Underground Economy Analysis 32
2.6. Ethical Considerations 33
2.6.1. Research on Private Communities 34
2.6.2. Circumventing Anti-crawling 35
2.6.3. Masking Researcher Identity 36
2.6.4. Cybercriminal-Researcher Interaction 36
2.7. Empirical Study 37
2.7.1. Background 38
2.7.2. Research Gaps and Questions 41
2.7.3. Research Design 41
2.7.4. Cybercriminal Forum Identification 42
2.7.5. Data Collection 44
2.7.6. Analytical Methods 44
2.7.8. Results & Discussion 46
2.8. Conclusion 48
3. ESSAY II: EXAMINING HACKER PARTICIPATION LENGTH WITHIN CYBERCRIMINAL IRC COMMUNITIES 50
3.1. Introduction 50
3.2. Literature Review 52
3.2.1. Cybercriminal Community Research 53
3.2.2. Virtual Community Analysis Techniques 61
3.2.3. Duration Modeling 63

3.3. Research Gaps and Questions 66
3.4. Research Design 67
3.5. Results & Discussion 77
3.6. Conclusion & Future Research 82
4. ESSAY III: DETECTING EMERGING THREATS IN CYBERCRIMINAL FORUMS USING VECTOR REPRESENTATIONS OF WORDS 84
4.1. Introduction 84
4.2. Literature Review 86
4.2.1. Hacker Community Research 86
4.2.2. Lexical Semantics 89
4.3. Research Gaps and Questions 92
4.4. Research Testbed and Design 93
4.4.1.Data Collection 93
4.4.2. Data Pre-processing 94
4.4.3. Lexical Semantics 95
4.4.4. Evaluation 97
4.5. Results and Discussion 99
4.6. Conclusion 106
5. ESSAY IV: TRACKING INFORMATION DISSEMINATION BETWEEN MULTILINGUAL CYBERCRIMINAL FORUM POPULATIONS: THE AZSCOUT RESEARCH FRAMEWORK 108
5.1. Introduction 108
5.2. Literature Review 111
5.2.1. Cybercriminal Community Research 111
5.2.2. Information Theory 114
5.2.3. Lexical Semantics 115
5.3. Research Gaps and Questions 118
5.4. Research Testbed and Design 119
5.4.1. Data Collection 120
5.4.2. Data Pre-processing 121
5.4.3. Lexical Semantics 121
5.4.4. Evaluation 124
5.5. Results and Discussion 125
5.6. Conclusion 128
6. CONCLUSIONS 130
6.1. Cybersecurity Contributions 130
6.2. Contributions to the IS Field 131
6.3. Future Research Directions 132
7. REFERENCES 134

LIST OF FIGURES

Figure 2.1 – Cybercriminal Forum Activity Examples 19
Figure 2.3 – The Hidden Wiki, http://zqktlwi4fecvo6ri.onion 25
Figure 2.4 – Russian Cybercriminal Forum Posting 40
Figure 2.5 – Cybercriminal Reputation Study Research Design 42
Figure 2.6 – Hackhound.org is the user *H**s* 48
Figure 3.1 – #OperationGreenRights Recruiting Advertisement 52
Figure 3.2 – IRC system architecture 55
Figure 3.3 – Example of IRC channel activity and contents. 57
Figure 3.4 – Recruitment video for #Optestet environmental hacktivist campaign 58
Figure 3.5 – Video tutorial shared in #Evilzone that depicts car hijacking using a NFC interception tool 59
Figure 3.6 – Research design 67
Figure 3.7 – *#OperationGreenRights* leaked e-mail and password list 74
Figure 3.8 – Wang-Chang Kaplan-Meier Estimate for Both IRC Communities 79
Figure 3.9 – Twitter account of #Anonops participant 'Strudalz' 82
Figure 4.1 – Example of a posted message on a hacker forum 87
Figure 4.2 – Research Design 93
Figure 4.3 – Evaluation Plan 98
Figure 4.4 – Average Word Embedding Precision-at-24 Curve 102
Figure 4.5 – Similarity of *Bifrost* and *Spygate* with the Term *RAT* Over Time 105
Figure 4.6 – Similarity of *Pony Stealer* and *Coin Stealer* with the Term *Stealer* Over Time 106
Figure 5.1 – Example of an English/Russian multilingual forum, *Crdclub.su*. 109
Figure 5.2 – Example of a posted message on a Cybercriminal forum 112
Figure 5.3 – The AZScout Research Framework 119
Figure 5.4 – Russian and English Subforum Activity by participant *Г***ь* within *Crdpro.su* .. 127
Figure 5.5 – Russian and English Subforum Activity by participant *N***n* within *Crdclub.su* 128

LIST OF TABLES

Table 1.1 – Overview of Four Essays 12
Table 2.1 – Summary of Recent Cybercriminal Community Studies 18
Table 2.2 – Cybercriminal Forum Anti-crawling Mechanisms 29
Table 2.3 – Cybercriminal Forum Ethical Research Heuristics 34
Table 2.4 – Research Test bed Summary 43
Table 2.5 – Forum Content and Usage Features 45
Table 2.6 – OLS Regression Results 47
Table 3.1 – A summary of recent cybercriminal community studies 54
Table 3.2 – Collection summary 69
Table 3.3 – Extract network- and content-based features 72
Table 3.4 – Example output of Wang-Chang Kaplan-Meier survival curve matrix for the #Anonops community 78
Table 3.5. – Results of extended Cox's model with recurrent events for #Anonops and #Evilzone 80
Table 4.1 – Research Testbed 94
Table 4.2 – Example of Embedding Divergence caused by Negative Sampling 96
Table 4.3 – Hacker Language Term Categories 97
Table 4.4 – Evaluation Objective 1 Results for *Botnet* and *Carder* 100
Table 4.5 – Example Results of Evaluation Objective 1 100
Table 4.6 – Boosted vs Non-boosted Training 101
Table 4.7 – Boosted vs Non-boosted Training per Hacker Language Category 102
Table 4.8 – Evaluation Objective 3 for Hacker Term *Botnet* 103
Table 4.9 – Evaluation Objective 3 for Hacker Term *Botnet* 103
Table 4.10 – Evaluation Objective 3 for Hacker Term *Botnet* 104
Table 5.1 – Research Testbed 120
Table 5.2 – Average SSE Per Feature Generation Method 125
Table 5.3 – PV-DBOW Generated Cluster Sizes 125
Table 5.4 – Identified Thread Entropy Statistics 126
Table 5.5 – Potential Key Actors 126

ABSTRACT

Cybersecurity has become one of the most pressing issues facing society today. In particular, cybercriminals often congregate within online communities to exchange knowledge and assets. As a result, there has been a strong interest in recent years in developing a deeper understanding on cybercriminal behaviors, the global cybercriminal supply chain, emerging threats, and various other cybersecurity-related activities. However, few works in recent years have focused on identifying, collecting, and analyzing cybercriminal contents. Despite the high societal impact of cybercriminal community research, only a few studies have leveraged these rich data sources in their totality, and those that do often resort to manual data collection and analysis techniques.

In this dissertation, I address two broad research questions: 1) In what ways can I advance cybersecurity as a science by scrutinizing the contents of online cybercriminal communities? and 2) How can I make use of computational methodologies to identify, collect, and analyze cybercriminal communities in an automated and scalable manner? To these ends, the dissertation comprises four essays. The first essay introduces a set of computational methodologies and research guidelines for conducting cybercriminal community research. To this point, there has been no literature establishing a clear route for non-technical and non-security researchers to begin studying such communities. The second essay examines possible motives for prolonged participation by individuals within cybercriminal communities. The third essay develops new neural network language model (NNLM) capabilities and applies them to cybercriminal community data in order to understand hacker-specific language evolution and to identify emerging threats. The last essay focuses on developing a NNLM-based framework for identifying information dissemination among varying international cybercriminal populations by examining multilingual cybercriminal forums. These essays help further establish cybersecurity as a science.

1. INTRODUCTION

Cybersecurity has become a critical issue in society today. Public, private, and governmental sectors are all facing increasing cyber-threats. It is estimated that cybercrime costs the global economy about $445 billion a year, mostly due to theft of intellectual property within developed countries and sale of stolen personal information (Sandle & Char, 2014). As a result of the growing threats and societal reliance on cyber infrastructure, there has been a strong interest in recent years in developing a deep understanding on hacker behaviors, the cybercriminal supply chain, emerging vulnerabilities, and various other cybersecurity related activities from several well-known institutions. The National Science and Technology Council (NSTC), for example, outlined a critical need to develop advanced methods for modeling cyber adversaries (NSTC, 2011). Similarly, the National Science Foundation (NSF) created the *Secure and Trustworthy Cyberspace* (SaTC) program designed to fund research that addressing cybersecurity issues by bridging the government and academic communities together to minimize the misuse of cyber technology, bolster cyber education and training, and establish the science of cybersecurity.

Despite these calls for more cybercriminal community research, few works in recent years identify, collect, and analyze a large scale of cybercriminal-generated data. For example, there are hundreds of web forums with tens of millions of postings created by cybercriminals. Such data sources span a variety of geo-political regions including the US, Russia, China, and Middle-East and are viable data sources to inform new perspectives on cybercrime (Motoyama et al., 2011; Benjamin & Chen, 2012). Despite the high societal impact of cybersecurity research, only a few studies have leveraged these rich data sources in their totality, and often resort to manual data collection and analysis techniques. For example, some early work has unveiled multiple international networks of credit card fraud among online hacker communities by using manual

data collection and analysis methods (Yip et al., 2013). However, due to reliance of manual methods, generally only a subset of data is scrutinized, or only high-level metrics are computed across data. As a result, the capability of employing more scalable techniques to collect and analyze hacker communications would be of great asset. Researchers and practitioners could better understand the scope of cybercrime, the global hacker supply chain, information dissemination between different hacker groups, and more.

These research gaps and limitations in current literature provide a unique opportunity for Information Systems (IS) researchers to contribute and help advance the science of cybersecurity with web and text mining perspectives. Given this context, this dissertation aims to address two broad research questions:

1. In what ways can I advance cybersecurity as a science by scrutinizing the contents of online cybercriminal communities?
2. How can I make use of computational methodologies to identify, collect, and analyze cybercriminal communities in an automated and scalable manner

This dissertation contains four essays. Key aspects of each essay are outlined in Table 1.1. I briefly summarize the key essence of each essay in the remainder of this section.

Essay	Research Objectives	Methodology
1	To develop a series of automated and scalable techniques designed for identifying, collecting, and analyzing cybercriminal; forums. Case study investigating hacker reputation within forums to exemplify developed methodology.	Snowball Procedure for cybercriminal forum identification; forum web crawling, anti-crawling counter-measures; content analyses; network analyses
2	To develop a technique for quickly identifying key participants within cybercriminal Internet-Relay-Chat (IRC)	Duration Modeling, Extended Cox's Proportional Hazards Model
3	To develop a language-independent, unsupervised machine learning approach for modeling hacker language and detecting emerging threats from cybercriminal community data	Skip-gram Negative Sampling Neural network Language Model with extended objective function for representing temporal aspects of data
4	To develop an automated, language-independent method for analyzing multilingual cybercriminal forums and identifying instances of information dissemination between different cybercriminal populations.	Neural Network Language Models, Paragraph Vector with Distributed Bag of Words, Information Theory

Table 1.1 – Overview of Four Essays

The first essay introduces a set of computational methodologies and research guidelines for conducting cybercriminal community research. To this point, there has been no literature establishing a clear route for non-technical and non-security researchers to begin studying such communities. This essay outline methodologies for identifying, collecting and analyzing cybercriminal community contents, and also discuss how to operationalize cybercriminal research in a safe and secure manner. This essay serves as a basis for the following three essays.

The second essay examines possible motives for prolonged participation by individuals within cybercriminal communities. I focus on Internet-Relay-Chat (IRC) communities run by cybercriminals. Methodology includes the application of the extended Cox's model. Findings help explain different possible community participation motives, and highlight the importance of interconnectedness among cybercriminals.

The third essay develops new neural network language model (NNLM) capabilities and applies them to cybercriminal community data in order to understand hacker-specific language evolution and to identify emerging threats. Specifically, I extend the skip-gram negative sampling NNLM to handle temporal elements of data. I also develop new capabilities to use existing knowledge base for boosted model training. With my new techniques, I am able to automatically identify changes in hacker language over time and evolving threats.

The last essay focuses on developing a framework for identifying information dissemination among varying international cybercriminal populations by examining multilingual cybercriminal forums. The framework utilizes the Paragraph Vector with Distributed Bag-of-Words NNLM, and borrows perspectives from information theory. The framework provides researchers and practitioners the capability to more closely scrutinize the global cybercriminal supply chain, as well as how information and assets transfer between different cybercriminal populations.

2. ESSAY I: DIVING INTO THE CYBERCRIMINAL WEB: A GUIDE FOR CONDUCTING LARGE-SCALE CYBERCRIMINAL FORUM RESEARCH

2.1. Introduction

Cybersecurity has become an imperative societal problem, with wide-spread implications for the public, industrial, and governmental sectors. Increasingly, critical infrastructures and complex systems have become reliant on computing technologies, and thus are vulnerable to cyber-attack. It is estimated that cybercrime costs the global economy about $445 billion a year, mostly due to theft of intellectual property within developed countries and sale of stolen personal information (Sandle & Char, 2014). Further, accessibility to technologies and methods for committing cybercrime has grown considerably. Cybercriminals routinely share cybercriminal assets with one another, making cybercrime more accessible to lesser skilled miscreants (Holt et al. 2012; Benjamin & Chen 2013). Often, such assets include malicious tools, source code examples, tutorials, and even instructions on using legitimate services to support cybercriminal operations. The increased reliance on cyber infrastructure, as well as an ever-increasing number of threats, presents challenging problems for researchers, practitioners, and society.

Increased threats and vulnerabilities have resulted in growing interest for advancing current cyber-defense capabilities. In particular, a report by the National Science and Technology Council (NSTC) outlined a critical need to develop advanced methods for modeling cyber adversaries (NSTC, 2011). Such research could result in deeper knowledge of the cybercriminal behaviors, the cybercriminal supply chain, emerging vulnerabilities, and so on. Similarly, the National Science Foundation created the *Secure and Trustworthy Cyberspace* (SaTC) program to fund research that addresses issues in cybersecurity. SaTC is a widely successful program that bridges government and the academic community together in order to minimize the misuse of cyber

technology, to bolster cyber education and training, and to establish the science of cybersecurity. Further, recent conversations within the IS community have specifically called for further research on "black hat" cybercriminals, or cybercriminals, in order to enrich my understanding on how to effectively combat cyber adversaries (Mahmood et al., 2010; Chen et al., 2012).

However, despite these calls for more research, few works in recent years have successfully performed large-scale identification, collection, and analysis of cybercriminal-generated data. In particular, many cybercriminal-operated web forums exist that can be studied to inform new perspectives on cybercrime, but these data sources have largely gone untapped by researchers. Though the lack of research seems paradoxical given the high societal impact of cybersecurity research, this shortcoming may be explained by understanding the numerous issues that face researchers. First, many who may be interested in conducting cybercriminal research may not know where to begin searching for cybercriminal-generated data. Many cybercriminals and cybercriminal forums take great care to obfuscate their online presence in order to protect their identities and to avoid legal repercussions (Martin, 2013). Second, cybercriminal-generated data is much more difficult to collect than traditional web data. Many cybercriminal forums employ sophisticated anti-crawling mechanisms that make comprehensive automated data collection difficult (Benjamin et al., 2015). A number of forums have even implemented "drive-by exploits" where JavaScript-based malware seeking to exploit web browser vulnerabilities is embedded within cybercriminal forum web pages in order to infect unsuspecting visitors' machines, including those of security researchers. Thus, there are more challenges and risks facing researchers who are otherwise capable of more typical virtual community research. Lastly, given the nontraditional nature of cybercriminal data and cybercriminal community contents, it may be necessary to

develop guidelines to help researchers formulate relevant research questions and to utilize appropriate analytical methods.

This paper is organized into the following sections: First, I provide some background information on cybercriminal forums and describe how they are both similar to and different from traditional web forums. Next, I detail how to operationalize a cybercriminal forum research project, including methods to identify cybercriminal forums and techniques to automatically collect forum contents. I also provide guidance on analytical directions and discuss concerns with conducting this type of research in an ethical manner. I include a case example of a cybercriminal forum study I conducted in order to demonstrate potential research that can be undertaken by utilizing the methods described in this paper. Lastly, I conclude by discussing the contributions of this paper. In sum, this work provides a comprehensive guide for researchers to successfully conduct cybercriminal forum research.

2.2. Background

Cybersecurity defense has traditionally focused on the development of technical solutions to mitigate cybercriminal threats. Some examples include patching known vulnerabilities, creating smarter anti-virus software, developing techniques to help networks remain resilient during cyber-attacks, and so on. However, these methods are generally reactionary and are only effective after cybercrime has occurred. In order to develop proactive cybersecurity capabilities, it is necessary to gain a deep understanding of cybercriminal behaviors and to observe cybercriminal communications for evidence of emerging threats.

Fortunately, cybercriminals congregate within online communities, typically in the form of web forums, creating a valuable repository of data relevant to advancing cybersecurity research. Web forums allow members to participate in numerous discussions simultaneously by posting

messages in topic threads. Many forums also allow for the sharing of hyperlinks, pictures, videos, and other resources. Traditional virtual community research has incorporated web forum data for studies relevant to many unique domains (Liu & Chen, 2013).

Unfortunately, research focused on cybercriminal forums has been limited despite the high societal importance of this domain (see Table 2.1 for a summary of recent work). The majority of this research is limited to using relatively small data sets for manual qualitative analyses, or automated analyses, but consisting of counting or basic network analysis metrics. For example, many studies make use of manual data collection methods, and thus limit themselves to a low-scale analysis of the total amount of cybercriminal forum data available for collection (Holt, 2013). As such, there is a large gap in research utilizing computational techniques that enable large-scale research, as commonly observed in more traditional virtual community or big data research.

Previous Studies	Data Sources	Research	Analytical Methods	Findings
Holt, 2013	Forums	Cybercriminal black markets	Manual qualitative analysis of Russian cybercriminal forum postings	Price, customer service, and trust influence relationships between black market actors
Yip et al., 2013	Forums	Cybercriminal black markets	Combination of manual analysis and automated network analysis of two cybercriminal carding forums	Underground trading facilitated by social networking, reputation, and quality control
Martin, 2013	Forums	Cybercriminal black markets	Manual analysis of the *Silk Road* cryptomarket and forums	*Silk Road* and similar cryptomarkets will assume greater share of global trade of illicit drugs
Benjamin & Chen, 2012	Forums	Reputation in cybercriminal forums	Automated content and network analysis to assess how cybercriminals gain reputation among peers	Contributions to the cognitive advancement of a community lead to reputation gains among cybercriminals
Holt & Kilger, 2012	Forums and other cybercriminal webpages	Cybercriminal skill in global hacking community	Manual qualitative analysis of contents and networks found within cybercriminal forums and other cybercriminal-related web pages	Global distribution of skill among cybercriminals is similar; few top-skilled cybercriminals, most are unskilled

Holt et al., 2012	Forums	Cybercriminal social networks	Manual qualitative analysis of Russian cybercriminal forum postings	Cybercriminals practice a meritocratic culture, majority of participants are unskilled
Motoyama et al., 2011	Forums and Internet-Relay-Chat	General exploration	Manual content analysis, some automated network analyses	General descriptions of cybercriminal interactions in forums and IRC, existence of meritocratic structure
Yip, 2011	Forums	Cybercriminal black markets	Manual analysis of two carding forums	Underground trading facilitated by reputation and trust
Fallman et al., 2010	Forums and Internet-Relay-Chat	Cybercriminal black markets	Implementation of collection system to gather information and measure usage of underground economies in cybercriminal forums and IRC	Present usage statistics of cybercriminal black markets, describe collection system
Holt & Lampke, 2010	Forums	Cybercriminal black markets	Manual qualitative analysis of 300 threads from 6 cybercriminal forums. Analysis focused on data thieves and the sale of stolen data	Cybercriminals sell stolen data at fraction of true value, prolific stolen data market exists
Radianti, 2010	Forums and Internet-Relay-Chat	Cybercriminal black markets	Manual analysis of contents and social interactions within cybercriminal forums and IRC channel	Formal and informal regulations and procedures influence black market participation. Black markets may impose their own rules, while an individual's reputation also impacts their relationships with other participants

Table 2.1 - Summary of Recent Cybercriminal Community Studies

Much of the existing cybercriminal-focused literature describes similar themes concerning social behaviors and community. Some notable findings are that cybercriminals will frequently post messages that include hacking tools, tutorials, malware, source code examples, and more (Benjamin & Chen, 2012; Holt et al., 2012). For example, on the left side of Figure 2.1, a member of Hackhound.org publishes the latest version of his hacking tool meant to help others steal cached passwords on victims' computers; on the right, a cybercriminal of the Chinese community Unpack.cn posts sample code for reverse engineering software written in the Microsoft .NET framework. Additionally, some participants share links to deep web hidden services and other underground communities. For example, the infamous *Silk Road* black market containing sellers of drugs, stolen data, hacking tools, and more exists as a hidden service on the Tor network

(Martin, 2013). Cybercriminal communities are known to exist across various geopolitical regions, and are especially common in the US, China, Russia, and the Middle-East (Motoyama et al., 2011; Benjamin & Chen, 2012). Existing literature demonstrates great value in further exploring cybercriminal forums for cybersecurity research.

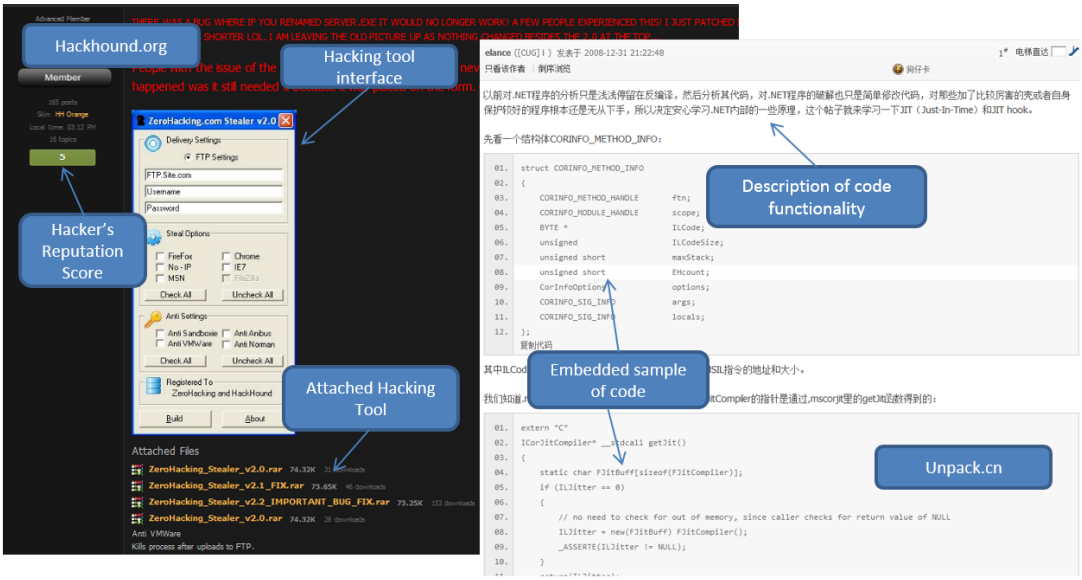

Figure 2.1 – Cybercriminal Forum Activity Examples

Unfortunately, many recent studies make use of manual or otherwise non-scalable identification, collection and analysis procedures. These shortcomings stem from the fact that cybercriminal forum data is often much more difficult to identify and collect than more traditional web forums. Further, researchers will expose themselves to numerous cyber-threats by collecting and viewing cybercriminal forum contents. Compared to more typical web forum research, a greater level of planning and technical sophistication is required to ensure researcher security and project success.

2.3. Conducting Cybercriminal Forum Research

To serve as a guide for operationalizing cybercriminal forum research, four different phases of research are described, including (see Figure 2.2): (1) techniques for identifying cybercriminal forums suitable for use in research studies; (2) techniques for collecting forum data that includes methods to circumvent anti-crawling mechanisms; (3) potential analytical methods for interpreting forum content; and (4) ethical considerations when conducting research within a criminal community. Overall, I provide a roadmap for undertaking a cybercriminal forum research project.

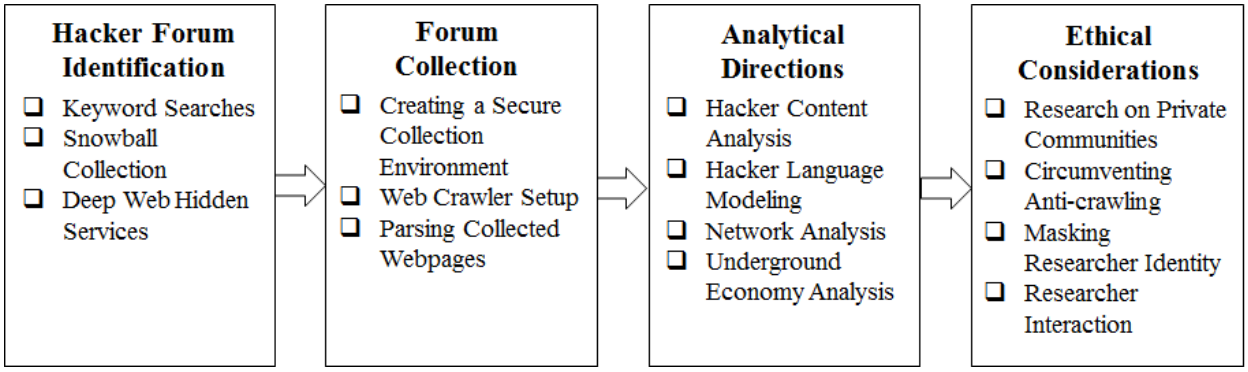

Figure 2.2 – Roadmap for Developing a Cybercriminal Forum Research Project

2.3.1. Cybercriminal Forum Identification

To successfully conduct a cybercriminal forum study, quality data sources must be identified. Fortunately, forums are used heavily by cybercriminals across the world, allowing researchers to examine cybercriminal activities in different geopolitical regions (Motoyama et al., 2011; Benjamin & Chen, 2015). Many of these forums utilize different languages depending on their origin, with the most frequently encountered languages being English, Chinese, and Russian. Further, forums will vary from one another in terms of size, activity level, and topical coverage (Holt et al., 2012). For example, some communities will include a wide variety of hacking-related discussion while others are focused for discussion on specific topics, such as carding (i.e., credit card fraud). By utilizing appropriate cybercriminal forum identification techniques, it is possible

to discover quality data sources in different languages and of diverse topical focus. In total, there are three primary techniques for identifying cybercriminal forums that can be utilized: (1) keyword searches, (2) snowball collection, and (3) deep web hidden services. Each technique differs and yields unique forums that may not otherwise be found with alternate forum identification procedures.

2.3.2. Keyword Searches

The first and most accessible method is to conduct keyword searches in attempt to find hacking forums (Holt & Lampke, 2010). For example, searching for "carding forum" may yield a cybercriminal community focused on credit card fraud, while "black hat forum" may just return a more general-topic cybercriminal forum. Additionally, as cybercriminal forums are an international phenomena, keyword searches can be tailored to find forums of a specific geopolitical region by searching for translated queries (e.g., 'хакер форум,'or Russian for 'cybercriminal forum'), or by including the country or language of interest as a keyword within the query (e.g., "Chinese cybercriminal forum"). The reason keyword searches are successful is simple; some cybercriminal forums value growing as large as they can become, and thus want to be found and are easily accessible. Unfortunately, forums that openly publicize themselves are notorious for attracting "script kiddies," or participants that possess little to no actual hacking skills and are entirely dependent on using hacking tools that more experienced cybercriminals release publicly (Benjamin & Chen, 2012; Holt et al., 2012). These participants, while contributing very little to no valuable information on their own, introducing noise within cybercriminal forums, making it more difficult for researchers to identify real threats, key participants, and more. To further exacerbate the issue, more knowledgeable cybercriminals are generally conscious to not reveal their ongoing activities or credible threats due to concern that law enforcement may be monitoring

forum activity (Motoyama et al., 2011). Thus, while many hacking forums may be identified through simple keyword searches, they do not yield the highest quality content for enhancing cybersecurity capabilities against credible threats. However, keyword searches may be particularly useful for finding cybercriminal forums of different languages; simply translated search terms may yield results.

2.3.3. Snowball Collection

After an initial set of forums are discovered through keyword searches, the next cybercriminal forum identification technique can be implemented. In many cybercriminal forum conversations, participants may reference or share hyperlinks to other cybercriminal communities or underground markets (Benjamin et al., 2015). Such discussions can be exploited by researchers in order to discover new cybercriminal forums, including more private and secretive forums that do not appear indexed by major search engines.

In order to operationalize this task, text parsers can be written to automatically scan through cybercriminal postings collected from previously identified forum. A simple scenario would be to create a text parser than scans cybercriminal forum postings for the string "*http://*" in order to automatically identify and extract hyperlinks shared among cybercriminals; such hyperlinks may lead to other cybercriminal communities. More complex text parsers could be developed utilizing regular expressions, the process of analyzing text by searching for pre-defined patterns. For example, a regular expression could be crafted to scan forum postings for strings that resemble credit card numbers; assuming a hypothetical credit card number consisted of 16 consecutive digits, a regular expression could be used to scan text for patterns of 16 consecutive digits. Cybercriminal posts found to contain such patterns may contain references to underground markets and carding communities. By utilizing a combination of different regular expression patterns,

researchers can exploit their existing data collections in effort to find brand new data sources of interest. Any URLs that are identified can be fed into a web crawler for downloading, which will be detailed in my section on cybercriminal forum collection.

2.3.4. Deep web Hidden Services

While most cybercriminal forums are accessible through the public Internet, there are some underground communities that exist in the deep Internet (or 'deep web'), and are not accessible through traditional means. In particular, much of the deep web exists as anonymized, peer-to-peer networks where network traffic is purposely obfuscated in attempt to protect user identity and conceal activity patterns (Martin, 2013; Benjamin & Chen, 2014). Many users of such networks will often privately host "hidden services," or web services, for use by other network participants. Potential applications of hidden services include benign services such as anonymized web and e-mail hosting, to more nefarious services including underground markets and cybercriminal forums.

Cybercriminal forums acting as hidden services may contain more advanced participants or more sensitive contents than more visible cybercriminal forums, explaining their need to be more secretive in nature (Martin, 2013). By extension, this means that such forums may be of great value to researchers for the purposes of understanding emerging cyber threats, or discovering potential targets of cybercrime. However, gaining access to such forums is nontrivial; a researcher generally requires special software or technical knowledge in order to connect to a deep web network and locate hidden services of interest. Thus, I will outline several steps researchers can take to connect and identify cybercriminal forums within the Tor anonymity network[1], one of the most active

[1]The Tor anonymity network was initially conceptualized in the mid-1990's by the United States Naval Research Laboratory, and was later further advanced by DARPA. The Naval Research Laboratory released Tor under a free license in 2004, where it was subsequently picked up by the open source community for continued development and support. Since then, the network has grown in a variety of directions, including becoming home to a variety of illicit underground communities (Martin, 2013; Benjamin et al., 2015).

networks at the time of this writing. The steps listed are also applicable to other similar anonymity networks.

Download a Network Client: To access a deep web anonymity network, a specialized software client must generally be used to establish a connection and communicate with the network. In the case of the Tor network, a public client can be downloaded from http://www.torproject.org. Options to download the Tor client in various forms exist, but perhaps the easiest to deploy is to download the "Tor Browser Bundle," where a Tor client is packaged as a plug-in into a stand-alone Mozilla Firebox browser. One can simply download the browser bundle and use the included browser to access and browse hidden services located within the Tor network. Additionally, since both Tor and Mozilla Firefox are open source projects, the browser bundle is cross-platform and available on a variety of operating systems. Other anonymity networks besides Tor may have their own custom software necessary for accessing the network.

Identify Hidden Service Directories: Identifying hidden services within the Tor network is a non-trivial task. First, the web addresses belonging to Tor hidden services are generally sequences of random alphanumeric characters. The web addresses thus do not indicate potential functionality or content of the hidden service they are assigned to. Second, Tor hidden services do not utilize traditional top-level domains such as '.com' or '.net.' Instead, Tor hidden services use the '.onion' nomenclature as a reference to the multi-layered network traffic encryption implemented in Tor. This multi-layered encryption is often conceptualized as layers of an onion, and is the reason that hidden services located within the Tor network are commonly referred to as "onion files."

These characteristics of Tor (and other similar anonymity networks) make it difficult for security researchers to identify relevant data sources. Fortunately, there are some hidden service directories that publicize themselves and can be discovered through keyword searches; simply

querying “Tor hidden service directory” on a major search engine will yield lists of various hidden services. For example, this technique yields one of the most well-known hidden service directories, the Hidden Wiki, located at http://zqktlwi4fecvo6ri.onion (Figure 2.3). These directories are generally public, open source efforts that are created and maintained by community members. Known hidden services are typically categorized by their topical relevance or intended use. Further, the directories are not representative of all hidden services in existence, However, they may sometimes include web addresses of some cybercriminal forums and other underground communities. After an initial set of cybercriminal forums are identified from hidden service directories, a snowball collection approach can be taken to find more data sources as described previously.

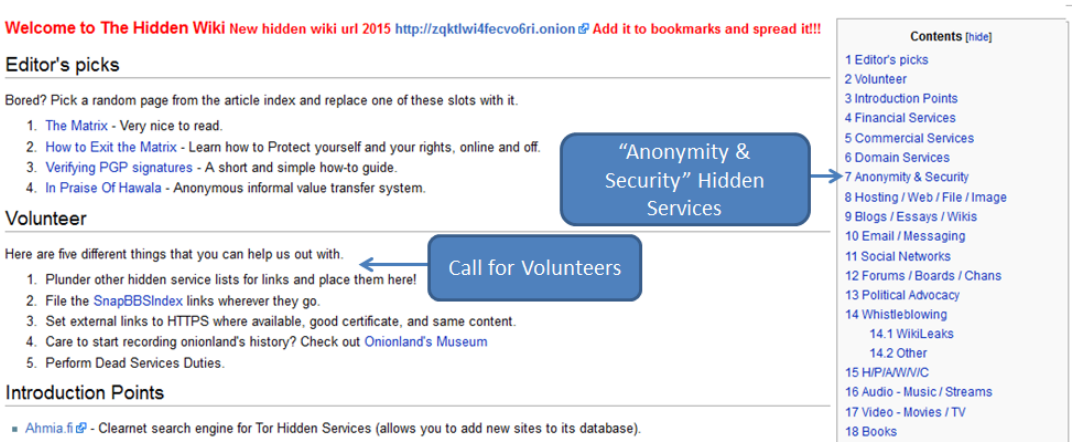

Figure 2.3 – The Hidden Wiki, http://zqktlwi4fecvo6ri.onion

2.4. Forum Collection

After identifying cybercriminal forums suitable for research purposes, forum contents must be downloaded for offline analysis. Previous studies have used manual collection methods such as downloading forum threads one at a time (Holt & Kilger, 2012). However, manual methods are not capable of capturing all forum contents nor are they scalable across multiple forums. Thus,

automated techniques are necessary for developing comprehensive, large-scale collections of cybercriminal forum data.

Typically, the use of web crawlers can be used to automate collection of websites and virtual communities, such as forums (Liu & Chen, 2013). To use a web crawler, one must specify a seed website for the crawler to start at. The crawler will automatically download webpages it encounters, while constantly discovering new webpages for collection by following encountered hyperlinks. Crawling behaviors can be altered depending on collection task requirements. For example, if targeting collection of a single forum, the crawler can be limited to only collecting web pages belonging to the forum's domain or originating IP address; this restriction would prevent the crawler from following hyperlinks to other websites unrelated to the cybercriminal forum. Such restrictions on crawler behavior can lead to more time-efficient crawling tasks while ensuring complete collection. Successfully downloaded cybercriminal data could then be processed and archived for long-term storage.

Unfortunately, using web crawlers to collect cybercriminal forums presents many unique challenges not encountered when crawling more traditional virtual communities and websites. Cybercriminal forums may employ various anti-crawling mechanisms that make automated data collection difficult (Benjamin et al., 2015). While the intention of such mechanisms is generally to prevent surveillance by law enforcement and to safeguard server resources from being abused by rival cybercriminal communities, they also severely affect the capability for security researchers to collect data. Further, cybercriminal forums may include malware or other threats that can harm researchers (Cova et al., 2010). Overall, the cybercriminal data collection process takes careful planning to safely gather and archive cybercriminal forum content.

2.4.1. Creating a Secure Forum Collection Environment

Due to unique challenges faced when attempting to collect cybercriminal forums, it is worthwhile for researchers to plan and create a computing environment for facilitating large-scale data collection. This environment will serve as a data collection pipeline that fulfills two purposes: (1) to aid researchers in safely downloading cybercriminal forum contents, and (2) to provide a method for sanitizing cybercriminal forum data and archiving it for long-term storage and retrieval. A properly set-up collection environment is invaluable to security researchers and will increase the likelihood of successfully executing cybercriminal forum collection tasks.

Data collection from cybercriminal forums can pose many threats to researchers. For example, many forums may embed malicious "drive-by" JavaScript code within web pages in attempt to exploit outdated and vulnerable browsers (Cova et al., 2010). Unsuspecting users, including researchers, would visit the forum and become infected with the malware through regular browsing Thus, it is important to make sure that if a computer used for data collection is infected with malware, that infection does not spread to other computing resources. One way to ensure safety is to take precautionary measures and perform all cybercriminal data collection on computers that are quarantined, or otherwise removed from local networks shared with other computers and devices. Renting virtual private servers from cloud services would also suffice, and provides the added benefit of easily being able to clone servers; one could simply set-up crawling tools on one server, and then clone it to scale collection for different forums.

Additionally, a separate computer should be setup for the sole purpose of hosting a database intended to archive cybercriminal forum data for long-term storage. A simple use-case would be to create a database table per forum collected, with each table record representing a forum post. The record could store forum message attributes such as the author's name, the post title, message

body, date, and other available information of interest. After such a database is created, data processing programs can be run against webpages collected by crawlers in order to extract relevant forum message attributes from surrounding HTML and JavaScript code. Extracted data could then be stored as plaintext within the database. This process also provides some additional security for researchers, as it allows forum data to be viewed without needing to view the original forum webpages that may contain JavaScript-based malware.

2.4.2. Web Crawler Setup

The collection procedure of each cybercriminal forum is unique. Some forums are pleasantly simple to collect, while others are frustratingly difficult, or even impossible without exposing oneself to severe risks or ethical concerns. Specifically, many cybercriminal forums will employ anti-crawling mechanisms, making collection difficult for researchers. However, many of these anti-crawling measures are easily circumvented with some technical wizardry. Table 2.2 contains descriptions and recommended counter-measures to seven of the most frequently encountered anti-crawling mechanisms.

Anti-crawling Measure	Description	Counter-measure
CAPTCHA Images	CAPTCHA images are a type of test used by many web services to determine if the user is human or an automated bot. Their purpose is to prevent bots from accessing contents.	Solve the CAPTCHA manually and bind the generated server session cookie with web crawling software. Requires advanced web crawling software.
Distributed Denial of Service (DDoS) Prevention	The forum detects scripted behavior, such as web crawling, and blocks the associated IP address. This is often done to prevent DDoS attacks against the forum.	Researchers can alter crawling rates and introduce random intervals between web page requests in order to mask crawling activity and avoid triggering DDoS prevention software.
IP Address Blacklists	The forum has blacklisted several IP addresses, including those of public proxy servers and Tor nodes that could be used for anonymity.	Setup a private, dedicated proxy server to reroute crawler network traffic. The proxy server can be deployed using cloud services so it is easy to spawn and destroy proxies for new IP addresses to use.
Paywalls	Forum content is locked behind a registration or access fee.	The only way to access forums with paywalls is to pay their fee. This carries a high risk as the researcher may be defrauded or encounter legal trouble. Recommended to consult law enforcement before pursuing.

User-agent Check	Forums verify HTTP requests come from common browser user-agents, and not web crawlers or other software.	Mimic accepted user-agents during crawling process. Many popular web crawlers possess this feature.
User/password Authentication	Forums require users to register and login before accessing the data.	Register an account with the forum. The registration process is generally completely automated and requires no interaction with cybercriminal forum participants.
Vouching	Gaining access to forum content requires receiving vouches from existing members.	Requires making connections within the cybercriminal community to receive vouchers to private communities. Not recommended for security researchers due to potential ethical problems and biasing data due to researcher manipulation.

Table 2.2 – Cybercriminal Forum Anti-crawling Mechanisms

While most of the anti-crawling mechanisms listed in Table 2 can be circumvented, there are two that I generally recommend against pursuing. Both forums that require registration fees and those that require vouching carry great risks to researchers in the form of fraud, legal risks, ethical concerns, and biasing data due to researcher manipulation. Thus, I strongly recommend that researchers avoid forums requiring payments or vouching, as the downside risks likely excide the benefits. Instead, researchers should focus on less-restricted forums.

For additional security, proxy servers and anonymity networks such as Tor can prove useful to researchers wishing to conceal their identity from cybercriminals. Specifically, whenever an individual or web crawler accesses cybercriminal forum servers, the server will generate log files revealing IP addresses that connected to the server. Thus, the origin IP address of researchers can be exposed, resulting in a significant security risk. Fortunately, proxy servers and anonymity networks can be utilized to re-route researcher web traffic through external connections, effectively concealing the identity of researcher machines from cybercriminal forum servers. Many popular web crawlers natively support proxy server usage, and researchers can simply search for public proxy servers in order to implement a web crawler. In the case of Tor, after a Tor client is installed on a computer, the web crawler can be bound to the client in order to communicate with the Tor network. To do this, the web crawler can be configured to forward

traffic to a SOCKS proxy located at the local network port that the Tor client is listening to for network traffic (by default, this is generally http://127.0.0.1:9051). After this step, the Tor client will automatically handle network communication, and the researcher can continue operating the web crawler as normal.

Overall, researchers willing to take the necessary precautions can safely collect cybercriminal forum data while avoiding many technical risks and without exposing their identity. However, additional steps are needed to extract relevant data from raw webpages downloaded directly from forums.

2.4.3. Parsing Collected Webpages

As web crawlers traverse through cybercriminal forums, they download web pages that must be processed in order to extract information of interest. Text parser programs utilizing regular expressions (similar to those used in snowball collection) can be used to accomplish this task. For example, text parser programs can be written to automatically extract cybercriminal forum postings, author names, thread titles, and other information by identifying patterns of HTML code that correspond with data of interest. Specifically, forums generally follow HTML design templates that contain unique HTML code patterns for encapsulating each forum message and associated author data. Such repeated patterns can be manually identified by researchers and subsequently used within text parsers for automated information extraction across all web pages for a given forum.

2.5. Analytical Directions

After cybercriminal forum contents are collected and parsed into a database, researchers can begin to formulate and test research questions by scrutinizing data through various analytical techniques. However, given the difficulty of identifying and collecting quality data sources within

the cybercriminal context, researchers are somewhat limited to exploring problems that are actually possible to solve with what cybercriminal data they are able to retrieve. I provide some brief suggestions of potential analytical directions that researchers can explore with their cybercriminal data test beds. However, this is not a definitive list of valuable directions to pursue, and I encourage interested security researchers to explore alternative ideas as well.

2.5.1. Cybercriminal Content Analysis

Cybercriminals routinely share cybercriminal assets with one another, including malware, hacking tools, written hacking tutorials, video tutorials, and source code examples (Motoyama et al., 2011; Holt et al., 2012). Such assets can be studied by researchers to understand more about cybercriminal trends, emerging threat patterns, the popularity and discussion volume surrounding different cybercriminal topics. Researchers can focus on developing new techniques to digest and understand the multitude of different assets shared within cybercriminal forums. Not only would the field of cybersecurity benefit from research in this area, but effort spent on creating techniques for analyzing cybercriminal assets could also result in benefits for other streams of research that rely on text processing, malware reverse engineering, video and image analysis, or source code analysis.

2.5.2. Cybercriminal Language Modeling

Practitioners face many challenges when attempting to study cybercriminal community contents. Specifically, cybercriminal communities possess content far removed from the scope of traditional virtual community research (Benjamin & Chen, 2015). Unfamiliar hacking terms, concepts, tools, and other cybercriminal-specific items are regularly discussed, presenting a challenge to researchers wishing to deeply understand community contents. Further, foreign language issues may also arise due to cybercriminal communities existing globally, presenting yet

another barrier to research. Due to these challenges, the research community would derive great value in research exploring methods for understanding the rapidly evolving cybercriminal terms and concepts. In particular, computational linguists could make substantial contributions in this area.

2.5.3. Cybercriminal Network Analysis

Network analyses are a staple component of many related streams of research, as they can reveal knowledge concerning the individuals their relationships within virtual communities. In the cybercriminal context, network analyses are important for understanding the relationships between forum participants, cybercriminal sub-groups, underground markets, and overall community structure. Additionally, network analyses would help cybersecurity researchers and practitioners better assess the credibility of threats emerging within cybercriminal forums. For example, literature suggests that there exists a variation of knowledge proficiency among forum participants (Radianti, 2010; Benjamin et al., 2012). By using network analysis techniques to identify key actors within forums, for example, the credibility of threats identified by researchers can be evaluated based on the associated participant or cybercriminal group.

2.5.4. Underground Economy Analysis

A number of cybercriminal forums possess underground markets where participants buy, sell, and trade cybercriminal assets and services (Benjamin et al., 2015). Thus, there are several opportunities to analyze data from cybercriminal markets, including analysis of underground economy participants, pricing mechanisms, goods exchanged, and more. In particular, there is a need to understand the currencies used by actors within underground markets as they are a potential point of Weakness that may be exploited by law enforcement agencies to disrupt financial transactions between buyers and sellers of cybercriminal assets. For example, cryptocurrencies

such as BitCoin are used because of their perceived security and anonymity, though it is unknown the volume of transactions making use of such payment mechanisms, or how such transactions actually occur.

Additionally, there is a need for quantitative assessments of relationships between underground economy participants by considering the number, shape, and composition of networks in cybercriminal markets (Motoyama et al., 2011; Yip et al., 2013). Research on cybercriminal market network relationships is extremely limited and exploratory, generally using a single forum or small samples of data from multiple forums. More effort in this area would increase my understanding of cybercriminal market dynamics.

2.6. Ethical Considerations

Conducting ethical research is a common concern when pursuing virtual community studies. Traditionally, virtual communities that offer open registration, and are publicly viewable, are typically considered acceptable to use in research (Chang & Chuang, 2011; Liu & Chen, 2013). However, cybercriminal-focused forums present unique ethics questions that are not encountered in traditional work. For example, consider the techniques described for circumventing anti-crawling measures, as well as the use of identity obfuscation to protect researcher identity. Should it be considered fair to deceive cybercriminal forum participants in order to potentially unveil new knowledge that may advance current cybersecurity capabilities? If not, what type of research can be done within the cybercriminal context? How can I strike a balance between ethical considerations and the need to mitigate potential risks that researchers face? Several issues are at hand, and each must be considered individually. I provide a summary of my arguments in Table 2.3, and provide detailed explanations below.

Research Activity	Ethics Violation?	Reasoning	Suggested Action

Research on Private Communities	Low Concern	-Commonly accepted practice in criminology and broader security informatics research	None
Circumventing Anti-crawling	Low Concern	-Anti-crawling mechanisms do not prevent researchers from manually collecting all contents -Crawlers simply allow automated collection for more efficient time usage	None
Masking Researcher Identity	Low Concern	-Necessary for researcher safety -Hidden services are only accessible on anonymity networks	None
Researcher Interaction	High Concern	-May inadvertently support cybercrime -May influence data and bias results	Avoid forums that necessitate researcher interaction, such as those requiring registration fees or referrals from existing members

Table 2.3 – Cybercriminal Forum Ethical Research Heuristics

2.6.1. Research on Private Communities

Ethical use of virtual community data for research is already prevalent across numerous domains. Specifically, researchers will often utilize data from public communities that provide unrestricted access, as discussions within such communities are considered to be open for anyone to freely view and participate in. There is generally little to no expectation for privacy among users participating within such communities.

Conversely, while some cybercriminal forums are public, the vast majority strive for privacy and secrecy. Thus, is there an ethical concern regarding research on cybercriminal forums that are private? To answer this, I can look to related security disciplines. In particular, there exist numerous studies focused on intercepting and stopping the activities of underground virtual communities that exhibit illicit behaviors, such as terrorism groups, pedophiles, traffickers, and so on (Leavitt, 2009; Martin, 2013). Further, studies in criminology and related domains will often utilize non-public data in order to explore important questions. By extension, I argue that it is of low ethical concern to study private and hidden cybercriminal communities.

2.6.2. Circumventing Anti-crawling

Second, I must consider whether it is unethical to collect data from cybercriminal forums that employ anti-crawling mechanisms, as this requires researchers to explicitly circumvent such mechanisms in order to collect cybercriminal forum data. In a sense, researchers are practicing a form of deception when customizing crawlers to navigate around technical hurdles put in place by self-protective cybercriminals. Circumventing anti-crawling mechanisms typically involves slightly altering the collection process to avoid triggering a series of server-side programs that can halt the collection progress. However, forums employing such defenses do not have any inherent characteristics that disallow researchers from simply visiting and downloading all contents manually. Unfortunately, the effort and time necessary to collect forums in this manner prevents researchers from assembling large-scale datasets.

The question thus becomes one of method of data collection rather than whether the data collection itself is ethical. Is it unethical for researchers to automate capture of forum data collection that is otherwise freely available to anyone with a web browser? I argue that such automation is of low ethical concern. Researchers employing web-crawling techniques benefit from automation in the same way that behavioral researchers benefit from automated capture and storage of survey data using online survey software. Using a more efficient method simply speeds up the process. Consider also that no human is being deceived in the collection process, but rather, a series of automated programs unintelligently running on the cybercriminal forum Webserver. Again, a researcher could manually download all forum contents, but automating the task allows for large-scale research that is important for advancing cybersecurity science. Clearly, the benefits overcome the concerns.

2.6.3. Masking Researcher Identity

Another potential concern regarding ethical conduct involves the use of proxy servers or anonymity networks to mask crawler network traffic and to protect researchers. First, as cybercriminal forums may embed JavaScript malware into webpages in attempt to exploit unsuspecting users with vulnerable web browsers, it is advisable to perform all collection on computers dedicated downloading cybercriminal forum data. One of the more practical methods to operationalize this strategy is to utilize virtual private servers provided by cloud providers, as researchers can easily manage computing resources and isolate at-risk machines. However, one consequence using virtual private servers is that researchers would be masking their true IP address affiliated with their organization, and instead would utilize addresses owned by their cloud provider. In this case, masking researcher identity is a side-effect of taking necessary precautionary security measures and using isolated computing resources to collect cybercriminal forum data.

Another method researchers can use to mask their identity is by utilizing the Tor network and other anonymity networks. However, use of such networks is often a requirement to simply access cybercriminal forums and related communities that are acting as hidden services. Such communities may possess unique data that could be important to understanding emerging cyber threats, identification of potential victims, or attributing cyberattacks. As development of better understanding in these areas is critical to improving cybersecurity capabilities, I argue that there is low ethnical concern for researchers to use anonymization networks in order to access hidden service contents.

2.6.4. Cybercriminal-Researcher Interaction

Some forums require potential members to pay registration fees or to possess referrals from existing participants before granting them access to view forum content and participate. Both of

these activities present concerns, as researchers risk being defrauded, may accidentally reveal their identity, or become otherwise compromised. Further both methods expose the researcher to direct interactions with cybercriminals, presenting additional ethics problems of high concern. In particular, researchers that interact directly with cybercriminal forum participants may inadvertently bias or manipulate the data they collect, presenting threats to research validity (Cook & Campbell, 1979). Additionally, legal concerns may arise from such interactions; for example, if a researcher encounters a cybercriminal forum that requires registration fees, any payments submitted by the researchers could in fact be used by cybercriminals to fund future cybercrime.

However, researcher interaction may in some cases be unavoidable and necessary in order to gain access to highly critical data directly related to ongoing cybercrime. In such scenarios, it is recommended to work with law enforcement for consultation and to stay within both legal and ethical boundaries. Further, law enforcement may assist in cases where an active cybercriminal operation is identified by researchers.

2.7. Empirical Study

In order to illustrate the aforementioned cybercriminal forum research guidelines, I conducted a study utilizing data from four forums located within the United States, China, Russia, and Iran. Prior work suggests that cybercriminal forum participants often collaborate by sharing cybercriminal knowledge and assets, or by participating in underground markets (Motoyama et al. 2011; Benjamin and Chen 2012). Underlying many of these cybercriminal behaviors is a strong focus on reputation and trust among community participants (Motoyama et al. 2011; Holt et al. 2012). For instance, many cybercriminal forums utilize internal reputation rating systems, allowing participants to rate the trustworthiness and contributions of others (Fallman et al. 2010;

Benjamin & Chen, 2012). Such systems appear to play important roles in facilitating black market transactions, cybercriminal group formation, community leadership, and so on.

However, despite the importance of reputation within cybercriminal communities, there has been little work investigating the exact mechanism in which cybercriminals gain reputation among peers. Better understanding of cybercriminal reputation may assist in assigning cybercriminal attribution, identifying emerging threats, mapping cybercriminal community topology, or studying the cybercriminal supply chain. Additionally, understanding how cybercriminals accumulate reputation would aid in the identification of key participants of cybercriminal communities that lack explicit reputation systems. Insights into these areas would have value for both research and practice.

Thus, the objective of this study is to better understand the phenomena of reputation within cybercriminal forums, with specific focus on identifying cybercriminal forum participation behaviors that may lead to reputation improvements. To achieve this, I utilize the same cybercriminal forum research guidelines that are presented earlier in this paper.

2.7.1. Background

In order to help guide my investigation of reputation within cybercriminal forums, I borrow perspectives from popular IS theory that may help explain cybercriminal behaviors that contribute to reputation gain. The selected theories are Social Capital Theory (SCT) and Media Synchronicity Theory (MST), as both may help explain cybercriminal forum participation behaviors that give rise to cybercriminal reputation. I provide a light review of each theory, and explain in what context they are helpful.

Social Capital Theory

Many consistent themes have been noted to exist across multiple cybercriminal forums. For example, one frequently documented behavior is that cybercriminal forum participants freely share cybercriminal knowledge and resources among their peers. While not much is known about why cybercriminals may choose to distribute such assets, knowledge and resource sharing has been a long-standing subject of interest in more traditional organizational and IS literature. Specifically, work investigating Social Capital Theory (SCT) may provide a strong theoretical basis for understanding why cybercriminals would openly share assets with one another (Nahapiet and Ghoshal 1998; Wasko & Faraj, 2005). SCT argues that individuals and organizations are motivated to create and share knowledge in order to obtain unique advantages in their relationships between their peers and community, including enhancement of one's reputation. In regards to the cybercriminal context, it may be that individuals share resources in order to build reputation with peers. Increased reputation may allow cybercriminals a number of increased benefits, such as access to more secretive and advanced hacking communities.

However, when studying what contents are shared between community participants, I also must consider differences in how content is shared, and not just what is shared. More specifically, different individuals will craft their forum messages in different ways to be interpreted by other forum users. For example, two different cybercriminals may release hacking tutorials that are similar in content, but if one cybercriminal chooses to stylize and format their tutorial in a way that is easier for others to understand, it may gain more attention. Thus, it may be that cybercriminals who are able to consistently communicate their forum messages more effectively than others may also be perceived as more reputable.

Media Synchronicity Theory

Media Synchronicity Theory (MST) describes how different characteristics of media can be used to help facilitate effectiveness when performing different communication tasks (Dennis et al, 2008). Different media types possess unique capabilities for conveying information or converging on shared meaning between individuals. In particular, each form of media possesses capabilities for message reprocessability, message rehearsability, symbol variety, and so on.

Regarding electronic text, information can be encoded in a variety of ways, such as font type, color, and size, bolding, italics, emoticons, and hyperlinks to name a few. A cybercriminal crafting a forum message could combine the aforementioned media symbols to improve understandability, highlight important information within their post, or otherwise improve the effectiveness of communicating the idea within message content. For example, the cybercriminal in Figure 2.4 includes a screenshot and makes use of different font styles to highlight different information. Effective usage of symbol sets may create an appearance that one cybercriminal's posted messages are more informative, or high-quality than others, potentially influencing their reputation.

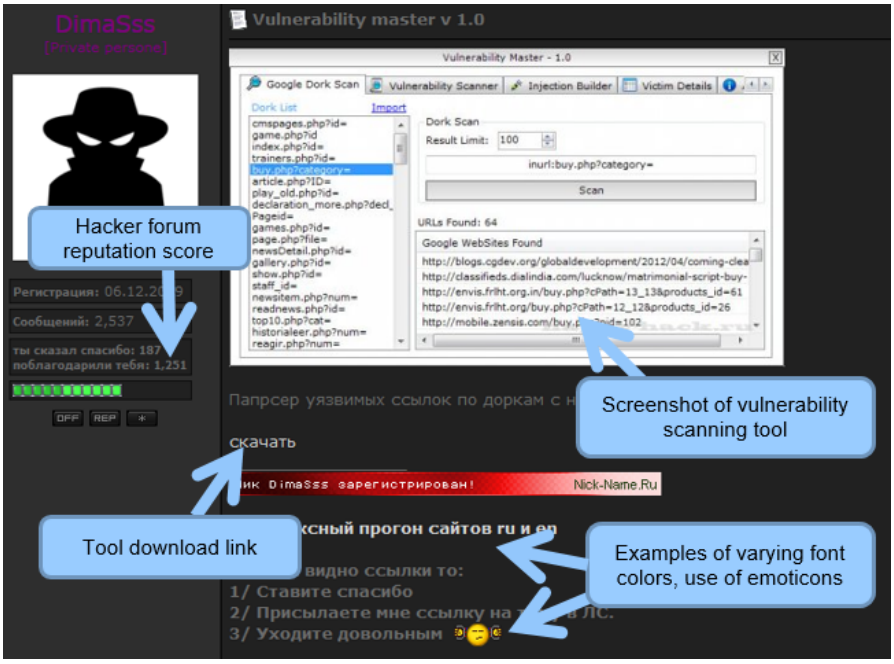

Figure 2.4 – Russian Cybercriminal Forum Posting

2.7.2. Research Gaps and Questions

While researchers have become interested in advancing current cybersecurity capabilities, few inquiries have focused on reputation within cybercriminal forums despite evidence of its importance. Insights concerning the process in which cybercriminals gain reputation among their peers would be beneficial to security researchers and practitioners; such knowledge would allow for identification of key actors within forums that lack reputation system, advancing current cybersecurity practitioner and researcher capabilities. Additionally, little work has been done to analyze and compare cybercriminal forums from differing geopolitical regions. Exploring the differences and similarities international cybercriminal communities across the globe can help develop a better picture of the global cybercriminal community, aiding efforts to improve cybercrime attribution.

From a theoretical perspective, it is worthwhile to identify the boundaries of theories commonly used in IS literature, such as SCT and MST. In particular, little work has been done to test traditional methods and theories on dark networks such as cybercriminal forums. Exploration in this area could yield contributions to better understand both cybercriminals and the boundaries of accepted theory. For these reasons, I am motivated to investigate several research questions. Specifically, through what processes do cybercriminals increase their reputation? What posting behaviors are most effective for increasing reputation? Are there differences in how cybercriminal forum participants behave in different geopolitical regions?

2.7.3. Research Design

My research design consists of data collection, feature set development, and cybercriminal reputation analysis (Figure 2.5). Cybercriminal forum webpages Ire collected using a suite of automated collection tools, including anti-crawling and identity obfuscation methods. Next,

collected webpages Ire processed by utilizing text parsers to extract relevant cybercriminal data, such as, cybercriminal names and messages. Using the extracted data, several features Ire developed that may help explain cybercriminal reputation. I scrutinize the relationship between such features and cybercriminal reputation through a regression analysis.

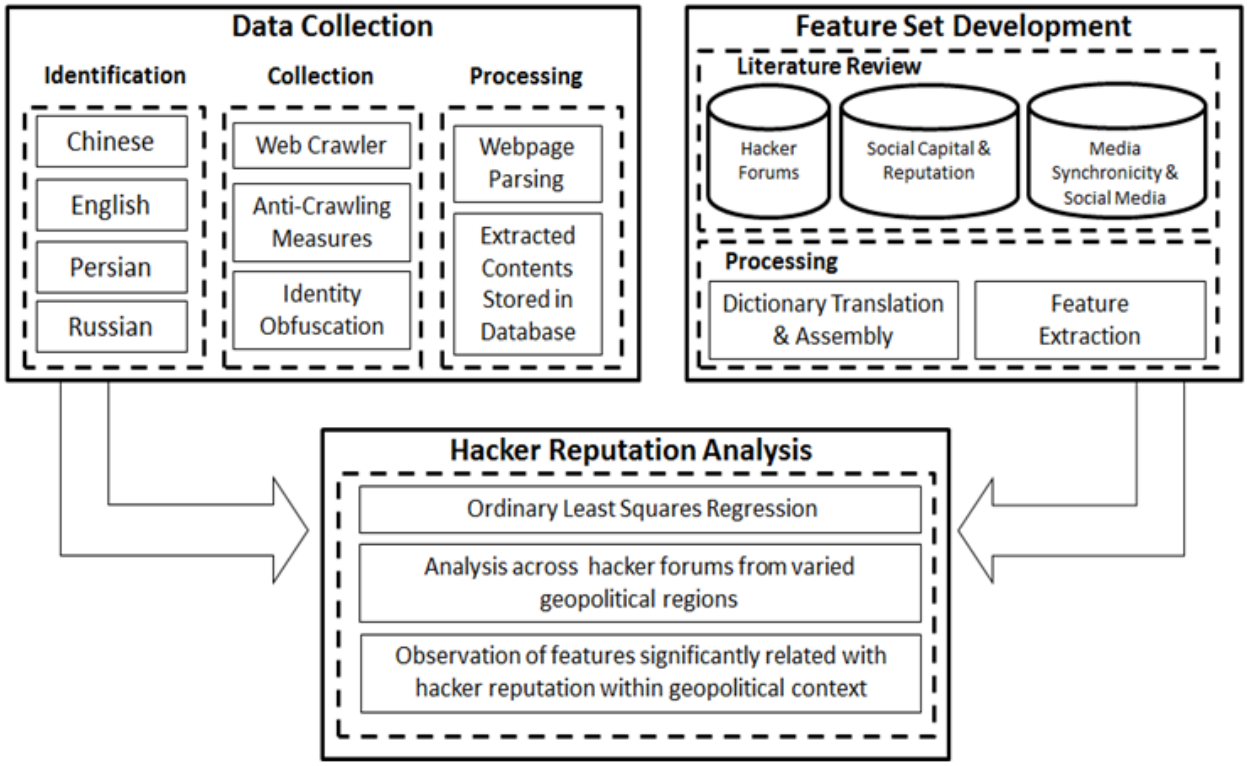

Figure 2.5 – Cybercriminal Reputation Study Research Design

2.7.4. Cybercriminal Forum Identification

To investigate the role of reputation within cybercriminal forums, I explored data from cybercriminal forums located in four different geopolitical regions: the United States, China, Iran, and Russia (Table 2.4). The selected forums are public cybercriminal forums that Ire discovered through a series of keyword searches as described in past research (Holt and Kilger 2012). Keywords such as "Russian cybercriminal forum" and "malware obfuscation" Ire used as search engine queries to reveal potential data sources. Additionally, while these forums are sufficient for

the context of this empirical study, they could be further scrutinized for hyperlinks to other potential cybercriminal communities in a snowball collection procedure for more large-scale research (as shown earlier in Figure 2).

Cybercriminal forums from the selected four geopolitical regions have often been the focus of past research due to growing societal relevance of cybersecurity (Zhuge et al. 2008; Motoyama et al. 2011; Holt and Kilger 2012). Additionally, each one of the identified forums possesses a reputation system that forum participants use to evaluate peers. The reputation system provides a form of ground truth for helping identify key cybercriminals within each forum, allowing us to investigate the mechanisms of how such cybercriminals become reputable among peers.

Forum Name	Language	# of Messages	# of Users	Data Collection Start Date	Data Collection End Date
Antichat.ru	Russian	232,920	11,865	01-01-2003	1-2-2015
Ashiyane.org	Persian	12,903	2,922	08-26-2008	1-2-2015
HackHound.org	English	6,011	817	10-12-2012	1-2-2015
Unpack.cn	Chinese	521,101	18,840	11-12-2004	1-2-2015

Table 2.4 - Research Test bed Summary

The forums I use in my analysis appear to be good representations of the overall cybercriminal community due to the contents and topical coverage they contain. For example, many of the cybercriminal forum contents observed in prior research appear within my dataset, such as hacking tools, tutorials, malware source code, and other cybercriminal assets. Additionally, the forums appear to have many discussions related various facets of cybercriminal culture, such as malware samples, programming, security news, and more. The contributions of this dataset are unique from prior work. Both MST and SCT are two popular theories studied widely within the IS discipline, but neither have been studied extensively within untraditional datasets that fall outside of business and other types of common organizations (cf. George et al, 2013). This study aims to explore the boundaries of these two theories by using them to scrutinize reputation within cybercriminal forums of differing geopolitical origins.

2.7.5. Data Collection

As described previously, I utilized customized web crawling programs to collect forum contents. Identity obfuscation as practiced by routing crawler traffic through the Tor network. After forum pages Ire collected, text parser programs utilizing regular expressions Ire written to extract relevant forum contents from downloaded webpages. Extracted data then was stored in a database for later retrieval and analysis.

2.7.6. Analytical Methods

The core of my experiment is a content analysis performed within a regression framework. I review prior literature and theory to identify a set of features that would be useful for predicting cybercriminal reputation. I developed two categories of features: message content features, based primarily on SCT, and forum usage features, borrowed from MST and other relevant social media studies. Message content features encompass keywords that pertain to technical and cybercriminal-specific knowledge. High frequency of these keywords in a participant's messages may indicate expertise. Additionally, shared cybercriminal tools, source code, and other assets are also included as message content features, as cybercriminals can accumulate social capital by disseminating such assets. Conversely, forum usage features capture user behaviors and characteristics such as posting frequency, forum threads (i.e., conversations) started, seniority/tenure, message symbols supported by the web forum software, and so on. Table 2.5 contains a comprehensive list of all features used.

Category	Feature	Description	Source
Message Content Features	Attachment of Cybercriminal Assets to Forum Posts	Forum participants sometimes attach cybercriminal assets to their forum posts. These include written & video tutorials, hacking tools, cracked software, etc.	Motoyama et al, 2011; Benjamn & Chen, 2012
	Embedding of Source Code within Forum Posts	Cybercriminals sometimes share source code for tools, malware, etc. by embedding them directly into forum posts	Holt & Kilger, 2012; Benjamn & Chen, 2012

	Discussion of Attack Vectors and Hacking Concepts	Demonstrates participant proficiency; Examples: *Rootkit, XSS*, *SQL Injection*, *DDoS*, *shellcode, PoC, drive-by*	Zhuge et al, 2008; Benjamn & Chen, 2012
	Discussion of Programming and other Technical Concepts	Demonstrates participant proficiency; Examples: *SQL*, *C++*, *ASM*, *.Net*, *XML*	Radianti et al, 2009; Holt & Kilger, 2012
	Reputation System Scores	Peer-evaluated cybercriminal reputation; reputation is not uniform across forums, i.e., a participant with good reputation in one forum may not necessarily have good reputation in another	Nahapiet & Goshal, 1998; Radianti, 2010; Benjamin & Chen, 2012
Forum Usage Features	Media Symbol Diversity	Total usage of media symbols such as font color, font style, text bolding, italics, etc. at a per-message level	Dennis et al, 2008
	Number of Threads Started	The number of threads a forum participant has started, normalized to the user's total number of posts.	Radianti, 2010; Motoyama et al, 2011; Holt et al, 2012
	Number of Posts Made	The number of posts a forum participant has made	Radianti, 2010; Motoyama et al, 2011; Holt et al, 2012
	Seniority	The number of days a forum participant has belonged to a forum	Radianti, 2010; Motoyama et al, 2011

Table 2.5 – Forum Content and Usage Features

I utilized feature dictionaries to operationalize a lexicon-based named entity recognition task for detecting discussion of hacking and technical concepts. I created dictionaries by utilizing two methods. First, I manually compiled lists of hacking and technical terms I identified from literature. Second, I performed some basic text analytics to extract potential hacking terms. For example, I used td-idf and word frequencies to identify potentially important cybercriminal terms that Ire not found to be mentioned in prior research.

Content and forum usage features Ire used as independent variables, while extracted cybercriminal reputation scores Ire the dependent variable. Additionally, along with features extracted from forum contents, I also must account for the different geopolitical context each forum exists within. It may be that the location a cybercriminal forum resides in may affect reputation mechanisms, and thus the relationship between location and reputation must be measured. To operationalize this, I utilize four dummy variables with each variable corresponding

to one of the geopolitical regions the forums are located within (China, Iran, Russia, and the United States). The use of such variables allow us to scrutinize any potential effects that geo-location may have on cybercriminal reputation. Additionally, measuring the impact of geo-location for each forum allows us to make more meaningful observations concerning the process in which cybercriminals gain reputation within different geopolitical contexts.

To conduct my experiment and verify my model, I utilized ordinary least squares (OLS) regression. OLS regression is an intuitive and commonly used maximum likelihood estimator, or method of estimating parameters within a statistical model (Pohlmann & Leitner, 2003). I perform the regression experiment separately on each set of forum data.

2.7.8. Results & Discussion

The results of my experiment are summarized in Table 2.6. First, it appears that contribution of knowledge and cybercriminal assets appears to be directly related to cybercriminal reputation. Specifically, cybercriminals that attach assets (e.g., hacking tools) or embed source code examples within their postings seem to have the highest reputation scores. Results are in agreement with SCT.

I also observe a positive relationship between reputation and frequent media symbol usage appears related in the Antichat (Russia) and Ashiyane (Iran) forums. Participants of these forums often use media symbols to format their messages and highlight pertinent information within postings, as suggested by MST. However, media symbol usage was not significant in the Hackhound (U.S.) and Unpack (China) communities; closer scrutiny of these forums revealed media symbols are not well-supported by forum software and thus not heavily used by participants. Analysis of additional forums would be beneficial.

	Antichat (Russia)		**Ashiyane (Iran)**		**Hackhound (U.S.)**		**Unpack (China)**	
	Estimate	*P-value*	*Estimate*	*P-value*	*Estimate*	*P-value*	*Estimate*	*P-value*

	H1 Contribution of Knowledge and Cybercriminal Assets							
Attachments	0.1275	**0.0062****	0.0235	**0.0472***	0.0204	**0.0074****	0.0173	**0.0181***
Embedded	0.0194	**0.0150***	0.0149	**0.0342***	0.0686	**0.0456***	0.0151	**0.0021****
Tech_Terms	0.0106	0.5517	-0.0106	-0.2023	0.0032	0.1657	-0.0005	0.2822
Hack_Terms	-0.0149	0.4928	0.0049	-0.173	0.0045	0.1804	0.0064	0.533
	H2. Forum Usage Behaviors							
Msg_Symbols	0.0323	**0.0291***	0.0224	**0.0321***	0.1105	0.1222	0.0018	0.1232
Threads_Started	-0.023	0.6956	-0.0049	-0.2549	-0.40201	0.3801	-0.0043	0.9054
Seniority	0.0056	0.8222	-0.0048	0.9976	-0.0903	0.8151	-0.0280	0.4190
Total_Posts	0.0145	**0.0024****	0.0249	**0.0082****	0.4864	**0.0023****	0.0343	**0.0026****
	H3. Forum Similarity Across Geopolitical Regions							
Geo-location	0.0032	0.8222	-0.0048	0.9976	-0.0143	0.8151	0.0218	0.4190
	Signif. Codes. ''<0.01, '*'<0.05**							
R^2	0.3184		0.4131		0.5299		0.3460	

Table 2.6 – OLS Regression Results

Lastly, none of the dummy variables representing geo-location appeared to be significant across the forums. This result implies that geopolitical origin has no significant influence on the relationship between cybercriminal behaviors and reputation.

To help demonstrate the value of my methods, I provide an example of a key cybercriminal identified from the U.S.-based forum Hackhound.org (Figure 2.6). The user *H**s* is one of the top sharers of cybercriminal tools in the Hackhack.org community, often participating in many different discussions simultaneously across the HackHound.org community. *H**s* has received good reputation rankings from his/her peers for contributions of hacking knowledge and assets that others have found useful. *H**s* is the top sharer of hacking tools, malware, etc. with nearly 30 unique hacking asset contributions. By more closely observing this user and others like them, researchers and security practitioners can better understand what threats and malware are becoming popular or more frequently distributed within cybercriminal communities, leading to better defense against potential attacks.

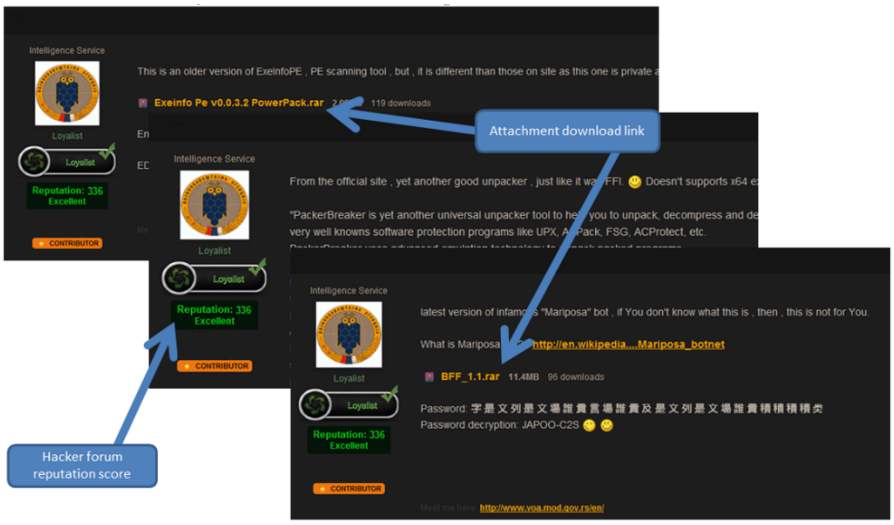

Figure 2.6 - Hackhound.org is the user *H**s*

2.8. Conclusion

As computing technologies become more ubiquitous within society, cybersecurity has become a problem of growing importance and concern. As a result, researchers have become increasingly interested in exploring cybercriminal social media in order to learn more about cybercriminal social behaviors, emerging threats, and the cybercriminal supply chain. However, until now, few works in recent years have successfully performed large-scale identification, collection, and analysis of cybercriminal-generated data. In particular, many cybercriminal-operated web forums exist that can be studied to inform new perspectives on cybercrime, but these data sources have largely gone untapped by researchers.

This research essay provides a set of guidelines for conducting large-scale cybercriminal forum research in order to support researchers wishing to enter the cybersecurity research stream. Four steps to conducting cybercriminal forum research are outlined, including (1) identification of data sources, (2) data collection procedures, (3) potential analytical directions, and (4) ethical concerns related to cybercriminal forum research. An empirical study is included to illustrate the suggested research guidelines. The study investigates the nature of reputation within cybercriminal forums by utilizing perspectives from two popular IS theories.

The main contribution of this work is in providing guidelines and methodological details to assist with operationalizing cybercriminal forum research. The included empirical study also provides some contributions to literature regarding the applicability of popular existing theory within the context of cybercriminal forums. I hope my guidelines help in increasing cybercriminal forum research programs, aiding both theoretical and practical contributions.

3. ESSAY II: EXAMINING HACKER PARTICIPATION LENGTH WITHIN CYBERCRIMINAL IRC COMMUNITIES

3.1. Introduction

Cybersecurity is one of the largest issues impacting the whole of society as individuals, industry, and government find themselves increasingly at risk of cyber-attack. News reports concerning cybercriminals stealing consumer data or cybercrime committed against high-profile targets have recently become of common occurrence. It is estimated that cybercrime costs the global economy about $445 billion a year, mostly due to theft of intellectual property within developed countries and sale of stolen personal information (Sandle & Char, 2014). In 2011, the National Security and Technology Council released a report claiming that "methods to model cyber adversaries" is one critical yet unfulfilled research need (NSTC, 2011). Further, many scientists consider that research which furthers my understanding of cybercriminals would greatly benefit the development of future cyber defenses (Mahmood et al., 2010). The security of my cyberspace will remain a problem of large magnitude for the foreseeable future.

Some of the early works investigating cybercriminals and their behaviors have identified that cybercriminals congregate within various online communities to share knowledge and form groups (Benjamin & Chen, 2012). The scrutiny of data from such communities can lead to actionable intelligence for security professionals. For example, in early 2014, analysis of the *Anonymous* hacking group's IRC community helped British cybersecurity analysts reveal botnet operators actively participating in cyber-attacks (Schone et al., 2014).

However, not all cybercriminal community members are equal. There exist different levels of cybercriminal capability, knowledge, and interest among participants (Holt & Kilger, 2012). Some possess little to no skills and may only have passing interest in cybercrime, while others engrain

themselves within a community and become long-term members. While expected, variance in participants can present a challenge for researchers and professionals. Researchers wanting to identify emerging cyber threats based on cybercriminal community data would have to scrutinize the credibility of observed participants. Threats made by long-term members and those who appear to be key participants of their community should be considered with more priority, as such cybercriminals may be more successful in achieving their cybercriminal goals. For example, Figure 3.1 contains a recruitment advertisement for *#OperationGreenRights*, a hacktivist campaign targeting corporations and organizations accused of significant environmental damage. This advertisement was propagated among cybercriminal IRC communities by many long-term participants, giving the hacktivist campaign more credibility, and thus potentially attracting more participants.

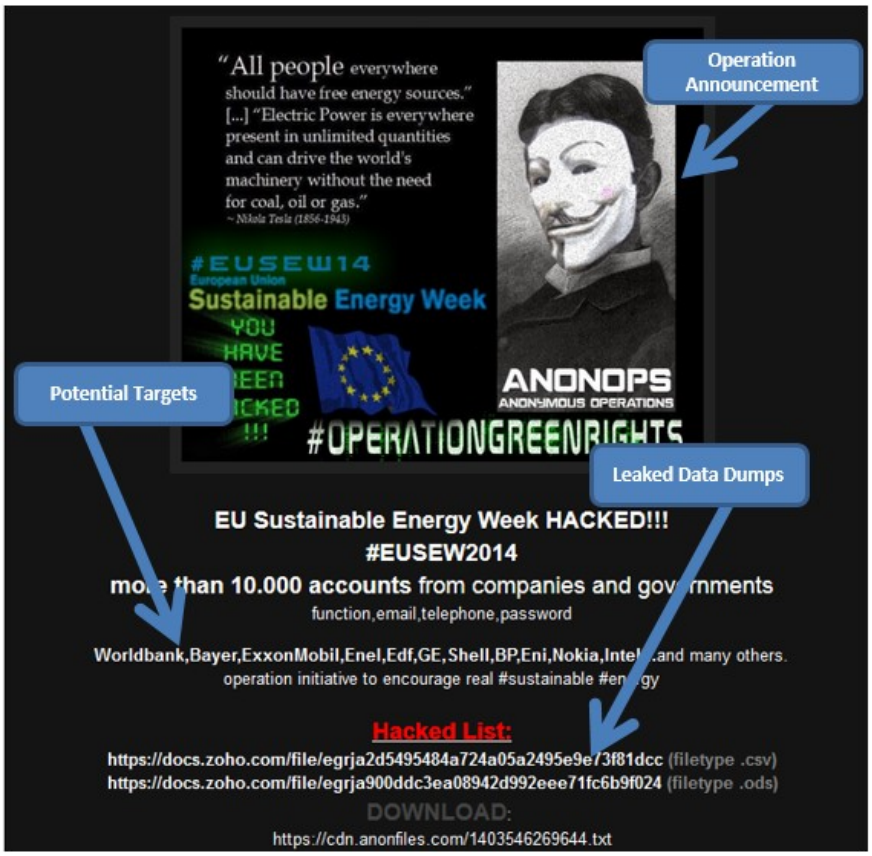

Figure 3.1 - #OperationGreenRights Recruiting Advertisement

For this reason, I am motivated to develop a system for collecting and analyzing cybercriminal IRC data in order to identify potential long-term and key cybercriminals. In particular, I scrutinize individual participants' networking and message content patterns to develop a set of features useful for identifying such cybercriminals. Specifically, I make use of the Wang-Chang Kaplan-Meier estimator and extended Cox's proportional hazards to model data from two popular cybercriminal IRC communities that appear representative of the greater cybercriminal community.

The remainder of this paper is organized into the following sections. First, I review relevant recent studies including literature on cybercriminal communities, traditional virtual community research, and studies utilizing duration modeling. A review of prior relevant research enables us to identify gaps in the existing research, helping to guide the formulation of research questions. Next, the research design is presented, followed by a presentation of the study's results. The paper concludes with a discussion of the study's implications for future research and practice.

3.2. Literature Review

To form the basis for this work, I review literature from related areas of research. First, while the cybercriminal community research stream is young, existing studies are helpful for motivating my exploration and identifying research gaps. In particular, prior work can provide contextual information on cybercriminal communities. Next, I look to prior studies focused on more traditional virtual communities. The abundance of literature in this area can provide information on community identification, collection, and analysis procedures that may be helpful for guiding the formulation of a research design for this study. Lastly, I look at methods for conducting temporal analysis to help with model selection for my analysis. Specifically, duration modeling is

one technique that appears to be used in prior virtual community studies for scrutinizing community participant behaviors.

3.2.1. Cybercriminal Community Research

Overview

To develop proactive cybersecurity capabilities, it is necessary to begin understanding cybercriminal behaviors and surveilling cybercriminal communities for knowledge of emerging threats. Unfortunately, current work on cybercriminal communities is limited. In Table 3.1, I provide a summary of some cybercriminal community literature from recent years. This list is by no means comprehensive, but is representative of the greater body of literature. Many of these works explore cybercriminal community contents, but utilize only manual qualitative analyses and basic metrics. Little work appears to make use of more scalable techniques despite a need to focus on such methods.

Previous Studies	Data Sources	Research	Analytical Methods	Findings
Holt, 2013	Forums	Cybercriminal black markets	Manual qualitative analysis of Russian cybercriminal forum postings	Price, customer service, and trust influence relationships between black market actors
Yip et al., 2013	Forums	Cybercriminal black markets	Combination of manual analysis and network analysis of two cybercriminal carding forums	Underground trading facilitated by social networking, reputation, and quality control
Martin, 2013	Forums	Cybercriminal Black markets	Manual analysis of the *Silk Road* cryptomarket and forums	*Silk Road* and similar cryptomarkets will assume greater share of global trade of illicit drugs
Benjamin & Chen, 2012	Forums	Reputation in Cybercriminal Forums	Content and network analysis to assess how cybercriminals gain reputation among peers	Contributions to the cognitive advancement of a community lead to reputation gains among cybercriminals
Holt et al., 2012	Forums	Cybercriminal social networks	Manual qualitative analysis of Russian cybercriminal forum postings	Cybercriminals practice a meritocratic culture, majority of participants are unskilled

Holt & Kilger, 2012	Forums and other cybercriminal webpages	Cybercriminal skill in global hacking community	Manual qualitative analysis of contents and networks found within cybercriminal forums and other cybercriminal-related web pages	Global distribution of skill among cybercriminals is similar; few top-skilled cybercriminals, most are unskilled
Motoyama et al., 2011	Forums and IRC	General exploration	Manual content analysis, some network analyses	General descriptions of cybercriminal interactions in forums and IRC, existence of meritocratic structure
Yip, 2011	Forums	Cybercriminal black markets	Manual analysis of two carding forums	Underground trading facilitated by reputation and trust
Fallman et al., 2010	Forums and IRC	Cybercriminal black markets	Implementation of collection system to gather information and measure usage of underground economies in cybercriminal forums and IRC	Present usage statistics of cybercriminal black markets, describe collection system

Table 3.1 – A summary of recent cybercriminal community studies

However, methodological limitations have not prevented researchers from observing similar findings across studies. First, cybercriminals communities serve as centers where participants seek collaborators and share cybercriminal assets such as hacking tools, malware, tutorials, and more (Radianti, 2010; Benjamin & Chen, 2012). Thus, cybercriminal communities can provide researchers with data directly relevant to identifying emerging cyber threats and evolving cybercriminal techniques. Second, cybercriminal community participants vary widely in skill and knowledge (Holt et al., 2013; Yip et al., 2013). Some expert participants exist, but many cybercriminal community members possess little skill and provide little value to researchers and practitioners wanting to identify credible, emerging threats. In order to advance current cybersecurity capabilities, effort is needed in the area of accurate and quick extraction of the most credible cybercriminals and threats. Lastly, cybercriminal communities are known to exist across various geopolitical regions, and are especially common in the US, China, Russia, and the Middle-East (Motoyama et al., 2011; Benjamin & Chen, 2012). Development of new research methods in this space can be potentially applied to cybercriminal communities on a global scale.

Cybercriminal Internet-Relay-Chat (IRC) Communities

Many cybercriminals congregate within various IRC communities (Radianti, 2010; Motoyama et al., 2011; Benjamin & Chen, 2014). IRC can support real-time chat among thousands of users simultaneously. It has traditionally been used for legitimate purposes, such as providing discussion and technical support groups for Linux distributions, but it has seen adoption by cybercriminals over time (Jones, 2002; Schone et al, 2014). To utilize this platform, an IRC-specific software client is needed to connect to existing IRC networks. Within an IRC network, several chat channels are open for users to freely join and participate within. Any messages that are posted by users are instantly broadcast to all other connected participants. This differs from forums, where individuals browsing messages may only do so by viewing one forum thread at a time. The broadcast nature of IRC ensures that all participants receive every single contributed message, while forum users will only see messages within threads they manually view. Figure 3.2 provides a visual summary of this architecture.

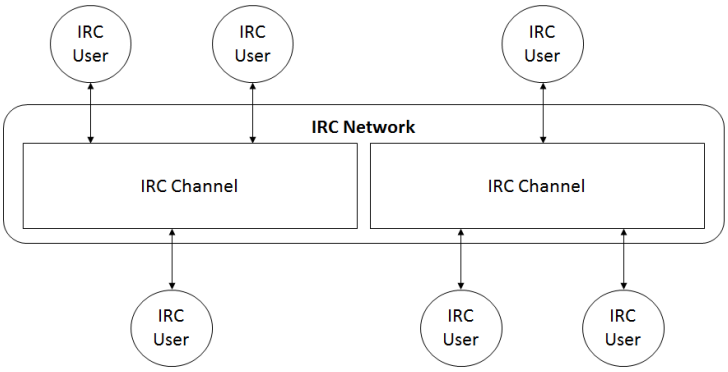

Figure 3.2 – IRC system architecture

IRC contains many unique features that make it a worthwhile platform to study. First, many traditional virtual communities, such as web forums, are divided into several sub-categories that guide focused discussion. IRC generally lacks such sub-categories, and instead has more of a free-flowing conversation. This leads to cybercriminals discussing a wider variety of topics within IRC,

some of which may potentially not be discussed within specialized forums. Second, some virtual communities act as natural archives of data, where threads and posts are stored and can be easily accessed years later. Conversely, IRC contents must be collected in real-time and are not normally archived anywhere for later retrieval. This difference also impacts data source identification, as archived forum data can be indexed by search engines, while IRC channel contents cannot be. Lastly, the purpose of IRC is to support real-time chat, and thus it is not uncommon to see hundreds or thousands of short (i.e., 1-2 sentences) messages per day among participants (Jones, 2002; Benjamin & Chen, 2014). Thus, IRC experiences vastly greater message volume and supports more fluid discourse among participants when compared to other forms of virtual communities (e.g., forums).

Example Cybercriminal IRC Contents

I show an example of a cybercriminal IRC in Figure 3.3. In this example, I view IRC participants within the channel #Anonops on irc.anonops.org server. The #Anonops channel is considered as one of the primary hacker IRC channels affiliated with the *Anonymous* hacking group. Here I witness one participant broadcasting a message that asks if any other participants have the URL to the underground black market Silk Road, which was temporarily taken offline by authorities in October, 2013. Another IRC participant responds with a '.onion' link, a type of deep web address, which leads to the Silk Road forum and new website.

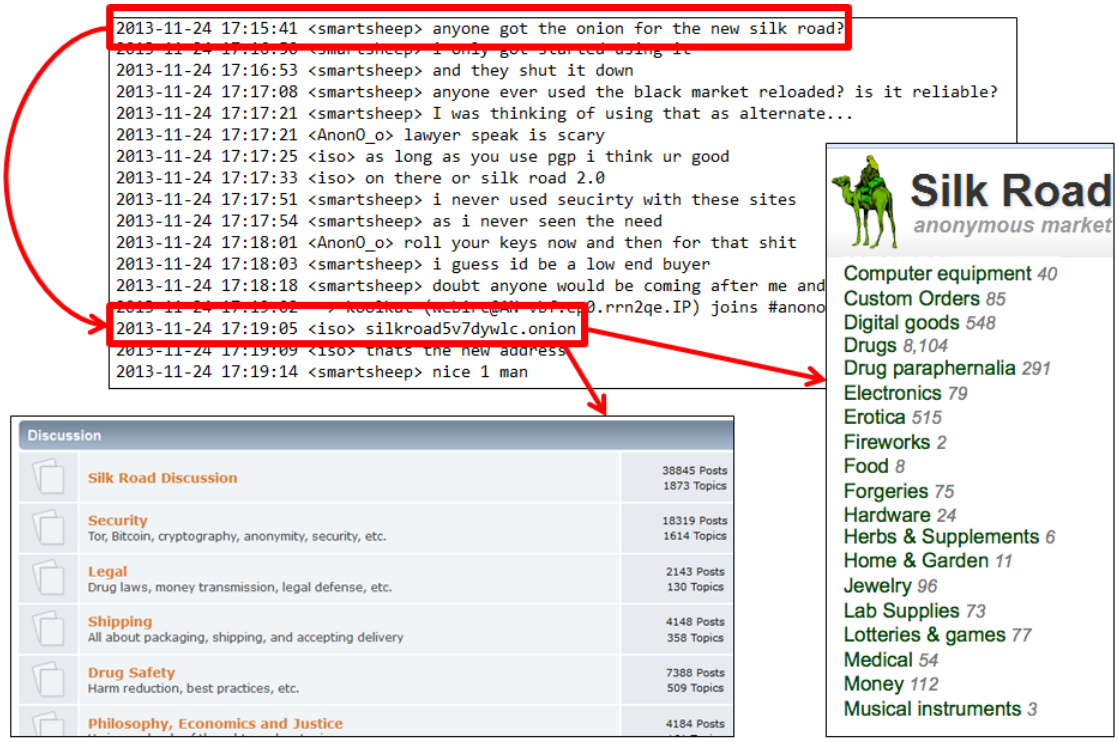

Figure 3.3 - Example of IRC channel activity and contents.

Another example from the #Anonops channel includes discussion of a hacktivist campaign and denial of service attacks. #Anonops members organized a hacktivist operation called *#Optestet,* which blames and targets the French Ministry of Defense for causing environmental damage and for causing the death of a young environmentalist protestor. In particular, hackers involved with the campaign attempt to recruit other individuals and have even posted a recruitment video on YouTube (Figure 3.4).

Figure 3.4 - Recruitment video for #Optestet environmental hacktivist campaign

Many other communities exist beyond just #Anonops. The #Evilzone channel is the primary channel used on the Evilzone IRC network, a self-proclaimed hacking and security community. While the #Evilzone community is not as large as #Anonops, there exist unique discussions occurring with the #Evilzone channel that warrant deeper inspection. In particular, the #Evilzone community appears to host discussions concerning the identification and exploitation of exploiting vulnerabilities related to near-field communication (NFC) technology used in development in products such as RFID-enabled credit cards, automobiles, mobile phones, and more. In particular, I have observed #Evilzone participants discussing deep technical details of how to conduct attacks against NFC devices, including suggestions of tools and tutorials. Figure 3.5 includes a video tutorial of how to use a $300 tool to conduct NFC cracking against vulnerable. This video was shared among the #Evilzone community. In the video, the automobile seemingly has all of its electrical systems and ignition powered on by simply using the NFC interception device to emulate a legitimate wireless key. It is suggested by the IRC participants that the same technique could be used to record and replay any RFID signals, leading to exploits such as fraudulent credit card

charges. Overall, the #Evilzone community appears to host some individuals knowledgeable in NFC cracking and other cybercrime techniques.

Figure 3.5 – Video tutorial shared in #Evilzone that depicts car hijacking using a NFC interception tool

As observed previously in Table 1, the majority of current cybercriminal community research experiments with forum data rather than IRC channel data. This may be perhaps due to easier accessibility of webpage-based forums than IRC channels requiring connection through a specific IRC client. Further, due to the lack of data archiving in IRC, search engines do not index IRC contents. Thus, researchers attempting to identify cybercriminal communities for study are more likely to come across cybercriminal forums than IRC channels. Nonetheless, the importance of IRC is evident as per real world apprehension of cybercriminals (Schone et al., 2014). More research is necessary on IRC-based cybercriminal communities.

IRC Identification & Collection Techniques

At the start of every cybercriminal community exploration, data sources must be identified and collected. There appear to be a few dominant methods used in past studies to identify cybercriminal communities. First, some researchers simply refer to third parties for information (Motoyama et al. 2011). Some security-related organizations or experts may be helpful in identifying cybercriminal communities. Second, some researchers conduct keyword searches in attempt to find forums on their own (Fallman et al. 2010; Holt & Lampke, 2010). This approach is useful for identifying contents indexed by search engines, but will miss contents within IRC channels and deep web hidden services. Lastly, snowball sampling is incredibly fruitful in producing results, including those not indexed by common search engines (Holt & Lampke, 2010). A snowball sampling procedure consists of scrutinizing known cybercriminal communities for hyperlinks and references to other potential communities or hidden services. This procedure is one of the few reliable methods to identify unindexed data.

After successful cybercriminal community identification, data can be collected through various means. As seen from Table 1, forum communities have often been collected manually. A computational approach can also be adopted by making use of automated crawlers to collect contents from webpage-based communities. For IRC contents, specialized listener programs can be developed to utilize the IRC protocol and sit-in on known cybercriminal IRC channels (Fallman, 2010; Benjamin & Chen, 2014). The listener programs can passively log all data transmitted between channel participants.

After data is collected, it can be processed and analyzed. As seen from Table 1, the majority of existing studies make use of manual qualitative techniques or metric-based approaches for analyzing cybercriminal community contents. Research using more traditional virtual community data is useful for identifying analyses for conducting research on cybercriminals.

3.2.2. Virtual Community Analysis Techniques

With the advent of web 2.0, researchers became interested in closely examining the behaviors of individuals in social media. The result is numerous studies that investigate the relationships between different virtual community participation behaviors, and thus may be useful for helping guide the formulation of a research design to study cybercriminal communities. In particular, I review a sample of literature from this research stream that may be helpful for providing methodology to ultimately conduct analyses of cybercriminal IRC contents.

Virtual community research methodology generally revolves around network and content analyses. Network analyses can be used to understand group structure and participant interactions. Content analysis is useful to understand conversations among participants. Such methods are often automated or semi-automated for conducting statistical analyses over large datasets (Abbasi & Chen, 2008; Balahur et al., 2010; Benjamin et al., 2014). Through automation, it is more feasible to scale analysis across multiple virtual communities, to account for temporal data, and perform other forms of analysis.

Automated network analysis are generally operationalized by capturing and measuring interaction among virtual community participants. In many cases, interaction may include participants directly messaging each other, posting in the same forum thread, or becoming "friends" in a social network (Zhang et al., 2012). Capturing ties is helpful for understanding community structure, identifying relationships between community participants, and for identifying key individuals. In the context of IRC, previous researchers have successfully built network ties based on direct addressing among users (Sinha & Rajasingh, 2014). Specifically, messages transmitted in IRC channels are broadcast to all channel participants, and thus a common etiquette was developed for channel participants to explicitly mention each other's names in direct

addresses. Additionally, networks evolve; tenure within a community can impact interaction with other participants (Benjamin & Chen, 2012). Individuals that remain active in a community may begin to be seen as more trustworthy or knowledgeable by other participants, which may drive more interaction.

For content analyses, researchers often rely on methodologies rooted in statistics and natural language processing to understand discussions in virtual communities (Garas et al., 2012). The most basic analyses methods include keyword counts, message volume calculation, or message length metrics. More advanced analyses include the use of machine learning classifiers to perform topic detection and categorization of virtual community contents (Liu & Chen, 2013). A combination of these techniques can be used to assemble a profile on the types of discussions that virtual community participants engage in. For example, such techniques can be used to detect how often a cybercriminal IRC participant discusses a particular type of cyber-attack.

By applying content and network analyses to cybercriminal IRC data, I can extract characteristics for each participant, such as social interaction patterns with other cybercriminals, types of content posted, participation frequency, and more. In a sense, a profile is developed for each cybercriminal IRC participant. The profiles of long-term participating and key cybercriminals can then be evaluated and compared against less interesting participants in order to gain better understanding of what characteristics are unique to key cybercriminals.

Additionally, network and content analyses can be enhanced when combined with a technique that enables temporal analysis of data. By accounting for time, network analyses can be extended to scrutinize how a cybercriminal's interactions with other IRC participants changes over time. For example, I can observe whether a cybercriminal becomes more embedded within their community over time by evaluating their interaction patterns. Similarly, content analyses can

provide new insights when incorporating the time, such as evaluating whether a cybercriminal's posting frequency increases or decreases over time. This type of analysis can help us determine between whether a particular cybercriminal is a key participant, or if they simply have passing interest and post less over time. A branch of statistics referred to duration modeling can help with such analyses.

3.2.3. Duration Modeling

Duration modeling, also known as "survival analysis," and "event history model," is useful for modeling of data that involves prediction of an event at a given point of time (Van Den Berg, 2001). In other words, the dependent variable in the model is duration or the time it takes for an event to happen, and it can be used to understand why a specific event occurs relative to time and other researcher-defined variables. It was used traditionally in the medical and health domains for multiple modeling problems of interest, for example, predicting when patient hospitalization may occur given age, weight, and smoking habits (Lin et al., 2014). Duration modeling is also popular in economic and social science contexts, where it has helped produce helpful findings that described virtual community participation behaviors in contexts such as online health support groups and the volunteering habits of Wikipedia editors (Wang et al., 2012; Zhang et al., 2012).

There are a few features necessary for operationalizing duration modeling. First, the time variable must be defined, which consists of the length of time until an event occurs (or the time between two events in recurrent event modeling). Next, the event variable must be defined, which is meant to code for whether an event happens for a particular record of data (e.g., patient experiences hospitalization in a longitudinal study on patient hospitalization rates). Some techniques for durational modeling can account for multiple different event variables, or multiple events of the same type for one record. Lastly, another important variable in duration modeling is

the censor variable, or whether a record of data drops out of your sample without the event occurring (e.g., a patient that experiences no side-effects for the entire duration of a clinical trial for a new pharmaceutical).

To help operationalize duration modeling, I review two of the most widely used event modeling techniques. Each of the two techniques provides slightly different outputs and insights, but both can be paired effectively to develop deeper understanding of a given data set. The first technique, the Kaplan-Meier model, is useful for developing an overall perspective of how a modeled event occurs over time to observed records (Bewick et al., 2004; Zhang et al., 2012). The second technique is the Cox's proportional hazards model, which is useful for diving deeper and understanding what covariates may positively or negatively influence the probability of an event occurring to a particular record (Van Den Berg, 2001; Wang et al., 2012).

Kaplan-Meier Estimator

The Kaplan-Meier model is most often used to calculate the "survival function" of records. That is, the Kaplan-Meier model is a decreasing step function useful for producing descriptive information about the average length of time a given record in your data will exist (i.e., "survive") without experiencing the defined event variable. In the context of cybercriminal community participation, I can understand how long the typical member participates for and what is the likelihood they stop participating at different time points.

Cox's Proportional Hazards Model

However, while the Kaplan-Meier estimator helps formulate generalized perspectives about data, it fails to describe underlying features that would explain the survival function's shape. The Kaplan-Meier estimator can help predict how long it takes for an event to occur on average, but

does not provide an explanation as to why. Instead, the Cox's proportional hazards model is useful as it helps provide explanation by scrutinizing the relationship of any number of covariates and the event variable. The model is used within a regression framework in order to evaluate the effect of various independent, explanatory variables and hazard (Cox, 1972). In other words, while the Kaplan-Meier is useful for exploration, the Cox's model helps develop a deeper understanding. Specifically the Cox's model will tell us what specific features positively and negatively affect event occurrence, providing some explanatory power to duration modeling. In the context of cybercriminal community research, it may help explain what cybercriminal behaviors may affect participation activity.

Extended Models

Traditional duration modeling is limited to capturing the occurrence of a given event only once per record. If wanting to track multiple events variables per record, the standard Kaplan-Meier and Cox's models are insufficient. For example, tracking multiple participation events per individual within a virtual community is not possible under the standard models. More recent work has expanded both the Kaplan-Meier estimator and the Cox's proportional hazards model to handle recurrent events. For the Kaplan-Meier, the Wang-Chang estimator is an extension that can handle multiple event variables per record (Wang & Chang, 1999). Similarly, there exists extensions of the standard Cox's model. For example, one extension allows the Cox's model to be used for modeling the effect of covariates on survival at multiple discrete points (Kalbfleisch and Prentice, 2002). Further, another extension allows the Cox's model to handle recurrent events per subject by manipulating the input covariate matrix (Ii et al., 1989; Van Den Berg, 2001). By combining these variations, an extended Cox's model can be used to handle multiple distinct records per subject, across multiple timespells.

Both the Kaplan-Meier and Cox's models are important tools for modeling virtual community participation data, as noted by previous work (Zhang et al., 2012). Each can present a unique perspective for understanding how participation occurs within cybercriminal communities. In particular, the Cox's model helps provide some explanatory power concerning what covariates (e.g., extracted features from IRC data) influence participation. Overall, duration modeling would provide great value when incorporated in researching investing cybercriminal community data.

3.3. Research Gaps and Questions

Based on my literature review, it appears that cybercriminal IRC communities can provide valuable information to security researchers and analysts concerning emerging security threats and cybercriminal behaviors. However, one of the largest problems with current research is that there is a lack of methods to quickly identify key cybercriminal community participants from more less interesting individuals that may only possess passing interests. Additionally, key individuals are a more credible source of data regarding emerging threats and cybercriminal trends. For this reason, I am motivated to develop a computational method for identifying key cybercriminals within IRC communities by scrutinizing networks between cybercriminals and the contents they share. Such method would be valuable for aiding researchers and practitioners in quickly identifying credible data sources within cybercriminal communities. I posit a series of questions related to my research goals. How can key cybercriminal identification be operationalized in a scalable fashion? Which cybercriminal network and content characteristics may be important indicators useful for differentiating between key cybercriminals and those with passing interest? How do I scrutinize cybercriminal IRC data through a temporal perspective?

3.4. Research Design

A summary of my research design can be viewed in Figure 3.6. Similar to previous cybercriminal community research, I utilize keyword searches to identify potential cybercriminal IRC channels. Two queries I used include "carding IRC" and "blackhat IRC." I scrutinize cybercriminal forums I discovered in prior work for links or mentions of existing cybercriminal IRC (Benjamin & Chen, 2012). After manually identifying a set of IRC networks, I join each network and issue an IRC "/list" command. The "/list" command returns a set of all active IRC channels within a particular network, including each channel's name and the number of active participants per channel. I use this feature to assess the activity level of a given IRC community, as well as to check which channels are the most populated.

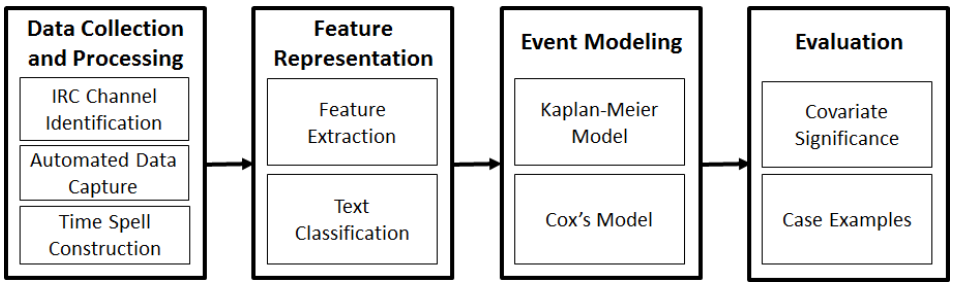

Figure 3.6 – Research design

When I identify an active cybercriminal IRC channels, I deploy automated IRC listening programs to identified IRC channels. The IRC listening programs are viewed as normal clients by the IRC network, but the programs are designed to passively log all IRC channel contents. Each IRC listening program connects to only one IRC network, but can connect to multiple channels simultaneously within each network. Since the nature of IRC necessitates for data to be collected in real-time, I deployed multiple listeners utilizing different IP addresses to the same IRC in order to ensure collection and to avoid gaps in data collection in case I experience dropped connections, bans, etc. Since each IRC listening program is viewed as a normal client by the network, they must

be assigned IRC nicknames; I change listener program nicknames upon every reconnection to help avoid removal from channel due to long-term idling and to reduce suspicions of collection behavior. This is consistent with IRC collection strategies outlined in prior work (Fallman et al., 2010). The listener programs passively sat within IRC channels and did not transmit any messages or attempt interaction with any other channel participants. This type of "lurker" behavior is common in IRC as many participants remain connected when idling, and thus IRC users have no expectations that all connected users will be active in channel discussions. IRC data collection and analysis operations can be executed easily on any modern computer. As IRC is completely plaintext, it has low bandwidth and processing requirements; hundreds of IRC channels could be collected simultaneously on one machine.

Further, Internet traffic generated by my chat listening programs is routed through the Tor peer-to-peer anonymity network and other proxy servers to hide researcher identity and university affiliation. Specifically, whenever an individual or listener program accesses cybercriminal IRC servers, the server will generate log files revealing IP addresses that are connected to the server. Thus, the origin IP address of researchers can be exposed, resulting in a significant security risk. Fortunately, proxy servers and anonymity networks can be utilized to re-route researcher web traffic through external connections, effectively concealing the identity of researcher machines from hacker forum servers. The Tor network is a peer-to-peer Internet traffic routing service with *specific* intentions to anonymize network packets generated by users. Internet packets that enter the Tor network are relayed to three or more volunteering peers that establish a circuit between the original sender and destination of the packet. Through this method, the IP address and details of the original sender are never relayed to the destination computer. Thus, I am able to utilize this service to anonymize the traffic I generate when connecting to cybercriminal communities. In

some cases where server belonging to the Tor network are blacklisted by a cybercriminal community, I route traffic through my own virtual private servers to conceal the origin of my network traffic. Making use of various cloud providers and data centers ensures I utilize IP addresses from unique subnets.

I observed captured data and selected my two largest IRC channels for this exploration, which can be viewed in Table 3.2. The two channels are largest in terms of both message volume as well as collection length. The #Anonops channel is part of the *Anonymous* hacking group's community infrastructure. The #Evilzone community is another prominent hacking IRC channel I identified that at times possesses hundreds of participants simultaneously.

Network	Channel	# of Users	# of Messages	Start Date	End Date
irc.anonops.org	#anonops	5,311	314,039	11/31/13	10/31/14
irc.evilzone.org	#evilzone	1,059	149,031	11/31/13	10/31/14

Table 3.2 – Collection summary

To develop a preliminary understanding of cybercriminal IRC contents, I performed some simple content analyses based on keyword frequencies. Prior research highlights many technical and hacking terms discussed in cybercriminal forums (Benjamin & Chen, 2012; 2014). For example, "trojan horse" and "keylogger" are two hacking terms of particular interest. I first computed the term frequency of each keyword used in prior works across my IRC collection. This provided us with a summary of how much each keyword was used by cybercriminal IRC participants, helping us identify the relevance of discussion towards hacking-related topics. This type of ranking is useful as it can provide a quick summary of overall conversation occurring within cybercriminal IRC communities. Additionally, I also tried to calculate the overall document frequency that each keyword appears in. This would provide us with information on how many distinct messages a keyword appears in, and not just overall frequency. However, IRC messages

in my dataset are typically short-length and rarely contain the same keyword more than once per message, thus resulting in term frequency and document frequency possessing similar values. Instead of document frequency, I found it to be more helpful to compute the number of different IRC participants that discuss each keyword. This provides us with some information on how widespread a keyword was discussed among all users. Overall, I found a substantial amount of content relevant to cybersecurity and emerging threats. I showcased highly-relevant examples in the results and discussion section of this paper.

After IRC chat data Ire logged and the relevance of collected contents Ire verified, data Ire pre-processed for further analysis with duration modeling. Specifically, I had to split my data into distinct timespells, and then extract data per individual participant for each timespell. I collected about 11 months of data collected for both IRC channels, and created two-week timespells for a total of 25 timespells. I did not find literature to help suggest a size for my timespells, and thus I experimented with a few different experiment configurations regarding timespell size. I chose two-week intervals as they seemed provide an appropriate level of granularity to assess changes in participation activity over time.

Within the context of this study, I only consider participants that appear within the first timespell. For example, if a new cybercriminal joins an IRC community within the second timespell, I do not include them in my analysis observing survival rates among cybercriminal IRC participants. However, I do still consider communications between participants featured in timespell 1 and those who joined after. I do this because I am interested in measuring network features of all participants featured in timespell 1, including those with new members. The IRC channels I observe possess no mechanism that exclude new members from interacting with more long-term participants, thus all communications are relevant for identifying potential key

members. Further, if an individual participant becomes inactive only to reappear in a later timespell, I do not once again consider them as part of the surviving population. This is because I am primarily interested in observing features correlated with constant long-term participation. It is possible a cybercriminal may switch their user name in the middle of the data collection period, but I do not control for this.

Next, features had to be extracted for each participant in each timespell, including the event variable, censor variable, network features, and text content features. I defined the event variable as participation; it is measured by checking whether a participant posted at least once during the length of an observed timespell. Since I modeled recurrent events, I could assign a new event variable per timespell for each individual. Further, as the recurrent event I Ire tracking was participation, I did not consider an explicit censor variable as traditional duration models may contain. Instead, I assumed right-censoring of data for all individuals in my analysis; that is, I did not assign a censor variable to subjects within the duration of the study, but rather I assumed the censor variable would apply to all subjects at the end of the last timespell. This form of censoring was considered normal practice in recurrent event duration analysis (Wang et al., 2012). I then extracted my network- and content-based features per individual per timespell. A summary of these time-variant features can be found in Table 3.3.

Category	Feature	Model Designation	Feature Justification	Source
Network Features	Total Direct Addresses	*DirectAddressOut*	Direct addressing is common in IRC channels and is an indicator of network ties. Individuals that commonly direct address others may feel interconnected and participate often.	Garas et al., 2012; Sinha & Rajasingh, 2014
	Total Times Addressed Directly	*DirectAddressIn*	Similarly, being addressed directly may increase feelings of interconnectedness within a network and lead to increased participation	Garas et al., 2012; Sinha & Rajasingh, 2014

	Total Different Individuals Directly Addressed	*UniqueAddressOut*	I also consider the total number of unique individuals addressed, helping measure the total social interconnectedness of each participant in my channel	Garas et al., 2012; Sinha & Rajasingh, 2014
	Total Times Directly Addressed by Different Individuals	*UniqueAddressIn*	The total number of times directly addressed by different individuals indicates the in-degree of social interconnectedness, which may lead to increased participation	Garas et al., 2012; Sinha & Rajasingh, 2014
	Days Participated	*DaysParticipated*	Number of days spent participating in community.	Radianti, 2010; Motoyama et al., 2011
Text Content Features	Total Message Volume	*MsgVolume*	Message volume is a commonly used indicator of participation rate, especially in the IRC context.	Motoyama et al., 2011; Benjamn & Chen, 2012
	Total Num. of Hacking Messages	*HackMsgs*	Demonstrates hacking proficiency, which may indicate increased participation; Examples: Rootkit, Cross-Scripting Attack, SQL Injection, Denial of Service, shellcode	Holt & Lampke, 2010; Benjamn & Chen, 2012
	Total Num. of Technical Messages	*TechMsgs*	Demonstrates technical proficiency, which may indicate increased participation; Examples: SQL, C++, ASM, .Net, XML	Holt & Lampke, 2010; Benjamn & Chen, 2012
	Total Num. of Black Market Messages	*MarketMsgs*	Demonstrates market activity, perhaps indicating to increased participation for black market purposes Examples: Bitcoin, E-Gold, IbMoney	Radianti et al., 2009; Holt & Kilger, 2012
	Hyperlinks Shared	*Hyperlinks*	Sharing of cybercriminal or technical resources, knowledge, or other information pertinent to community participants. May indicate to greater investment of time and participation.	Radianti, 2010; Benjamin & Chen, 2012
	Deep web Hidden Services Shared	*HiddenServices*	Sharing of deep web hidden services, pertinent to community participants. May indicate to greater investment of time and participation.	Martin, 2013

Table 3.3 – Extract network- and content-based features

To identify features for use as covariates for my Cox's model, I scrutinized previous cybercriminal and virtual community studies. However, as I am focusing on IRC data, features identified from past works must be considered for their suitability within the IRC context. Overall,

I identified a subset of both network- and content-based features that allow us to capture a complete picture of each participants' activities within the #Anonops and #Evilzone communities.

Concerning network features, past IRC research has utilized direct addressing between community participants in order to model network ties (Garas et al., 2012; Sinha & Rajasingh, 2014). Similar to past work, I considered both the total number of direct addresses, as well as the number of different distinct ties among participants to capture how each participant interacts with his or her peers. Due to the plaintext nature of IRC chat, direct addressing appears to be the only way to capture ties between participants with a high degree of certainty. Additionally, I captured the number of days each cybercriminal spent participating within their community, as seniority appears to be a notable characteristic of cybercriminals observed in prior studies (Radianti, 2010; Motoyama et al., 2011). I measure tenure by keeping an incrementing count of the total number of timespells a cybercriminal is active for.

I omitted certain network features that many previous studies make use of, such as centrality measures like betweeness centrality and closeness centrality. I did not consider centrality measures to be suitable in my context as they are often utilized to abstract and understand how information may flow throughout a given network, rather than to model the explicit interaction activity between network nodes (Freeman, 1979). This form of modeling information flow was not helpful in the IRC context, as each participant message was broadcasted to all other users regardless of intended recipients. Thus, centrality measures became less informative than observing direct addressing among IRC participants.

Concerning content features, the literature indicates that key cybercriminal community participants often contribute to the cognitive advancement of their community (Holt & Lampke, 2010; Benjamin & Chen, 2012). Such behavior includes sharing hacking tools, malicious source

code, tutorials, malware, stolen data, etc. Cybercriminals contribute knowledge through frequent discussion of relevant topics without explicitly sharing external resources such as aforementioned hacking assets. Overall, these assets can be used for education on general topics or for malicious attacks. Should cybercriminals gain insight into systems and applications in an organization, they can identify vulnerabilities and potentially exploit them with assets found in IRC. For example, in Figure 3.7, I showcase a stolen data asset found in the *#Anonops* community. The stolen data is a product of the hacktivist campaign *#OperationGreenRights*, and it contains the account names and passwords for multiple e-mail addresses belonging to various victim organizations. A cybercriminal IRC participant could make use of such assets to conduct their own crime, motivated either by financial or political reasons.

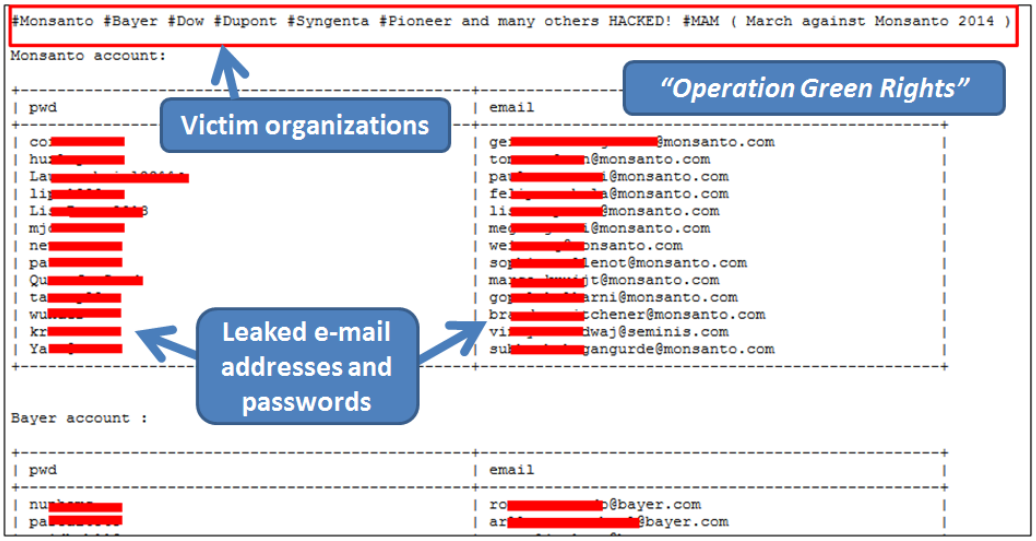

Figure 3.7 - *#OperationGreenRights* leaked e-mail and password list

Given the plaintext nature of IRC, it is not possible to embed images or attach files to messages. Instead, hyperlinks to external resources are the only way IRC participants can share resources with one another. Normal hyperlinks and hyperlinks to hidden services (for example, a '.onion' link to the Silk Road black market) may be shared to fulfill this need. Thus, I included counts of

both normal hyperlinks as well as hyperlinks to '.onion' hidden services as measures to assess how much different participants contribute to their community.

Beyond sharing assets through hyperlinks, cybercriminals may contribute to their community by participating in relevant discussions and passing along knowledge and relevant information to their peers. In particular, past research has identified some common topic categories that appear important and popular across multiple cybercriminal communities. First, much discussion revolves around sharing information hacking techniques (Holt & Lampke, 2010; Benjamin & Chen, 2012, Yip et al., 2013). This includes discussion of exploits, malware, cryptology, and other cybercriminal behaviors. Another one of the major topic categories identified in literature is more general discussion about technology (Holt & Lampke, 2010; Benjamin & Chen, 2012). Conversations of this theme include discussing programming languages, operating systems, network technologies, and more within a non-hacking context. The last major topic category referenced by past research involves electronic black markets (Radianti et al., 2009; Motoyama et al., 2011; Holt & Kilger, 2012). Overall, frequent participation in discussions of relevant themes may be indicative of more key participants. Regular contributions in these topic categories may lead individuals to interact with other interesting and key cybercriminals, which could potentially feedback into more active participation.

For these reasons, it may be useful to develop features that represent how often each cybercriminal participates in discussions of relevant topic categories. Specifically, I observe the amount of hacking, general technology, and black market related messages a participant contributes. As noted previously, machine learning classification is often useful for automatic topic categorization tasks, including categorization of virtual community data.

To perform text classification for categorizing messages based on their relevance to hacking, technical, or black market topics, I must first train the classifier as stated in previous research (Abbasi & Chen, 2008; Benjamin et al., 2014). Training requires manual coding of messages based on their topic for the classifier to learn from. I defined four classes for IRC messages: hacking messages, technical messages, black market messages, and general chatter. Later in my Cox's model, I do not measure the relationship between messages classified general chatter and participation activity, as I am more specifically concerned with understanding the influence messages that are underground in nature and contain hacking, technical, or black market relevance. Additionally, I already accounted for total message volume, which captures general chatter behavior.

In order to select messages for manual coding, 500 messages containing keywords of interest Ire chosen. I first built three separate dictionaries containing keywords related to hacking, technical, and black market discussion respectively. Keywords Ire extracted from cybersecurity literature and suggested by peers in the cybersecurity community. These dictionaries Ire used to extract messages that are of interest to scrutinize true relevance (i.e., whether a message containing "Apple" concerns fruit or computers). Two manual coders Ire involved and each coded all 500 messages, with inter-rater reliability of 98.2% (492/500 messages). The rest of the participant messages Ire then categorized using the trained classifier. I made use of a SVM classifier with a linear kernel, as this configuration has been frequently adopted in prior topic classification studies using virtual community data with success (Liu & Chen, 2013). I then used the trained classifier to categorize messages of each participant per time period to gather topic feature counts. These counts are useful for my duration model.

To conduct duration modeling, I make use of the Cox's proportional hazard model. The Cox's model allows us to measure the effect of my explanatory variables in a statistically-grounded test. To interpret a Cox model, I examined coefficients for each feature much like a regression. However, interpretation of Cox's model results is slightly unique; with the Cox's model, the presence of positive variable coefficients indicate that a particular variable contributes towards experiencing the modeled event in a shorter duration of time than normal.

After features Ire extracted per participant, I organized my data into a matrix for use with the Wang-Chang Kaplan-Meier and Cox's models. All modeling work performed in *R* using the *Survrec* package for Wang-Chang Kaplan-Meier estimate and the *Survival* package for extended Cox's model with recurrent events. I performed a two-step analysis using the two models. First, I executed the extended Wang-Chang Kaplan-Meier model to gain generalized perspective of the IRC channel population's survival curve. This model helps provide the "big picture" on event occurrence over time within my observed IRC channels. Next, I then utilized the extended Cox's proportional hazards model to test the explanatory power of the various extracted content and network features. The Cox model would help us identify behavior differences among participant behaviors that would result in different magnitudes of participation.

3.5. Results & Discussion

I first performed the Wang-Chang Kaplan-Meier model develop a high-level perspective of my dataset (Wang & Chang, 1999). Through the Wang-Chang estimator, I produced a matrix containing information on hazard and survival rates per timespell. An example of this matrix can be viewed in Table 3.4. When values from the resulting matrix Ire plotted, I Ire presented with a decreasing step-function that visualizes cybercriminal "survival", or participation over time for the IRC channels (Figure 3.8). For the #Anonops community, I observe that 60% of my participants

continue their participation after timespell 1, while only about 15% continue until timespell 25. One could deduce that the 15% of users that participate for almost a full year are the most engrained within the hacking community, and thus are more likely to become potential cyber adversaries than their peers that participated for much shorter lengths of time. For #Evilzone, a greater number of participants appear active up until the end of my recorded data (~20%). It may be that the #Anonops community is more popular since it has received media coverage, thus leading to more curious visitors with passing interests that do not remain participate in the community for very long. Some data supporting this conclusion is that the total average messages per participant is 140 messages in #Evilzone, but only 59 messages in #Anonops.

Time Period	Survival Rate	Std. Err	LoIr 95% CI	Upper 95% CI
0	0.5939	0.1279	0.5694	0.6195
1	0.4699	0.0130	0.4451	0.4961
…	…	…	…	…
22	0.0736	0.0084	0.582	0.932

Table 3.4 – Example output of Wang-Chang Kaplan-Meier survival curve matrix for the #Anonops community

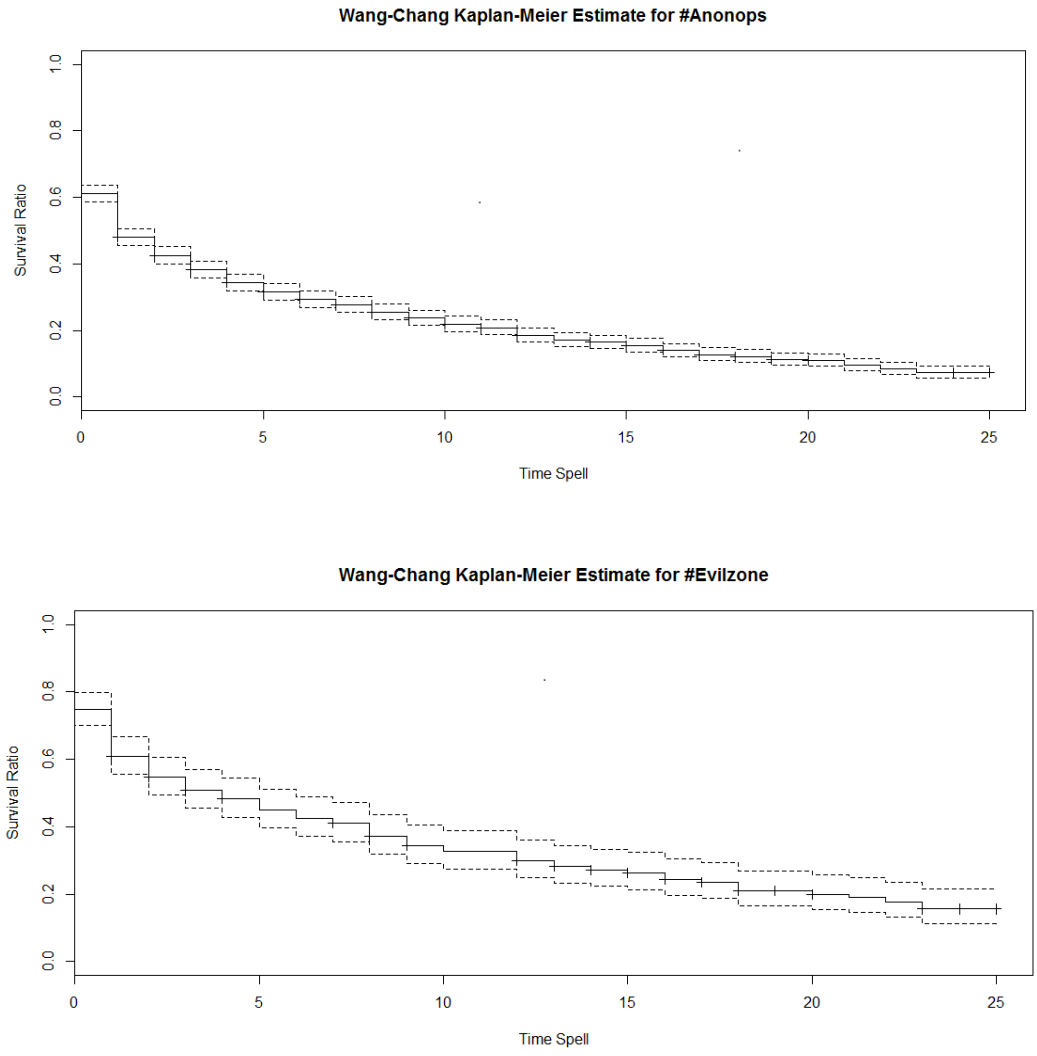

Figure 3.8 – Wang-Chang Kaplan-Meier Estimate for Both IRC Communities

After exploring my data with the Wang-Chang Kaplan-Meier estimate and producing a better understanding of the IRC channel's survival curve, I seek to further my understanding of which explanatory variables significantly affect survival over time. To do this, I made use of the Cox's proportional hazard model to measure the effect of my explanatory variables in a statistically-grounded test. To interpret a Cox model, I examined coefficients for each feature, much like a regression. However, the interpretation of Cox's model results is slightly unique; with the Cox's model, the presence of positive variable coefficients indicates that a particular variable contributes

towards experiencing the modeled event variable in a shorter average duration than normal. Conversely, a negative coefficient indicates a feature would promote a cybercriminal to participate within IRC communities for a longer amount of time than normal. The Cox's model would help us identify behavior differences among users that would result in different degrees of participation. My results using an extended Cox's model with recurrent events for both the #Anonops and #Evilzone community can be seen in Table 3.5.

Feature	**#Anonops**		**#Evilzone**	
	Coef	**P-Value**	**Coef**	**P-Value**
MsgVolume	0.0005	0.356	-0.00004	0.895
HackMsgs	0.0043	0.813	0.01913	0.232
TechMsgs	-0.0106	0.493	-0.00626	0.552
MarketMsgs	0.0049	0.147	0.04713	0.527
Hyperlinks	0.0018	0.419	-0.00002	0.998
HiddenServices	-0.0043	0.905	-0.07220	0.836
DirectAddressOut	-0.0003	0.815	0.00077	0.575
DirectAddressIn	0.00001	0.180	0.0005	0.131
UniqueAddressOut	-0.0164	0.0001 **	-0.02985	0.00342 **
UniqueAddressIn	-0.0005	0.022 *	-0.01092	0.0472 *
DaysParticipated	0.0045	0.818	-0.00645	0.561

n = 2,712, Signif. codes: '**' 0.01, '*' 0.05, #Anonops R^2 = 0.219, #Evilzone R^2 = 0.265

Table 3.5. – Results of extended Cox's model with recurrent events for #Anonops and #Evilzone

After observing the results of my model, it appears that only a small subset of covariates Ire helpful predictors of cybercriminal participation within IRC. For network features, both distinct in-degree and out-degree ties Ire significant. Participants who create many distinct ties are characterized by longer periods of active participation. It may be that such individuals are increasing their social interconnectedness by taking the time to engage numerous individuals in conversations, and thus increasing their length of stay in the cybercriminal IRC communities.

To my surprise, no content features Ire found to be significant. I have performed robustness checks to confirm my results, and have manually scrutinized data to better understand this phenomena. It is plausible that many individuals with fleeting interest may focus on discussing technical- or hacking-related topics upon joining a cybercriminal IRC channel, but then fail to become active for any significant length of time. Thus, simply talking about cybercrime, black markets, and related technical concepts is not necessarily exclusive to those who are long-term participants or key members of their community. Further, in the case of market transactions, there appears to be no 'honor among thieves'; that is, I see instances where some cybercriminal IRC participants may attempt to steal from others through fraudulent trades or by providing false information. These reasons could attribute as to why content features appear to be less important than network features in this context.

The ability to quickly identify participants is crucial to security researchers and practitioners tasked with detection of emerging cyber threats. This capability can lead to better evaluation of threat credibility. While the results of my analysis appear simple, they provide a powerful capability of one may quickly highlight cybercriminal IRC participants who may be interesting for researchers and practitioners to examine more closely.

For example, one of the most highly networked individuals in my #Anonops data is a participant that goes by the username "Strudalz." After manually scrutinizing messages broadcast by Studalz, I discovered they maintain a highly-followed Twitter account used to spread propaganda and hacktivist recruitment advertisements (Figure 3.9). The Twitter account has nearly 1,500 folloIrs and is used frequently to spread hacktivist propaganda and recruitment advertisements. My investigation revealed Strudalz appears to have been previously involved in DDoS attacks against China in support of Hong Kong protestors. They are notable #Anonops

community participant that was identified by observing participant networking features, highlighting the value of results.

Figure 3.9 - Twitter account of #Anonops participant 'Strudalz'

Network analyses have been conducted in many virtual community studies, but few have applied them to dark web communities. Overall, the techniques used and analysis results may be applicable to traditional IRC communities. However, there is value in applying these techniques to untraditional data sets, such as cybercriminal IRC. I have conducted an explanatory study and contribute to my greater understanding of cybercriminals, as well as to the science of cybersecurity.

3.6. Conclusion & Future Research

In this research, I attempt to explain differences in cybercriminal IRC participation activity based on participants' behaviors. By extracting user participation behaviors and incorporating them as covariates in duration modeling, I Ire able to measure the relationship between extracted covariates and participation. This provides us a method that enables deeper understanding of activity within cybercriminal IRC communities. Further, this can be utilized to quickly help

identify the most long-term participating or key users within a community, filtering out noise generated from more benign users that may only possess passing interest. This work also contributes to my overall understanding of cybercriminal IRC communities, and is of great value to security researchers and practitioners.

This research, and future related works, help contribute to my understanding of cybercriminal community participation behavior. With additional testing and model validation, cybersecurity researchers and practitioners could use the results of this research to better predict the participation of cybercriminal IRC users and to quickly identify the most long-term and key participants in any given IRC channel. This can help us better identify credible cyber threats and better prepare cyber defenses. Future work can expand in multiple directions to extend this work. Additional cybercriminal IRC channels, or even forums, can be analyzed to observe participation trends across multiple cybercriminal communities. Examining IRC channels acting as hidden services and comparing them to more public channels may yield interesting results. Lastly, incorporating cybercriminal communities from other geopolitical regions could help researchers draw new conclusions.

4. ESSAY III: DETECTING EMERGING THREATS IN CYBERCRIMINAL FORUMS USING VECTOR REPRESENTATIONS OF WORDS

4.1. Introduction

Cybersecurity is one of the largest issues impacting society, affecting individuals, businesses, and government alike. News of advanced cybercrime and major data theft has become a common occurrence. It is estimated that cybercrime costs the global economy about $445 billion a year, mostly due to theft of intellectual property within developed countries and sale of stolen personal information (Sandle & Char, 2014). Overall, cybersecurity will remain a problem of great relevance for the foreseeable future.

As a result, the need for more research on hackers is a common suggestion in recent years. Specifically, the development of methods to model cyber adversaries is one of the critical but unfulfilled research need outlined in a 2011 report on cybersecurity by the National Science and Technology Council (NSTC, 2011). More research on "black hat hackers", i.e. cybercriminals, would offer new knowledge on securing cyberspace against those with malicious intent, leading to the development of more effective countermeasures against security threats (Mahmood et al., 2010).

In particular, many online hacker communities exist that are of interest to cybersecurity researchers. Hackers congregate within such communities to share cybercriminal assets and knowledge, such as hacking tools, malware, hacking tutorials, and more (Benjamin & Chen, 2012). Some communities contain underground economies where participants may buy, sell, and trade for cybercriminal assets and services (Holt & Kilger, 2012). However, researchers and practitioners face many challenges when attempting to study hacker community contents, as hacker communities contain unique data and characteristics not encountered in more traditional

virtual community studies. For example, researchers may encounter hacking terms, concepts, tools, and other hacker-specific items that are unknown and present challenges in understanding hacker contents. Further, foreign language issues may also arise due to hacker communities existing globally, presenting another layer of challenge facing researchers attempting to understand hacker contents.

For these reasons, I am motivated to develop an automated method for understanding hacker language. Specifically, I seek to automatically identify relations between hacker-specific words, to track changes in hacker language over time, and to reveal potential emerging threats. To operationalize this, I make significant contributions to the state-of-the-art in neural network language models (NNLMs) by developing novel techniques for training NNLMs to represent temporal attributes of data, as well as to boost model training by incorporating information from existing knowledgebase. This technique would provide great value to security researchers and professionals wanting to better understand hacker contents.

This paper is organized into the following sections: First, I provide some background information concerning (1) past hacker community research, and (2) the state-of-the-art in NNLMs. The review of past work is followed by a discussion of research gaps and questions. Next, I describe my research design, with specific details on how I extend current work in NNLMs. I then run some experiments using my extended NNLM on hacker data, and follow with a discussion of my results. Lastly, I conclude by discussing the contributions of this paper. In sum, this work provides a new method for researchers to quickly identify new hacker concepts and emerging threats. My contributions to the most recent works in NNLMs is also applicable to other domains outside of the cybersecurity context.

4.2. Literature Review

To conduct this study and successfully develop an automated way for understanding hacker language, I must look to the latest research and borrow relevant perspectives. First, I review prior investigations of hacker forums in order to provide contextual understanding relevant to this study. Additionally, a review of hacker forum literature can help identify research gaps in need of attention. Next, I review recent works in lexical semantics, particularly in the areas of neural network language models (NNLMs). NNLMs are the state-of-the-art in unsupervised machine learning for developing understanding of language, and can be especially helpful for aiding researchers in understanding untraditional and unknown contents found within hacker forums.

4.2.1. Hacker Community Research

Hackers make extensive use of online communities to support cybercriminal activity. In particular, hackers use such communities to share cybercriminal assets and hacking knowledge with each other (Motoyama et al., 2011; Benjamin & Chen, 2014). It is not uncommon to witness hacking tools, malware samples, hacking tutorials, and more to be freely shared among community members. An example of such activity can be seen in Figure 4.1. In this example, the message author shares a video tutorial for configuring a popular botnet tool. The message also contains some text describing the video's contents. Such text can be used to build language models that help researchers better understand the role of different hacker terms. For example, here I can observe that "Zeus" refers to a botnet tool. Additionally, many hackers will share links to other communities, underground economies, and deep web hidden services (Matin, 2013). Such communities are not limited to a specific geopolitical region, and have been found to exist globally, including areas such as the United States, China, Russia, and the Middle-East (Benjamin & Chen, 2012; Holt et al., 2012).

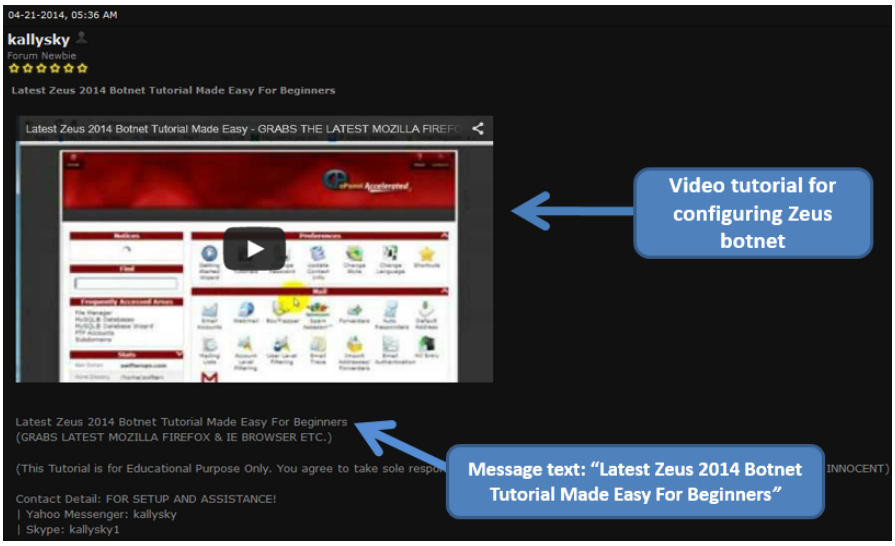

Figure 4.1 - Example of a posted message on a hacker forum

As a result, recent years have seen security researchers and practitioners develop increased interest in analyzing data from such communities. Past work provides useful methods for identifying and collecting hacker community contents. Additionally, literature provides context and insights for future hacker studies.

Some common methods to identify hacker communities exist throughout literature. Primarily, past studies resort to keyword searches for finding public hacking communities (Fallman et al., 2010). After an initial set of seed communities are identified, they can be scrutinized for hyperlinks and references to other hacker communities, resulting in a snowball collection procedure (Holt & Lampke, 2010). After identification, data can be collected through various means. Forum can be collected with web crawlers; However, anti-crawling measures are sometimes put in place by hacker forums to detect and halt crawling activity (Spencer, 2008; Fallman et al., 2010). Thus, it may be necessary to use proxy servers and identity obfuscation techniques to avoid detection of crawling activities (Benjamin et al., 2015). For example, adjusting crawling rates and alternating between IP addresses used for crawling hacker contents may help conceal researcher identity and prevent hacker communities from discovering crawling activity.

The majority of previous hacker community research can be categorized within a few major themes. First, much existing work utilizes qualitative analyses to observe and describe hacker community activities (Motoyama et al., 2011; Holt et al., 2012). The second branch of work generally involves counting procedures and high-level statistical analyses of underground economy and carding community contents (Martin, 2013; Yip et al., 2013). Lastly, many recent works have focused effort on identifying key participants within hacker communities (Benjamin & Chen, 2012; Yip et al., 2013). These three categories of prior work are useful for describing ongoing activity within hacker communities, as they reveal commonly discussed topics, provide better understanding of hacker social dynamics, and help develop techniques to quickly identify key hacker community participants.

However, one underdeveloped research area is the construction of language models to better interpret hacker contents. Advancements in this area could help boost capabilities for identifying the meaning of hacker-specific terms. Additionally, an understanding of hacker language could help reveal role and functionality of existing and emerging hacker tools, malware, and threats. Lastly, better understanding of hacker language could be used to guide feature generation for future research.

Fortunately, methodology from computational linguistics is useful in text analysis applications. In particular, many prior virtual community studies utilize natural language processing for analyzing web contents. Specifically, methodology from the lexical semantics domain is useful for developing understanding of words and phrases. Such techniques may prove useful for analyzing hacker language.

4.2.2. Lexical Semantics

Lexical semantics is a subfield of linguistics that focuses on: (1) the study of lexical units such as words, affixes, and phrases, (2) lexical relations, or how different lexical units relate to each other, and (3) how lexical units map into different concepts. Relationships between different lexical units (or words in this case) can be mapped. Higher abstractions of meaning can be inferred by scrutinizing lexical relationships; e.g., if Chicago is related to the words city and Illinois, and Illinois is related to the words state and Chicago, then Chicago can be understood as a city within the state of Illinois.

Literature on lexical semantics is far too broad to be discussed in full here, and thus I focus only on the most recent, relevant stream of work. This includes work on scalable, automated techniques that are suitable for large-scale virtual community research. Further, I limit my review to research utilizing unsupervised learning, as identifying informative and useful features in untraditional datasets (e.g., hacker communities) in order to guide supervised learning is a difficult challenge. Supervised learning is also often times language-specific, presenting problems for generalizing hacker language modeling to the global scale.

In particular, neural network language models (NNLMs) have captured much attention in recent years (Levy & Goldberg, 2014; Mikolov et al, 2013a; Mnih & Kavukcuoglu, 2013; Pennington et al., 2014). NNLMs have gained vast popularity due to recent advancements in computing continuous vector representations of words, resulting in high performance and low computational cost relative to previous techniques. Additionally, neural networks possess node layers that pass their output to previous layers; this is unlike more traditional neural networks, where processed data is strictly fed forward and cannot be forwarded to previous layers. The extended capability allows neural networks to develop an internal state, causing the model's

training process to become dependent on previous model history during training. Such behavior is advantageous for learning tasks such as language modeling.

Recent works focus on using NNLMs to build word embeddings through unsupervised learning of word meaning by scrutinizing the local context that each word is used within. At a conceptual level, word embeddings simply amount to vectors that contain values representing the local contexts a given word is found within. These vectors, or embeddings, can be used for further computational analyses to extract meaning from unstructured text.

Word embeddings have been researched heavily in recent literature (Jansen et al., 2014; Mikolov et al., 2013a). One major research application involving word embeddings is to use them for computing the similarity/distance between any two words that are part of the same vocabulary. This process can reveal the conceptual similarity or distance between two words, based on how those words are used within natural language. Since embeddings are vectors, a similarity metric such as cosine similarity is applicable for such cases. A second major application for word embeddings for learning analogy tasks such as "hat is to head as shoe is to ___ (foot)." Word embeddings are successful at this task as they are able to be used to infer abstracted meaning between relationships of different words. Overall, embeddings can be useful for developing conceptual understanding of unfamiliar terms, which is useful for advancing my understanding of hacker language.

NNLMs that utilize Skip-gram learning with Negative Sampling (SGNS) have generated much excitement in the computational linguistics community due to their ability to generate state-of-the-art word embeddings (Mikolov et al., 2013b; Pennington et al., 2014). Specifically, SGNS NNLMs have been benchmarked across multiple studies and shown to be a leading performer for generating word embeddings (Jansen et al., 2014; Levy & Goldberg, 2014). The SGNS NNLM is composed

of two separate parts. The first part is skip-gram learning, which is a model training algorithm that is useful for building word embeddings. When given a particular word, a Skip-gram trained model can predict surrounding context words, effectively resulting in multi-class classification of words. For example, given the center word $w_{0,}$ the skip-gram algorithm will try to predict the surrounding context words w_{-2}, w_{-1}, w_1, and $w_{2.}$

More formally, skip-gram predicts surrounding words in a window of length c for every word. The objective function is to maximize the log probability of any context word given the current center word:

$$J(\theta) = \frac{1}{T}\sum_{t=1}^{T} \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{t+j} \mid w_t)$$

Where T is the total number of words in the corpus, t holds the position of the current center word, c is the defined context window (e.g., a window size of 2 includes the preceding and proceeding two words surrounding the current center word), j denotes the specific position of a context word within c, relative to t, and w_t is the actual word at the center position t.

However, the Skip-gram objective function is not scalable as it would train slowly on large corpus due to its design of iterating through each word individually. To increase scalability, data could be instead sampled from a large corpus in a statistically sound manner. This is precisely what the negative sampling portion of the SGNS NNLM enables (Mikolov et al., 2013b; Benjamin & Chen, 2015). Negative sampling approximates a probability distribution of words within a given corpus. This probability distribution is used to sample data and assist Skip-gram learning during word embedding construction. Overall, negative sampling is an important part of significantly reducing the time complexity of model training, and thus enabling scalable analysis across large data sets.

Overall, NNLMs are useful for developing understanding of relationships between lexical units, as they can measure of how words are conceptually related to one another. In the hacker context, NNLMs can be used to help researchers better understand hacker language by identifying relationships between known and unknown hacker terms. Further, NNLMs can be used to potentially reveal emerging threats such as new hacker tools, malware, and more.

4.3. Research Gaps and Questions

Unfortunately, current work on NNLMs is still at its infant stages and suffers from many limitations. First, the latest work on NNLMs is currently only applicable to a static corpus with unchanging vocabulary. Thus, the state-of-the-art methods cannot be used to observe how language changes over time, which is a necessary capability for detecting emerging threats and hacker trends. Further, current SGNS NNLMs do not possess the capability to utilize existing knowledge for boosting training; for example, using known hacker words to boost identification of unknown terms. Since known hacker terms are directly relevant to my understanding of hacker language, it makes sense that I want to pay extra attention to known terms during model training. In this way, I can make use of existing knowledge to find new, previously unknown hacker-specific language.

I am thus motivated to extend the latest work in NNLM.s Specifically, I focus on the SGNS NNLM and aim to introduce new capabilities that are useful for developing understanding of hacker language and detecting emerging cyber threats. I am guided by a series of questions in pursuit of this research. H how can I develop the capability to automatically digest hacker community contents and develop understanding of hacker-specific language? In what ways can I model changes in hacker language over time? Can tracking such changes help us identify emerging

cyber threats? How may I I use pre-existing knowledge about hackers to boost the performance of automated techniques?

4.4. Research Testbed and Design

My research design (Figure 4.2) consists of a series of steps involving automated data processing and analysis. First, I identify and collect hacker forums for this study. Next, I scrub and process collected data into a form ready for analysis. I then construct my NNLMs and execute experiments. Finally, experiment results are evaluated and conclusions are drawn based on my findings. I provide details for each component separately.

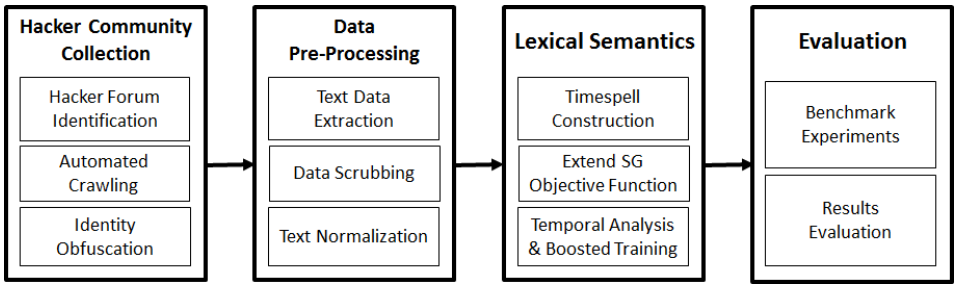

Figure 4.2 – Research Design

4.4.1.Data Collection

I identified a subset of public hacker forums by utilizing keyword searches such as “black hat forum” and “carder community” to identify hacker forums. From this initial set, I scrutinize hyperlinks that participants share for potential linkage to other hacker communities. I identify and collect three hacker forums for this study (Table 4.1). Forums Ire chosen based on several factors. First, all three forums are English-speaking hacker communities. While the technique I use is language independent, I chose to test my models on English forums as I can more easily interpret results than other languages. Additionally, the forums contain consistent forum activity over time with recent activity from multiple forum participants. Lastly, I observe abundant discussion of

hacking concepts, tools, and black market activity that would be interesting to study in this research.

Forum	Members	Threads	Posts	Time Span
Cardersforum	3,359	2,111	7,740	12/30/2012 - 12/30/2014
HackFive	947	1,108	5,334	1/24/2013 – 12/30/2014
HackHound	633	507	3,621	12/10/2012 – 12/30/2014

Table 4.1 – Research Testbed

Automated crawlers deployed to collect identified forums. Several steps Ire taken to customize crawlers in order to ensure successful collection of hacker data while protecting researcher security (Benjamin et al., 2015). For example, I route Internet traffic generated by my crawlers through the Tor anonymity network and personal proxy servers in order to protect researcher identity and university affiliation. The Tor network is a peer-to-peer Internet traffic routing service that effectively anonymizes Internet communications. Further example of crawler customization includes altering crawling rates in order to avoid triggering server-side anti-crawling mechanisms.

After successful collection, data of interest must be extracted from raw HTML pages that are downloaded by the web crawlers. Extracted data must then be further processed in order to be readied for analysis. Overall, this includes several steps.

4.4.2. Data Pre-processing

After collection, I extract message text from collected hacker forum web pages. Regular expressions are written to extract data embedded within HTML, specifically thread titles and message bodies. I then scrub data by removing duplicate messages and removing instances where a forum participant quotes the message of another participate, which would result in duplicate text. I then normalize extracted messages in preparation for analysis. First, I convert all text to loIrcase in order to avoid having different cases of the same word become treated as two separate words in my model. Second I strip punctuation from words to again avoid duplication of words.

4.4.3. Lexical Semantics

As mentioned previously, the SGNS NNLM suffers from a lack of capability for modeling temporal attributes of data. Instead, the model assumes a corpus to be static with an unchanging vocabulary, and is thus not able to aid us in tracking hacker language evolution and emerging threats. I seek to create this capability by extending the Skip-gram objective function and by designing a novel model training scheme that can model language over multiple timespells.

The use of timespells enables assessment of language evolution as it enables for the comparison of models trained across different timespells. Researchers can scrutinize the relationship between hacker terms in adjacent timespells in order to identify hacker trends such as whether a particular malware is rising or falling in popularity. To operationalize this capability within the SGNS NNLM, a two-step procedure is necessary: (1) the corpus must be split evenly across multiple timespells, and (2) a separate model utilizing an extended SG objective function will be trained for each timespell.

I first split my dataset into timespells of equal length, while keeping in mind that the SGNS NNLM is designed to perform best on larger datasets (Mikolov et al., 2013a; 2013b). Thus, there is a need to choose a timespell length that balances between model performance while also providing granularity for assessing language evolution. I evaluate model performance with different timespell lengths and find 3 months to be sufficient. With two years of data being split into 3 month timespells, I generate 8 total non-overlapping timespells.

Next, an extended objective function for SG is needed to handle models trained in different timespells. Recall that the Skip-gram learning algorithm is not scalable, and relies on negative sampling to approximate a probability distribution of corpus words, and drives sampling of data for training purposes based on this distribution. If separate models are trained across timespells,

embeddings learned in adjacent timespells are at risk of diverging due to sampling differences rather than actual drift in language. Consider the example in Table 4.2. Two different embeddings for the botname named *Zeus* can be generated from the same original sentence. These differences are caused by the negative sampling process, and must be corrected with an extended objective function for use during embedding construction.

Original Sentence	**Embedding 1**	**Embedding 2**
"The popular botnet software *Zeus* presents a challenge for security experts"	"the botnet *Zeus* challenge security"	"popular software *Zeus* presents experts"

Table 4.2 – Example of Embedding Divergence caused by Negative Sampling

As models are trained across all timespells jointly, I can make use of a modified Skip-gram objective function that possesses an additional term to minimize. The term is an attempt to minimize distance between models in adjacent timespells, and can be defined as:

$$V = \sum (dist(w_j^{t_i}, w_j^{t_{i+1}}) + dist(w_j^{t_i}, w_j^{t_{i-1}}))$$

Where $w_j^{t_i}$ is the embedding of word w_j in time spell t_i, $dist$ is a function that computes the distance between two word embeddings from adjacent time spells. As stated previously, since word embeddings are vectors, I can make use of cosine similarity to determine the distance between two vectors. Overall, the extended objective function can be read as:

$$J(\theta) = \frac{1}{T}\sum_{t=1}^{T}\left(\left(\sum_{-c\le j\le c, j\neq 0}\log p(w_{t+j} \mid w_t)\right) * \frac{1}{V}\right)$$

The objective function term has two main effects: (1) models in adjacent timespells will be encouraged to sample the same words during word embedding construction, and (2) the extended objective function minimizes divergence caused by sampling different words. Overall, it allows for more meaningful comparison between different models in adjacent timespells.

Beyond minimizing divergence, I am also interested in utilizing existing knowledgebase to boost training. In particular, recall that because of negative sampling of data, known hacker words relevant to understanding hacker language may be missed and not used for training purposes. However, such known words are important as they directly contain information relevant to understanding hacker language. Such words can include tool names, malware, hacking techniques, and other hacker terms which should always be sampled in order to better identify other hacker terms and language evolution, thus increasing relevancy of results. Table 4.3 contains some information containing four unique categories of known hacker terms. My existing knowledge base contains terms relevant to cybercriminal attack techniques, black markets, malware features, and malware names (Benjamin & Chen, 2012; Holt & Kilger, 2012). I maintain a list of 25 words per hacker language category, totaling 100 words for this analysis.

Hacker Language Category	Example Words	Number of Words
Attack Techniques	SQL Injection, XSS, Drive-by, DDoS	25
Cybercriminal Black Markets	Agora Market, Blackcoin, Darkcoin, Silk Road	25
Malware Features	Crypter, FUD, Injection, Reverse Connection	25
Malware Names	Bifrost, Citadel, Spyeye, Zeus	25

Table 4.3 – Hacker Language Term Categories

To remedy this problem, I force sampling of such words when they are encountered during the SG learning task. Normally, the probability that a word is sampled is a function of its distance from the center word w_0. However, in my case, I can simply ensure sampling of words that appear in existing knowledgebase concerning hacker language.

4.4.4. Evaluation

Evaluation of unsupervised learning is a common challenge in research. Evaluation usually occurs by comparing performance of new algorithms and techniques against clearly established benchmarks (Jansen et al., 2014). Further, evaluation typically will occur on traditional data sets that are widely available to the greater research community. However, I face several difficulties

that complicate evaluation for my extended SGNS NNLM: (1) the standard SGNS NNLM is new, few established benchmarks exist, (2) SGNS NNLM performance varies depending on application context (Mikolov et al., 2013a; 2013b), (3) I use an untraditional data set as I have specific interests in using my work to detect emerging cyber threats, (4) this appears to be the first study extending the SGNS NNLM to handle temporal aspects of data, and thus no direct benchmarks exist, and (5) there is a lack of studies observing the effect of boosted training on model performance. In attempt to address the aforementioned evaluation challenges, I devised an evaluation plan containing three separate evaluation objectives to measure the performance of different capabilities my extended SGNS NNLM provides (Figure 4.3).

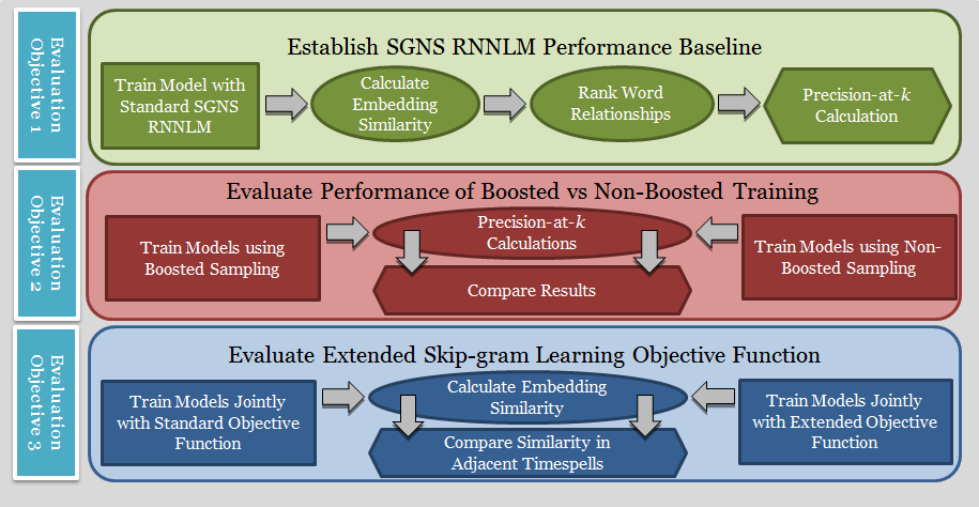

Figure 4.3 – Evaluation Plan

The first evaluation objective is to measure the performance of the standard SGNS NNLM on hacker data in order to establish a baseline for which to compare my extended SGNS NNLM. Additionally, I am assessing the ability for the standard SGNS NNLM to construct word embeddings that are useful for identifying hacker terms. I can treat this task as a standard information retrieval problem by using precision-at-*k* evaluation, a common technique for benchmarking unsupervised learning algorithms (Agichtein et al., 2006). Since word embeddings

are word vectors, I can rank the relationships between term pairs in my corpus based on embedding similarity (i.e., cosine similarity). After ranking term pairs, I can score the relevancy of top-ranked *k* embeddings for known hacker terms. For example, for the embedding of the term *botnet*, I could score the top *k* most similar embeddings for relevancy towards the hacker term *botnet.*

The second evaluation objectives entails comparing performance of boosted vs non-boosted training. Boosted training will include forced sampling of hacker terms provided through a manually compiled list. Non-boosted training will consist of the default training scheme (Mikolov et al., 2013b). To operationalize this I can simply compare the precision-at-*k* scores for boosted and non-boosted training within the same timespell.

In the third evaluation, I must evaluate divergence of models with and without using my extended Skip-gram objective function. I can comparing performance by running two simultaneous experiments and compare model precision. Specifically, I train two sets of models for each timespell, with one set utilizing the standard objective function and the other set using my extension. I can then compute similarity of word embeddings in adjacent timespells for both sets. Embeddings produced with the extended SG objective function should be more similar to adjacent timespells than those produced with the standard objective function.

4.5. Results and Discussion

For my first experiment (i.e., Evaluation Objective 1), I calculate an average precision-at-*k* of the standard SGNS NNLM. To do this, I select 100 known hacker and black market terms (from Table 3) and use them for testing. By using known hacker terms to evaluate precision-at-*k*, I am demonstrating performance within real-world context. Two examples can be seen in Table 4.4. For the terms *botnet* and *carder*, I rank the top 10 most similar embeddings, and then score rankings for precision. I score precision by evaluating whether a term is related or relevant to the

input test term. For example, for the term *botnet*, the top two results are *Citadel* and *Zeus*, which are both botnet tools.

	Input Term: *Botnet*		**Input Term: *Carder***	
	Word	*Similarity Score*	*Word*	*Similarity Score*
1	**Citadel**	**0.561456**	**Ccv**	**0.611420**
2	**Zeus**	**0.554653**	**Dumps**	**0.603473**
3	Partners	0.548900	**Fulls**	**0.691825**
4	**Pandemiya**	**0.545221**	**Paypal**	**0.583072**
5	**Mailer**	**0.540075**	Email	**0.564231**
6	Panel	0.524557	**Logins**	**0.55939**
7	Linksys	0.498224	**Bins**	**0.557148**
8	**Cythosia**	**0.480465**	**Amex**	**0.547302**
9	**Phase**	**0.464738**	Rules	**0.520016**
10	**Spyeye**	**0.459695**	**Accounts**	**0.505419**
P@10	70%		80%	

Table 4.4 – Evaluation Objective 1 Results for *Botnet* and *Carder*

I conduct a similar procedure for the remaining 98 hacker terms used for testing the standard SGNS NNLM. Some example results can be seen in Table 4.5. Overall, I average a 64% P*recision-at-10* using the standard SGNS NNLM. The standard SGNS NNLM does not include boosted training nor my extended Skip-gram objective function.

Test Word	**P@10**
RAT	80%
Logins	40%
Keylogger	60%
Crypter	70%
Rootkit	70%
Salt	60%
Binder	60%
Dork	70%
Vulnerability	70%

Table 4.5 – Example Results of Evaluation Objective 1

For my second experiment (i.e., Evaluation Objective 2), I evaluate performance of boosted vs non-boosted training. I repeat my first evaluation using boosted training, and compare results with the non-boosted model. Results are listed in Table 4.6. Overall, I observe a ~4% increase in performance by incorporating existing knowledgebase during sampling. My reasoning for the performance increase is that known hacker terms contribute to my understanding of overall hacker language; by ensuring sampling of these terms, constructed word embeddings will generally contain information more relevant to understanding hacker language and identifying previously unknown hacker words, emerging threats, and more.

	Average P@10
Standard SGNS NNLM	64%
SGNS NNLM with Boosted Training	68%

Table 4.6 – Boosted vs Non-boosted Training

I observe different levels of performance across the hacker language categories (i.e., attack techniques, black markets, malware features, and malware names). Additionally, boosted training provides different levels of improved performance per category. Table 4.7 contains a summary of results concerning the impact of boosted training on P@10 scores across all four hacker language categories. Overall, I see boosted training improve the performance of each hacker language category. However, the categories including malware names and features improved the most. I speculate that, because these two categories include many unique nouns and pronouns not found in normal English, they can benefit the most from existing knowledgebase. It would be interesting to examine this phenomena in other domains and contexts outside of security, and through perspectives supported by linguistic theory (e.g., systemic functional language theory).

Hacker Language Category	**Standard SGNS NNLM Average P@10**	**SGNS NNLM with Boosted Training Average P@10**	**Percent Change**
Attack Techniques	66.0%	68.8%	4.24%

Cybercriminal Markets	67.2%	69.6%	3.57%
Malware Features	62.4%	67.2%	7.69%
Malware Names	60.4%	66.4%	9.93%

Table 4.7 – Boosted vs Non-boosted Training per Hacker Language Category

To give better context of my results, I observe the average word embedding precision-curve for each NNLM (Figure 4.4). Recall that my knowledgebase of hacker language consists of 100 words split between four language categories, with 25 words per category. Thus, when evaluating the embedding of any single word, there are 24 other words in the same language category. Ideally, a word's embedding should be highly similar to other words in its category. For this reason, I evaluate precision curves at $k = 24$ in order to remain consistent with language category size.

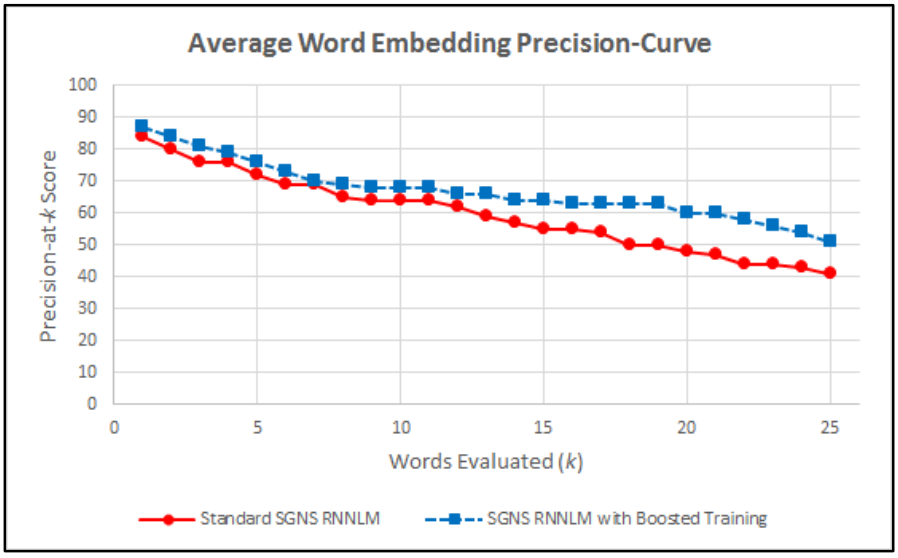

Figure 4.4 – Average Word Embedding Precision-at-24 Curve

For my last experiment (i.e., Evaluation Objective 3), I evaluate divergence of models when trained over multiple timespells. To operationalize this evaluation, I compare model divergence between models in adjacent timespells, with & without using my extended objective function. I compute divergence across all timespells for all 100 over my test words mentioned previously. An

example of this evaluation for the term *botnet* can be viewed in Table 4.8. The similarity of embeddings generated with the extended Skip-gram objective function is greater than the similarity of embeddings generated using the standard SGNS NNLM.

	Cosine Similarity of Adjacent Timespells							
	Timespells 1 & 2	Timespells 2 & 3	Timespells 3 & 4	Timespells 4 & 5	Timespells 5 & 6	Timespells 6 & 7	Timespells 7 & 8	**Average**
Standard SG Objective Function	0.5767	0.5077	0.5953	0.5174	0.4829	0.5095	0.5897	0.5399
Extended SG Objective Function	0.5813	0.6440	0.6698	0.5906	0.5702	0.6692	0.6582	0.6262

Table 4.8 – Evaluation Objective 3 for Hacker Term *Botnet*

Table 4.9 contains results for average similarity between embeddings in adjacent timespells. Overall, I see a ~7.7% increase in word embedding similarity when using the extended SG objective function. The extended objective function seems to successfully reduce model divergence caused by differences in data sampling. This leaves us with a more accurate representation of language evolution across timespells.

	Average Similarity
Standard SG Objective Function	0.5495
Extended SG Objective Function	0.5917

Table 4.9 – Evaluation Objective 3 for Hacker Term *Botnet*

Similar to evaluation objective 1, I examine performance across each hacker language category in order to develop deeper insights (Table 4.10). The fluidity of embeddings in adjacent timespells appears to differentiate between hacker language categories. Further, I again see different levels of improvement for each hacker language category. Malware features and malware names again seem to improve the most, indicating differences in linguistic evolution across language categories. It may be that language describing attack techniques and cybercriminal markets remains stable over time, while malware names and the features those malware contain are more frequently

changing. Overall, I observe interesting results worth scrutinizing further through theoretical perspectives and within different domains.

Hacker Language Category	Standard SG Objective Function	Extended SG Objective Function	Percent Change
Attack Techniques	0.5771	0.6017	4.263%
Cybercriminal Markets	0.6239	0.6329	2.452%
Malware Features	0.5134	0.5681	10.654%
Malware Names	0.4836	0.5579	15.363%

Table 4.10 – Evaluation Objective 3 for Hacker Term *Botnet*

To demonstrate the value of this research, I provide a sample use case. Many security practitioners and researchers are interested in studying remote administration tools (RATs). RATs are malicious programs that give hackers backdoor access and control over infected computers. Many RATs are shared on hacker forums and see widespread usage. I can use my research to determine what RATs are emerging or losing popularity among hackers. To demonstrate this, I compare the similarity of the embedding for *RAT* with two well-known hacking tools (Figure 4.5). *Bifrost* and *Spygate* are both RATs that grant hackers backdoor access to victim computers. The similarity between the term pair *Spygate* and *RAT* outgrows that of term pair *Bifrost* and *RAT* from 12/30/2012 - 12/30/2014 (3 month timespells). After manually scrutinizing my data, it appears *Spygate* is first introduced to the hacker community in Spring 2013, which corresponds to Timespell 2. Over time, it has grown in popularity and become more strongly associated with the term *RAT*. I infer this to mean that *Spygate* is becoming a more popular threat than *Bifrost.* This form of analysis is useful for researchers and practitioners to better understand growing threats.

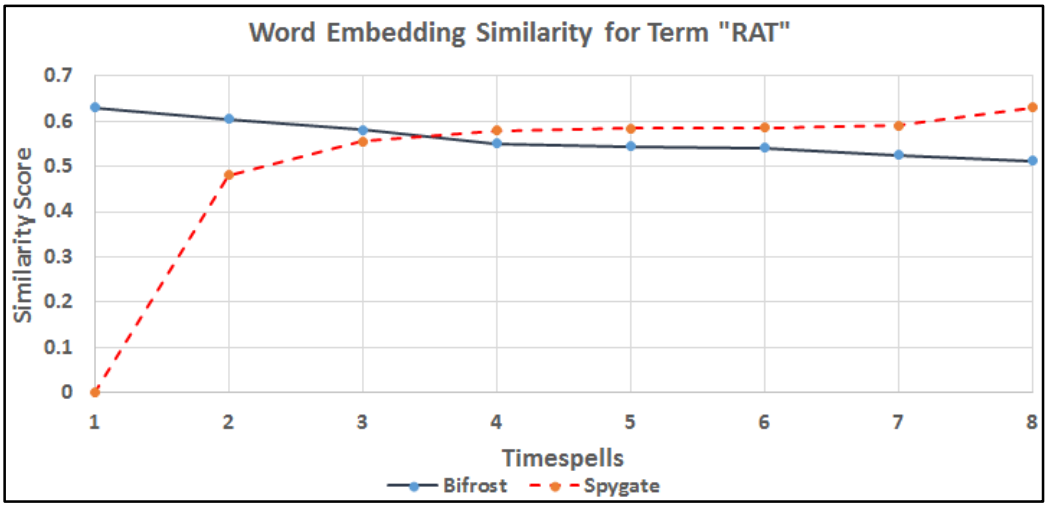

Figure 4.5 – Similarity of *Bifrost* and *Spygate* with the Term *RAT* Over Time

A similar example over the same timeframe can be seen in Figure 6. *Pony Stealer* and *Coin Stealer* are both 'stealer' malware, or programs used to identify and steal critical data from infected machines (Figure 4.6). *Pony Stealer* is a traditional 'stealer'-type malware that focuses on theft of usernames, passwords, and credit card data. Conversely, *Coin Stealer* is a new-generation type malware that attempts to locate and steal BitCoins from infected machines, rather than more traditional data. *Coin Stealer* is particularly interesting as it appears to be one of the first BitCoin-focused malware, and first appears towards the end of timespell 4 (i.e., fourth quarter 2013) and quickly gains popularity.

Through these examples, I demonstrate how my research allows for early detection of new and evolving threats. I can also learn more about general trends and changes with hacker language, and potentially unveil targets or victims of cybercrime through such linguistic analyses. Overall, developing understanding of hacker language is of great asset to security professionals and researchers wanting to better understand hacker communities and cyber threats.

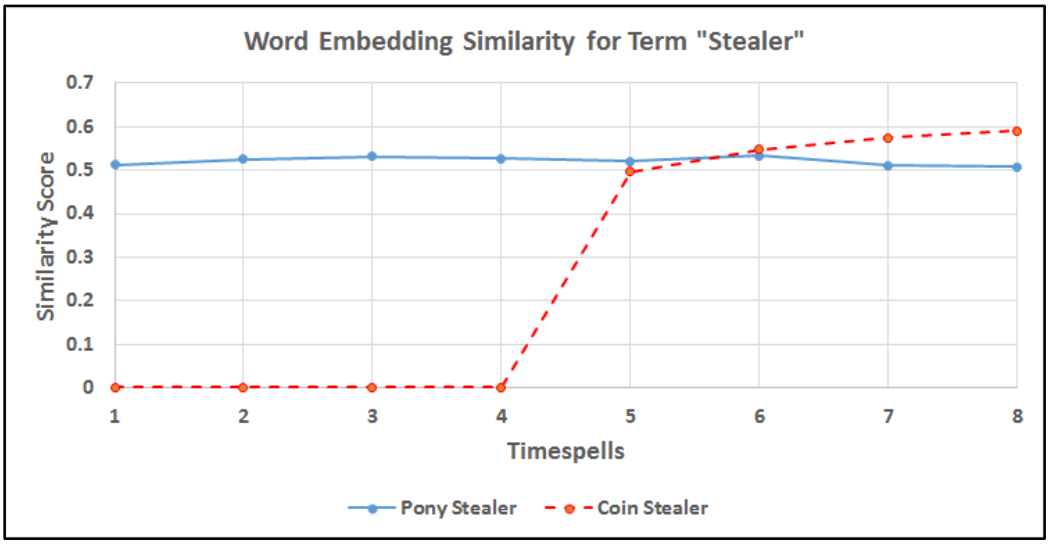

Figure 4.6 – Similarity of *Pony Stealer* and *Coin Stealer* with the Term *Stealer* Over Time

4.6. Conclusion

I use the SGNS NNLM to develop understanding of hacker language. The SGNS NNLM is a state-of-the-art technique in computational linguistics for generating word embeddings. It is an unsupervised, scalable learning technique that can be used to identify hacker terms, concepts, and relationships between them. I extend the SG objective function to include capability for handling temporal aspects of data, and also to boost training by enforcing sampling of known hacker terms. I also implement a new model training method to train multiple models jointly across timespells. The extended objective function helps ensures models do not diverge too much due to random sampling.

This research has numerous contributions. First, I contribute to methodology by expanding current state-of-the-art NNLMs to handle temporal data. Second, I contribute to cybersecurity by developing a new method to automatically identify potentially unknown hacker terms, and to develop understanding of hacker language trends over time. Lastly, my technique for boosted training can be applied to utilize terms from other domains beyond the hacker context. Future

research can seek to generalize this work on other application contexts, or seek to make additional advancements to the SGNS NNLM.

5. ESSAY IV: TRACKING INFORMATION DISSEMINATION BETWEEN MULTILINGUAL CYBERCRIMINAL FORUM POPULATIONS: THE AZSCOUT RESEARCH FRAMEWORK

5.1. Introduction

Cybersecurity is one of the most pressing issues facing society as it affects individuals, businesses, and government alike. News of advanced cybercrime and major data theft has become a common occurrence. It is estimated that cybercrime costs the global economy about $445 billion a year, mostly due to theft of intellectual property within developed countries and sale of stolen personal information (Sandle & Char, 2014). Overall, cybersecurity will remain a problem of great relevance for the foreseeable future.

As a result, the need for more research on cybercriminals is a common suggestion in recent years. Specifically, the development of methods to model cyber adversaries is one of the critical but unfulfilled research needs outlined in a 2011 report on cybersecurity by the National Science and Technology Council (NSTC, 2011). More research on "black hat cybercriminals" would offer new knowledge on securing cyberspace against those with malicious intent, leading to the development of more effective countermeasures against security threats (Mahmood et al., 2010).

In particular, many online cybercriminal forums exist that are of interest to cybersecurity researchers. Cybercriminals congregate within such forums to share cybercriminal assets, such as hacking tools, malware, hacking tutorials, and more (Holt & Kilger, 2012). Many forums support underground economies, where participants may buy, sell, and trade such assets. Cybercriminal forums are an international phenomena, with many communities possessing origins in America, China, Russia, and other geopolitical regions (Motoyama et al., 2011; Benjamin & Chen 2012).

In some instances, multilingual forums exist that are host to cybercriminals from various geopolitical regions (Holt & Kilger, 2012; Benjamin & Chen, 2015). Many different languages are used within such forums, however, forum participants may often participate within just one language population. Figure 5.1 has an example of a multilingual forum, *Crdclub.su.* This forum is host to English- and Russian-speaking cybercriminals, and possesses two distinct subforums dedicated to discussions within these languages. Individual users will typically limit themselves to subforums that contain discussions in their native language, and only interact to participants of different language populations in limited instances. However, instances where such cross-population activity does occur may lead to information and asset dissemination between different cybercriminal populations. An example of this phenomena is the hacking tool *Try2DDOS*. This tool was meant to assist cybercriminals with launching denial-of-service attacks; it was originally created by a French cybercriminal, but the tool was later found adopted and modified by communities in Argentina, China, Ecuador, Guatemala, and Russia (Holt & Kilger, 2012).

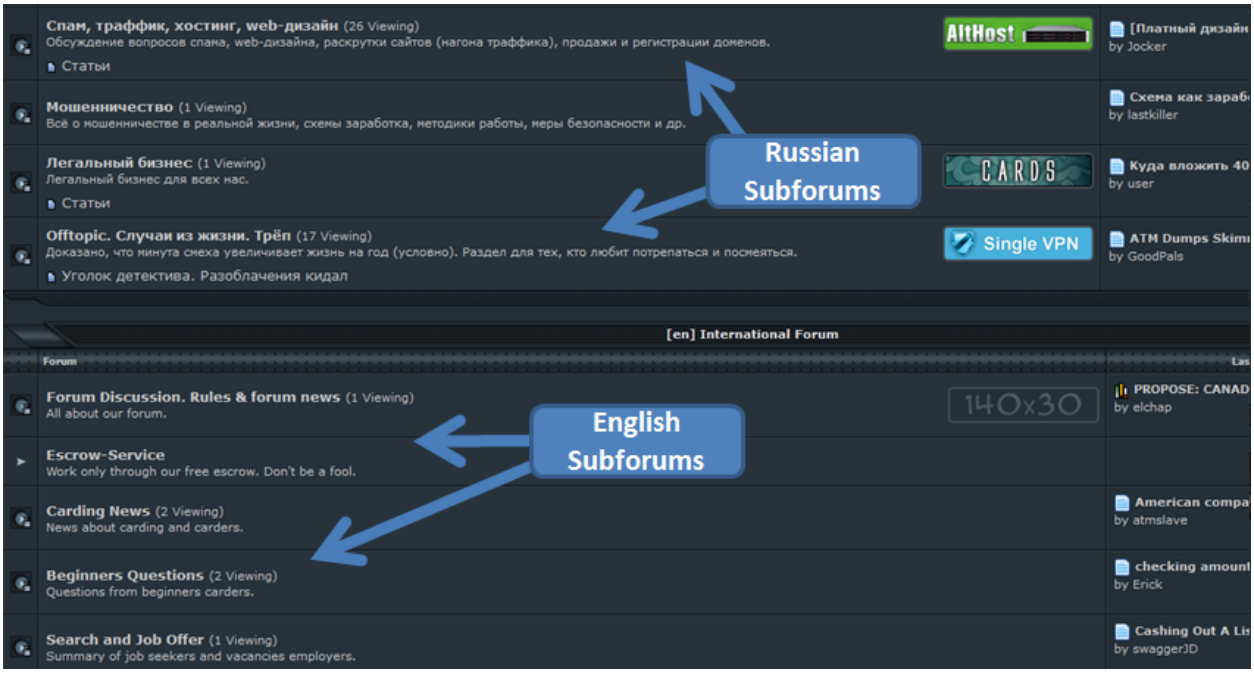

Figure 5.1 - Example of an English/Russian multilingual forum, *Crdclub.su*.

Unfortunately, little work has focused on studying transfer of knowledge and assets between cybercriminal populations. This line of research may be inhibited in part by difficulties identifying appropriate data sources, and challenges that arise with multilingual text processing. However, the benefits of work in this area are many; researchers and practitioners can gain insight into the global cybercriminal supply chain, better understand how information flows between distinct cybercriminal populations, and gain capability to assess the strategies that a specific cybercriminal group of interest may learn new techniques or acquire cybercriminal tools. Thus, I am motivated to develop a framework for identifying and analyzing instances of information dissemination between differing language populations within multilingual cybercriminal forums. I approach this unique research problem with the perspective that any effective design should include automated, scalable, and language-independent techniques. Additionally, I make use of ideas from information theory, such as entropy, to develop my framework upon a theoretical foundation. This framework would provide great value to security community by supporting research focused on international cybercrime and related areas.

This essay is organized into the following sections: First, I provide some background information concerning (1) past cybercriminal community research, (2) information theory, and (3) the state-of-the-art in NNLMs. My literature review is followed by a discussion of identified research gaps and questions. Next, I describe my research framework design, including data identification, collection, pre-processing, and analysis. I discuss in detail the steps necessary to identify potential forum discussions where information dissemination may occur between different cybercriminal populations. I then perform some experiments and discuss results while showcasing examples of forum discussions my framework identifies. Lastly, I conclude by discussing the contributions of this essay. In sum, this work provides a new framework for researchers to better

scrutinize information dissemination and asset flow among international cybercriminal communities.

5.2. Literature Review

To conduct this study and successfully develop an automated way for understanding Cybercriminal language, I must look to the latest research and borrow relevant perspectives. First, I review prior investigations of cybercriminal forums in order to provide contextual understanding relevant to this study. Additionally, a review of cybercriminal forum literature can help identify research gaps in need of attention. Next, I review information theory in order to develop a theoretical foundation to build my framework upon. Perspectives borrowed from information theory may also help guide research framework design. Lastly, recent works in lexical semantics, particularly in the areas of NNLMs. NNLMs are a state-of-the-art approach that can analyze textual contents in an automated, scalable approach. All three areas of literature are vital for achieving research goals.

5.2.1. Cybercriminal Community Research

Cybercriminals make extensive use of online web forums to support cybercriminal activity. In particular, cybercriminals use such forums to share cybercriminal assets and hacking knowledge with each other (Motoyama et al., 2011; Benjamin & Chen, 2014). It is not uncommon to witness hacking tools, malware samples, hacking tutorials, and more to be freely shared among forum members. Additionally, many cybercriminals will share links to other communities, underground economies, and deep web hidden services (Martin, 2013). Forums are not limited to a specific geopolitical region, and have been found to exist globally, including areas such as the United States, China, Russia, and the Middle-East (Benjamin & Chen, 2012; Holt et al., 2012). Black market activity regularly occurs within such forums. An example of such activity from the

multilingual English/Russian forum *Crdpro.su* can be seen in Figure 5.2. In this example, the message offers a phishing webpage creation service. The message author lists pricing and target companies that they are capable of creating scam webpages for.

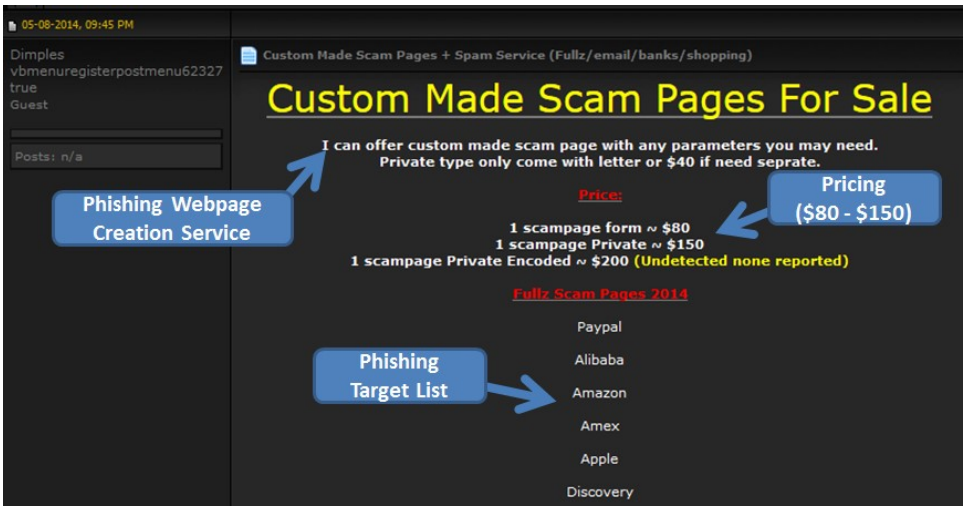

Figure 5.2 - Example of a posted message on a cybercriminal forum

As a result, recent years have seen security researchers and practitioners develop increased interest in analyzing data from such communities. Past work provides useful methods for identifying and collecting cybercriminal community contents. Additionally, literature provides context and insights for future cybercriminal studies.

Some common methods to identify cybercriminal communities exist throughout literature. Primarily, past studies resort to keyword searches for finding public hacking communities (Fallman et al., 2010). After an initial set of seed communities are identified, they can be scrutinized for hyperlinks and references to other cybercriminal communities, resulting in a snowball collection procedure (Holt & Lampke, 2010). After identification, data can be collected through various means. Forums can be collected with web crawlers; however, anti-crawling measures are sometimes put in place by cybercriminal forums to detect and halt crawling activity (Spencer, 2008; Fallman et al., 2010). Thus, it may be necessary to use proxy servers and identity

obfuscation techniques to avoid detection of crawling activities (Benjamin et al., 2015). For example, adjusting crawling rates and alternating between IP addresses used for crawling Cybercriminal contents may help conceal researcher identity and prevent Cybercriminal communities from discovering crawling activity.

The majority of previous Cybercriminal community research can be categorized within a few major themes. First, much existing work utilizes qualitative analyses to observe and describe Cybercriminal community activities (Motoyama et al., 2011; Holt et al., 2012). The second branch of work generally involves counting procedures and high-level statistical analyses of underground economy and carding community contents (Martin, 2013; Yip et al., 2013). Lastly, many recent works have focused effort on identifying key participants within Cybercriminal communities (Benjamin & Chen, 2012; Yip et al., 2013). These three categories of prior work are useful for describing ongoing activity within Cybercriminal communities, as they reveal commonly discussed topics, provide better understanding of Cybercriminal social dynamics, and help develop techniques to quickly identify key Cybercriminal community participants.

However, one underdeveloped research area is the identification of knowledge and asset dissemination between differing cybercriminal populations. Cybercriminals possess varied language, cultural, and geopolitical backgrounds; these differences present challenges for information dissemination and asset exchange between different cybercriminal communities. Fortunately, multilingual cybercriminal forums exist and can be explored in research to better understand the cybercriminal supply chain and how information may flow from one population to another. Thus, a research framework for scrutinizing multilingual cybercriminal forums is of value.

In particular, perspectives from information theory are useful for designing such a framework. Information theory provides a foundation on which to develop my framework upon. Concepts such

as entropy can be borrowed and operationalized within my framework to aid with identification of information dissemination between different cybercriminal populations.

5.2.2. Information Theory

Information theory, the study of the coding of information and how information can be transmitted, is useful in a number of applications involving information transmission, such as signal processing and data compression (Shannon, 1948). Additionally, information theory has strong influence on the natural language processing (NLP) community. Prior to the advent of modern information theory, most methods of coding and transmitting information implicitly assumed that all possible communication events were of equal probability. For example, suppose a multilingual virtual community; early methods would assume the probability of a discussion occurring that contains only single-language participants to be the same as the probability of a discussion occurring that contains multilingual participants. However, this assumption generally does not hold in real virtual communities.

Modern information theory introduces the concept of entropy, allowing for assumption that varying communication events can have unequal probability. Thus, it can be thought of as a measure that quantifies uncertainty or anomalies involved in communication. This is useful in language-related tasks, and within the context of a multilingual virtual communities, one may assume that information dissemination across language populations occurs when discussion threads contain multilingual participants.

Entropy could then be used to measure the uncertainty (i.e., probability) that information dissemination occurs between individuals from different language populations within a single discussion thread. Threads with mostly single-language participants would have low entropy, as they possess low chance of information disseminating between populations. Conversely, threads

with high entropy may contain a more diverse mix of participants, allowing for greater probability information dissemination occurs. Formally, entropy can be defined as follows, where p is the probability of discussions in thread x being language i:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i)$$

However, to operationalize this entropy measure within a computational framework for detecting information dissemination across language populations in multilingual forums, it is necessary to have the capability for handling multilingual text. Specifically, I must develop the capability to automatically identify what language population different forum participants belong too, which would allow us to potentially identify information dissemination between specific individuals. An automated technique that scales to many languages with minimal effort is ideal. The latest advancements in lexical semantics can help us with this task.

5.2.3. Lexical Semantics

Lexical semantics is a subfield of linguistics that focuses on: (1) the study of lexical units such as words, affixes, and phrases, (2) lexical relations, or how different lexical units relate to each other, and (3) how lexical units map into different concepts. Relationships between different lexical units (or words in this case) can be mapped. Higher abstractions of meaning can be inferred by scrutinizing lexical relationships; e.g., if Phoenix is related to the words city and Arizona, and Arizona is related to the words state and Phoenix, then Phoenix can be understood as a city within the state of Arizona.

Literature on lexical semantics is far too broad to be discussed in full here, and thus I focus only on the most recent, relevant stream of work. This includes work on scalable, automated techniques that are suitable for large-scale virtual community research. Further, I limit my review

to research utilizing unsupervised learning techniques that are scalable across multiple languages with minimal effort.

Recent advancements in NNLMs have captured much attention recently for NLP tasks (Mikolov et al., 2013a; Jansen et al., 2014; Le & Mikolov, 2014; Pennington et al., 2014). This recent rise in popularity is due to advancements in computing continuous vector representations of words and documents (i.e., word and document embeddings) that allow for high performance at low computation cost. In particular, a newly developed class of unsupervised two-layer NNs have generated much excitement within the NLP community (Mikolov et al., 2013b; Pennington et al., 2014). This new NN generates state-of-the-art word embeddings (i.e., vector representations of words) and has been benchmarked as a leading performer across multiple studies (Mikolov et al., 2013a; Le & Mikolov, 2014). Further, the technique is language-independent, and thus suitable for multilingual forum analysis.

NNLMs that utilize the Skip-gram learning model have become quite popular as they are one of the best performers among recently developed techniques (Mikolov et al., 2013b; Pennington et al., 2014). When given a particular word, a Skip-gram trained model can predict surrounding context words. For example, given the center word $w_{0,}$ the skip-gram algorithm will try to predict the surrounding context words w_{-2}, w_{-1}, w_1, and $w_{2.}$ More formally, skip-gram predicts surrounding words in a window of length c for every word. The objective function is to maximize the log probability of any context word given the current center word:

$$J(\theta) = \frac{1}{T}\sum_{t=1}^{T} \sum_{-c \le j \le c, j \ne 0} \log p(w_{t+j} \mid w_t)$$

Where T is the total number of words in the corpus, t holds the position of the current center word, c is the defined context window (e.g., a window size of 2 includes the preceding and proceeding

two words surrounding the current center word), j denotes the specific position of a context word within c, relative to t, and w_t is the actual word at the center position t.

It was later discovered that by concatenating word embeddings, one can create vector representations of documents (i.e., document embeddings). The idea is as follows: with word embeddings, each word within a corpus is mapped to a unique vector containing other words. This technique can be extended so that document embeddings can be constructed by mapping each document to a unique vector of word embeddings. The document embedding then becomes a concatenation of these word embeddings.

The Paragraph Vector model is popular for building state-of-the-art document embeddings (Le & Mikolov, 2014). Paragraph Vector is an unsupervised algorithm that learns fixed-length vector representations of variable-length documents, where each document is represented by a dense vector which is trained to predict words in the document. Although named paragraph vector, all different types and lengths of documents can be processed with this technique, including sentences, forum messages, book chapters, and more. A specific implementation of Paragraph Vector that has gained much traction is the Paragraph Vector Distributed Bag-of-Words (PV-DBOW) model (Le & Mikolv, 2014). Its popularity is largely in part due to conceptual similarities with skip-gram learning for word embeddings. More formally, there is only one term in addition to the standard skip-gram model:

$$J(\theta) = \frac{1}{T}\sum_{t=1}^{T}\sum_{-c\le j\le c, j\neq 0}\log p(w_{t+j} \mid w_t, d_i)$$

Where d_i refers to the document where the current word w_t is included, every document is mapped to a unique vector in matrix *D,* and every word is mapped to a unique vector in matrix *W*. Document embeddings are composed by vectors in *D* that are concatenated with multiple vectors from *W*.

The advantages of this technique is that embeddings can be learned from unlabeled data that can scale across languages. Additionally, it contain properties that are able to inherit semantics of words and consider word order, while the standard bag-of-words model does not.

Document embeddings can be used for many NLP-related applications. For example, embeddings can be used directly in machine learning classification and clustering tasks, or paired with other engineered features for multi-faceted analyses. In contexts where other text features do not need to be generated, embeddings can be easily utilized for multilingual analyses.

When considering the virtual community context, document embeddings can be powerful tools for understanding community discussions. For example, recall that the previously discussed concept of entropy and how it can be used to help identify threads that contain potential information dissemination between different cybercriminal language populations may occur. However, for entropy to be useful, community discussions and/or participants must be categorized by language usage and potential geopolitical origin. Document embeddings generated by unsupervised NN models, such as PV-DBOW, could be useful for categorizing texts by language category. The two combined may help develop an automated and scalable capability for scrutinizing information dissemination within multilingual cybercriminal communities.

5.3. Research Gaps and Questions

One underdeveloped research area is the identification and analysis of knowledge and asset dissemination between differing cybercriminal populations. It is known that cybercriminals possess varied language, cultural, and geopolitical backgrounds, and that they may form groups based on such identities. Further, there is substantial evidence of cybercriminal asset flow between such groups, even those separated by language barriers. e.g., Try2DDOS malware (Holt & Kilger, 2012). Many multilingual cybercriminal forums exist that can be explored in research to better

understand such activities. However, no method or set of guidelines appear to exist for scrutinizing such forums and identifying potential information dissemination. Thus, additional work is necessary to develop this capability.

I am motivated to address these issues by developing a framework for identifying and analyzing instances of information dissemination between differing language populations within multilingual cybercriminal forums. The framework is composed of an automated, scalable, and neural network-based approach. The framework is also language-independent and can handle a multitude of cybercriminal forums. Lastly, I build the framework upon the foundations of information theory. This leads us to the following research questions: how can I develop a scalable, language-independent approach for categorizing multilingual texts? Is the Paragraph Vector model better suited for this task than more traditional bag-of-words models? Can I operationalize perspectives borrowed from information theory (e.g., entropy) to help identify potential information dissemination among varying cybercriminal populations?

5.4. Research Testbed and Design

My research design (Figure 5.3) consists of a series of steps involving automated data processing and analysis. First, I identify and collect cybercriminal forums for this study. Next, I scrub and process collected data into a form ready for analysis. I then construct my NNLMs and execute experiments. Finally, experiment results are evaluated and conclusions are drawn based on my findings. I provide details for each component separately.

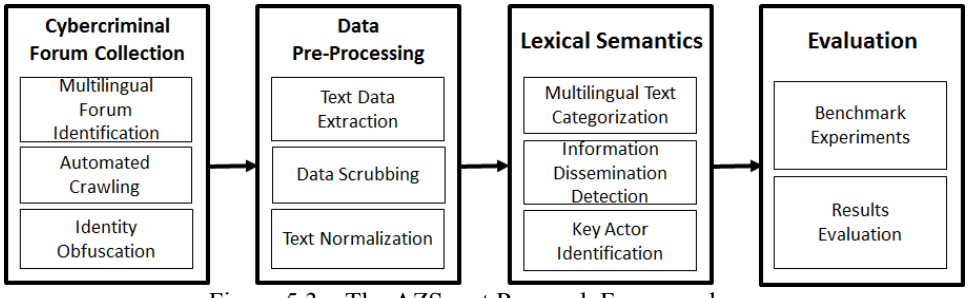

Figure 5.3 – The AZScout Research Framework

5.4.1. Data Collection

To develop a testbed for this research, I identify and collect two popular cybercriminal forums frequently visited by English- and Russian-speaking participants I identified a subset of public Cybercriminal forums by utilizing keyword searches such as "black hat forum" and "carder community" to identify cybercriminal forums. From this initial set, I scrutinize hyperlinks that participants share for potential linkage to other cybercriminal communities. Forums were chosen based on several factors: their activity level and message volume over time, consistent multilingual activity over time, abundant discussions of various cybercriminal concepts, including black market activity, intellectual property theft, fraud, hacking tools, malware, and more

I identify and collect two cybercriminal forums for this study (Table 5.1). Forums were chosen based on several factors. First, both forums are multilingual communities that contain English- and Russian-speaking participants. Additionally, the forums contain consistent forum activity over time with recent activity from multiple forum participants. Lastly, I observe abundant discussion of hacking concepts, tools, and black market activity that would be interesting to study in this research.

Forum Name	Members	Threads	Posts	Testbed Time Span
Crdclub.su	2,401	2,895	12,428	1/1/2014 – 1/20/2015
Crdpro.su	6,301	6,278	23,523	1/1/2014 – 1/20/2015

Table 5.1 – Research Testbed

Automated crawlers deployed to collect identified forums. Several steps were taken to customize crawlers in order to ensure successful collection of Cybercriminal data while protecting researcher security (Benjamin et al., 2015). For example, I route Internet traffic generated by my crawlers through the Tor anonymity network and personal proxy servers in order to protect researcher identity and university affiliation. The Tor network is a peer-to-peer Internet traffic routing service that effectively anonymizes Internet communications. Further example of crawler

customization includes altering crawling rates in order to avoid triggering server-side anti-crawling mechanisms.

After successful collection, data of interest must be extracted from raw HTML pages that are downloaded by the web crawlers. Extracted data must then be further processed in order to be readied for analysis. Overall, this includes several steps.

5.4.2. Data Pre-processing

After collection, I extract message text from collected Cybercriminal forum web pages. Regular expressions are written to extract data embedded within HTML, specifically thread titles and message bodies. I then scrub data by removing duplicate messages and removing instances where a forum participant quotes the message of another participate, which would result in duplicate text. I then normalize extracted messages in preparation for analysis. First, I convert all text to lowercase in order to avoid having different cases of the same word become treated as two separate words in my model. Second I strip punctuation from words to again avoid duplication of words.

5.4.3. Lexical Semantics

The proposed research framework is intended to identify potential instances of information dissemination between cybercriminals of different geopolitical origin. The framework enables researchers and practitioner the capability to map the global cybercriminal supply chain, and to track the spread of information, hacking tools, malware, and more across varying cybercriminal populations. Additionally, my framework can help identify key actors involved in such dissemination.

After acquiring multilingual forum data, there are four steps necessary to reach research goals. Step 1 includes categorizing all forum participants into groups based on language usage, potential geopolitical origin. In step 2, I must identify discussion threads containing messages from forum

participants of different groups. During step 3, I rank identified discussion threads for potential of information dissemination between groups occurring. I end with step 4, where I can extract participant activity and identify key *actors involved.*

Step 1: Within multilingual forums, participants can be categorized into different potential geopolitical origins based on their language usage. To do this, I can generate a series of lexical features for each forum participant, thus creating feature vectors or 'author embeddings,' and using these feature vectors to cluster participants. Clusters are likely to represent different language groups due to n-gram exclusivity to specific languages; for example, in a forum with English- and Russian- speaking cybercriminals, grouping all forum members into two clusters would likely result in categorizing participants by their language usage.

However, one challenge with clustering forum participants is the need for balanced feature generation that represents all languages used within a forum. Additionally, I must consider the need for unsupervised models that can scale across multiple languages effortlessly. One popular unsupervised text feature generation method is to select n-grams (i.e. words or phrases) as features based on their frequency. For example, imagine a corpus of 10,000 documents. The top 1,000 n-grams shared between these documents can be used as a global feature for which to evaluate all documents in clustering tasks. However, one concern with this method is whether selecting top n-grams can result in good representations of multilingual forum data. For example, imbalances in language usage within a forum may result in the top n-grams belonging predominantly to one language, making generated features less useful for less popular languages.

I previously discussed the recent development of neural network-based PV-DBOW method, and its usefulness in generating vector representations of documents. One key difference that separates it from the traditional n-gram model is that it samples and generates features from each

individual document within a corpus, rather than simply selecting the most frequent n-grams. This difference indicates that features can be better generated for less-popular languages by PV-DBOW instead of the traditional n-gram frequency model.

After feature vectors are generated, they can be used to categorize participants by language usage and potential geopolitical origin. Russian-speakers are likely to come from Russia, many English-speakers appear to focus their discussions on stolen data and payment methods relevant to America, and so on. To categorize participants, I utilize an automated method that can scale across large datasets effortlessly. Thus, I rely on machine learning clustering; the k-means is a transparent and easily interpretable clustering algorithm that can group participants (Hartigan & Wong, 1979). With k-means, I can generate two clusters for each of my forums, with one cluster representing English-speaking participants and the other for Russian-speakers.

*Step 2: F*orum participants are categorized into two clusters with k-means, I can identify discussion threads that contain participation from both groups. Threads containing participants from each cluster are representative of interactions between different cybercriminal populations, and thus contain potential information or asset dissemination between populations. However, discussion threads will vary in the amount of information they contain regarding such activity. Threads that contain a large volume of messages from numerous participants of both clusters are more likely to contain information dissemination than threads containing participants mostly from one cluster, with the other population being underrepresented. Thus, a ranking mechanism is necessary to highlight the most promising threads

Step 3: Each thread will have a unique ratio of participants that belong to different clusters. Some threads will only have participants from one cluster, while other threads will possess greater participant diversity. Threads with greatest potential for information dissemination between groups

are those that have the most participants from differing clusters. Recall that entropy can then be used to measure the probability that information dissemination occurs between individuals from different language groups. Discussion threads containing more participants from varying clusters will have higher entropy. I can use entropy as a ranking measure to find most pertinent threads.

Step 4: After ranking threads, it is also of interest to identify participants who are the most active in disseminating information between groups. Thus, I must consider each participant's message volume in high-entropy threads. Greater participation in high-entropy threads indicates more active role in the information dissemination process. Thread participants can be ranked by their activity within the top-*n* high-entropy threads with the following metric, where *v* is a participant's message volume within thread *t* and *p* is the entropy of thread *t*:

$$R = \sum_{t=1}^{n} v_t * p_t$$

5.4.4. Evaluation

I must evaluate the capability for PV-DBOW to successfully generate feature vectors representative of all multilingual forum participants. Features are used to categorize participants using k-means clustering; feature generation that better represents participants will result in superior categorization of potential geopolitical origin. I can measure how well features represent underlying data by calculating the sum of squared errors (SSE) for generated clusters using varied feature generation methods (PV-DBOW vs n-grams) (Hartigan & Wong, 1979). For each participant within a cluster, SSE represents "closeness" to cluster center. SSE indicates variance among participants within each cluster. The feature generation method that best represents participants will result in the smallest SSE.

5.5. Results and Discussion

The first experiment was to cluster participants based on generated feature sets. I generate features with PV-DBOW and n-gram models, with the n-gram models including unigram, bigram, and trigram configurations. I use k-means clustering to categorize forum participants into English and Russian groups based on their message contents. I perform clustering across all feature sets and evaluate the average SSE after 100 iterations (Table 5.2). PV-DBOW appeared to produce features resulting in the smallest SSE for both the *Crdclub* and *Crdpro* forums. This coincides with my previous idea that PV-DBOW will be able to better represent less frequently used languages than traditional n-gram models, thus producing a smaller SSE. I further explored the clusters generated by PV-DBOW feature vectors and describe them in Table 5.3.

	Within Cluster SSE	
	Crdclub	***Crdpro***
Unigrams	389.573	512.754
Bigrams	427.129	563.842
Trigrams	484.432	618.540
PV-DBOW	357.936 *	491.801 *

Table 5.2 – Average SSE Per Feature Generation Method

	PV-DBOW Cluster Size Information	
	Crdclub	***Crdpro***
English	745 (31.036%)	4,863 (77.182%)
Russian	1,656 (68.964%)	1,438 (22.818%)
Total	2,401 Participants	6,301 Participants

Table 5.3 – PV-DBOW Generated Cluster Sizes

After categorizing forum participants within groups, I can identify discussion threads that carry the most potential for information dissemination between groups. To do this, I can calculate the entropy of each thread as a measure of potential information dissemination. Threads containing

more messages from participants of different groups will have the highest entropy. I highlight findings in Table 5.4.

	Thread Entropy Statistics	
	Crdclub	*Crdpro*
Total Number of Threads with Between-Cluster Discussion	523 (18.063%)	710 (11.307%)
Average Thread Entropy	0.2648	0.2137
Maximum Thread Entropy	0.3631	0.3593
Minimum Thread Entropy	0.1722	0.1175

Table 5.4 – Identified Thread Entropy Statistics

After identifying threads of interest through my entropy measure, I can extract participant activity and identify key actors involved in information dissemination between clusters. I rank participants by utilizing my previously described rating metric, i.e., top-ranked participants contribute multiple messages to high-entropy threads. I extract the top-5 participants based on my metric from each forum (Table 5.5).

	Top 5 Participants in High-Entropy Threads			
	Crdclub		***Crdpro***	
	Name	*Score*	*Name*	*Score*
1	M***k	11.541	T***s	9.543
2	B***0	8.836	Г***ь	7.464
3	M***d	8.593	B***0	5.872
4	N***n	8.211	D***t	4.364
5	C***u	7.974	J***e	4.273

Table 5.5 – Potential Key Actors

Using these results, I showcase some examples of top participants. The first case example is of top *Crdpro.su* participant *Г***ь: This user has an en*tropy score of 7.464 and is notable for running a Russian-centered carding service (i.e., credit card fraud) (Figure 5.4, top). They also distribute malware in English and Russian discussions (Figure 5.4, bottom). Forum participants such as

*Г***ь* appear to be the main drivers of discourse and asset exchange between the English and Russian subcommunities within the *Crdpro* forum.

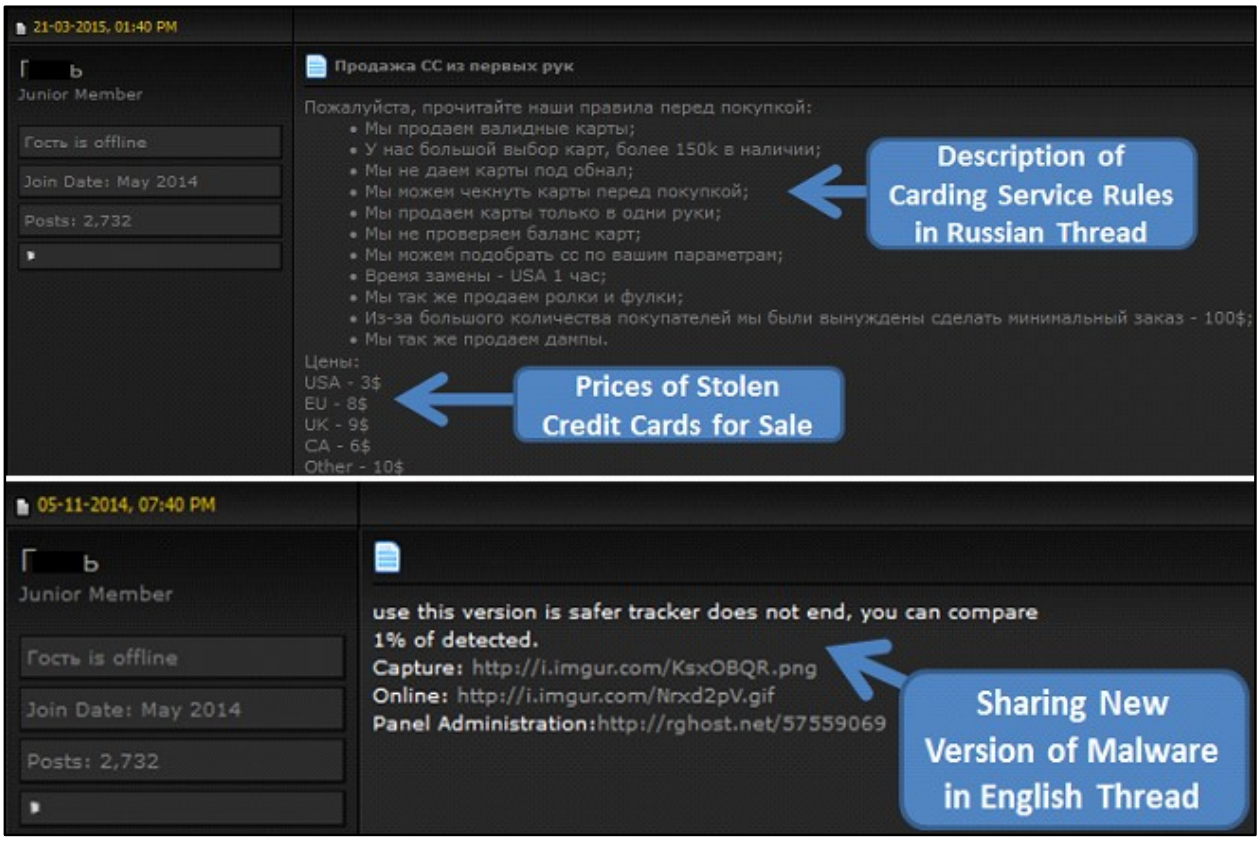

Figure 5.4 –Russian and English Subforum Activity by participant *Г***ь* within *Crdpro.su*

In the *Crdclub.su* forum, one of the top participants involved in information dissemination between populations is *N***n.* This user has an entropy score of 8.211. Among Russian cybecriminals, this user is involved with trading and sharing stolen credit card data (Figure 5.5, top). Within English subforums, *N***n* is often seeing discussing the latest malware or hacking techniques, and will sometimes openly share assets with others (Figure 5.5, bottom). Like *Crdpro*, multilingual participants of *Crdclub* facilitate much cybercriminal knowledge and asset exchange between English- and Russian-speaking populations.

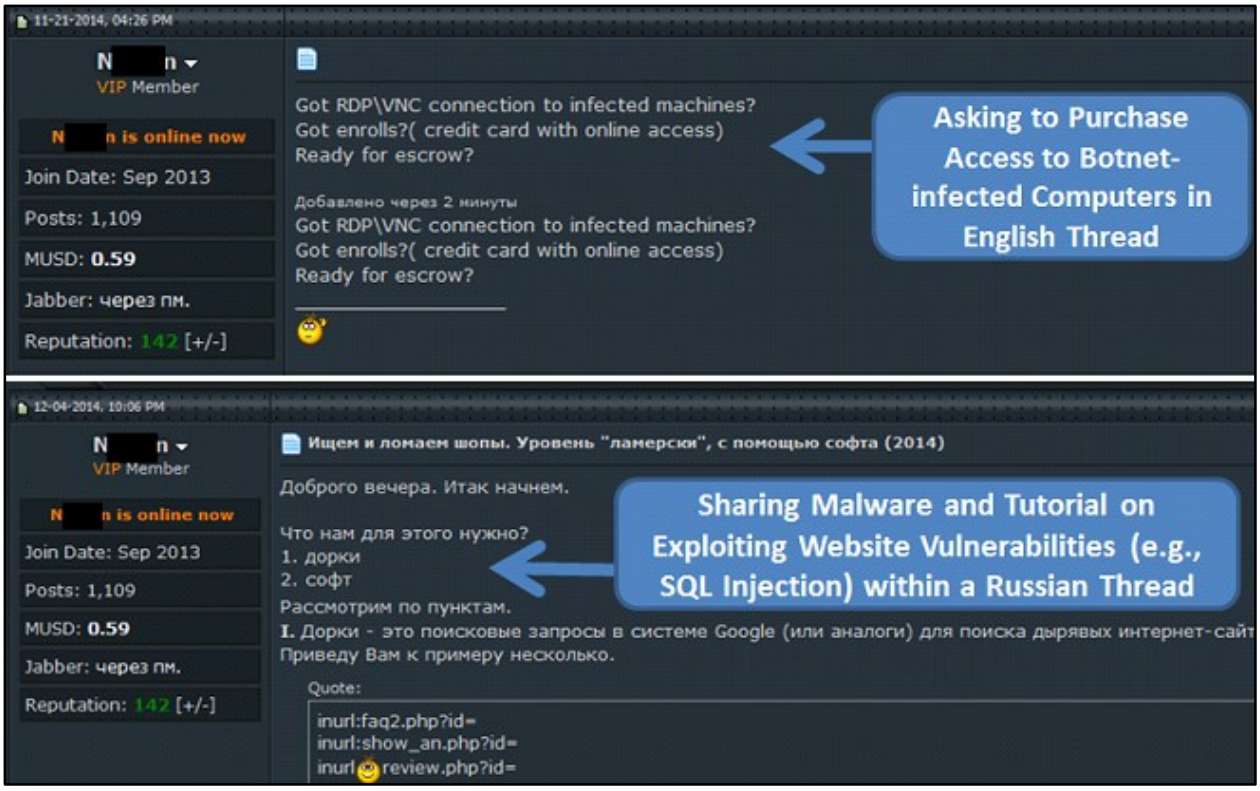

Figure 5.5 –Russian and English Subforum Activity by participant *N***n* within *Crdclub.su*

5.6. Conclusion

In this work, I develop a framework for identifying and analyzing instances of information dissemination between differing populations within multilingual cybercriminal forums. Little work has focused on studying transfer of knowledge and assets between cybercriminal populations thus far, perhaps due to challenges with identifying data and processing multilingual text. However, such a capability would provide great insights on the global cybercriminal supply chain, as well as information flow between different cybercriminal populations.

The framework employs an automated, scalable, state-of-the-art neural network approach while leveraging perspectives of information theory and entropy. Additionally, it is language-independent due to design choices of utilizing unsupervised feature generation. The described research framework is suitable for studying many different types of cybercriminal communities across varying geopolitical regions.

This research has many contributions. First, I contribute to NLP literature by benchmarking and creating application for new paragraph vectors technique, as well as benchmarking the PV-DBOW's ability to generate features that closely represent underlying data. I compare performance against a traditional n-gram frequency approach. Further, I contribute to cybersecurity literature by providing new capabilities for studying multilingual cybercriminal forums and scrutinizing information dissemination across cybercriminal populations.

6. CONCLUSIONS

As computing technologies become more ubiquitous within society, cybersecurity has become a problem of growing importance and concern. As a result, researchers have become increasingly interested in exploring cybercriminal social media in order to learn more about cybercriminal social behaviors, emerging threats, and the cybercriminal supply chain. However, until now, few works in recent years have successfully performed large-scale identification, collection, and analysis of cybercriminal-generated data. In particular, many cybercriminal-operated communities exist that can be studied to inform new perspectives on cybercrime, but these data sources have largely gone untapped by researchers.

I have presented four essays in this dissertation that center on exploring cybercriminal community contents through automated web and text mining perspectives. The first essay outlines a series of guidelines for conducting cybercriminal community research, and the three subsequent essays build upon this foundation. Each essay expands upon my knowledge of cybercriminals, while also contributing to methodological development. The focus on creating practical systems that can solve real-world problems and impact the capabilities of other security researchers and practitioners, I showcase the role of IS and how it may contribute to cybersecurity.

6.1. Cybersecurity Contributions

The first essay introduces a set of computational methodologies and research guidelines for conducting cybercriminal community research. This essay outline methodologies for identifying, collecting and analyzing cybercriminal community contents, and also discuss how to operationalize cybercriminal research in a safe and secure manner. There have been no similar studies that provide guidelines or steps for other security researchers until this point.

In the second essay, I examine possible motives for prolonged participation by individuals within cybercriminal IRC communities. I identify challenges and solutions for collecting cybercriminal IRC data. I also provide description of feature generation from such data. Findings underline the importance of interconnectedness within cybercriminal communities.

The third essay has numerous contributions. First, I contribute to methodology by expanding current state-of-the-art NNLMs to handle temporal data. Second, I contribute to cybersecurity by developing a new method to automatically identify potentially unknown hacker terms, and to develop understanding of hacker language trends over time. Lastly, my technique for boosted training can be applied to utilize terms from other domains beyond the cybercriminal context.

The last essay focuses on developing a framework for identifying information dissemination among varying international cybercriminal populations by examining multilingual cybercriminal forums. The framework utilizes the Paragraph Vector with Distributed Bag-of-Words NNLM, and borrows perspectives from information theory. The AZScout research framework provides researchers and practitioners the capability to more closely scrutinize the global cybercriminal supply chain, as well as how information and assets transfer between different cybercriminal populations.

6.2. Contributions to the IS Field

This dissertation has several contributions to expanding the scope of the IS field, particularly due to its focus on unique problems not traditionally pursued by the greater IS community. First, I showcase the important role that IS researchers can have in advancing fields of science outside of the immediate business context. By borrowing from the latest advancements in computer science and computational linguistics, this dissertation helps solve high-impact, real-world problems in the cybersecurity space. Additionally, the first essay of this dissertation outlines

several steps that IS researchers can utilize for their own explorations into cybercriminal communities. This allows for more individuals within our community to pursue this important problem domain, leading to impactful research with practical and real-world applications. By working on high-impact problems of great societal relevance, the IS field will invite a greater audience to our research.

Further, this dissertation also shows that IS researchers are capable of contributing back to methodology borrowed from reference disciplines. Specifically within the third essay, I am able to make algorithmic contributions back to the NLP community by extending the skip-gram NNLM to have the capability for modeling temporal attributes of data. Such algorithmic contributions are not typical of IS research, but such work will enable the IS community to begin forming stronger relationships with computer scientists, linguists, and researchers from other disciplines.

The IS field is unique due to its role as the business discipline most closely-aligned with science, technology, engineering, and mathematics (STEM) disciplines We possess the greatest potential of any discipline for bridging business scholarship with the greater STEM community. This dissertation showcases how IS research can have relevance to STEM-related fields while also focusing on high-impact problems of great societal relevance.

6.3. Future Research Directions

The work described in this dissertation can be extended in several directions for future research.

First, there is a need to explore more deep web-based communities, such as those on the Tor anonymity network. While efforts have been made in this space, there are many communities that remain unknown to researchers which may contain critical data for identifying emerging threats, cybercriminal trends, and more. Additionally, many forms of black markets exist within anonymity

networks that may be useful for developing deeper understanding of the global cybercriminal supply chain. Investigations of cybercriminal communities may also provide researchers with some leads on how cybercrime will affect the growing Internet-of-Things and Internet-enabled medical devices.

A second area of future research is to continue application of state-of-the-art machine learning and text mining methodologies. Many of these methods introduced and benchmarked on traditional datasets, but their effectiveness in more real-world applications are often times untested. The application of such methodologies may also enable new, unique research perspectives on cybercriminal data. For example, my work on NNLMs enabled unique modeling of hacker language. Future methodologies may improve performance for automating emerging threat detection and identification of other critical information. Advances with new methodologies are also generalizable to other domains.

Perhaps the most critical future direction is to continue identifying how cybercriminals adopt new platforms to build communities upon. Forums and IRC are both two traditional platforms to build web communities upon. With the ubiquity of mobile devices and emergence of Internet-of-Things devices, cybercriminals will be sure to evolve how they choose to communicate with each other, as well as what they choose to communicate. It is important for researchers and practitioners to remain aware of such trends and to adapt new technologies to stay current with cyber adversaries.

7. REFERENCES

Abbasi, A., & Chen, H. (2008). CyberGate: A Design Framework and System for Text Analysis of Computer-Mediated Communication. *MIS Quarterly*. *32*(4), 811–837.

Abbasi, A., Li, W., Benjamin, V., Hu, S., & Chen, H. (2014). Descriptive Analytics: Investigating Expert Cybercriminals in Web Forums. *Proceedings of the IEEE Joint Intelligence and Security Informatics Conference.* 55-63. The Hague, Netherlands. September 24-26.

Abfalter, D., Zaglia, M. E., & Mueller, J. (2012). Sense of virtual community: A follow up on its measurement. *Computers in Human Behavior.* 28(2), 400–404.

Adamic, L. A., Zhang, J., Bakshy, E., Ackerman, M. S., and Arbor, A. (2008). Knowledge Sharing and Yahoo Answers : Everyone Knows Something. *Proceedings of the 17th International Conference on World Wide Web*. 665–674. Beijing, China. April 21-25.

Alfaro, L. De, and Kulshreshtha, A. (2011). Reputation Systems for Open Collaboration. *Communications of the ACM*. 54(8), 81–87.

Balahur, A., Hermida, JM., Montoyo, A. (2010). Detecting implicit expressions of emotion in text: a comparative analysis. *Decision Support Systems*. 53(4). 742-753.

Benjamin, V., & Chen, H. (2012). Securing Cyberspace : Identifying Key Actors in Cybercriminal Communities. *Proceedings of the IEEE Joint Intelligence and Security Informatics Conference.* 24-29. Washington, D.C. June 11-14.

Benjamin, V. A., & Chen, H. (2013). Machine Learning for Attack Vector Identification in Malicious Source Code. Proceedings of the *IEEE Intelligence and Security Informatics Conference.* 21–23. Seattle, Washington. June 4-7.

Benjamin, V., & Chen, H. (2014). Time-to-event Modeling for Predicting Hacker IRC Community Participant Trajectory. *Proceedings of the IEEE Joint Intelligence and Security Informatics Conference.* 25-32. The Hague, Netherlands. September 24-26.

Benjamin, V., and Chen, H. (2015). Developing Understanding of Cybercriminal Language through the use of Lexical Semantics. *Proceedings of the IEEE Joint Intelligence and Security Informatics Conference.* 79-84. Baltimore, MD. May 27-29.

Benjamin, V., Chung, W., Abbasi, A., Chuang, J., Larson, C. a, & Chen, H. (2014). Evaluating text visualization for authorship analysis. *Security Informatics*. Springer. 3(1), 1-13.

Benjamin, V., Chung, W., Abbasi, A., Chuang, J., Larson, C., and Chen, H. (2013). Evaluating text visualization: An experiment in authorship analysis. *Proceedings of the IEEE Intelligence and Security Informatics Conference.* 16–20. Seattle, Washington. June 4-7.

Benjamin, V., Li, W., Holt, T., and Chen H. (2015). Exploring Threats and Vulnerabilities in Cybercriminal Web Forums, IRC, and Carding Shops. *Proceedings of the IEEE Joint Intelligence and Security Informatics Conference.* 85-90. Baltimore, MD. May 27-29.

Bewick, V., Cheek, L., & Ball, J. (2004). Statistics review 12: Survival analysis. *Critical Care.* 8(5), 389–94.

Carlson J. R., and Zmud, R. W. (1999). Channel Expansion Theory and the Experiential Nature of Media Richness Perceptions. *Academy of Management Journal*. 42(2), 153-170.

Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly*. 36(4), 1165–1188.

Chiu, C.-M., Hsu, M.-H., & Wang, E. T. G. (2006). Understanding knowledge sharing in virtual communities: An integration of social capital and social cognitive theories. *Decision Support Systems*. 42(3), 1872–1888.

Cook, T.D., & Campbell, D.T. (1979). Quasi-experimentation: Design and analysis issues for field settings. Boston, MA. Houghton Mifflin Company.

Council, N. S. and T. (2011). Trustworthy Cyberspace: Strategic Plan for the Federal Cybersecurity Research and Development Program. 1–19.

Cova, M., Kruegel, C., and Vigna, G. (2010). Detection and analysis of drive-by-download attacks and malicious JavaScript code. *Proceedings of the 19th International Conference on World Wide Web*. 281-290. Raleigh, North Carolina. April 26-30.

Cox, D. R., Society, S., & Methodological, S. B. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society*. 34(2), 187–220.

Daft, R. L., and Lengel, R. H. (1986). Organizational Information Requirements, Media Richness and Structural Design. *Management Science.* 32(5), 554-571.

Dennis, A. R., Fuller, R. M., and Valacich, J. S. (2008). Media, Tasks, and Communication Processes: A Theory of Media Synchronicty. *MIS Quarterly.* 32(3), 575–600.

Dholakia, U. M., Bagozzi, R. P., and Pearo, L. K. (2004). A social influence model of consumer participation in network- and small-group-based virtual communities. *International Journal of Research in Marketing*. 21(3), 241–263.

Fallman, H., Wondracek, G., and Platzer, C. (2010). Covertly Probing Underground Economy Marketplaces. *Proceedings of the 7th International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment (DIMVA).* 101–110. Bonn, Germany. July 8-9.

Freeman, L. C. (1979). Centrality in Social Networks Conceptual Clarification. *Social Networks*. 1(3), 215–239.

Garas, A., Garcia, D., Skowron, M., & Schweitzer, F. (2012). Emotional persistence in online chatting communities. *Scientific Reports*. 2, 1–34.

George, J.F., Carlson, J., & Valacich, J.S. (2013). Media Selection as a Strategic Component of Deceptive Communication. *MIS Quarterly*. 37(4), 1233-1251.

Holt, T. J. (2013). Examining the Forces Shaping Cybercrime Markets Online. *Social Science Computer Review*. 31(2), 165–177.

Holt, T. J., & Kilger, M. (2012). Know Your Enemy : The Social Dynamics of Hacking. *The Honeynet Project*. 1–17.

Holt, T. J., & Lampke, E. (2010). Exploring stolen data markets online: products and market forces. *Criminal Justice Studies: A Critical Journal of Crime, Law, and Society*. 23(1), 33–50.

Holt, T. J., Strumsky, D., Smirnova, O., & Kilger, M. (2012). Examining the Social Networks of Malware Writers and Cybercriminals. *International Journal of Cyber Criminology*. 6(1), 891–903.

Jansen, P., Surdeanu, M., & Clark, P. (2014). Discourse Complements Lexical Semantics for Non-factoid Answer Reranking. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics.* 977–986. Baltimore, Maryland. June 22-27.

Jones, S. (2002). Internet Relay Chat. *Encyclopedia of New Media: An Essential Reference to Communication and Technology.* SAGE Publications. 1, 257.

Kalbfleisch, J. D., & Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data (2nd ed.)*. Wiley-Interscience.

Kim, H.S., and Sundar, S. (2011). Using interface cues in online health community boards to change impressions and encourage user contribution. *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems.* 599–608. Vancouver, BC. May 7-12.

Koh, J., Kim, Y.-G., Butler, B., & Bock, G.-W. (2007). Encouraging Participation in Virtual Communities. *Communications of the ACM*. 50(2), 68–73.

Le, Q., Mikolov, T. (2014). Distributed Representations of Sentences and Documents. *arXiv preprint arXiv:1405.4053*.

Leavitt, N. (2009). Anonymization Technology Takes a High Profile. *IEEE Computer Society*. 15–18. November.

Levy, O., & Goldberg, Y. (2014). Linguistic Regularities in Sparse and Explicit Word Representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning Conference* 171. Baltimore, Maryland. June 26-27.

Lin, Y.-K., Chen, H., Brown, R. A., Li, S.-H., & Yang, H.-J. (2014). Time-to-Event Predictive Modeling for Chronic Conditions using Electronic Health Records. *IEEE Intelligent Systems*. 29(3), 14-20.

Liu, L., and Munro, M. (2012). Systematic analysis of centralized online reputation systems. *Decision Support Systems*. 52(2), 438–449.

Liu, X., & Chen, H. (2013). AZDrugMiner : An Information Extraction System for Mining Patient-Reported Adverse Drug Events. *Smart Health*. 134–150. Springer Berlin Heidelberg.

Mahmood, A. M., Siponen, M., Straub, D., Rao, H. R., & Raghu, T. S. (2010). Moving Toward Black Hat Research in Information Systems Security: An Editorial Introduction to the Special Issue. *MIS Quarterly*. 34(3), 431–433.

Martin, J. (2013). Lost on the Silk Road: Online drug distribution and the "cryptomarket." *Criminology and Criminal Justice*. 14(3), 351-367.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv*:1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*. 26. 3111–3119.

Mnih, A., & Kavukcuoglu, K. (2013). Learning word embeddings efficiently with noise-contrastive estimation. *Advances in Neural Information Processing Systems*. 26. 2265–2273.

Moore, T., and Clayton, R. (2009). Evil Searching: Compromise and Recompromise of Internet Hosts for Phishing. *Financial Cryptography and Data Security.* 256–272.

Motoyama, M., McCoy, D., Levchenko, K., Savage, S., & Voelker, G. M. (2011). An analysis of underground forums. *Proceedings of the ACM SIGCOMM Conference on Internet Measurement Conference.* 71-80. Berlin, Germany. November 2-4.

Nahapiet, J., & Ghoshal, S. (1998). Social Capital, Intellectual Capital, and the Organizational Advantage. *Academy of Management Review*. 23(2), 242–266.

National Science and Technology Council (NSTC) (2011). Trustworthy Cyberspace: Strategic Plan for the Federal Cybersecurity Research and Development Program. *Report of the National Science and Tehcnology Council, Executive Office of the President.* 1–19.

Onds, I. N. B., Ren, Y., Harper, F. M., Drenner, S., Terveen, L., Kiesler, S., Kraut, R. E. (2012). Building Member Attachment in Online Communities: Applying Theories of Group Identity and Interpersonal Bonds. *MIS Quarterly*. 36(3), 841–864.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe : Global Vectors for Word Representation. *Proceedings of the Empirical Methods in Natural Language Processing*. 14, 1532–1543. Doha, Qatar. October 25-29.

Pohlman, J. T., and Leitner, D. W. (2003). Comparison of Ordinary Least Squares and Logistic Regression. *The Ohio Journal of Science*. 103(5), 118–125.

Qassrawi, M. T., and Zhang, H. (2010). Client Honeypots: Approaches and Challenges. *Proceedings of the IEEE International Conference on New Trends in Information Science and Service Science (NISS).* 19–25. Gyeongju, South Korea. May 11-13.

Radianti, J. (2010). A Study of a Social Behavior inside the Online Black Markets. *Proceedings of the International Conference on Emerging Security Information, Systems and Technologies*. 88–92. Nice, France. July 24-28.

Sandle, P., & Char, P. (2014). Cyber crime costs global economy $445 billion a year: report. *Reuters.* http://www.reuters.com/article/2014/06/09/us-cybersecurity-mcafee-csis idUSKBN0EK0SV20140609

Schone, M. Esposito, R. Cole, M. & Greenwald, G. (2014). War on Anonymous: British Spies Attacked Cybercriminals, Snowden Docs Show. *NBC News*. http://www.nbcnews.com/news/investigations/war-anonymous-british-spies-attacked-cybercriminals-snowden-docs-show-n21361.

Shriver, S., Nair, H., & Hofstetter, R. (2013). User-Generated Content and Social Ties: Evidence from an Online Social Network. *Management Science.* 59(6), 1425-1443.

Sinha, T., & Rajasingh, I. (2014). Investigating substructures in goal oriented online communities: Case study of Ubuntu IRC. *Proceedings of the IEEE International Advance Computing Conference (IACC)*. 916–922. New Delhi, India. February 21-22.

Spencer, J. F. (2008). Using XML to map relationships in hacker forums. *Proceedings of the 46th Annual Southeast Regional Conference.* 487-489. Kennesaw, Georgia. March 28-29.

Sun, Y., Fang, Y., & Lim, K. H. (2012). Understanding sustained participation in transactional virtual communities. *Decision Support Systems*. 53(1), 12–22.

Van Den Berg, G. J. (2001). Duration Models: Specification, Identification, and Multiple Durations. *Handbook of Econometrics.* 5, 3383-3420.

Wang, H., Chung, J. E., Park, N., McLaughlin, M. L., & Fulk, J. (2011). Understanding Online Community Participation: A Technology Acceptance Perspective. *Communication Research*. 39(6), 781–801.

Wang, M.C., & Chang, S.H. (1999). Nonparametric Estimation of a Recurrent Survival Function. *Journal of the American Statistical Association.* 94(445), 146-153.

Wang, Y., Kraut, R., & Levine, J. M. (2012). To Stay or Leave? The Relationship of Emotional and Informational Support to Commitment in Online Health Support Groups. *Proceedings of the ACM Conference on Computer Supported Cooperative Work*. 833–842. Seattle, Washington. February 11-15.

Wei, LJ., Lin, DY., Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American statistical association*. 84(408), 1065-1073.

Yip, M. (2011). An Investigation into Chinese Cybercrime and the Applicability of Social Network Analysis. *ACM Web Science Conference*. 1-4. Koblenz, Germany. June 14-17.

Yip, M., Shadbolt, N., & Webber, C. (2013). Why Forums ? An Empirical Analysis into the Facilitating Factors of Carding Forums. *ACM Web Science*. 453-462. Paris, France. May 2-4.

Zhang, D., Prior, K., & Levene, M. (2012). How long do Wikipedia editors keep active? *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration.* 4-7. Linz, Austria. August 27-29.

Zhang, Q., Wang, F.-Y., Zeng, D., & Wang, T. (2012). Understanding crowd-powered search groups: a social network perspective. *PloS One*. 7(6), e39749.

Zhuge, J., Holz, T., Song, C., Guo, J., and Han, X. (2008). “Studying Malicious Websites and the Underground Economy on the Chinese Web”. *Managing Information Risk and the Economics of Security.* 225–244.