List of Suggested Reviewers or Reviewers Not To Include (optional)

SUGGESTED REVIEWERS:

Not Listed

REVIEWERS NOT TO INCLUDE:

Not Listed

COVER SHEET FOR PROPOSAL TO THE NATIONAL SCIENCE FOUNDATION

PROGRAM ANNOUNCEMENT/SOLICITATION NO./CLOSING DATE/if not in response to a program announcement/solicitation enter NSF 11-1

NSF 12-596 11/30/12

FOR NSF USE ONLY

NSF PROPOSAL NUMBER

1314631

FOR CONSIDERATION BY NSF ORGANIZATION UNIT(S) (Indicate the most specific unit known, i.e. program, division, etc.)

CNS - Secure &Trustworthy Cyberspace

DATE RECEIVED	NUMBER OF COPIES	DIVISION ASSIGNED	FUND CODE	DUNS# (Data Universal Numbering System)	FILE LOCATION
11/30/2012	2	05050000 CNS	8060	806345617	07/27/2018 4:44pm S

EMPLOYER IDENTIFICATION NUMBER (EIN) OR TAXPAYER IDENTIFICATION NUMBER (TIN)

742652689

SHOW PREVIOUS AWARD NO. IF THIS IS
☐ A RENEWAL
☐ AN ACCOMPLISHMENT-BASED RENEWAL

IS THIS PROPOSAL BEING SUBMITTED TO ANOTHER FEDERAL AGENCY? YES ☐ NO ☒ IF YES, LIST ACRONYM(S)

NAME OF ORGANIZATION TO WHICH AWARD SHOULD BE MADE

University of Arizona

AWARDEE ORGANIZATION CODE (IF KNOWN)

0010835000

ADDRESS OF AWARDEE ORGANIZATION, INCLUDING 9 DIGIT ZIP CODE

University of Arizona
888 N Euclid Ave
Tucson, AZ. 857194824

NAME OF PRIMARY PLACE OF PERF

University of Arizona Artificial Intelligence Lab

ADDRESS OF PRIMARY PLACE OF PERF, INCLUDING 9 DIGIT ZIP CODE

University of Arizona Artificial Intelligence Lab
1130 East Helen Street
Tucson ,AZ ,857210108 ,US.

IS AWARDEE ORGANIZATION (Check All That Apply) (See GPG II.C For Definitions)
☐ SMALL BUSINESS
☐ FOR-PROFIT ORGANIZATION
☐ MINORITY BUSINESS
☐ WOMAN-OWNED BUSINESS

☐ IF THIS IS A PRELIMINARY PROPOSAL THEN CHECK HERE

TITLE OF PROPOSED PROJECT SBE TTP: Medium: Securing Cyber Space: Understanding the Cyber Attackers and Attacks via Social Media Analytics

REQUESTED AMOUNT	PROPOSED DURATION (1-60 MONTHS)	REQUESTED STARTING DATE	SHOW RELATED PRELIMINARY PROPOSAL NO. IF APPLICABLE
$ 1,195,722	36 months	09/01/13	

CHECK APPROPRIATE BOX(ES) IF THIS PROPOSAL INCLUDES ANY OF THE ITEMS LISTED BELOW

☐ BEGINNING INVESTIGATOR (GPG I.G.2)
☐ DISCLOSURE OF LOBBYING ACTIVITIES (GPG II.C.1.e)
☐ PROPRIETARY & PRIVILEGED INFORMATION (GPG I.D, II.C.1.d)
☐ HISTORIC PLACES (GPG II.C.2.j)
☐ EAGER* (GPG II.D.2) ☐ RAPID** (GPG II.D.1)
☐ VERTEBRATE ANIMALS (GPG II.D.6) IACUC App. Date ________________
PHS Animal Welfare Assurance Number ________________

☐ HUMAN SUBJECTS (GPG II.D.7) Human Subjects Assurance Number ________________
Exemption Subsection ________ or IRB App. Date ________________
☐ INTERNATIONAL COOPERATIVE ACTIVITIES: COUNTRY/COUNTRIES INVOLVED (GPG II.C.2.j)
____
☐ HIGH RESOLUTION GRAPHICS/OTHER GRAPHICS WHERE EXACT COLOR REPRESENTATION IS REQUIRED FOR PROPER INTERPRETATION (GPG I.G.1)

PI/PD DEPARTMENT

Management Information Systems

PI/PD FAX NUMBER

520-621-2433

PI/PD POSTAL ADDRESS

1130 E Helen St
McClelland Hall, Rm 430Z
Tucson, AZ 85721
United States

NAMES (TYPED)	High Degree	Yr of Degree	Telephone Number	Electronic Mail Address
PI/PD NAME Hsinchun Chen	PhD	1989	520-621-4153	hchen@eller.arizona.edu
CO-PI/PD Salim Hariri	PhD	1986	520-621-4378	hariri@ece.arizona.edu
CO-PI/PD				
CO-PI/PD				
CO-PI/PD				

CERTIFICATION PAGE

Certification for Authorized Organizational Representative or Individual Applicant:

By signing and submitting this proposal, the Authorized Organizational Representative or Individual Applicant is: (1) certifying that statements made herein are true and complete to the best of his/her knowledge; and (2) agreeing to accept the obligation to comply with NSF award terms and conditions if an award is made as a result of this application. Further, the applicant is hereby providing certifications regarding debarment and suspension, drug-free workplace, lobbying activities (see below), responsible conduct of research, nondiscrimination, and flood hazard insurance (when applicable) as set forth in the NSF Proposal & Award Policies & Procedures Guide, Part I: the Grant Proposal Guide (GPG) (NSF 11-1). Willful provision of false information in this application and its supporting documents or in reports required under an ensuing award is a criminal offense (U. S. Code, Title 18, Section 1001).

Conflict of Interest Certification

In addition, if the applicant institution employs more than fifty persons, by electronically signing the NSF Proposal Cover Sheet, the Authorized Organizational Representative of the applicant institution is certifying that the institution has implemented a written and enforced conflict of interest policy that is consistent with the provisions of the NSF Proposal & Award Policies & Procedures Guide, Part II, Award & Administration Guide (AAG) Chapter IV.A; that to the best of his/her knowledge, all financial disclosures required by that conflict of interest policy have been made; and that all identified conflicts of interest will have been satisfactorily managed, reduced or eliminated prior to the institution's expenditure of any funds under the award, in accordance with the institution's conflict of interest policy. Conflicts which cannot be satisfactorily managed, reduced or eliminated must be disclosed to NSF.

Drug Free Work Place Certification

By electronically signing the NSF Proposal Cover Sheet, the Authorized Organizational Representative or Individual Applicant is providing the Drug Free Work Place Certification contained in Exhibit II-3 of the Grant Proposal Guide.

Debarment and Suspension Certification

(If answer "yes", please provide explanation.)

Is the organization or its principals presently debarred, suspended, proposed for debarment, declared ineligible, or voluntarily excluded from covered transactions by any Federal department or agency? Yes □ No ☒

By electronically signing the NSF Proposal Cover Sheet, the Authorized Organizational Representative or Individual Applicant is providing the Debarment and Suspension Certification contained in Exhibit II-4 of the Grant Proposal Guide.

Certification Regarding Lobbying

The following certification is required for an award of a Federal contract, grant, or cooperative agreement exceeding $100,000 and for an award of a Federal loan or a commitment providing for the United States to insure or guarantee a loan exceeding $150,000.

Certification for Contracts, Grants, Loans and Cooperative Agreements

The undersigned certifies, to the best of his or her knowledge and belief, that:
(1) No federal appropriated funds have been paid or will be paid, by or on behalf of the undersigned, to any person for influencing or attempting to influence an officer or employee of any agency, a Member of Congress, an officer or employee of Congress, or an employee of a Member of Congress in connection with the awarding of any federal contract, the making of any Federal grant, the making of any Federal loan, the entering into of any cooperative agreement, and the extension, continuation, renewal, amendment, or modification of any Federal contract, grant, loan, or cooperative agreement.
(2) If any funds other than Federal appropriated funds have been paid or will be paid to any person for influencing or attempting to influence an officer or employee of any agency, a Member of Congress, an officer or employee of Congress, or an employee of a Member of Congress in connection with this Federal contract, grant, loan, or cooperative agreement, the undersigned shall complete and submit Standard Form-LLL, ''Disclosure of Lobbying Activities,'' in accordance with its instructions.
(3) The undersigned shall require that the language of this certification be included in the award documents for all subawards at all tiers including subcontracts, subgrants, and contracts under grants, loans, and cooperative agreements and that all subrecipients shall certify and disclose accordingly.

This certification is a material representation of fact upon which reliance was placed when this transaction was made or entered into. Submission of this certification is a prerequisite for making or entering into this transaction imposed by section 1352, Title 31, U.S. Code. Any person who fails to file the required certification shall be subject to a civil penalty of not less than $10,000 and not more than $100,000 for each such failure.

Certification Regarding Nondiscrimination

By electronically signing the NSF Proposal Cover Sheet, the Authorized Organizational Representative is providing the Certification Regarding Nondiscrimination contained in Exhibit II-6 of the Grant Proposal Guide.

Certification Regarding Flood Hazard Insurance

Two sections of the National Flood Insurance Act of 1968 (42 USC §4012a and §4106) bar Federal agencies from giving financial assistance for acquisition or construction purposes in any area identified by the Federal Emergency Management Agency (FEMA) as having special flood hazards unless the:
(1) community in which that area is located participates in the national flood insurance program; and
(2) building (and any related equipment) is covered by adequate flood insurance.

By electronically signing the NSF Proposal Cover Sheet, the Authorized Organizational Representative or Individual Applicant located in FEMA-designated special flood hazard areas is certifying that adequate flood insurance has been or will be obtained in the following situations:
(1) for NSF grants for the construction of a building or facility, regardless of the dollar amount of the grant; and
(2) for other NSF Grants when more than $25,000 has been budgeted in the proposal for repair, alteration or improvement (construction) of a building or facility.

Certification Regarding Responsible Conduct of Research (RCR)
(This certification is not applicable to proposals for conferences, symposia, and workshops.)

By electronically signing the NSF Proposal Cover Sheet, the Authorized Organizational Representative of the applicant institution is certifying that, in accordance with the NSF Proposal & Award Policies & Procedures Guide, Part II, Award & Administration Guide (AAG) Chapter IV.B., the institution has a plan in place to provide appropriate training and oversight in the responsible and ethical conduct of research to undergraduates, graduate students and postdoctoral researchers who will be supported by NSF to conduct research.
The undersigned shall require that the language of this certification be included in any award documents for all subawards at all tiers.

AUTHORIZED ORGANIZATIONAL REPRESENTATIVE	SIGNATURE	DATE
NAME Mary Gerrow	Electronic Signature	Nov 30 2012 2:01PM

TELEPHONE NUMBER 520-626-6433	ELECTRONIC MAIL ADDRESS maryg@u.arizona.edu	FAX NUMBER 520-626-4130

* EAGER - EArly-concept Grants for Exploratory Research
** RAPID - Grants for Rapid Response Research

SBE TTP: Medium: "Securing Cyber Space: Understanding the Cyber Attackers and Attacks via Social Media Analytics" [PI: Hsinchun Chen, University of Arizona; Salim Hariri, University of Arizona; Tom Holt, Michigan State University].

***Keywords*: cyber security, hacker community analysis, social media analytics, malware attribution, hacker forums and IRC, hacker web research portal**

Cyber Security is an important challenge in today's world, as corporations, governments, and individuals have increasingly become victims of cyber attacks and hacking. Such attacks exploit weaknesses in technical infrastructures and human behavior. Understanding the motivations and incentives of individuals and institutions, both as attackers and defenders, can aid in creating a more secure and trustworthy cyberspace. Developing "methods to model adversaries" is one of the critical but unfulfilled research needs recommended in the "Trustworthy Cyberspace" report by the National Science and Technology Council (2011). Demand for knowledge and tools to conduct cyber crime has grown so widespread that entire international virtual communities and black markets have spawned across the Internet to help facilitate trade between cyber criminals scattered in different parts of the world. Black market participants often offer expertise, snippets of code, or fully-developed applications in exchange for other virtual goods or financial gain. Despite a high relevance to our society, cyber criminal communities and related activities have remained largely unexplored. Existing web social media content presents a rich opportunity for various research opportunities, as virtual communities often maintain large stores of useful data digestible through many forms of computational analyses. The discussions and interactions occurring on such communities allow high-impact, data-driven research; researchers are able to empirically test hypotheses and discover new, unprecedented phenomena. Online anonymity, multilingual challenges, hacker community culture, and the sheer volume of online messages contributed by the diverse cyber citizens all make cyber content analysis an essential yet strenuous research endeavor.

To address these challenges, we are motivated to develop an integrated and scalable computational social media collection and analytics framework in support of the cyber attacker community analysis. Our research team will address important social science research questions of relevance to hacker skills, community structure and ecosystem, contents and artifacts, and cultural differences. We will develop automated hacker forums and IRC (Internet Relay Chat) collection techniques for the international (U.S., Russian and Chinese) hacker communities. We will also deploy scalable honeypot platforms to collect malware in the wild and generate feature representation for malware attribution. The proposed integrated computational framework and the resulting algorithms and software will allow researchers and practitioners to: (1) detect, classify, measure and track the formation, development and spread of topics, ideas, and concepts in cyber attacker social media communication; (2) identify important and influential cyber criminals and their interests, intent, sentiment, and opinions in online discourses; and (3) induce and recognize hacker identities, online profiles/styles, communication genres, and interaction patterns. We will leverage our highly successful computational Dark Web research in terrorism informatics. In this SBE/TTP project, we will develop open source tools, a large longitudinal research testbed, and a web-based Hacker Research Portal in support of cyber attacker community investigation and research. These resources will be introduced to the inter-disciplinary community of social, computing, and cyber security researchers and practitioners. The PI, Dr. Hsinchun Chen, is a leader in security informatics research, with his highly successful projects of COPLINK for crime data mining and Dark Web for open source terrorism social media analytics, both funded by NSF. Our research team consists of experts in hacker community research (Dr. Tom Holt, School of Criminal Justice, with current funding from National Institute of Justice) and cybersecurity and autonomic computing research (Dr. Salim Hariri of Electrical and Computer Engineering Department, with current funding from NSF and Dept. of Defense).

The primary intellectual merit of our research resides in: (a) methodological contributions to SBE by developing automated multilingual content analysis and social media analytics techniques and open source tools to assist SBE scholars in studying strategic communication in critical social media; (b) providing a rich, large-scale, longitudinal, open source collection of hacker community field data to support timely and data-driven SBE research exploration and hypothesis testing; (c) exploring hacker community structure and ecosystem across different international communities. The broader impacts of this research include: (a) transitioning research into practice (TTP) by leveraging our previous research to create a sustainable testbed supporting research modeling cyber security adversaries; (b) assisting researchers and practitioners in detecting interesting and important phenomena in strategic communication in cyber security related social media; (c) supporting analysts and decision makers in understanding the motivation, incentives, dynamics, ecosystems, and trends associated with the cyber attacker community.

TABLE OF CONTENTS

For font size and page formatting specifications, see PAPPG section II.B.2.

Appendix Items:

*Proposers may select any numbering mechanism for the proposal. The entire proposal however, must be paginated. Complete both columns only if the proposal is numbered consecutively.

SBE TTP: Medium: "Securing Cyber Space: Understanding the Cyber Attackers and Attacks via Social Media Analytics" [PI: Hsinchun Chen, University of Arizona; Salim Hariri, University of Arizona; Tom Holt, Michigan State University].

1. Introduction

Cyber security is an important challenge in today's world as corporations, governments, and individuals have increasingly become victims of cyber attacks. Such attacks exploit weaknesses in technical infrastructures and human behavior. Understanding the motivation and incentives of individuals and institutions, both as attackers and defenders, can aid in creating a more secure and trustworthy cyberspace. Instead of taking a reactive approach to infrastructure protection and damage control, proactive cyber security attribution and situational awareness of the sources of attacks will allow researchers and practitioners to better understand the community of cyber attackers (and the closely affiliated hacker community), their profiles and incentives, and the associated vast underground cyber criminal networks and ecosystems. Developing "methods to model adversaries" is one of the critical but unfulfilled research needs recommended in the "Trustworthy Cyberspace" report by the National Science and Technology Council (2011).

Demand for knowledge and tools to conduct cyber crime has grown so widespread that entire international virtual communities and black markets have spawned across the Internet to help facilitate trade between cyber criminals scattered in different parts of the world. Black market participants often offer expertise, snippets of code, or fully-developed applications in exchange for other virtual goods or financial gain. Yet despite their high impact, cyber criminal communities and related activities have remained largely unexplored. Existing web social media content presents a rich opportunity for research and analysis, as virtual communities often maintain large stores of useful data digestible through many forms of computational analyses. The discussions and interactions occurring on such communities allow scientists to conduct high-impact, data-driven research; researchers are able to empirically test various existing hypotheses, and discover new, unprecedented phenomena. Online anonymity, multilingual challenges, hacker community culture, and the sheer volume of online messages contributed by the diverse cyber citizens all make cyber content analysis an essential yet strenuous research endeavor.

To address these challenges, we are motivated to develop an integrated and scalable computational collection and analytics framework in support of the cyber attacker community analysis. Our research team will address important social science research questions of relevance to cyber attacker or hacker skills, community structure and ecosystem, contents and artifacts, and cultural differences. We will develop automated hacker forums and IRC (Internet Relay Chat) collection techniques for the international (US, Russian and Chinese) hacker communities. We will also deploy scalable honeypot platforms to collect malware in the wild and generate feature representation for malware attribution. The proposed integrated computational framework and the resulting algorithms and software will allow social science researchers and security practitioners to: (1) detect, classify, measure and track the formation, development and spread of topics, ideas, and concepts in cyber attacker social media communication; (2) identify important and influential cyber criminals and their interests, intent, sentiment, and opinions in online discourses; and (3) induce and recognize hacker identities, online profiles/styles, communication genres, and interaction patterns.

We will leverage our highly successful computational Dark Web research in terrorism informatics. In this SBE/TTP project, we will develop open source tools, a large longitudinal research testbed, and a web-based Hacker Research Portal to support cyber attacker community investigation and research. These resources will be introduced to the inter-disciplinary community of social science and cyber security researchers and practitioners. The PI, Dr. Hsinchun Chen, is a leader in security informatics research who has directed highly successful NSF-funded projects including COPLINK (public safety and crime data mining) and Dark Web (open source terrorism social media analytics). Our research team consists of experts in hacker community research (Dr. Tom Holt, School of Criminal Justice, with current funding from the National Institute of Justice) and cyber security and autonomic computing research (Dr. Salim Hariri of the Electrical and Computer Engineering Department, with current funding from NSF and the Department of Defense).

2. Securing Cyber Space: Modeling Cyber Security Adversaries

In this section we review the cyber security concerns facing the world and the research opportunities for understanding cyber attackers and attacks.

2.1 Cyber Security Concerns and the Hacker Community

As computers and the Internet become more ubiquitous within our society, the security of networks and information technologies remains a growing concern. Critical infrastructures such as smart power grids and communication networks are facing an increasing number of cyber-based threats, many of which are capable of causing serious disruption of services and permanent harm to systems (Meserve, 2009). Theft of sensitive data has been an emerging problem experienced by a rising number of individuals, businesses, and governments. Technologies enabling cybercriminals to hijack machines from across the Internet have also become more widely available, leading to a rise in botnets and malicious Internet traffic. In the U.S., the Department of Defense has recognized that cyber space is unsecure and a need exists for more cyber security research, education and training (Ackerman, 2011).

Alarmingly, as our society becomes more dependent on cyber infrastructure, the availability of technologies and methods to commit cybercrime have also become more readily available. Moore and Clayton (2009) and Abbasi et al. (Abbasi, Chen, et al., 2010) discovered in their research that cyber attacks and e-commerce phishing sites are becoming more and more common, often performed with the assistance of legitimate tools. For example, cyber criminals will often gain unauthorized access to legitimate web servers in order to host phishing websites and plant malware (Abbasi, Chen et al., 2010).

Demand for knowledge and tools to conduct cybercrime has grown so widespread that entire virtual communities and black markets have been spawned across the Internet to help facilitate trade between cybercriminals. Black market participants often offer expertise, snippets of code, or fully-developed applications in exchange for other virtual goods or financial gain (Holt & Lampke, 2010; Radianti & Gonzalez, 2009). Some even offer themselves as mercenaries, assisting other cybercriminals with projects such as customizing standard malware packages to avoid anti-virus detection or to disable firewalls and other security software suites (Chu et al., 2010; Motoyama, et al., 2011). There seems to be no shortage of willing individuals to assist each other in committing cybercriminal activities. Much collaboration seems to stem from discussions and trade originating in black markets.

With these and other threats recognized, recent research has focused on exploring how and where cybercrime occurs. Past research has observed that cybercriminals often congregate in virtual communities that they create, most commonly in Internet Relay Chat (IRC) networks or online forums (Holt et al., 2012; Holt & Kilger, 2012). These communities appear to be central to many cybercriminal operations; technical expertise, pirated software, code examples, malicious tools, and collaboration can all be found within such communities. Thus, they serve an important role in identifying the nature of cybercrime and should be closely scrutinized.

2.2 Social Science Research on the Hacker Community

Social science research on the hacker community suggests that actors vary in their skill and practical ability to apply hacking techniques against various targets (Holt 2007; Jordan & Taylor 1998; Taylor 1999). It is thought that a top tier of hackers have the complex skills needed to create software and tools to facilitate complex automated attacks against a variety of systems (Holt 2007; Holt & Kilger 2012; Jordan & Taylor 1998). Below this group lies a larger proportion of hackers with less technical skill, but who can apply tools and techniques to engage in attacks with or without authorization from system owners (Holt & Kilger 2012). Finally, the largest proportion of the hacker community has limited knowledge of computers, but learns techniques and acquires resources from the two groups above in order to expand their knowledge further (Holt 2007; Jordan & Taylor 1998; Taylor 1999).

The distribution of skill is thought to be consistent across hacker populations in industrialized nations including China, Russia, and the United States. There is substantive evidence that a number of actors in these nations are involved in the creation and distribution of malware and sophisticated zero day exploits targeting financial institutions and government agencies (Symantec 2012). There is, however, no real empirical research that provides a direct comparison between hackers in these nations to validate these claims. In fact, there is some evidence of differences in regional preferences for tools and communications methods. For instance, Chinese hackers appear to prefer to communicate via QQ or use social networking sites in Baidu while Russian hackers prefer ICQ and Vkontakte (Holt & Kilger 2012). Additionally, there is some evidence of differences in the attack tools preferred by actors by country, such as the use of certain trojans like Pinch in the Russian community (Chu et al., 2010). Thus, there is a need for substantive research to document these variations and test the validity of skill distributions within and across these nations.

The generally small body of empirical research on the hacker community in the social sciences is due in part to difficulties in accessing active hackers through interviews and survey methodologies (Holt 2007; Gilboa 1996; Jordan & Taylor 1998; Taylor 1999). The hacker subculture encourages and values information sharing with other hackers, but stresses secrecy when communicating with outsiders due to prospective legal risks for the admission of criminal hacks (Holt 2007; Jordan & Taylor 1998). As a consequence, many malicious hackers appear to be absent from voluntary survey research or qualitative interviews (Bachmann 2010).

An alternative and very rich data source lies in the use of on-line data from various forms of Computer Mediated Communications (CMCs) where actors discuss and exchange information about serious forms of illegal hacking, such as the creation and sale of malware, management and leasing of active botnets, and the sale of stolen financial information (Chu et al. 2010; Franklin et al., 2007; Decary-Hetu & Dupont 2012; Holt & Lampke 2010; Motoyama et al. 2011; Thomas & Martin 2006). For instance, the content of forums can be used to assess various aspects of the hacker community, such as the distribution of users' skills based on user comments, rank or status labels (Chu et al., 2010; Decary-Hetu & Dupont 2012; Holt, 2007; Holt & Lampke 2010). Additionally, information can be generated on the scope of illicit markets, such as botnet sales and the economic factors affecting the cost of products (Chu et al., 2010; Franklin et al., 2007; Motoyama et al., 2011). In some cases forums can even be used as malware repositories that allow users to download tools freely for their own use (Chu et al., 2010).

While forums provide significant insights into the hacker community, social science researchers who utilize forum data sets typically employ qualitative methods due to the difficulties in automating large scale data collection and quantitative analyses (e.g. Chu et al., 2010; Holt, 2007, 2009; Holt & Lampke, 2010). The ability to leverage techniques to capture large quantities of data in a distributed fashion from multiple forms of CMC could greatly expand the analysis capability of researchers and engender the identification of hidden networks of actors as well as variations in hacker communities across place and over time. In addition, this could be used to address fundamental questions related to computer hacking and cybercrime generally, such as the prevalence and costs of stolen data and malware services (Chu et al. 2010; Franklin et al. 2007; Holt & Lampke 2010; Thomas & Martin 2006), the flow of information between participants in on-line communities (Decary-Hetu & Dupont 2012; Decary-Hetu, Morselli, & Leman-Langlois 2012; Holt 2012; Holt et al. 2008; Motoyama et al. 2011), and the factors that affect individual reputation and social status within large scale communities (Holt 2009; Holt & Lampke 2010; Mann & Sutton 1998; Motoyama et al. 2011).

Based on our review, it appears there are many reasons to and opportunities for extending cyber security research by combining social science methodology, computational techniques, and security analysis. There are numerous critical and emerging questions social science and security researchers should ask themselves when considering pursuing new studies:

- What types of hierarchies and skill composition exist within hacker communities? Are meritocracies and reputation-based social structures the only ones that exist?
- Do hacking cultures vary across nationalities? How are English, Chinese, and Russian hacking groups different and similar? In which geopolitical regions is cyber security a growing problem?
- How do hacking groups evolve over time? What can temporal data analysis reveal in this context?
- How else do hackers make use of peer-to-peer networks? What types of content are hidden within such networks, such as Tor and I2P?
- Many studies appear to utilize manual techniques to collect and analyze data, perhaps due to crawling counting-measures that many hacker communities employ. How can we develop more useful automated or semi-automated techniques for this application context?

Dr. Holt, who is a leading expert in hacker community research will lead the social science research (SaTC SBE Perspective), with the aid of the computational tools and network infrastructure developed by Drs. Chen and Hariri. Dr. Holt's research will increase understanding of the motivation, incentives, dynamics, ecosystems, and trends associated with the cyber attacker community. Our proposed computational framework and techniques for hacker community collection and analytics will provide significant methodological contributions to SBE by making available automated multilingual content analysis and social media analytics techniques and open source tools to assist SBE scholars in studying strategic communication in critical social media. Our research will also provide a rich, large-scale, longitudinal, open source collection of hacker community field data to support timely and data-driven SBE research exploration and hypothesis testing.

3. Hacker Community Analysis: Overview

In this section, we review existing hacker community literature and identify several areas for future research. Specifically, we summarize past work focused on identifying, collecting, and analyzing hacker communities. We include implementation details of research methodologies when possible. Multiple research gaps are identified and discussed for possible exploration and research extensions.

3.1 Identification Strategies

Hacker forums and IRC channels are the two major avenues for identifying the international hacker community. Much hacker community domain knowledge is required for this task.

3.1.1 Hacker Forums

A review of related literature reveals that the majority of researchers often refer to third parties for information about hacker forums. Others may conduct keyword searches in an attempt to find forums on their own, or scrutinize known forums for hyperlinks and references to unknown communities.

Public Sources: The simplest method used to identify hacker forums is to refer to third-party sources for data and information. Motoyama et al (2011) was able to acquire the full SQL server data dumps of multiple English and German hacker forums. Other researchers look to third parties for information on hacker forums, rather than raw data, as it is generally more available to the security community. Radianti et al (2007) found a hacker forum cited in cybercrime covered in news media and was able to directly visit the forum. Similarly, other researchers have utilized other non-traditional sources, such as the Google Safe Browsing API, to acquire data on malicious, cybercrime related websites (Cova et al., 2010).

Keyword Searches: Another method commonly used by researchers to identify hacker forums is to conduct a series of keyword searches. For example, Holt & Lampke (2010) crafted the keyword search “carding dump purchase sale cvv” to identify hacker black markets where stolen credit card information is sold. They were able to create the keyword based on their domain knowledge and explorations in past studies. This approach seems to be common in many similar studies on hacker forums (Fallman et al., 2010; Holt, 2010).

Link Identification: It is common practice to scrutinize known forums for links to other hacker forums and communities. Many studies found that hacker forum participants often cite or refer to other hacker communities (Radianti et al., 2009b; Fallman et al., 2010; Holt et al., 2012). Thus, a snowball approach using one forum to identify many others could be promising.

3.1.2 IRC Channels

Cyber security research conducted on IRC channels often focuses on both hacker communities and botnets. Hacker community IRC research is similar to forum studies, as researchers attempt to locate hacker discussions and cybercriminal black markets. Conversely, botnet related research is more focused on identifying botnet command & control (C&C) channels, which are chat rooms often used by cybercriminals to control large groups of malware-infected computers with malicious intent.

Hacker IRC Identification: As stated in aforementioned studies, the participants of a hacker community will often cite and provide URLs of other hacker communities. This includes IRC channels (Radianti et al., 2009b; Radianti, 2010). Scrutinizing known communities is key to finding new IRC channels to study. Some researchers make use of special IRC commands and functionality to find new hacker IRC channels. Some IRC servers support the IRC *list* command, which allows a user to query an IRC server for a list of existing channels (Fallmann et al., 2010).

Botnet C&C Identification: A different research focus for some security researchers is to identify botnet command and control (C&C) channels. These channels are used by cybercriminal “botmasters” to give commands to collections of malware-infected computers that covertly join the IRC channel and wait for instruction. C&C identification techniques generally attempt to observe common botnet behaviors across multiple IRC channels, or utilize a “honeypot” approach, which involves the voluntary capture of malware and analysis of execution behaviors in attempt to identify hidden network connection attempts to botnet C&C channels (Chu et al., 2010). Many open source honeypot clients (“honeyclients”) exist which can be used by researchers to quickly set-up a honeypot environment on any normal computer. Both Mielke & Chen (2008) and Wang et al. (2009) report that zombie computers connected to a botnet are often times nicknamed in a standardized, sequential order for the botnet operator's convenience. Research suggests that broad monitoring of multiple channels and observations of a regularly migrating population of users between channels may reveal botnet operations (Tsai et al., 2011).

3.2 Collection Procedures

3.2.1 Hacker Forums

Similar to forum identification methods, collection procedures in reviewed literature take both manual and automated approaches. Manual approaches are the simplest and most accessible for researchers, but are time-consuming and may not yield complete coverage of a forum. Automated techniques are more difficult to employ, but are much quicker and can collect much more data in a given time. However, many hacker communities utilize anti-crawling measures in order to protect themselves from researchers and law enforcement; these counter-measures must be circumvented.

Manual Collection: Most of the reviewed literature resorted to manual collection or simple observation of live hacker communities. Some researchers simply observe live forums without attempting any sort of collection (Holt, 2010; Yip, 2011). Holt (2010) states this method is valuable because it helps avoid researcher contamination of studies. Yip (2011) follows the same methodology by manually browsing Chinese hacker communities and draws conclusions after observations. Other researchers have manually downloaded observed threads in order to build a collection and preserve information (Radianti et al., 2007; Radianti et al., 2009b; Holt & Lampke, 2010; Radianti, 2010). It is important to store data intended for research offline, as hacker forums may sometimes spontaneously disappear due to hosting provider changes, attacks from other hackers, or simply to reduce visibility of participants (Radianti, 2010).

Automated Collection: Other researchers utilize more automated data collection methods. For example, Benjamin & Chen (2012) used a web crawler to automatically collect all publicly available content from America and Chinese hacker forums. In some cases, it may be necessary to use proxy servers and other identity obfuscation software to evade detection and attacks from other hackers (Goel, 2011). Crawling counter-measures are sometimes put in place by hacker communities, such as limiting bandwidth consumption or page views over a certain period of time. Configuring crawlers to make use of proxy servers for multiple IP addresses can be used to circumvent such limitations (Fallman et al., 2010; Fu et al., 2010). Others have paired a web crawler with a honeypot system that can detect when a web page attempts to execute malicious code through browser vulnerabilities. This helps with automated detection of additional cybercriminal web contents when using automated crawlers (Zhuge et al., 2008, Cova et al., 2010).

3.2.2 IRC Channels

There are two common techniques used to collect IRC chat data; both involve logging of real-time chat. Some researchers visit IRC channels and log data manually or use automated bots from within channels. Others who have captured malware with honeypots can simply log network packets related to the IRC protocol. In either case, researchers can assemble complete records of chat data and IRC channel activity.

In-channel Logging: If a researcher gains access to an IRC channel, they can simply log all IRC channel activities and chat interactions. Some researchers prefer manual collection and coding of observed data, as it is the easiest way to build a data set (Radianti et al., 2009b). Others use software bots to automatically join IRC channels and log data in real-time; automated collection can often lead to more complete coverage of a channel, with logged data sometimes spanning over several months (Franklin et al., 2007; Fallmann et al., 2010). Some research also exploits various IRC commands to collect additional data (Fallmann et al., 2010).

Honeypot Observation: Researchers using honeypots can simply monitor network traffic related to the IRC protocol (Lu & Ghorbani, 2008; Mielke & Chen, 2008; Lu et al., 2009; Paxton et al., 2011). In a honeypot environment, researchers can observe all inbound and outgoing network traffic. IRC protocol packets contain information such as who connected to and left a channel, what connected users are saying in their messages, and all other visible activity one would normally observe as if they were actually logged into a channel; the data encapsulated in network packets would effectively reconstruct the actual activity occurring in IRC channels. Honeypot implementation for IRC collection will be discussed in more detail in the following subsection.

Automated hacker community forum and IRC channel collection are critically needed for SBE researchers to better model and understand the hacker community structure and ecosystem.

3.3 Analytical Methods

After hacker community content is collected, it can be analyzed using various methodologies. In particular, content analyses, network analyses, and botnet C&C discovery appear to be the most common goals of analysis. Both manual and automated techniques are commonly used. Automated, multilingual social media analytics techniques and tools can provide significant methodological contribution for hacker community SBE research.

3.3.1 Content Analysis

Content analysis studies often look to document existing content and activity occurring within hacker communities, including forums and IRC channels. There appears to be little difference in how content analysis is performed in either type of community after data has been collected. In the literature we reviewed, these studies tend to employ manual collection and analytical methods, and generally conduct simple counting and statistical work when making quantitative analyses. Computer-aided keyword and topic extraction will assist in content analysis of hacker community forums.

Manually conducted content analyses are often performed by scrutinizing manually coded hacker community contents or by simply observing live systems, in both the forum and IRC context (Holt & Lampke, 2010; Radianti, 2010; Imperva, 2012). For example, after visiting an English-language forum serving as a hacker black market and coding sampled contents, Holt & Lampke (2010) were able to make several conclusions regarding black market activities. In particular, they observed an active stolen data market where trades between forum participants were largely facilitated by hacker reputation and trustworthiness. In fact, the importance of reputation in such communities is a common finding in reviewed literature; multiple researchers conducting separate analyses arrived at the same conclusions regarding the role reputation plays in hacker communities (Radianti et al., 2009b; Holt & Lampke, 2010; Motoyama et al., 2011).

3.3.2 Social Network Analysis

Social network analyses often aim to observe the relationships between participants within a network, as well as the structure of a given network. In the hacker community context, understanding more about how the participants of such communities interact with one another and what social structures exist among hackers would be useful. Both manual observations and automated techniques have been utilized to better understand the relationships between hacker community participants.

Additionally, some researchers are interested in observing how hackers behave across multiple communities. Motoyama et al. (2011) was able to track hackers across multiple forums by looking for e-mail addresses used for forum participation. They found significant membership overlap over several hacker communities. Holt et al. (2012) also followed this methodology, but searched by username instead of e-mail addresses. By looking for the prominence and involvement of hackers across multiple communities, Holt et al. were able to create a perceived threat level for each hacker.

3.3.3 Peer-to-Peer Technologies

In recent years, many hackers have been making use of peer-to-peer (P2P) communication technologies for their cybercriminal operations (Lu et al., 2009; Fu et al., 2010). Such technologies include the popular Tor anonymization network, and also the similar, yet less popular I2P network. Unfortunately, there appears to be a significant lack of existing papers attempting to explore hacker communities buried within such networks.

Peer-to-peer communication technologies help hackers eliminate problems stemming from a single failure point common in traditional client-server botnets (Lu et al., 2009; Wang, 2012). In traditional botnets, if the server fails, the botmaster loses control of their botnet. Hackers are moving towards botnets built on top of HTTP (infected webservers), where any webserver can become the command and control used to instruct other bots. More hackers are also taking advantage of creating P2P-based botnets, giving them more flexibility and control than if they were to rely on established protocols such as IRC. Furthermore, peer-to-peer anonymization technologies, such as the Tor network, can serve as a conduit to Internet communication channels for cyber criminals. Hackers can easily obfuscate their identity (and thus protect their anonymity) by routing their network packets through such networks (Fu et al., 2010). Although online forums and IRC are still the major avenues for hacker community building, in our research we plan to explore emerging issues related to hacker P2P networks.

4. Proposed Research

Based on our literature review we present our proposed research framework and summarize the key collection and analytics approaches for understanding the hacker community.

4.1 Research Framework

The proposed research framework is shown in Figure 1.

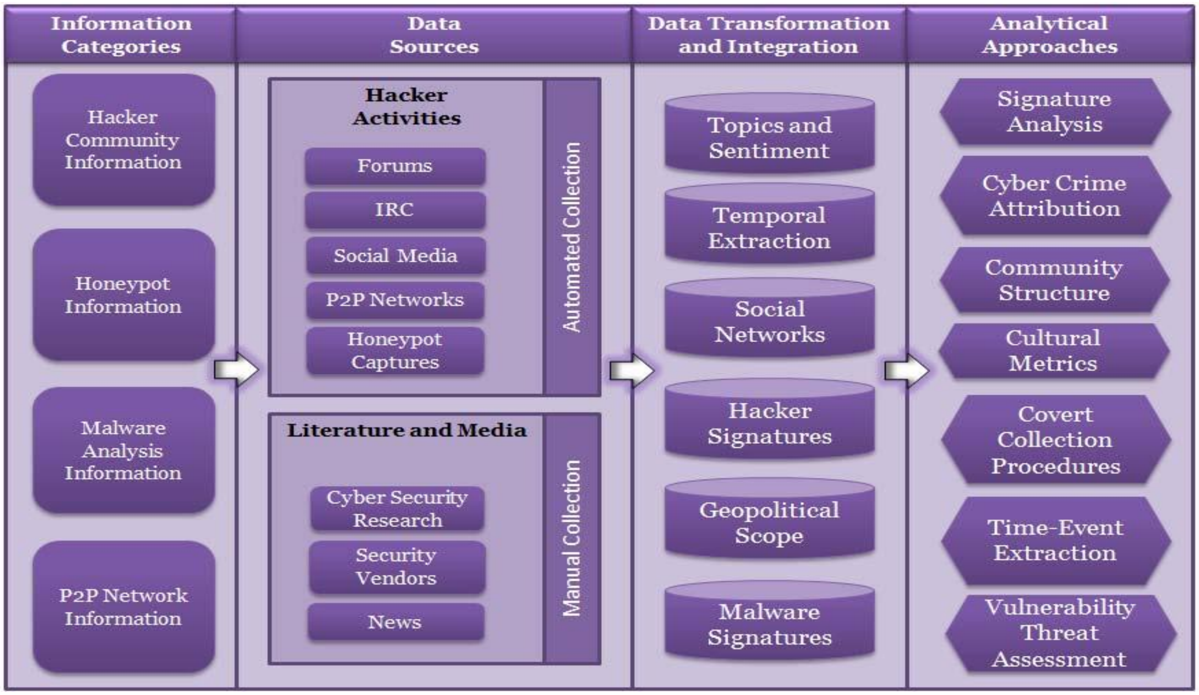


Figure 1. Proposed Research Framework

We start by identifying several important categories of information necessary for cyber security investigation. Our research will focus on hacker community information (the actors) and honeypot information (malware output), to be supplemented by further malware analysis and selected emerging P2P network information. Then, data sources for each information category are identified and collected to assist in our hacker community analysis. We plan to develop automated techniques for collecting major U.S., Russian and Chinese hacker forums and IRC contents. We will also explore additional social media, P2P networks, and honeypot captures. In addition, manual collection methods will be deployed for emerging cyber security research and news and other security vectors based on our social science and security analysis research questions. Next, collected data is scrubbed and transformed for usage in various analyses. We will leverage our extensive experience in social media analytics for from our NSF funded Dark Web research (more on this later) for topics and sentiment, temporal extraction, and social networks. Additional hacker and malware signatures (e.g., programming languages used, attack targets, source code used) and other geopolitical information (e.g., locations) will be identified to assist in hacker community analysis. Lastly, numerous types of social science and security analyses will allow us to gain new perspectives and knowledge from the acquired data: hacker signature analysis (profile), cyber crime attribution (linking malware to actors), hacker community structure (and skills), and cultural metrics identification (for US, Russian, and Chinese groups). In addition, our research will help with time-event extraction, covert hacker community content collection, and vulnerability threat assessment.

4.2 Identification and Collection: Forums, Honeypots, and IRC Channels

In this section we present our key technologies for collection and analytics based on our past and ongoing research from our team of experts in criminology, computational science, and network research. Our research aims to provide significant methodological contributions to hacker community SBE research by introducing tools for capturing valuable hacker community field data and assisting in more automated content and social network analysis of rich multilingual hacker communication.

4.2.1 Hacker Forums Collection

The challenges in collecting and managing large-scale, longitudinal hacker social media contents from various international data sources are numerous. However, our project will leverage many of the techniques and methods developed in the course of our Dark Web work. Dark Web is an internationally recognized long-term terrorism research program that examines international terrorism and extremism via a computational, data-centric approach (Chen, 2012). It is being funded in part by the National Science Foundation and the Defense Thread Reduction Agency. We collect web content generated by international terrorist and extremist groups, including web sites, forums, chat rooms, blogs, social networking sites, videos, virtual worlds, etc. (Chen, et al., 2004; Zhou et al., 2006; Fu et al., 2010; Chen et al., 2011b). We have also developed various multilingual data mining, text mining, and web mining techniques to perform link analysis (Zhou, et al., 2005), sentiment analysis (Abbasi and Chen, 2008a, 2008b) authorship analysis (Abbasi & Chen, 2005; 2006;2008a; 2008b; Zheng et al., 2006) and video analysis (Huang et al., 2010) in our research. Selected forum spidering and related technical components that could be adopted for hacker forums collection are shown in Figure 2.

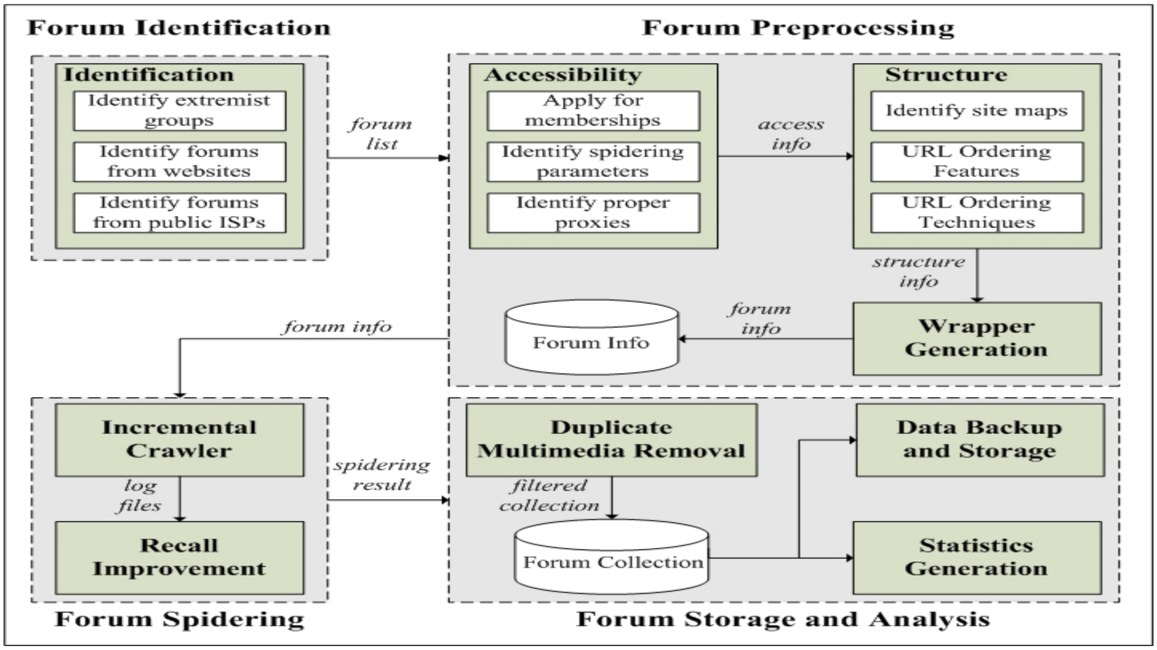


Figure 2. Dark Web Forums Spidering (Fu, Abbasi & Chen, 2010); to be adopted for Hacker Forum Spidering

The spidering tools were adapted based on the SpidersRUs Digital Library Toolkit developed in earlier projects (Qin et al., 2004; Chau et al., 2005; Chau and Xu, 2006). The system consists of four components: Forum Identification, Forum Preprocessing, Forum Spidering, and Forum Storage and Analysis. Relevant extremist forums are identified by collaborating terrorism research experts. Forum preprocessing design takes great care in identifying spidering parameters (e.g., number of bots used to connect to a target forum and the connection timeout setting) to avoid overloading the forum server. We have also developed sophisticated opaque proxies to disguise our bots based on network reliability and latency (Fu, Abbasi & Chen, 2010). Major open-source forum hosting software (e.g., vBulletin) are carefully studied to identify forum-specific syntax, site maps, and page URL ordering patterns. After the initial batch mode of spidering for all historical forum contents, incremental spidering is incorporated for future updating. Both textual and multimedia attachments collected are stored in the Dark Web database using a relational database design. Our systematic spidering approach has been proven successful and invaluable in developing our large-scale and longitudinal Dark Web testbed.

We believe much of our Dark Web forum spidering technology can be adopted for hacker forums collection (as a requirement of SaTC Transition to Practice, TTP). Dr. Tom Holt, who is one of the leading experts in hacker community research, will lead the effort in identifying key U.S., Russian and Chinese hacker forums in various web sites and public ISPs. For selected forums that require membership, Dr. Holt's group will gain access through their extensive contacts. The AI Lab, headed by

Dr. Chen, will lead the effort in adapting Dark Web spidering tools for hacker community forums. Careful spidering and proxy setting will be developed to avoid detection and bypass anti-crawling mechanisms. Forum contents collected will be stored in a local cached relational database for research purposes. Selected malware source code and attachments will be collected and analyzed with the help of Dr. Hariri's cyber security research group.

4.2.2 Honeypots and IRC Channels Collection

In addition to hacker forums collection, there is also value in applying data collected through honeypot technologies based on their ability to mimic an unpatched vulnerable end user computer. Honeypots are systems that are configured to simulate computer environments with software vulnerabilities; the idea is to have wild malware exploit honeypot vulnerabilities so that the malware can be captured and studied in a sandboxed environment. All code execution, system changes, and network traffic are tracked and logged within a honeypot, letting security researchers understand the nature of some particular malware (Mielke & Chen, 2008; Zhu et al., 2008). For this project we propose combining the analysis of forum data with analyses of active malware acquired through the use of various collection devices. Once acquired, we will then infect honeypot systems managed by the research team and analyze the traffic and behavior of malware through the use of honeywall software (The Honeynet Project, 2003). Honeypot approaches towards C&C identification will be implemented. By observing outbound network connections attempted by captured malware, researchers may potentially reveal botnet C&C channels and other hacker-related web addresses.

Based on the Honeynet Project of Dr. Holt (2012) and Dr. Hariri's extensive cyber security research (Luo, Hariri et al., 2010; Viswanathan, Hariri, et al., 2011), we plan to develop a honeypot environment for collecting hacker IRC data. Figure 3 represents an overview of the proposed environment, which consists of the following components. *IRC Server* will support the interaction with the rest of the IRC network and also log all IRC messages. *Autonomic Bot Generator* will be responsible for generating bots that provide interaction mechanism with the environment. The bot behaviors, types, and number are enforced based on a preset policy. *Autonomic Monitoring* is responsible for picking up all the IRC packets, and it will have Network policies that define which ports to monitor and when. *IRC Message Extraction* will extract IRC messages from IRC packets, and categorize them into different IRC message types. *File Extraction* will detect file transfer and extract files from communications; *Conversation Historian* will build conversations from the IRC messages and storing those for further analysis. *Feature Extraction and Reduction* will extract all the features needed to perform the analysis from the IRC messages. *Malware Analysis* and *Social Media Analyzer* are the analytical engines for the environment; one focusing on detecting malware and the other on detecting, classifying, measuring, and tracking the formation, development, and spread of topics, ideas, and concepts in cyber attacker social media communication (more details to follow in the next subsection).

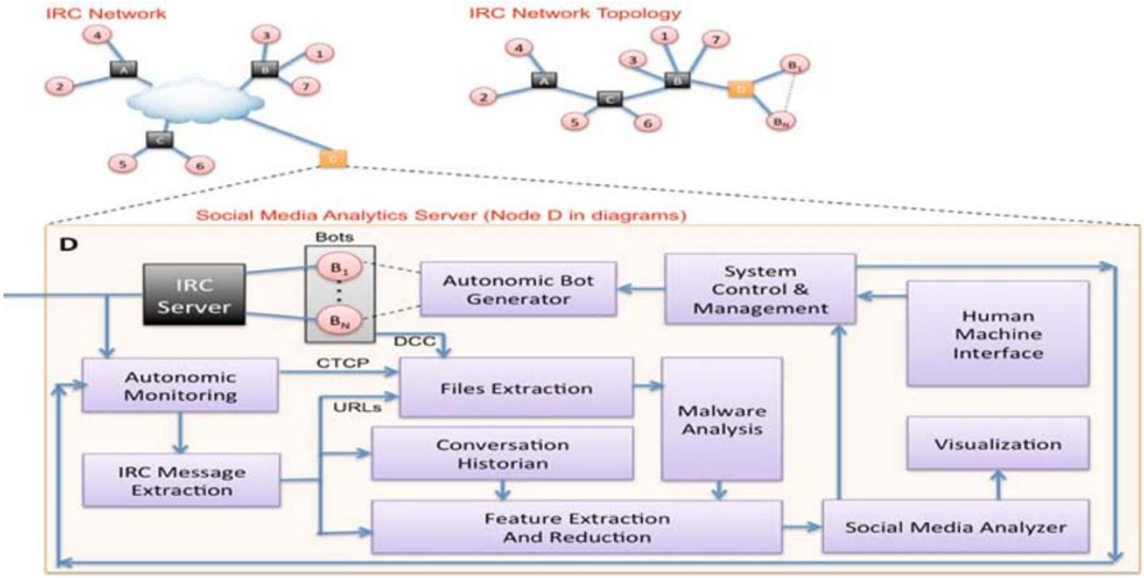


Figure 3. Honeypot IRC Collection and Analytics Environment

The environment will be supported by three additional modules on *Visualization*, *Human Machine Interaction*, and *System Control and Management*. A similar network infrastructure was developed previously for intrusion detection research at Dr. Hariri's NSF Cloud and Autonomic Computing Center.

4.3 Social Media Analytics for CMC

Computer mediated communication (CMC) has grown tremendously due to the explosive growth of the Internet and social media. Text-based modes of CMC include email, listservs, web forums, chat rooms, instant messaging, video-conferencing, blogs, social networking sites, and media-sharing technology. These CMC modes have redefined the fabric of organizational culture and social/political interaction. CMC text analytics is the analysis of text-based modes of CMC in various social media. There is a need for analysis techniques that can evaluate, summarize, and present multilingual CMC text in a holistic and comprehensive manner. CMC text in social media is rich in social cues including emotions, opinions, style, and genres. Improved CMC text analysis capabilities based on rich and comprehensive text representations are necessary. In this research we propose a computational framework and the associated techniques for multilingual CMC text analytics and visualization, with a focus on English, Russian, and Chinese (languages commonly used by major international hacker and cyber criminal groups).

An important characteristic of CMC is the language complexities it introduces as compared to other forms of text. Effective analysis of CMC text entails the utilization of a language theory that can provide representational guidelines. Grounded in Functional Linguistics, Systemic Functional Linguistic Theory (SFLT) provides an appropriate mechanism for representing CMC text information (Halliday, 2004). SFLT states that language has three meta-functions: ideational, interpersonal, and textual. The three meta-functions are intended to provide a comprehensive functional representation of language meaning by encompassing the physical, mental, and social elements of language.

Figure 4 shows a computational framework diagram for our proposed research. It involves the use of a rich set of features capable of representing various forms of information found in multilingual CMC text.

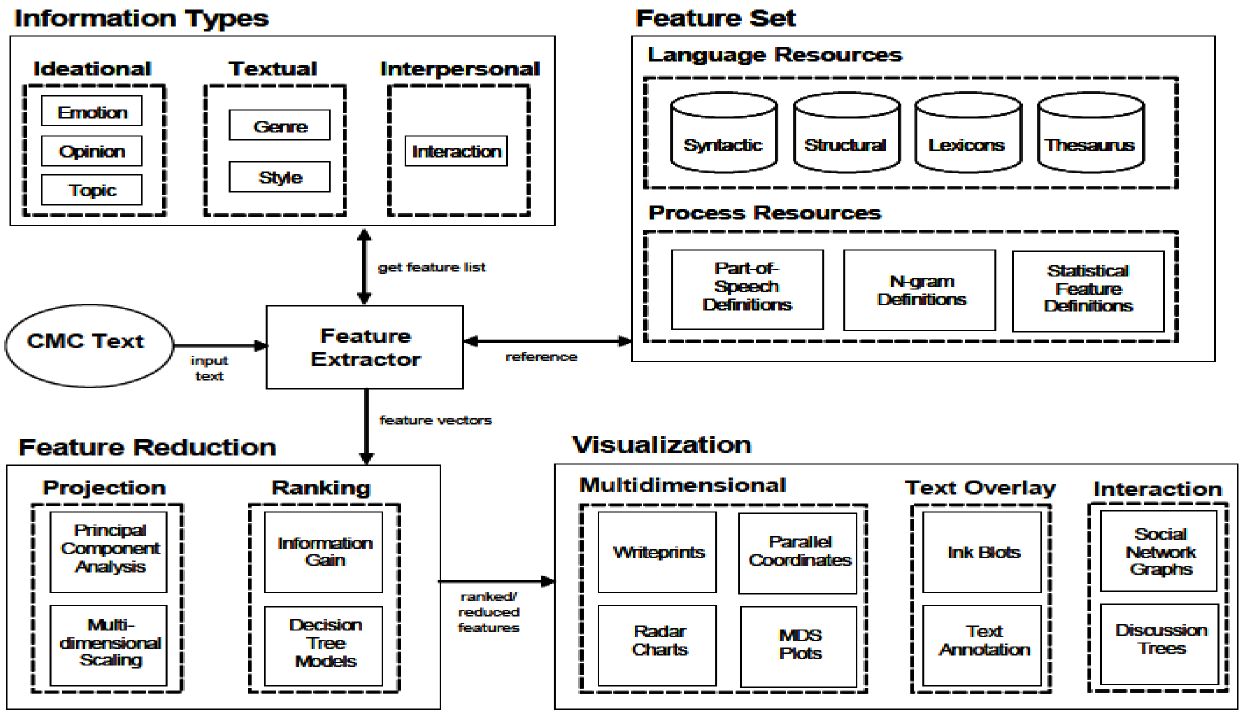


Figure 4. Proposed System Design for Social Media Analytics for CMC

We propose the use of feature selection and visualization methods to transform large, noisy feature spaces into focused and refined target representations and message signatures (profiles) (Abbasi & Chen, 2008a 2008b; Abassi et al., 2010; Chen 2010, 2012). Our design supports several information types for representing the ideational, textual, and interpersonal meta-functions, including: emotion, opinion, topic, genre, style, and interaction. In order to capture such a wide array of information types, several language and processing resources have been incorporated into the design, including various syntactic, structural, lexical, and thesaurus resources and state-of-the-art statistical NLP techniques such as part-of-speech definitions, n-gram definitions, and other statistical feature definitions. Robust statistical and data mining feature reduction techniques are proposed to reduce the feature complexity and to extract salient member CMC characteristics, including Principal Component Analysis and Multi-Dimensional Analysis from statistical analysis, and Information Gain and Decision Tree Models from data mining. After feature reduction, selected multidimensional, text overlay and interaction visualization techniques are proposed to highlight key CMC features that may represent concepts, ideas, intent, identities, style, genres, and opinions of the hacker community.

Our proposed research will benefit greatly from our previous stylometric analysis research in English, Arabic, and Chinese web forums and newsgroups. In this previous work we extended traditional *lexical features* (e.g., words per sentence, word length distribution, vocabulary richness) and *syntactic features* (e.g., punctuations, function words) to include *structural features* which deal with the CMC text's organization and layout (e.g., greetings, signatures, font size, font color) and *content-specific features* which are words that are important within a specific topic domain (e.g., Jihad, crusader, heaven). Using the four feature sets and selected classification techniques (C4.5 and Support Vector Machines), we were able to uniquely identify anonymous authors with an accuracy level of 97% for English messages and 94.83% for Arabic messages. We have obtained similar results for Chinese web forums. In light of the prevalence of English, Russian and Chinese contents in various cyber hacker communities, from general security discussions to possible cyber terrorism and cyber warfare materials, we believe our prior and ongoing multilingual social media analytics research will provide a sound academic foundation and useful insights into these communities. The ability to leverage our previous research will allow us to accomplish what may seem to be an ambitious amount of work.

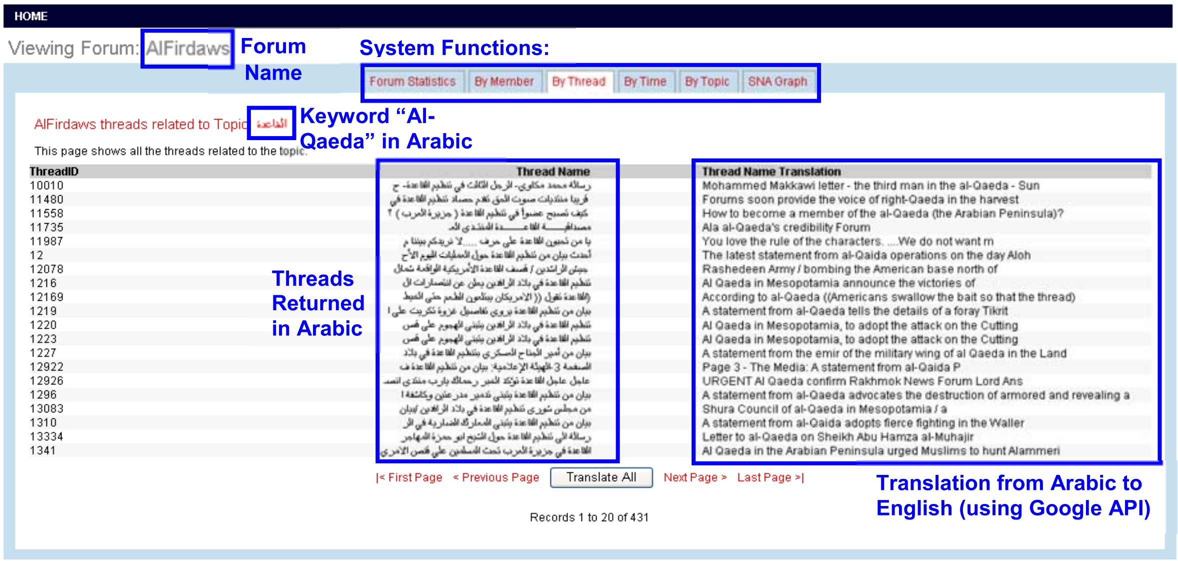
Evaluation of the proposed features, selection, and visualization methods will entail assessing their ability to represent CMC text. For the proposed CMC text analysis design framework, a suitable implementation must incorporate features, feature selection, and visualization techniques capable of effectively characterizing and discriminating information types used to represent the ideational, textual, and interpersonal meta-functions. Prior CMC systems used application examples or case studies to illustrate their systems' data characterization capabilities. In contrast, the effectiveness of data discrimination is generally assessed using rigorous text categorization experiments and metrics (e.g., accessing accuracy, error rate, recall, precision) for various information types. We intend to use similar methods to evaluate our features' and techniques' ability to represent CMC text: case studies and user evaluations for data characterization and CMC text categorization experiments for data discrimination (Chen, 2012).

5. Developing a Research Test Bed: The Hacker Web Research Portal

Based on our significant past Dark Web research and the proposed social media analytics research for the online hacker community, our project will address the challenge of moving from research to capability (Transition to Practice, TTP). We will leverage successful results from our previous and current NSF funded basic research and focus on later-stage activities, including applied research, development, prototyping, testing, and experimental deployment to help address critical cyber security concerns of relevance to the hacker community. We plan to create a large research testbed, known as the Hacker Web Research Portal, and a research infrastructure for use by computer and information scientists as well as social and political scientists studying a wide range of computational problems and social and organizational phenomena of relevance to the international hacker community. The archive will ultimately comprise testbed data containing millions of multilingual social media contents (e.g., forums, newsgroups, botnet C&C channels) collected from major U.S., Russian and Chinese hacker virtual communities. A methodology and set of spidering (collection building) tools for time-based automated capture of relevant social media and multimedia resources will be developed through this project; this approach will then support automated monthly updates of the entire collection. In addition, the infrastructure will include tools supporting search, browse, summary and analysis capabilities. Our proposed social media analytics functionalities will be developed as open source tools to assist researchers and practitioners in identifying critical topics, opinions, styles, genres, and interactions of relevance to the international hacker community.

5.1 Leveraging Dark Web Research (TTP): Developing the Hacker Web Research Portal

Our research will leverage our internationally acclaimed Dark Web terrorism informatics research. Funded by the NSF and DOD since 2003, the Dark Web archive and analytics project has been collecting and analyzing web pages, blogs, multimedia files, and forum postings pertaining to U.S. extremist groups (in English) and international Jihadist (in Arabic) terrorist groups from more than 10,000 web sites (Zhou, et al., 2005; Qin, et al., 2007; Chen et al., 2008a; 2008b; Abbasi, Chen, et al., 2010). It is believed to be the largest such collection in the academic world (holding open source web contents only) and has become an invaluable longitudinal academic resource for cyber crime, extremism, and terrorism research. As a research and analytics tool, the Dark Web Forum Portal (Figure 5) is currently in use by more than 500 information/computer scientists, social/political scientists, and military/intelligence analysts (Zhang et al., 2009).



AlFirdaws threads related to Topic القاعدة

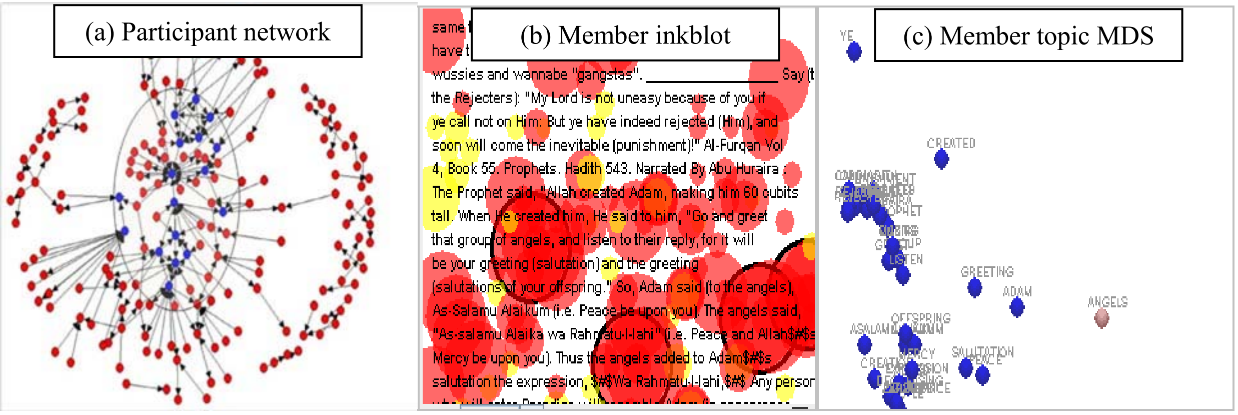
This page shows all the threads related to the topic.

ThreadID	Thread Name	Thread Name Translation
10010	رسالة محمد مكاوي- الرجل الثالث في تنظيم القاعدة- ح	Mohammed Makkawi letter - the third man in the al-Qaeda - Sun
11480	قريبا منتديات صوت الحق تقدم حصاد تنظيم القاعدة في	Forums soon provide the voice of right-Qaeda in the harvest
11558	كيف تصبح عضواً في تنظيم القاعدة (جزيرة العرب) ؟	How to become a member of the al-Qaeda (the Arabian Peninsula)?
11735	مصداقيـــــة القاعـــــدة المنتدى الع	Ala al-Qaeda's credibility Forum
11987	يا من تحبون القاعدة على حرف ......لا نريدكم بيننا م	You love the rule of the characters. ....We do not want m
12	أحدث بيان من تنظيم القاعدة حول العمليات اليوم الأح	The latest statement from al-Qaida operations on the day Aloh
12078	جيش الراشدين / قصف القاعدة الأمريكية الواقعة شمال	Rashedeen Army / bombing the American base north of
1216	تنظيم القاعدة في بلاد الرافدين يعلن عن انتصارات ال	Al Qaeda in Mesopotamia announce the victories of
12169	القاعدة تقول ((الامريكان يبتلعون الطعم حتى الخيط)	According to al-Qaeda ((Americans swallow the bait so that the thread)
1219	بيان من تنظيم القاعدة يروي تفاصيل غزوة تكريت على ا	A statement from al-Qaeda tells the details of a foray Tikrit
1220	تنظيم القاعدة في بلاد الرافدين يتبنى الهجوم على قص	Al Qaeda in Mesopotamia, to adopt the attack on the Cutting
1223	تنظيم القاعدة في بلاد الرافدين يتبنى الهجوم على قص	Al Qaeda in Mesopotamia, to adopt the attack on the Cutting
1227	بيان من أمير الجناح العسكري بتنظيم القاعدة في بلاد	A statement from the emir of the military wing of al Qaeda in the Land
12922	الصفحة 3- الهيئة الإعلامية: بيان من تنظيم القاعدة ف	Page 3 - The Media: A statement from al-Qaida P
12926	عاجل عاجل القاعدة تؤكد الخبر رحمالك يارب منتدى انص	URGENT Al Qaeda confirm Rakhmok News Forum Lord Ans
1296	بيان من تنظيم القاعدة يتبنى تدمير مدرعتين وكاشفة ا	A statement from al-Qaeda advocates the destruction of armored and revealing a
13083	من مجلس شورى تنظيم القاعدة في بلاد الرافدين /بيان	Shura Council of al-Qaeda in Mesopotamia / a
1310	بيان من تنظيم القاعدة يتبنى المعارك الضارية في الر	A statement from al-Qaida adopts fierce fighting in the Waller
13334	رسالة إلى تنظيم القاعدة حول الشيخ ابو حمزة المهاجر	Letter to al-Qaeda on Sheikh Abu Hamza al-Muhajir
1341	القاعدة في جزيرة العرب تحث المسلمين على قنص الامري	Al Qaeda in the Arabian Peninsula urged Muslims to hunt Alammeri

Records 1 to 20 of 431

Figure 5. Forum Search Using the Keyword "Al-Qaeda" in the Al Firdaws Forum

Currently, the portal contains about 13M total multilingual messages (in English, Arabic, German, and Russian) from approximately 350,000 online members. As shown in the top display panel in Figure 5, the portal currently offers four basic types of functions: browsing and searching of single and multiple forums (by member, thread, time, and topic), forum statistics, multilingual translation (in any supported language, using Google Translation API), and social network graph visualization (using JUNG API). The system supports searching, browsing, social network visualization, and multilingual translation of terrorism social media contents.



same t
have t
wussies and wannabe "gangstas". ________________ Say (t
the Rejecters): "My Lord is not uneasy because of you if
ye call not on Him: But ye have indeed rejected (Him), and
soon will come the inevitable (punishment)!" Al-Furqan Vol
4, Book 55. Prophets. Hadith 543. Narrated By Abu Huraira :
The Prophet said, "Allah created Adam, making him 60 cubits
tall. When He created him, He said to him, "Go and greet
that group of angels, and listen to their reply, for it will
be your greeting (salutation) and the greeting
(salutations of your offspring." So, Adam said (to the angels),
As-Salamu Alaikum (i.e. Peace be upon you). The angels said,
"As-salamu Alaika wa Rahmatu-l-lahi" (i.e. Peace and Allah$#$s
Mercy be upon you). Thus the angels added to Adam$#$s
salutation the expression, $#$Wa Rahmatu-l-lahi,$#$ Any person

Figure 6. ClearGuidance.com Terrorist Suspect Participant Network and Selected Member Profile

In Figure 6 we present an illustrative example of social media analytics performed on the

ClearGuidance.com web forum that was used by selected (Toronto) terrorist suspects. Figure 6(a) shows the partial participant social network, where each node represents one online member and key members are highlighted in the center in blue. Two members who participated in the same forum discussion thread are connected via a link. One key member's message text was analyzed using selected feature extraction and text visualization techniques. The member was found to discuss predominantly religious topics (justifying Jihad), e.g., angels, Adam, and Allah, as highlighted by the large red blot in inkblot overlaying key text in Figure 6(b) and the scattered bag-of-word nodes in the Multi-Dimensional Scaling (MDS) visualization based on text distance in Figure 6(c).

Hacker Web Research Portal: Collection, Curation, and Access

We believe these kinds of social media analytics tools will be invaluable for both computational and social scientists in support of their cyber security research. The proposed SaTC Transition to Practice (TTP) research will help us develop a robust and scalable Hacker Web Research Portal by leveraging our significant past research. Ms. Cathy Larson (project senior personnel), MS in Library Science and the associate director of the AI Lab, has significant experience in hosting digital content and in managing the NSF-funded (under the Computational Research Infrastructure Program) Dark Web collection. Based on our Dark Web experience, we will develop a curated and secured data collection and access policy. Only open-access hacker community contents will be collected and will be updated on a monthly basis. Such open-access web contents with no identifiable information are considered IRB-exempt due to their anonymous nature (verified previously with the UA IRB committee). Contents collected will only be used for research purposes, thus not creating human subjects or copyright issues. Once contents are collected, they will be stored in a secured password-required card-access environment within the AI Lab of the University of Arizona; this kind of secured lab space has been previously created for the COPLINK (scrubbed Arizona criminal records) and Dark Web projects (anonymous terrorism social media contents). Our collection will be made available mostly to the SBE, computational, and security research communities. We have developed a web form for access request and qualification. Email and telephone follow-ups may also be implemented to verify the identity and need of the requesters before granting access privileges. The portal system will include a bulk download mechanism based on various database formats. Special requests can be made to obtain a large cross-sectional data dump based on specific forums, IRCs, groups, events, or timeline. In our Dark Web experience, we have found many excellent uses of our collection from mostly SBE and computational scholars, research groups, and graduate students. In addition to granting access to our collection, we also find it valuable to provide a user feedback mechanism on our portal to help improve our search interface and identify additional data sources for future collection and update. As a TTP effort, our collection and analytics tools will also be released under an open source license listed by on the Open Source Initiative.

5.2 Preliminary Exploration of the Online Hacker Community

In this section we summarize preliminary hacker community research conducted by our group based on our research framework.

5.2.1 5.2.1 Hacker Forum Exploration

Two publicly accessible online hacker communities (see Table 1) were identified and collected recently using our Dark Web forum crawling system (Fu et al., 2010). Our spidering and proxy setup was able to successfully avoid forum server overload or detection. Forum content was processed and data relevant to forum users and messages was parsed into a SQL database. The two communities discuss a range of technical topics, including cybercrime-related items in two different languages: English and Chinese. We adopted selected social media analytics proposed earlier to extract key forum features for analysis relating to member reputation in the forum.

We explored the mechanisms in which some hackers become key actors within their communities (Benjamin & Chen, 2012). The relationships between various hacker posting behaviors (six features) and reputation (acquired by each member and displayed as reputation point) were observed through the use of Ordinary Least Squares (OLS) regression.

$$\begin{aligned} Reputation = {} & \beta_1 Average\_Message\_Length + \beta_2 Number\_Of\_Replies + \beta_3 Number\_Of\_Threads\_Involved \\ & + \beta_4 Tenure + \beta_5 Sum\_Of\_Attachments + \beta_6 Total\_Messages + \varepsilon \end{aligned}$$

Three hypotheses were tested:

H1: Discussion intensity (average message length and number of replies) is a significant contributor towards hacker reputation as it relates to the cognitive advance of a community.

H2: Community involvement (number of threads, seniority, attachments, message volume) is a significant contributor of hacker reputation as it encompasses user activity levels and contributes towards the cognitive advancement of the community.

H3: Results of the regression model are consistent across both the English and Chinese forums, as both forums share aspects of hacker culture.

Our analysis showed that involvement in various threads, the sum of attachments, and total message volume all appeared to be significant contributors of reputation. This demonstrated support for theories tying higher user reputations to individuals who are active and contribute to cognitive advancement of their communities. Average post length, number of replies per thread, and tenure did not appear to be significant contributors towards reputation. Both communities shared similar patterns in regards to how reputation was built by members, supporting H3. Despite obvious cultural differences between the English and Chinese forums, an overriding form of hacker culture appeared to be experienced by both communities.

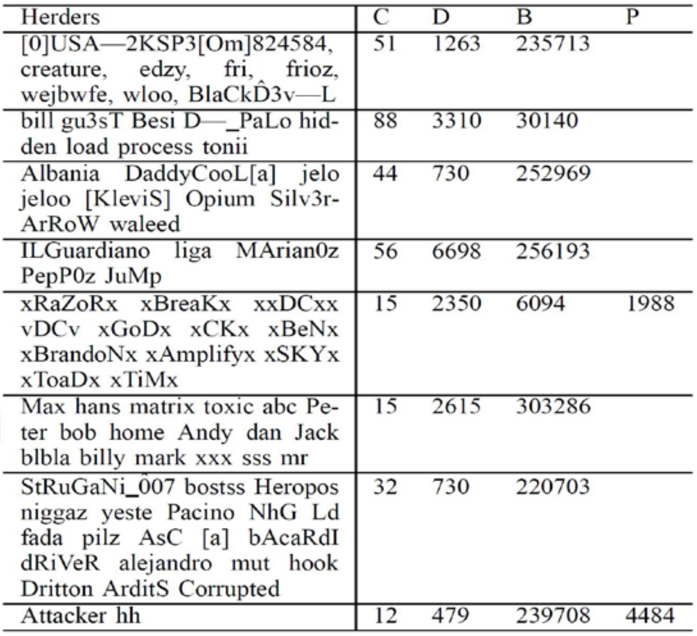
Table 1. Sample Statistics for Selected Hacker Community Forums

Forum Name	Language	# of Messages	# of Users	Forum Start Date
Hackhound.org	English	77,061	5,794	October 9, 2008
Unpack.cn	Chinese	646,494	22,743	October 12, 2004

5.2.2 Botnet IRC Exploration

In an earlier study (Mielke & Chen, 2008), we investigated the IRC command & control (C&C) signatures of major botmasters using honeypot IRC data collected from the ShadowServer Foundation, a non-profit research group for botnet research.

Table 2. Several cyber criminal groups were found, based on concentrations of criminal events and membership size. The columns, from left to right, represent the criminal nicknames, the number of C&C channels controlled by the gang (C), the number of unique DDoS targets (D), estimated number of bots (B), and number of PSTORE compromises of victim passwords (P).



Herders	C	D	B	P
[0]USA—2KSP3[Om]824584, creature, edzy, fri, frioz, wejbwfe, wloo, BlaCkD̂3v—L	51	1263	235713	
bill gu3sT Besi D—_PaLo hidden load process tonii	88	3310	30140	
Albania DaddyCooL[a] jelo jeloo [KleviS] Opium Silv3rArRoW waleed	44	730	252969	
ILGuardiano liga MArian0z PepP0z JuMp	56	6698	256193	
xRaZoRx xBreaKx xxDCxx vDCv xGoDx xCKx xBeNx xBrandoNx xAmplifyx xSKYx xToaDx xTiMx	15	2350	6094	1988
Max hans matrix toxic abc Peter bob home Andy dan Jack blbla billy mark xxx sss mr	15	2615	303286	
StRuGaNi_0̂07 bostss Heropos niggaz yeste Pacino NhG Ld fada pilz AsC [a] bAcaRdI dRiVeR alejandro mut hook Dritton ArditS Corrupted	32	730	220703	
Attacker hh	12	479	239708	4484

We also performed exploratory population modeling of the bots and cluster analysis of selected cyber criminal groups. Social network analysis was performed on the community structure of the underlying IRC communication network. All pre-filtered “human” nicknames were taken as nodes in a large social network. Links were defined between any two nodes found collaborating in a single C&C channel. Weights were assigned to each link with a simple Jaccard metric measuring the percentage of channels the nicknames shared in common divided by the total number of channels occupied by either nickname. A hierarchical agglomerative clustering algorithm was used to cluster vertices together as groups of "herders" and suggest their key characteristics, e.g., the number of C&C channels controlled by the group, the number of unique DDoS targets, estimated number of bots controlled, and number of PSTORE compromises of victim passwords (as shown in Table 2).

In this ambitious, yet achievable, project, we plan to collect and maintain most if not all of the critical international online hacker community social media contents, from forums to IRC channels, in our proposed Hacker Web Research Portal. Our advanced social media analytics tools will also make large-scale and systematic cyber security related content and criminal network analysis possible for researchers and practitioners in different parts of the world.

6. Prior Support and Related Works

The PIs have extensive experience in security informatics, autonomic computing, and criminology research. Selected prior support and related works of project PIs are summarized below.

NSF COPLINK Center for Homeland Security Research (NSF-Digital Government/NIJ, IIS-0429364, 2000-2007, $3.1M): Dr. Chen is PI of a major project funded by the NSF Digital Government Program, and National Institute of Justice for developing information sharing and criminal analysis technologies for law enforcement and homeland security community, and founder/developer of COPLINK, cross-jurisdictional information sharing, analysis, and visualization software for the law enforcement and intelligence communities. In the summer of 2009, COPLINK was acquired by i2 for integration with its popular i2 crime analysis and visualization toolkit, the Analyst Notebook; i2 was acquired by IBM in July 2011, making COPLINK one of the most financially and programmatically successful UA start-ups in the university's history. *Two Selected Papers:* 1) D. Hu, S. Kaza, and H. Chen, "Identifying Significant Facilitators of Dark Network Evolution," *Journal of the American Society for Information Science and Technology*, 60:4, pp. 655-665, 2009. 2) S. Kaza, J. Xu, B. Marshall, and H. Chen, "Topological Analysis of Criminal Activity Networks: Enhancing Transportation Security," *IEEE Trans. on Intelligent Trans. Sys.* 10:1, 2009.

NSF Dark Web Research Program (NSF/DOD, 2007-present, CBET-0730908, $2.5M): Dr. Chen is PI of several NSF and DOD projects that aim to develop computational approaches to understanding global extremism and terrorism phenomena on the Internet. The Dark Web collection is one of the largest open-source terrorism research testbeds in the academia. *Two Selected Publications:* 1) H. Chen, W. Chung, J. Qin, E. Reid, M. Sageman, and G. Weinmann, "Uncovering the Dark Web: A Case Study of Jihad on the Web," *Journal of the American Society for Information Science and Technology*, 59:8, pp. 1347-1359, 2008. 2) H. Chen, *Exploring and Mining the Dark Side of the Web: The Dark Web Project*, Springer, 2012.

NSF Cloud and Autonomic Computing Center @ UofA (NSF/AFOSR/ARL, IIP-0758579, 2008-present; $1.5M): Dr. Hariri is director for the NSF Center which broadly encompasses cloud computing systems and applications and the use of autonomic computing methods for the management of these and other IT systems. CAC activities on cloud computing cut across several layers of IT systems, including: hardware platforms for computing, storage and networking; cyber-security; design of data centers that aggregate platforms to provide cloud services; and systems software and distributed computing middleware. *Two Selected Papers:* 1) Y. Luo, F. Szidarovszky, Y. Al-Nashif, and S. Hariri, "A Game Theory Network Security," *J. Inform. Security*, 2010, Published Online, July 2010, pp. 41-44, 2) Y. Alnashif, A. Kumar, S. Hariri, G. Qu, Y. Luo, and F. Szidarovsky, "Multi-Level Intrusion Detection System (ML-IDS)," in International Conference on Autonomic Computing, pp. 131-140, 2008.

NIJ Hacker Community Research (National Institute of Justice, 2011-2013, $270K): Dr. Holt is PI of an NIJ project examining the structure, organization, and processes of the international market stolen data online. *Two Selected Papers:* 1) Holt, T. J., 2012, "Examining the Forces Shaping Cybercrime Markets Online," *Social Science Computer Review*. 2) Holt, T. J., and Kilger, M., 2012, "The Social Dynamics of Hacking," *Know Your Enemy* Series, The Honeynet Project, viewed 18 August, 2012 from https://honeynet.org/papers/socialdynamics.

7. Collaboration Plan

7.1 Project Personnel and Roles

Dr. Hsinchun Chen is project PI. He is the director of the Artificial Intelligence (AI) Lab and the NSF COPLINK Center at the University of Arizona. Chen and his lab of 20+ staff and scientists have received major funding from the NSF, NIJ, DARPA, CIA, DHS, and NIH, among others over the past 20 years, and have extensive research experience in the areas of digital government, digital libraries, intelligence and security informatics, biomedical informatics, and knowledge management systems. Dr. Chen founded the IEEE International Conference on Intelligence and Security Informatics (ISI), the premiere meeting in national security IT research, and has served as the Steering Committee chair since 2003. Dr. Chen is Editor-in-Chief of the *ACM Transactions on MIS* and the Springer *Security Informatics* journal. He is one of the leaders in developing the science of Security Informatics. Dr. Chen will supervise the overall progress of the proposed research program and contribute his extensive research expertise in data, text, and web mining, and security informatics. He will lead the social media analytics research and the development of the Hacker Web Research Portal. He will also disseminate and promote our research within the IEEE ISI community and solicit publications for his journals from our project researchers and graduate students.

Dr. Salim Hariri, Professor in the Department of Electrical and Computer Engineering, is Co-PI and brings an exceptionally strong research program in network and cyber security. Dr. Hariri will lead the development of the proposed honeypot IRC collection and analytics platform based on the infrastructure developed earlier from his NSF Cloud and Autonomic Computing Center. Dr. Hariri has significant federal research support for cyber security research and development. His team has recently transitioned the cyber security technologies supported by Airforce Research Lab (ARL) and the Air Force Office of Scientific Research (AFOSR) into commercial products. Dr. Hariri's research team transitioned their anomaly behavior analysis and self-management tools to a cyber security startup (Avirtek) that received significant STTR and SBIR (Phase I and II) funding. In addition to technical honeypot research in our project, he will also actively engage in various IEEE networking and cloud computing conferences through publications and workshop activities.

Dr. Thomas J. Holt (Co-PI) is an Associate Professor in the School of Criminal Justice at Michigan State University specializing in computer crime, cybercrime, and technology. He has been funded by NSF and NIJ and has an extensive focus on researching computer hacking, malware, carding, stolen data markets, and the role that technology and the Internet play in facilitating all manner of crime and deviance. He has been published in a variety of academic journals, including *Crime and Delinquency*, *Deviant Behavior* and the *Journal of Criminal Justice*; co-authored the book *Digital Crime and Digital Terror;* edited the book *Crime On-Line*; and co-edited *Corporate Hacking and Technology-Driven Crime.* He is the project lead for the Spartan Devils Honeynet Project, a joint project of Michigan State University, Arizona State University, and private industry, to examine the technical and social dynamics of hacking and cybercrime using Honeynet technologies and open source data. Dr. Holt is also the director of the Open Source Research Laboratory at Michigan State University which utilizes the Internet to examine the hidden communities driving cybercrime and attacks against critical infrastructure. He will facilitate the identification and mining of websites from various hacking and cybercrime communities, as well as establish and manage Honeynet technologies that will be used in this research. He will also assist in the development of keyword searches, text, and mining techniques to facilitate this project. Finally, he will promote the findings of relevance to the social sciences through publication and presentation activities.

Ms. Catherine Larson (senior personnel), MSLIS, is associate research scientist and associate director of the AI Lab, University of Arizona. Her areas of expertise include digital library and content management, user studies, and information privacy. As project coordinator, she will help with tracking milestones, creating documentation, overseeing data collection and curation, and coordinating user studies.

Graduate Research Assistants and Associates: Graduate (masters and Ph.D.) students will formally participate in all aspects of the project, from project kick-off and meeting participation, through research and development, and project dissemination including writing for publication and for conference presentations. More junior students will be paired with senior students when possible to optimize the learning experience.

Recruiting Under-represented Groups: When hiring, we will make special efforts to reach out to under-represented student groups (including women, Hispanic Americans, African Americans, and military veterans) through outreach to university student centers, relevant classes, and nearby bases. Situated in the Southwest, the University of Arizona has a higher proportion of Hispanic students. In addition, Tucson is adjacent to two large military bases – Fort Huachuca (Amy) and David-Monthan (Air Force), which produce a significant number of veteran students at the university. Both the Hispanic and veteran student groups will be heavily recruited for our research.

7.2 Project Management Approach and Coordination Mechanisms

Project management approaches will be a combination of agile and traditional project management. The PI will establish an overall detailed project plan, based on the timeline and milestones in Table 1, below, and share that with all project personnel prior to project kick-off. Research is a process of discovery, however, and because we cannot predict how those discoveries might affect our future tasks and development efforts, agile project management processes will also be used as a means of allowing us to respond and adapt appropriately to what we learn.

7.2.1 Kick-Off and Other Regular Project Meetings

Work between the PI, Co-PIs, and other project personnel will begin with a kick-off meeting to be attended by all onsite UA personnel (PI Chen, Co-PI Hariri, and all UA research associates), with Co-PI Holt and MSU research associates attending virtually. At project kick-off, the plan will be reviewed in detail with specific assignments, deliverables, and timelines set for each person. These milestones will serve as the agendas for the monthly project meetings. The research assistants and associates will also meet regularly with their respective PI, beginning weekly for project initiation and start-up, then biweekly. Offsite personnel will report progress regularly at the monthly project meetings, which will also be used to generate monthly written reports (which will enable us to more speedily produce accurate annual reports for NSF).

7.2.2 Agile Project Management

As our research unfolds, it may become necessary to adapt our actions to new discoveries or a changing environment. We may need to, for example, radically change our data collection methods. We may need to change personnel to ensure we have a needed but unexpected skill. To help identify the requirements for the tools we will be developing, we will be actively soliciting input from the research community, and will therefore need to adopt an iterative, flexible process that allows for discoveries to be readily integrated into our decision-making. While the PI and Co-PIs share ultimate authority for ensuring project success, this kind of project management requires, to some extent, a more "flattened" project hierarchy to allow team members some latitude in decision-making. Toward that end, at project kick-off, this approach will on the meeting agenda for "ratification" to ensure that we all share a common understanding of the personal responsibility needed for success. Project personnel will be introduced to each other and a complete contact list compiled and shared via a wiki or other collaborative tool(s). Project personnel will be encouraged to communicate and share ideas, results, and problem-solve directly with each other.

7.2.3 Community Input

As described earlier in this proposal, an overarching goal is to develop an integrated computational framework that will allow social science researchers and security practitioners to: (1) detect, classify, measure and track the formation, development and spread of topics, ideas, and concepts in cyber attacker social media communication; (2) identify important and influential cyber criminals and their interests, intent, sentiment, and opinions in online discourse; and (3) induce and recognize hacker identities, online profiles/styles, communication genres, and interaction patterns. To support this kind of investigation and research, our plan includes the development of open source tools with embedded analytical algorithms and integrated visualization, a large longitudinal research testbed, and a web-based portal system.

To ensure that the tools, testbed, and portal are responsive to user needs, throughout the project we will seek information and input from the research community about their research questions, relevant data sources, and preferred methods of analysis. During our annual software releases, input will also be sought regarding potential tool and interface designs, functions, and usability.

We will employ a variety of methods to seek this input. We will regularly submit papers and posters to conferences and use our presentation time in part as an opportunity to seek community input (this was a very successful method of gathering input and information in our Dark Web project). When affordable, we will also request table or exhibit space. Conferences with a focus on cybercrime and/or cyber security and protection to which we will submit include but are not limited to the International Conference on Information Warfare and Security, the Cyber Infrastructure Protection Conference, the International

Conference on Digital Forensics & Cyber Crime, and the DOD Cyber Crime Conference (we are eligible through our department's National Center of Academic Excellence in Information Assurance). Conferences with a broader audience to which we will submit include but are not limited to the IEEE International Conference on Technologies for Homeland Security, the IEEE International Conference on Intelligence and Security Informatics (founded by PI Chen ten years ago), the American Political Science Association annual meeting, and the Academy of Criminal Justice Sciences annual meeting, for a few examples.

We will also actively reach out to researchers and practitioners through their organizations. We will first comprehensively identify relevant centers, organizations and the like, and follow up by making contact with them through email, phone calls, and/or targeted mailings. In the case of teaching faculty, we will also send announcements for class distribution. Example organizations include, the Center for Cybercrime Studies (College of Criminal Justice, CUNY); the Smith School of Business Cybersecurity Leadership program; and Terrorism Studies (University of St. Andrews), to name only a few in a field of many.

At each development phase we will conduct user studies. And prior to project completion, of course, we will focus our efforts on informing the community of the availability of the system.

7.2.4 Timeline and Milestones

The following timeline and milestones will guide project management. Two tracks of development will be implemented in parallel. In I, Hacker Community Research, we will explore and identify relevant hacker communities and develop our social media analytics (text and visualization) research of relevance to hacker community exploration and analysis. In II, selected contents and proven techniques will, with community input, be further developed into open source tools and testbed. Our proposed research tools and testbed will be made available to the researchers and practitioners starting Year 2.

Table 3. Project Tasks and Timeline

Research Tasks	**Year 1**		**Year 2**		**Year 3**		**Year 4**	
	1st Half	*2nd Half*	*1st Half*	*2nd Half*	*1st Half*	*2nd Half*	*1st Half*	*2nd Half*
I. Hacker Community Research								
-- Hacker Community Exploration	●	━	●		●	━	●	
-- Forums/IRC Collection Research		●	━	━	●			
-- Social Media Analytics Research			●	━	━	●		
II. Research Tools and Testbed								
-- Collection Tools Development			●	━	●			
-- Analytics Tools Development					●	━	●	
-- Hack Web Portal Development			●	━	━	━	━	●

7.2.5 Coordination Mechanisms

To summarize, our coordination mechanisms will include: 1) a project kick-off meeting to review and ratify our timeline, goals, approaches, milestones, and deliverables; 2) immediate communication and feedback regarding questions, issues, problems, etc. though email, phone, tele-, or virtual conferencing; 3) regular (monthly) project meetings backed up by written reports; 4) and a project wiki, SharePoint, or other similar online method for communicating about and documenting our work. During portal, testbed, and tool development, we will seek input and use a bug tracker to track requirements/development. Finally, we will host our testbed and portal back them up internally and externally. We will make our research tools available via an open source license through an appropriate repository/directory (one example is SourceForge; another is GitHub); the exact flavor of license and the specific repository are TBD.

7.3 Budget line items Supporting Collaboration and Coordination Mechanisms

Travel funds requested will in part support our efforts to obtain community input and project dissemination. Some of the funding in our supplies and materials request may also be used for the creation of "marketing" materials related to project dissemination and obtaining community input, and for fees that may be associated with the use of online collaboration tools. Funding requested in the data licensing fee category may be used in part for any fees associated with our affiliation with an open source repository (although free ones may be preferred).

References

Abbasi, A., and Chen, H. "Identification and Comparison of Extremist-Group Web Forum Messages using Authorship Analysis," *IEEE Intelligent Systems* (20:5), pp. 67-75, 2005.

Abbasi, A., and Chen, H. "Visualizing Authorship for Identification." *In* Proceedings of the 4th IEEE Symposium on Intelligence and Security Informatics (ISI 2006), San Diego, U.S.A. Springer, pp. 60-71, May 23-24, 2006.

Abbasi, A. and Chen, H. "Writeprints: A Stylometric Approach to Identify-Level Identification and Similarity Detection in Cyberspace," *ACM Transactions on Information Systems*, (26:2), 2008a.

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

Abbasi, A., Z. Zhang, D. Zimbra, H. Chen, and J. F. Nunamaker, "Detecting Fake Websites: The Contribution of Statistical Learning Theory," *MIS Quarterly*, 34:3, pp. 435-461, 2010.

Ackerman, Spencer. DARPA Begs Hackers: Secure Our Networks, End 'Season of Darkness'.*Wired* [online]. November 7, 2011. Available at: http://www.wired.com/dangerroom/2011/11/darpa-hackers-cybersecurity, 2011.

Bachmann, M. "The Risk Propensity and Rationality of Computer Hackers," *The International Journal of Cyber Criminology,* 4*,* 643-656, 2010.

Benjamin, V., and Chen, H. "Securing Cyberspace□: Identifying Key Actors in Hacker Communities," *IEEE Intelligence and Security Informatics*, 2012.

Chau, M., Qin, J., Zhou, Y., Tseng, C., and Chen, H."SpidersRUs: Automated Development of Vertical Search Engines in Different Domains and Languages," in Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'05), Denver, Colorado, USA, June 7-11, 2005.

Chau, M., and Xu, J. (2006). "A Framework for Locating and Analyzing Hate Groups in Blogs," in *Proceedings of the Pacific-Asia Conference on Information Systems*, Kuala Lumpur, Malaysia, July 6-9, 2006.

Chen, H. (2012). *Dark Web: Exploring and Data Mining the Dark Side of the Web.* Springer, 2012.

Chen, H. "AI and Security Informatics," *IEEE Intelligent Systems,* Volume 25, Number 5, Pages 82-83, September/October, 2010.

Chen, H., Qin, J., Reid, E., Chung, W., Zhou, Y., Xi, W., Lai, G., Bonillas, A. and Sageman, M., "The Dark Web Portal: Collecting and Analyzing the Presence of Domestic and International Terrorist Groups on the Web," Proceedings of the 7th International Conference on Intelligent Transportation Systems (ITSC), Washington D.C., October 3-6, 2004.

Chen, H., E. Reid, J. Sinai, A. Silke, and B. Ganor (Eds.), *Terrorism Informatics: Knowledge Management and Data Mining for Homeland Security*, Springer. 2008a.

Chen, H., Chung, W., Qin, J., Reid, E., Sageman, M., and Weimann, G. "Uncovering the Dark Web: A Case Study of Jihad on the Web," *J. of the Am. Soc. Info. Sci. and Tech.*, 59:8, pp. 1347-1359. 2008b.

Chen, H., D. Denning, N. Roberts, C. Larson, X. Yu, and C. Huang, "The Dark Web Forum Portal: From Multi-lingual to Video," *in* Proc. of the IEEE International Conference on Intelligence and Security Informatics, ISI 2011, Beijing, China, July 2011a.

Chen, H., C. Larson, T. Elhourani, D. Zimbra, and D. Ware, "The GeoPolitical Web: Assessing Societal Risk in an Uncertain World," Proceedings of 2011 IEEE International Conference on Intelligence and Security Informatics, ISI 2011, Beijing, China, July 2011b.

Chu, B., Holt, T. J., and Ahn, G. J. "Examining the Creation, Distribution, and Function of Malware On-Line," Washington, DC, National Institute of Justice. August, 2010.

Cova, M., Kruegel, C., and Vigna, G. "Detection and analysis of drive-by-download attacks and malicious JavaScript code," *Proceedings of the 19th international conference on World wide web - WWW '10*. 2010.

Decary-Hetu, D., and Dupont, B. "The social network of hackers", *Global Crime* 13, 160-175. 2012.

Decary-Hetu, D., Morselli, C., and Leman-Langlois, S. “Welcome to the scene: A study of social organization and recognition among Warez Hackers,” *Journal of Research in Crime and Delinquency* 49, 350-382. 2012.

Fallmann, Hanno; Wondracck, Gilbert; Platzer, Christian. *Covertly Probing Underground Economy Marketplaces.* Vienna University of Technology. 2010.

Franklin, J.; Paxson, V., Perrig., Savage, S. “*An Inquiry into the Nature and Causes of Wealth of the Internet Miscreants.* Proceedings of the 14th ACM conference on Computer and Communications Security, 2007.

Fu, T., Abbasi, A., and Chen, H. “A Hybrid Approach to Interaction Coherence Analysis in Web Forums,” *Journal of the American Society for Information Science and Technology*, 2008.

Fu, T.J., Abbasi, A., and Chen, H. "A Focused Crawler for Dark Web Forums," *Journal of the American Society for Information Science and Technology*, 61:6, 2010.

Gilbert, E. and Karahalios, K. “Using Social Visualization to Motivate Social Production,” *IEEE Trans. on Multimedia* (11:3), pp. 413-421, 2009.

Gilboa, N., 1996, “Elites, lamers, narcs, and whores: Exploring the computer underground,” In L. Cherny and E. R. Weise (eds.), *Wired Women*, pp. 98-113, Seattle: Seal Press.

Goel, S. “Cyberwarfare Connecting the Dots in Cyber Intelligence,” *Communications of the ACM*, *54*(8), 132. 2011.

Halliday, M.A.K. *An Introduction to Functional Grammar*, 3rd (ed). Revised by Christian Matthiessen, London: Hodder Arnold, 2004.

Holt, T.J. “Subcultural evolution? Examining the influence of on- and off-line experiences on deviant subcultures,” *Deviant Behavior*, 28, 171-198, 2007.

Holt, T. J. “Lone Hacks or Group Cracks: Examining the Social Organization of Computer Hackers,” In F. Schmalleger and M. Pittaro (eds.), *Crimes of the Internet*, pp. 336-355, Upper Saddle River, NJ: Pearson Prentice Hall, 2009.

Holt, T. J. “Exploring Strategies for Qualitative Criminological and Criminal Justice Inquiry Using OnLine Data,” *Journal of Criminal Justice Education*, *21*(4), 466–487, 2010.

Holt, T. J. “Examining the forces shaping cybercrime markets on-line,” *Social Science Computer Review*, 2012.

Holt, T. J., and Kilger, M. “Know Your Enemy: The Social Dynamics of Hacking,” *The Honeynet Project*, 1–17. 2012.

Holt, T. J., and Lampke, E. “Exploring stolen data markets on-line: Products and market forces,” *Criminal Justice Studies,* 23, 33-50, 2010.

Holt, T.J., Soles, J., and Leslie, L. “Characterizing malware writers and computer attackers in their own words,” paper presented at the *3rd International Conference on Information Warfare and Security*, April 24-25, 2008.

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

The Honeynet Project. *Know your enemy*, 2nd Edition. New York: Addison Wesley Professional, 2003.

Huang, C.-N., Fu, T.J., and Chen, H. (2010). “Text-based Video Content Classification for Online Video-Sharing Sites.” *Journal of the American Society for Information Science and Technology*, 61:5, pp. 891-906, 2010.

Imperva. “Imperva Hacker Intelligence Intitiative.” *Monthly Trend Report #13*, 2012.

Jordan, T., and Taylor, P. “A sociology of hackers”, *The Sociological Review*, 46, 757-780, 2009.

Lu, W., and Ghorbani, A. “Botnets Detection Based on IRC-Community,” *IEEE GLOBECOM 2008 - 2008 IEEE Global Telecommunications Conference*, (1), 1–5, 2008.

Lu, W., Tavallaee, M., and Ghorbani, A. “Automatic discovery of botnet communities on large-scale communication networks,” *Proceedings of the 4th International Symposium on Information, Computer, and Communications Security - ASIACCS ’09*, 2009.

Luo, Y., F. Szidarovszky, Y. Al-Nashif, and S. Hariri, "A Game Theory Network Security," J. Inform. Security, 2010, Published Online, pp. 41-44, 2010.

Manjikian, M. M., “From Global Village to Virtual Battlespace: The Colonizing of the Internet and the Extension of the Realpolitik,” *International Studies Quarterly* (54:2), pp. 381-401, 2010.

Mann, D. and Sutton, M. “Netcrime: More Change in the Organization of Thieving,” *The British Journal of Criminology*, 38, 201-229, 1998.

Meserve, Jeanne. “‘Smart Grid’ may be vulnerable to hackers.” *CNN* [online]. March 20, 2009. Available from: http://articles.cnn.com/2009-03-20/tech/smartgrid.vulnerability_1_smart-grid-power-grid-blackout?_s=PM:TECH

Mielke, Clinton J., Chen, Hsinchun. *Botnet and the Cybercriminal Underground.* IEEE International Conference on Intelligence and Security Informatics, 2008.

Moore, Tyler; Clayton, Richard. *Evil Searching: Compromise and Recompromise of Internet Hosts for Phishing*. Lecture Notes in Computer Science, 5628, pp. 256-272, 2009.

Motoyama, M., McCoy, D., Levchenko, K., Savage, S., and Voelker, G. M. “An analysis of underground forums,” *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference – IMC’11*, 2011.

National Science and Technology Council, “Trustworthy Cyberspace: Strategic Plan for the Federal Cybersecurity Research and Development Program,” 2011.

Paxton, N. C., Ahn, G.-J., and Shehab, M. “ MasterBlaster: Identifying Influential Players in Botnet Transactions,” *2011 IEEE 35th Annual Computer Software and Applications Conference*, 413–419, 2011.

Qin, J., Zhou, Y., and Chau, M. “Building Domain-Specific Web Collections for Scientific Digital Libraries: A Meta-Search Enhanced Focused Crawling Method,” in Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'04), Tucson, Arizona, USA, pp. 135-141, June 7-11, 2004.

Qin, J., Zhou, Y., Reid, Ed., Lai, G., and Chen, H. “Analyzing Terror Campaign on the Internet: Technical Sophistication, Content Richness, and Web Interactivity,” *International Journal of Human-Computer Studies,* special issue on Information Security in the Knowledge Economy, vol. 65, pp. 71-84, 2007.

Radianti, J. “A Study of a Social Behavior inside the Online Black Markets,” *2010 Fourth International Conference on Emerging Security Information, Systems and Technologies*, 88–92, 2010.

Radianti, Jaziar; Gonzalez, Jose J. *A Preliminary Model of the Vulnerability Black Market.* System Sciences, 2009.

Radianti, J., Rich, E., and Gonzalez, J. J. “Using a Mixed Data Collection Strategy to Uncover Vulnerability Black Markets.,” *Workshop for Information Security and Privacy*, 2007.

Radianti, J., Gonzalez, J. J., and Rich, E. “A Quest for a Framework to Improve Software Secu□r:ity Vulnerability Black Markets Scenario,” *International Conference of the System Dynamics Society*, 2009a.

Radianti, J., Rich, E., and Gonzalez, J. J. “Vulnerability Black Markets□: Empirical Evidence and Scenario Simulation,” *42nd Hawaii International Conference on*, 1–10, 2009b.

Symantec Corporation. *Symantec Internet security threat report, Volume 17*, Symantec Corporation, viewed 25 May, 2012, from http://www.symantec.com/threatreport/, 2012.

Taylor, P. “Hackers: Crime in the Digital Sublime,” London: Routledge, 1999.

Thorne, Steven. “Computer Mediated Communication.” In Hornberger, Nancy H. (Ed.), *Encyclopedia of Language and Education*, 2nd ed. Springer, 2008.

Thomas, R. and Martin, J. “The underground economy: Priceless,” *The Usenix Magazine,* 31, 7-17, 2006.

Tsai, M.-H., Chang, K.-C., Lin, C.-C., Mao, C.-H., and Lee, H.-M. “CandC tracer: Botnet command and control behavior tracing,” *2011 IEEE International Conference on Systems, Man, and Cybernetics*, 1859–1864, 2011.

Viswanathan, R. P., Y. Al-Nashif, S. Hariri "Application Attack Detection System (AADS): An Anomaly Based Behavior Analysis Approach", Accepted in the The 9th ACS/IEEE International Conference On Computer Systems and Applications, 2011.

Wang, R. Malware B-Z□: Inside the Threat From Blackhole to ZeroAccess. Sophos Whitepaper, 2012.

Wang, W., Fang, B., Zhang, Z., and Li, C. "A Novel Approach to Detect IRC-Based Botnets," *2009 International Conference on Networks Security, Wireless Communications and Trusted Computing*, 408–411, 2009.

Yip, M. "An Investigation into Chinese Cybercrime and the Applicability of Social Network Analysis," *ACM Web Science Conference,* 2011.

Zhang, Y., S. Zeng, L. Fan, Y. Dang, C. Larson, and H. Chen, "Dark Web Forums Portal: Searching and Analyzing Jihadist Forums," Proceedings of 2009 IEEE International Conference on Intelligence and Security Informatics, ISI 2009, Dallas, Texas, June 2009.

Zheng, R., Qin, Y., Huang, Z., and Chen, H. "A Framework for Authorship Analysis of Online Messages: Writing-style Features and Techniques," *Journal of the American Society for Information Science and Technology* (57:3), pp.378-393, 2006.

Zhou, Y., Reid, E., Qin, J., Chen, H., and Lai, G. "U.S. Extremist Groups on the Web: Link and Content Analysis," *IEEE Intelligent Systems*, 20(5), 44-5, 2005.

Zhou, Y., Qin, J., Lai, G., Reid, E., and Chen, H. (2006). "Exploring the Dark Side of the Web: Collection and Analysis of U.S. Extremist Online Forums," in Proceedings of the Intelligence and Security Informatics: IEEE International Conference on Intelligence and Security Informatics (ISI 2006), San Diego, CA, USA, May 23-24, 2006.

Zhu, Z., Lu, G., Chen, Y., Fu, Z. J., Roberts, P., and Han, K. "Botnet Research Survey." *2008 32nd Annual IEEE International Computer Software and Applications Conference*, 967–972, 2008.

Zhuge, Jianwei; Holz, Thorsten; Song, Chengyu; Guo, Jinpeng; Han, Xinhui; Zou, Wei. *Studying Malicious Websites and the Underground Economy on the Chinese Web.*Workshop on the Economics of Information Security (WEIS). May, 2008.

Hsinchun Chen McClelland Professor of Management Information Systems; Director, Artificial Intelligence Lab, University of Arizona, Tucson, AZ 85721; http://ai.arizona.edu

A. Professional Preparation

New York University, Information Systems: Ph.D., 1989
New York University, Information Systems: M.S., 1987
State University of New York at Buffalo, MIS, Mgmt Science: MBA, 1985

B. Positions and Honors:

- Full Professor and McClelland Endowed Professor, University of Arizona, 1998-present.
- IEEE Fellow and AAAS Fellow; IEEE Computer Society Technical Achievement Award, 2006; MISQ Best Paper Award, 2010; IEEE ISI Research Achievement Award, 2011.
- Scientific Counselor/Advisor, National Library of Medicine (USA), Academia Sinica (Taiwan), and National Library of China (China).
- EIC, ACM *Transactions on Management Information Systems* & Springer *Security Informatics* Journal; Associate EIC, IEEE *Intelligent Systems*; AE, *Journal of the American Society for Information Science and Technology*, IEEE *Transactions on Systems, Man, and Cybernetics*, and *Decision Support Systems*.
- Steering Committee Chair and Conference Co-Chair for the IEEE Intelligence and Security Informatics Conference (IEEE ISI), 2003-2011. The ISI conference, which has been sponsored by NSF, CIA, DHS, and NIJ, is the premiere meeting for international and homeland security IT research.
- Founding director of Artificial Intelligence Lab and Hoffman E-Commerce Lab. The UA Artificial Intelligence Lab, which houses 20+ researchers, has received more than $30M in research funding from NSF, DOD, NIH, NLM, DOJ, CIA, and other agencies over the past 20 years.
- Author/editor of 20 books and more than 250 SCI journal articles covering security informatics, biomedical informatics, data/text/web mining, digital library, knowledge management, and Web computing. Author of several research books in information systems: *Sports Data Mining* (2010); *Infectious Disease Informatics* (2010); *Terrorism Informatics* (2008); *Mapping Nanotechnology Knowledge and Innovation (2008*), *Digital Government (2007); Intelligence and Security Informatics for International Security (2006)*; and *Medical Informatics (2005),* all published by Springer.
- Ranked #8 in publication productivity in Information Systems (CAIS 2005); #1 in Digital Library research (IP&M 2005); h-index of 54, among the top three in MIS (TMIS 2011).
- Dr. Chen's COPLINK system, which has been quoted as a national model for public safety information sharing and analysis, has been adopted in more than 3,500 law enforcement and intelligence agencies. The COPLINK research had been featured in New York Times, Newsweek, Washington Post, Boston Globe, among others. The COPLINK project was selected as a finalist by the prestigious International Association of Chiefs of Police (IACP)/Motorola 2003 Weaver Seavey Award for Quality in Law Enforcement in 2003. The COPLINK company was acquired by i2 in 2009 and the combined company acquired by IBM in 2011 for $500M.
- Dr. Chen's Dark Web research, funded by NSF and DOD, has been featured in BBC, Associated Press, National Public Radio, Fox News, Discover Magazine, NSF Press, USA Today, Washington Post, among others, as a model of advanced computational approach for countering terrorism in cyberspace.
- Received numerous awards in information technology education and industry research including: AT&T Foundation Award, SAP Award, the Andersen Consulting Professor of the Year Award, the University of Arizona Technology Innovation Award, and the National Chiao-Tung University Distinguished Alumnus Award.

C. Selected Publications *(from over 20 books, 250 peer-reviewed journal articles, and 140 refereed conference articles or book chapters):*

1) **H. Chen.** *Dark Web: Exploring and Data Mining the Dark Side of the Web.* Springer, 2012.
2) A. Abbasi, S. France, Z. Zhang, and **H. Chen**, "Selecting Attributes for Sentiment Classification using Feature Relation Networks," *IEEE Transactions on Knowledge and Data Engineering*, 23(3), pp. 447-462, 2011.

3) A. Abbasi, Z. Zhang, D. Zimbra, **H. Chen**, and J. F. Nunamaker, "Detecting Fake Websites: The Contribution of Statistical Learning Theory," *MIS Quarterly*, 34(3), pp. 435-461, September 2010. (Winner, MISQ Best paper 2010)
4) T. J. Fu, A. Abbasi, and **H. Chen**, "A Focused Crawler for Dark Web Forums," *Journal of the American Society for Information Science and Technology,* 61(6), pp. 1213-1231, 2010.
5) A. Abbasi, **H. Chen**, S. Thoms, and T. J. Fu, "Affect Analysis of Web Forums and Blogs using Correlation Ensembles," *IEEE Transactions on Knowledge and Data Engineering*, 20(9), pp. 1168-1180, September 2008.
6) D. Hu, S. Kaza, and **H. Chen**, "Identifying Significant Facilitators of Dark Network Evolution," *Journal of the American Society for Information Science and Technology,* 60(4), pp. 655-665, 2009.
7) **H. Chen** and M. Roco. *Mapping Nanotechnology Innovations and Knowledge: Global and Longitudinal Patent and Literature Analysis.* Springer, 2009.
8) **H. Chen**, "AI and Global Science and Technology Assessment," *IEEE Intelligent Systems,* 24(8), pp. 68-71, July/August, 2009.
9) Y. Dang, **H. Chen**, Y. Zhang, and M. Roco, "Knowledge Sharing and Diffusion Patterns: International Patent and Patent Family Analysis," *IEEE Nanotechnology,* 3(3), pp. 16-21, 2009.
10) **H. Chen** and M. Roco, *Mapping Nanotechnology Innovations and Knowledge: Global and Longitudinal Patent and Literature Analysis,* Springer, 2008.

D. Synergistic Activities:

1) Dark Web Terrorism Research (NSF, DTRA, LOC, DHS, $2.5M). Principal investigator for the Dark Web Project, a long-term scientific research program that aims to study modern international terrorism via a computational, data-centric approach; developed the Dark Web Forum Portal which provides searchable access to and social network analysis of terrorist/extremist forums.
2) GeneScene: A Toolkit for Genomic Pathway Analysis (NIH/NLM, 2002-2006, $1.4M). PI for GeneScene, a system designed to utilize information derived from medical literature, genomic ontologies, and microarray data to help suggest interactions between genomic and biochemical pathways.
3) A National Center of Excellence for Infectious Disease Informatics (NSF/ITR, 2004-2009, $2.2M): PI for a collaborative initiative (BioPortal) to explore the development of an integrated and scalable information sharing, monitoring and analysis environment across jurisdictions and for different infectious diseases (e.g., west Nile virus, foot-and-mouth disease).
4) COPLINK Center for Homeland Security Research (NSF-Digital Government/DHS/CIA, 2000-2007, $3.1M): PI of a major NSF Digital Government project, which develops information sharing and criminal analysis technologies for law enforcement and homeland security community.

E. Collaborators (Selected):

Atabakhsh, Homa – Raytheon
Chen, Ching-chih – Simmons College
Chen, Su-Shing – The University of Florida
Cox, James – University of Arizona
Demchak, Chris – U.S. Naval War College
Eidson, Millicent - NYSDOH
Gotham, Ivan - NYSDOH
Gupta, Harsh – University of Arizona
Hendriawan, David – University of Arizona
Hu, Paul – University of Utah
Hubbard, Susan – NIH National Cancer Institute
Lally, Ann – University of Washington
Lynch, Cecil – CADHS
Martinez, Jesse – University of Arizona
Nunamaker, Jay – University of Arizona
Petersen, Timothy – Tucson Police Dept.
Ramsey, Marshall – Microsoft
Roco, Mihail – National Science Foundation
Romano, Nicolas – University of Tulsa
Scanlon, Pamela – ARJIS (San Diego, Calif.)
Schatz, Bruce – University of Illinois
Schroeder, Jenny – Tucson Police Dept.
Sewell, Robin – University of Arizona
Thurmond, Mark – UC Davis
Violette, Charles – Tucson Police Dept.
Wilson, Pete – Pima County Sheriff's Dept.
Wyzga, Wojciech – Knowledge Computing Corp.
Yang, Christopher – Chinese Univ. of Hong Kong
Zeng, Daniel – University of Arizona
Zhao, Leon – University of Arizona
Ph.D. Thesis Advisor: Vasant Dhar – New York University

Biographical Sketches

Salim Hariri, Co-Director
NSF CENTER FOR AUTONOMIC COMPUTING
The University of Arizona
Tucson, Arizona, 85721-0104
email: hariri@ece.arizona.edu,
Phone: (520) 621-4378, Fax: (520) 621-8076,
http://nsfcac.arizona.edu

EDUCATION

- Ph.D. (Computer Engineering), University of Southern California, Los Angeles, CA, 1986.
- M.Sc., Electrical Engineering, The Ohio State University, Columbus, Ohio, 1982.
- B.Sc., Electrical Engineering, Damascus University, Damascus, Syria, 1977.

PROFESSIONAL EXPERIENCE

- Professor in the Department of Electrical and Computer Engineering, University of Arizona, 2004 to present.
- Associate Professor in the Department of Electrical and Computer Engineering, University of Arizona, 1998 to 2003.
- Associate Professor in the Department of Electrical Engineering and Computer Science, Syracuse University, 1993 to 1998.
- Assistant Professor in the Department of Electrical and Computer Engineering, Syracuse University, 1989 to 1993.

SELECTED PUBLICATIONS

1. Venkata Krishn Nimmagadda, Ali Akoglu, Salim Hariri, and Talal Moukabary, "Cardiac Simulation on Multi-GPU Platform," *The Journal of Supercomputing, 2010, Online First, 22 pages.*
2. Arjun Hary, Ali Akoglu, Youssif AlNashif, Salim Hariri, and Darrel Jenerette, "Design and Evaluation of a Self-healing Kepler for Scientific Workflow," ACM International Symposium on High Performance Distributed Computing (HPDC 2010), Chicago, Illinois, June 20-25, 2010.
3. M. Parashar and S. Hariri, ***Autonomic Computing: Concepts, Infrastructure, and Applications,*** CRC Press, Taylor & Francis Group, ISBN 0-8493-9367-1, 2007.
4. Y. Jaraweh, A. Hary, Y.B. Al-Nashif, S. Hariri, A. Akoglu, and D. Jenerette, "Accelerated Discovery Through Integration of Kepler with Data Turbine for Ecosystem Research," the IEEE/ACS International Conference on Computer Systems and Applications, 2009, pp. 1005-1012.
5. S. Hariri, AUTONOMIA: An Autonomic Computing Environment, IEEE 22nd International Performance, Computing, and Communication Conference, April 2003.

OTHER SIGNIFICANT PUBLICATIONS

1. ***Tools and Environments for Parallel and Distributed Computing***, S. Hariri and Manish Parashar, Wiley Book Series on Parallel and Distributed Computing, Series, 2004.
2. Bithika Khargharia, Haoting Luo, Youssif Al-Nashif and Salim Hariri, "AppFlow-based Autonomic Performance-per-Watt Management of Large-Scale Data Centers, in *Proc. 2010 IEEE/ACM International Conference on Green Computing and Communications (GreenCom-2010)*, Hangzhou, China, December 2010.
3. B. Khargharia, S. Hariri and M.S. Yousif, "An Adaptive Interleaving Technique for Memory Performance-per-Watt Management," IEEE Transactions on Parallel and Distributed Systems, Vol. 20, Issue 7, July 2009, pp. 1011-1022.

4. Huoping Chen, Salim Hariri, and Fahd Rasal; "An Innovative Self-Configuration Approach for Networked Systems and Applications"; The 4th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA-06).
5. S. Hariri. Bithika Khargharia, and M. Parashar, "The Foundations of Autonomic Computing," in "Handbook of Bioinspired Algorithms and Applications," Editor: A. Zomaya and Olariu, CRC Press LLC, 2005.

SYNERGISTIC ACTIVITIES

- **Editor-in-Chief** Cluster Computing: The Journal of Networks, Software and Applications, http://www.wkap.nl/journalhome.htm/1386-7857.
- Co-general, Autonomic Computing Workshop, In conjunction with the 12th International Symposium on Grid Computing (HPDC 12), Seattle WA, June 24th 2003**, http://www.caip.rutgers.edu/ams2003**
- Co-Founder of The IEEE International Symposium on High Performance Distributed Computing, http://www.hpdc.org.
- Director of the NSF Center for Autonomic Computing, University of Arizona Site.
- Developed software tools that include Net-HPCC, ADViCE, PAMS, CATALINA, Pragama, and AUTONOMIA that have been used in teaching and/or research projects to help students better understand the operation control and management of parallel and distributed computing applications.

Ongoing Research Projects

- **Autonomic Programming Paradigm:** Large scale scientific applications generally experience different execution phases at runtime and each phase has different computational, communication and storage requirements as well as different physical characteristics. Autonomic Programming (AP) paradigm enables application developers to identify the appropriate solution methods to exploit the heterogeneity and the dynamism of the application execution states.
- **Autonomic Defense System:** We are designing and implementing a multilevel anomaly-based Autonomic Defense System (ADS), which is capable of detecting any type of attacks targeting resources, with low false alerts and high detection rates.
- **Autonomic Application Management:** We are developing anomaly-based approach to analyze the behavior of applications and detect and protect against any anomalous behaviors that can be triggered application malicious attacks, failures or performance degradations due to changes in workloads.
- **Autonmia: An Autonomic Control and Management Environment:** Autonomia provides dynamic programmable control and management services to support the development and deployment of autonomic applications. It can automate the dynamic allocation of resources to achieve high performance, secure and fault tolerant operations. It provides a secure, open computing environment with automated deployment, registration and discovery of application components.

OTHER COLLABORATORS

The following have been co-authors in the last five years: M. Parashar, C. S. Raghavendra, Mazin Yousif, Guangzhi Qu, Bernard Zeigler, I. Ra, Y. Kim, M. Djunadi, S. Park, J. Lee, M. Ladan, E. Al-Hajarey, S. Al-Kasabi, G. Fox, K. Kwiat, W. Debani,.

GRADUATE AND POST-GRADUATE ADVISEES IN LAST 5 YEARS

The following are my former Ph.D Students: Bithika Khargharia, Huoping Chen, Byoung Kim, Samer Fayssal, Guangzhi Qu, Yeliang Zhang, Ilkeyun Ra, Yoonhee Kim, Dongmin Kim, Manish Parashar, JongBaek Park, Mohamad Ladan, Eyas Al-Hajarey, and Saad Al-Kasabi.

MY OWN GRADUATE AND POSTDOCTORAL ADVISORS

My Ph.D. advisor was C.S. Raghavendra.

BIOGRAPHY – THOMAS J. HOLT

A. Professional Preparation

University of Missouri-Saint Louis	Criminology and Criminal Justice	B.A., 2000
University of Missouri-Saint Louis	Criminology and Criminal Justice	M.A., 2003
University of Missouri-Saint Louis	Criminology and Criminal Justice	Ph.D., 2005

B. Appointments

Associate Professor, School of Criminal Justice, Michigan State University, since 2011
Assistant Professor, School of Criminal Justice, Michigan State University, 2008-2011
Assistant Professor, Criminal Justice, UNC Charlotte, 2005-2009.

C. Relevant Publications

Holt Thomas J. 2012. "Examining the Forces Shaping Cybercrime Markets Online." *Social Science Computer Review*. DOI: 10.1177/0894439312452998

Holt, Thomas J., and Max Kilger. 2012. "Examining Willingness to Attack Critical Infrastructure On and Off-line." *Crime & Delinquency*, 58(5): 798-822.

Holt, Thomas J., Deborah Strumsky, Olga Smirnova, and Max Kilger. 2012. "Examining the social networks of malware writers and hackers." *The International Journal of Cyber Criminology,*6: 891-903.

Holt, Thomas J., and Max Kilger. 2012. "The Social Dynamics of Hacking." *Know Your Enemy* Series, The Honeynet Project. Available online: https://honeynet.org/papers/socialdynamics

Holt, Thomas J., and Eric Lampke. 2010. "Exploring stolen data markets on-line: Products and market forces." *Criminal Justice Studies,* 23: 33-50.

Significant Publications

Holt, Thomas J. and Bernadette Schell. (Editors). 2011. *Corporate Hacking and Technology Driven Crime: Social Dynamics and Implications.* Hershey, PA: IGI Global Publishers

Chu, Bill, Thomas J. Holt, and Gail Joon Ahn. 2010. *Examining the creation, distribution, and function of malware on-line.* Washington, DC: National Institute of Justice. www.ncjr.gov/pdffiles1/nij/grants/230112.pdf

Bossler, Adam M. and Thomas J. Holt. 2010. "On-line Activities, Guardianship, and Malware Infection: An Examination of Routine Activities Theory." *The International Journal of Cyber Criminology,* 3: 400-420.

Holt, Thomas J. 2009. "Lone Hacks or Group Cracks: Examining the Social Organization of Computer Hackers." Pp. 336-355 in *Crimes of the Internet,* Frank Schmalleger and Michael Pittaro, eds. Upper Saddle River, NJ: Pearson Prentice Hall.

Holt, Thomas J. 2007. "Subcultural Evolution? Examining the influence of on and off-line subcultural experiences on deviant subcultures." *Deviant Behavior*, 28: 171-198.

D. Synergistic Activities

1. I have developed and served as the PI on multiple research projects funded by the National Institute of Justice on the social dynamics of malware writers, hackers, and data thieves, combining social and technical sciences to understand these offenses.

2. Founded and serve as the project lead for the multi-university, multi-disciplinary research chapter of the International Honeynet Project to better integrate social science research techniques to the problem of hacking and malware. Our research has led to several substantive publications in this area in both the social and technical sciences.

3. Led the establishment of the Carolinas Cyber Defender Program, receiving over $5 million funding from NSF and NSA (2001-2012). This program funds full tuition and stipend for their graduate studies at either UNC Charlotte or NC A&T State University.

E. Collaborators & Other Affiliations

i. Collaborators and Co-Editors

Gail-Joon Ahn (Arizona State University), Kristie Blevins (Eastern Kentucky University), Adam Bossler (Georgia Southern University), George Burruss (Southern Illinois University at Carbondale), Heith Copes (University of Alabama-Birmingham), Max Kilger (The Honeynet Project), Bernadette Schell (Laurentian University), Michael G. Turner (University of North Carolina-Charlotte).

ii. Graduate and Postdoctoral Advisors

G. David Curry (University of Missouri-Saint Louis), Scott Decker (Arizona State University), Jody Miller (Rutgers University)

iii. Thesis Advisor and Postgraduate-Scholar Sponsor

Lev Fejes (Michigan State University), Sarah Fitzgerald (Michigan State University), Hakan Hekim (University of North Carolina-Charlotte), Byung Lee (Michigan State University)

Graduate Students

Kristine Denholm (Arizona State University), Lev Fejes (Michigan State University), Sarah Fitzgerald (Michigan State University), Karie Gregory (Michigan State University), Byung Lee (Michigan State University), Amelia Levitt (Michigan State University)

Total Students: 7
Postdoctoral Fellows:0

SUMMARY PROPOSAL BUDGET

YEAR 1

ORGANIZATION University of Arizona	FOR NSF USE ONLY			
	PROPOSAL NO.	DURATION (months)		
		Proposed	Granted	
PRINCIPAL INVESTIGATOR / PROJECT DIRECTOR Hsinchun Chen	AWARD NO.			

A. SENIOR PERSONNEL: PI/PD, Co-PI's, Faculty and Other Senior Associates (List each separately with title, A.7. show number in brackets)	NSF Funded Person-months CAL	ACAD	SUMR	Funds Requested By proposer	Funds granted by NSF (if different)
1. Hsinchun Chen - PI	0.00	0.36	1.00	36,688	
2. Salim Hariri - Co-PI	0.00	0.06	0.95	14,142	
3.					
4.					
5.					
6. (0) OTHERS (LIST INDIVIDUALLY ON BUDGET JUSTIFICATION PAGE)	0.00	0.00	0.00	0	
7. (2) TOTAL SENIOR PERSONNEL (1 - 6)	0.00	0.42	1.95	50,830	
B. OTHER PERSONNEL (SHOW NUMBERS IN BRACKETS)					
1. (0) POST DOCTORAL SCHOLARS	0.00	0.00	0.00	0	
2. (1) OTHER PROFESSIONALS (TECHNICIAN, PROGRAMMER, ETC.)	0.60	0.00	0.00	3,501	
3. (4) GRADUATE STUDENTS				80,000	
4. (0) UNDERGRADUATE STUDENTS				0	
5. (0) SECRETARIAL - CLERICAL (IF CHARGED DIRECTLY)				0	
6. (0) OTHER				0	
TOTAL SALARIES AND WAGES (A + B)				134,331	
C. FRINGE BENEFITS (IF CHARGED AS DIRECT COSTS)				68,951	
TOTAL SALARIES, WAGES AND FRINGE BENEFITS (A + B + C)				203,282	
D. EQUIPMENT (LIST ITEM AND DOLLAR AMOUNT FOR EACH ITEM EXCEEDING $5,000.) Server and storage $ 8,423					
TOTAL EQUIPMENT				8,423	
E. TRAVEL 1. DOMESTIC (INCL. CANADA, MEXICO AND U.S. POSSESSIONS)				4,400	
2. INTERNATIONAL				2,200	
F. PARTICIPANT SUPPORT COSTS 1. STIPENDS $ 0 2. TRAVEL 0 3. SUBSISTENCE 0 4. OTHER 0					
TOTAL NUMBER OF PARTICIPANTS (0) TOTAL PARTICIPANT COSTS				0	
G. OTHER DIRECT COSTS					
1. MATERIALS AND SUPPLIES				4,500	
2. PUBLICATION COSTS/DOCUMENTATION/DISSEMINATION				0	
3. CONSULTANT SERVICES				0	
4. COMPUTER SERVICES				0	
5. SUBAWARDS				71,469	
6. OTHER				10,000	
TOTAL OTHER DIRECT COSTS				85,969	
H. TOTAL DIRECT COSTS (A THROUGH G)				304,274	
I. INDIRECT COSTS (F&A)(SPECIFY RATE AND BASE) MTDC (Rate: 51.5000, Base: 202742)					
TOTAL INDIRECT COSTS (F&A)				104,412	
J. TOTAL DIRECT AND INDIRECT COSTS (H + I)				408,686	
K. RESIDUAL FUNDS				0	
L. AMOUNT OF THIS REQUEST (J) OR (J MINUS K)				408,686	

M. COST SHARING PROPOSED LEVEL $ 0	AGREED LEVEL IF DIFFERENT $		
PI/PD NAME Hsinchun Chen	FOR NSF USE ONLY INDIRECT COST RATE VERIFICATION		
ORG. REP. NAME* Mary Gerrow	Date Checked	Date Of Rate Sheet	Initials - ORG

*ELECTRONIC SIGNATURES REQUIRED FOR REVISED BUDGET

SUMMARY PROPOSAL BUDGET

YEAR 2

ORGANIZATION University of Arizona	FOR NSF USE ONLY PROPOSAL NO.	DURATION (months) Proposed	 Granted
PRINCIPAL INVESTIGATOR / PROJECT DIRECTOR Hsinchun Chen	AWARD NO.		

A. SENIOR PERSONNEL: PI/PD, Co-PI's, Faculty and Other Senior Associates (List each separately with title, A.7. show number in brackets)	NSF Funded Person-months CAL	ACAD	SUMR	Funds Requested By proposer	Funds granted by NSF (if different)
1. Hsinchun Chen - PI	0.00	0.36	1.00	36,688	
2. Salim Hariri - Co-PI	0.00	0.06	0.95	14,142	
3.					
4.					
5.					
6. (0) OTHERS (LIST INDIVIDUALLY ON BUDGET JUSTIFICATION PAGE)	0.00	0.00	0.00	0	
7. (2) TOTAL SENIOR PERSONNEL (1 - 6)	0.00	0.42	1.95	50,830	
B. OTHER PERSONNEL (SHOW NUMBERS IN BRACKETS)					
1. (0) POST DOCTORAL SCHOLARS	0.00	0.00	0.00	0	
2. (1) OTHER PROFESSIONALS (TECHNICIAN, PROGRAMMER, ETC.)	0.60	0.00	0.00	3,501	
3. (4) GRADUATE STUDENTS				80,000	
4. (0) UNDERGRADUATE STUDENTS				0	
5. (0) SECRETARIAL - CLERICAL (IF CHARGED DIRECTLY)				0	
6. (0) OTHER				0	
TOTAL SALARIES AND WAGES (A + B)				134,331	
C. FRINGE BENEFITS (IF CHARGED AS DIRECT COSTS)				68,951	
TOTAL SALARIES, WAGES AND FRINGE BENEFITS (A + B + C)				203,282	
D. EQUIPMENT (LIST ITEM AND DOLLAR AMOUNT FOR EACH ITEM EXCEEDING $5,000.)					
TOTAL EQUIPMENT				0	
E. TRAVEL 1. DOMESTIC (INCL. CANADA, MEXICO AND U.S. POSSESSIONS)				5,000	
2. INTERNATIONAL				2,200	
F. PARTICIPANT SUPPORT COSTS 1. STIPENDS $ 0 2. TRAVEL 0 3. SUBSISTENCE 0 4. OTHER 0					
TOTAL NUMBER OF PARTICIPANTS (0) TOTAL PARTICIPANT COSTS				0	
G. OTHER DIRECT COSTS					
1. MATERIALS AND SUPPLIES				4,500	
2. PUBLICATION COSTS/DOCUMENTATION/DISSEMINATION				0	
3. CONSULTANT SERVICES				0	
4. COMPUTER SERVICES				0	
5. SUBAWARDS				75,633	
6. OTHER				10,000	
TOTAL OTHER DIRECT COSTS				90,133	
H. TOTAL DIRECT COSTS (A THROUGH G)				300,615	
I. INDIRECT COSTS (F&A)(SPECIFY RATE AND BASE) MTDC (Rate: 51.5000, Base: 178342)					
TOTAL INDIRECT COSTS (F&A)				91,846	
J. TOTAL DIRECT AND INDIRECT COSTS (H + I)				392,461	
K. RESIDUAL FUNDS				0	
L. AMOUNT OF THIS REQUEST (J) OR (J MINUS K)				392,461	

M. COST SHARING PROPOSED LEVEL $ 0	AGREED LEVEL IF DIFFERENT $		
PI/PD NAME Hsinchun Chen	FOR NSF USE ONLY INDIRECT COST RATE VERIFICATION		
ORG. REP. NAME* Mary Gerrow	Date Checked	Date Of Rate Sheet	Initials - ORG

*ELECTRONIC SIGNATURES REQUIRED FOR REVISED BUDGET

SUMMARY PROPOSAL BUDGET

YEAR 3

ORGANIZATION **University of Arizona**	FOR NSF USE ONLY PROPOSAL NO.	DURATION (months)		
		Proposed	Granted	
PRINCIPAL INVESTIGATOR / PROJECT DIRECTOR **Hsinchun Chen**	AWARD NO.			

A. SENIOR PERSONNEL: PI/PD, Co-PI's, Faculty and Other Senior Associates (List each separately with title, A.7. show number in brackets)	NSF Funded Person-months CAL	ACAD	SUMR	Funds Requested By proposer	Funds granted by NSF (if different)
1. **Hsinchun Chen - PI**	0.00	0.36	1.00	**36,688**	
2. **Salim Hariri - Co-PI**	0.00	0.06	0.95	**14,142**	
3.					
4.					
5.					
6. (**0**) OTHERS (LIST INDIVIDUALLY ON BUDGET JUSTIFICATION PAGE)	0.00	0.00	0.00	**0**	
7. (**2**) TOTAL SENIOR PERSONNEL (1 - 6)	0.00	0.42	1.95	**50,830**	
B. OTHER PERSONNEL (SHOW NUMBERS IN BRACKETS)					
1. (**0**) POST DOCTORAL SCHOLARS	0.00	0.00	0.00	**0**	
2. (**1**) OTHER PROFESSIONALS (TECHNICIAN, PROGRAMMER, ETC.)	0.60	0.00	0.00	**3,501**	
3. (**4**) GRADUATE STUDENTS				**80,000**	
4. (**0**) UNDERGRADUATE STUDENTS				**0**	
5. (**0**) SECRETARIAL - CLERICAL (IF CHARGED DIRECTLY)				**0**	
6. (**0**) OTHER				**0**	
TOTAL SALARIES AND WAGES (A + B)				**134,331**	
C. FRINGE BENEFITS (IF CHARGED AS DIRECT COSTS)				**68,951**	
TOTAL SALARIES, WAGES AND FRINGE BENEFITS (A + B + C)				**203,282**	
D. EQUIPMENT (LIST ITEM AND DOLLAR AMOUNT FOR EACH ITEM EXCEEDING $5,000.)					
TOTAL EQUIPMENT				**0**	
E. TRAVEL 1. DOMESTIC (INCL. CANADA, MEXICO AND U.S. POSSESSIONS)				**5,000**	
2. INTERNATIONAL				**2,200**	
F. PARTICIPANT SUPPORT COSTS 1. STIPENDS $ **0** 2. TRAVEL **0** 3. SUBSISTENCE **0** 4. OTHER **0**					
TOTAL NUMBER OF PARTICIPANTS (**0**) TOTAL PARTICIPANT COSTS				**0**	
G. OTHER DIRECT COSTS					
1. MATERIALS AND SUPPLIES				**4,500**	
2. PUBLICATION COSTS/DOCUMENTATION/DISSEMINATION				**0**	
3. CONSULTANT SERVICES				**0**	
4. COMPUTER SERVICES				**0**	
5. SUBAWARDS				**77,747**	
6. OTHER				**10,000**	
TOTAL OTHER DIRECT COSTS				**92,247**	
H. TOTAL DIRECT COSTS (A THROUGH G)				**302,729**	
I. INDIRECT COSTS (F&A)(SPECIFY RATE AND BASE) **MTDC (Rate: 51.5000, Base: 178342)**					
TOTAL INDIRECT COSTS (F&A)				**91,846**	
J. TOTAL DIRECT AND INDIRECT COSTS (H + I)				**394,575**	
K. RESIDUAL FUNDS				**0**	
L. AMOUNT OF THIS REQUEST (J) OR (J MINUS K)				**394,575**	

M. COST SHARING PROPOSED LEVEL $ **0**	AGREED LEVEL IF DIFFERENT $		
PI/PD NAME **Hsinchun Chen**	FOR NSF USE ONLY INDIRECT COST RATE VERIFICATION		
ORG. REP. NAME* **Mary Gerrow**	Date Checked	Date Of Rate Sheet	Initials - ORG

***ELECTRONIC SIGNATURES REQUIRED FOR REVISED BUDGET**

SUMMARY PROPOSAL BUDGET

Cumulative

	FOR NSF USE ONLY		
ORGANIZATION University of Arizona	PROPOSAL NO.	DURATION (months)	
		Proposed	Granted
PRINCIPAL INVESTIGATOR / PROJECT DIRECTOR Hsinchun Chen	AWARD NO.		

A. SENIOR PERSONNEL: PI/PD, Co-PI's, Faculty and Other Senior Associates (List each separately with title, A.7. show number in brackets)	NSF Funded Person-months			Funds Requested By proposer	Funds granted by NSF (if different)
	CAL	ACAD	SUMR		
1. Hsinchun Chen - PI	0.00	1.08	3.00	110,064	
2. Salim Hariri - Co-PI	0.00	0.18	2.85	42,426	
3.					
4.					
5.					
6. ( ) OTHERS (LIST INDIVIDUALLY ON BUDGET JUSTIFICATION PAGE)	0.00	0.00	0.00	0	
7. (2) TOTAL SENIOR PERSONNEL (1 - 6)	0.00	1.26	5.85	152,490	
B. OTHER PERSONNEL (SHOW NUMBERS IN BRACKETS)					
1. (0) POST DOCTORAL SCHOLARS	0.00	0.00	0.00	0	
2. (3) OTHER PROFESSIONALS (TECHNICIAN, PROGRAMMER, ETC.)	1.80	0.00	0.00	10,503	
3. (12) GRADUATE STUDENTS				240,000	
4. (0) UNDERGRADUATE STUDENTS				0	
5. (0) SECRETARIAL - CLERICAL (IF CHARGED DIRECTLY)				0	
6. (0) OTHER				0	
TOTAL SALARIES AND WAGES (A + B)				402,993	
C. FRINGE BENEFITS (IF CHARGED AS DIRECT COSTS)				206,853	
TOTAL SALARIES, WAGES AND FRINGE BENEFITS (A + B + C)				609,846	
D. EQUIPMENT (LIST ITEM AND DOLLAR AMOUNT FOR EACH ITEM EXCEEDING $5,000.) $ 8,423					
TOTAL EQUIPMENT				8,423	
E. TRAVEL 1. DOMESTIC (INCL. CANADA, MEXICO AND U.S. POSSESSIONS)				14,400	
2. INTERNATIONAL				6,600	
F. PARTICIPANT SUPPORT COSTS 1. STIPENDS $ 0 2. TRAVEL 0 3. SUBSISTENCE 0 4. OTHER 0					
TOTAL NUMBER OF PARTICIPANTS (0) TOTAL PARTICIPANT COSTS				0	
G. OTHER DIRECT COSTS					
1. MATERIALS AND SUPPLIES				13,500	
2. PUBLICATION COSTS/DOCUMENTATION/DISSEMINATION				0	
3. CONSULTANT SERVICES				0	
4. COMPUTER SERVICES				0	
5. SUBAWARDS				224,849	
6. OTHER				30,000	
TOTAL OTHER DIRECT COSTS				268,349	
H. TOTAL DIRECT COSTS (A THROUGH G)				907,618	
I. INDIRECT COSTS (F&A)(SPECIFY RATE AND BASE)					
TOTAL INDIRECT COSTS (F&A)				288,104	
J. TOTAL DIRECT AND INDIRECT COSTS (H + I)				1,195,722	
K. RESIDUAL FUNDS				0	
L. AMOUNT OF THIS REQUEST (J) OR (J MINUS K)				1,195,722	

M. COST SHARING PROPOSED LEVEL $ 0	AGREED LEVEL IF DIFFERENT $		
PI/PD NAME Hsinchun Chen	FOR NSF USE ONLY INDIRECT COST RATE VERIFICATION		
ORG. REP. NAME* Mary Gerrow	Date Checked	Date Of Rate Sheet	Initials - ORG

C *ELECTRONIC SIGNATURES REQUIRED FOR REVISED BUDGET

Budget Justification

I. Personnel:

PIs and Co-PIs:

Hsinchun Chen, Ph.D.: PI. Dr. Chen is McClelland Professor of Management Information Systems and Director of the Artificial Intelligence Lab at the University of Arizona. He will serve as project PI and supervise the overall progress of the proposed research program. He will contribute his extensive research expertise in social media analytics, information systems, data analysis, and visualization. He also has an understanding of the privacy and security implications of the proposed project and will uphold the standard of academic excellence and the open source nature of the proposed research. He will hold regular update meetings with project personnel to ensure that research milestones are met. Dr. Chen will commit 11.3% time each project year.

Salim Hariri, Ph.D.: Dr. Hariri is Professor of Electrical and Computer Engineering and the Director of the NSF Autonomic Computing Laboratory at the University of Arizona. As Co-PI, he will be responsible for overseeing the design and engineering of the IRC honeypot collection and analytics environment (which will support the autonomic monitoring component of the project). He will also provide guidance in the analysis of the data it collects, and supervise the Engineering and Computing Science graduate student(s). Dr. Hariri will commit 8.46% time each project year.

Other Personnel:

Catherine Larson (senior personnel), MSLIS, is associate research scientist and associate director of the Artificial Intelligence Lab at the University of Arizona. Her areas of expertise include digital library and content management, user studies, and information privacy. As project coordinator, she will help with establishing and tracking milestones, creating project documentation, overseeing data collection efforts, and coordinating user studies; 5% time.

Graduate Students (Research Assistants and Associates): Graduate students (3 to 5 students for a total of 1.3 FTE per project year) with expertise in cyber security social media analytics, social network analysis, and visualization will be assigned. They will conduct all research and development activities directly under the supervision of Drs. Chen and Hariri. More junior graduate students will be assigned to the project to assist the senior Ph.D. students, especially with tasks such as data collection and parsing, programming, and coding, and to be mentored by them. We expect that several part-time students will work on this project, rotating in as their expertise is needed, for a total of 1.63 FTE each project year for the University of Arizona. Situated in the Southwest, the University of Arizona has a higher proportion of Hispanic students. In addition, Tucson is adjacent to two large military bases – Fort Huachuca (Army) and Davis-Monthan (Air Force) – which produce a significant number of veteran students at the university. When hiring, we will recruit heavily and make special efforts to reach out to the Hispanic, veteran, and other under-represented student groups (e.g., women; African Americans) through outreach to university student centers and to relevant classes.

II. Fringe Benefit Rates:

Faculty and Appointed Staff: Chen, Hariri, Larson: 31.2%

Graduate Students (all levels): Benefits: 6.7%, Tuition remission: 58.3%

Actual rates in place during the time of the award will be charged.

III. Capitalized equipment:

Funding is requested for a storage server for data collection. Significant storage will be needed throughout the project to support data collection activities all three years. To enhance the storage capacity, funding from the Materials and Supplies budget will also be used. The spidering (collection) activities will be carried out on individual workstations, which per regulations are not charged to the project. Estimates are approximate, based on currently available costs for comparable equipment. The specific make and models to be purchased will be specified at the time of award in order to take advantage of new capacities and any price specials available at the time.

IV. Travel:

Travel funds are requested to support attendance at the first PI meeting (required) scheduled after the award is made. Funding is also requested for the PI, Co-PIs, and research associates (when appropriate) to travel for project dissemination purposes such as paper and poster presentation at relevant conferences. Funding is requested for both domestic and foreign travel, as we are frequently invited to present at international meetings.

V. Other Direct Costs:

Supplies and materials: Funding is requested for project-specific storage media, backup tapes, reference manuals applicable to this project, software licenses, and miscellaneous small equipment (such as memory upgrades, hard drives, and the like) which are specific to this project. Considerable data storage may be needed. Some funding may also be used for the creation of “marketing” materials related to project dissemination and obtaining community input.

Data licensing fee: This funding will cover the costs of data subscription and licensing fees for data related to IRC channels for the forums that will be identified and collected, as well as for any fees associated with establishing a honeypot environment, and if needed, for our affiliation with an open source repository.

Subcontract: Dr. Tom Holt, Michigan State University, will serve as the main lead for the social/behavioral (SBE) aspects of the research. He will provide guidance and direction in the analysis of the information collected through the honeypots. A complete subcontract budget and statement of work has been provided separately in this proposal.

VI. Indirect Costs:

Indirect Costs are calculated in accordance with the university’s federally negotiated indirect cost rate. The indirect cost rate is charged against the Modified Total Direct Cost (MTDC), which includes all direct costs except capital equipment and the tuition remission portion of fringe benefits, and also includes only the first $25,000 per subcontract. The indirect cost rate is 51.5% for all years.

SUMMARY PROPOSAL BUDGET

YEAR 1

	FOR NSF USE ONLY		
ORGANIZATION **Michigan State University**	PROPOSAL NO.	DURATION (months)	
		Proposed	Granted
PRINCIPAL INVESTIGATOR / PROJECT DIRECTOR **Thomas Holt**	AWARD NO.		

A. SENIOR PERSONNEL: PI/PD, Co-PI's, Faculty and Other Senior Associates (List each separately with title, A.7. show number in brackets)	NSF Funded Person-months			Funds Requested By proposer	Funds granted by NSF (if different)
	CAL	ACAD	SUMR		
1. **Thomas Holt - PI**	0.00	0.90	0.99	**18,602**	
2.					
3.					
4.					
5.					
6. (**0**) OTHERS (LIST INDIVIDUALLY ON BUDGET JUSTIFICATION PAGE)	0.00	0.00	0.00	**0**	
7. (**1**) TOTAL SENIOR PERSONNEL (1 - 6)	0.00	0.90	0.99	**18,602**	
B. OTHER PERSONNEL (SHOW NUMBERS IN BRACKETS)					
1. (**0**) POST DOCTORAL SCHOLARS	0.00	0.00	0.00	**0**	
2. (**0**) OTHER PROFESSIONALS (TECHNICIAN, PROGRAMMER, ETC.)	0.00	0.00	0.00	**0**	
3. (**1**) GRADUATE STUDENTS				**14,235**	
4. (**1**) UNDERGRADUATE STUDENTS				**1,700**	
5. (**0**) SECRETARIAL - CLERICAL (IF CHARGED DIRECTLY)				**0**	
6. (**0**) OTHER				**0**	
TOTAL SALARIES AND WAGES (A + B)				**34,537**	
C. FRINGE BENEFITS (IF CHARGED AS DIRECT COSTS)				**15,125**	
TOTAL SALARIES, WAGES AND FRINGE BENEFITS (A + B + C)				**49,662**	
D. EQUIPMENT (LIST ITEM AND DOLLAR AMOUNT FOR EACH ITEM EXCEEDING $5,000.)					
TOTAL EQUIPMENT				**0**	
E. TRAVEL 1. DOMESTIC (INCL. CANADA, MEXICO AND U.S. POSSESSIONS)				**0**	
2. INTERNATIONAL				**0**	
F. PARTICIPANT SUPPORT COSTS					
1. STIPENDS $ **0**					
2. TRAVEL **0**					
3. SUBSISTENCE **0**					
4. OTHER **0**					
TOTAL NUMBER OF PARTICIPANTS (**0**) TOTAL PARTICIPANT COSTS				**0**	
G. OTHER DIRECT COSTS					
1. MATERIALS AND SUPPLIES				**0**	
2. PUBLICATION COSTS/DOCUMENTATION/DISSEMINATION				**0**	
3. CONSULTANT SERVICES				**0**	
4. COMPUTER SERVICES				**0**	
5. SUBAWARDS				**0**	
6. OTHER				**0**	
TOTAL OTHER DIRECT COSTS				**0**	
H. TOTAL DIRECT COSTS (A THROUGH G)				**49,662**	
I. INDIRECT COSTS (F&A)(SPECIFY RATE AND BASE) **MTDC (Rate: 53.5000, Base: 40760)**					
TOTAL INDIRECT COSTS (F&A)				**21,807**	
J. TOTAL DIRECT AND INDIRECT COSTS (H + I)				**71,469**	
K. RESIDUAL FUNDS				**0**	
L. AMOUNT OF THIS REQUEST (J) OR (J MINUS K)				**71,469**	

M. COST SHARING PROPOSED LEVEL $ **0**	AGREED LEVEL IF DIFFERENT $		
PI/PD NAME **Thomas Holt**	FOR NSF USE ONLY INDIRECT COST RATE VERIFICATION		
ORG. REP. NAME* **Mary Gerrow**	Date Checked	Date Of Rate Sheet	Initials - ORG

***ELECTRONIC SIGNATURES REQUIRED FOR REVISED BUDGET**

SUMMARY PROPOSAL BUDGET

YEAR 2

ORGANIZATION **Michigan State University**	**FOR NSF USE ONLY** PROPOSAL NO.	DURATION (months) Proposed	Granted
PRINCIPAL INVESTIGATOR / PROJECT DIRECTOR **Thomas Holt**	AWARD NO.		

A. SENIOR PERSONNEL: PI/PD, Co-PI's, Faculty and Other Senior Associates (List each separately with title, A.7. show number in brackets)	NSF Funded Person-months CAL	ACAD	SUMR	Funds Requested By proposer	Funds granted by NSF (if different)
1. **Thomas Holt - PI**	0.00	0.90	0.99	**18,974**	
2.					
3.					
4.					
5.					
6. (**0**) OTHERS (LIST INDIVIDUALLY ON BUDGET JUSTIFICATION PAGE)	0.00	0.00	0.00	**0**	
7. (**1**) TOTAL SENIOR PERSONNEL (1 - 6)	0.00	0.90	0.99	**18,974**	
B. OTHER PERSONNEL (SHOW NUMBERS IN BRACKETS)					
1. (**0**) POST DOCTORAL SCHOLARS	0.00	0.00	0.00	**0**	
2. (**0**) OTHER PROFESSIONALS (TECHNICIAN, PROGRAMMER, ETC.)	0.00	0.00	0.00	**0**	
3. (**1**) GRADUATE STUDENTS				**14,662**	
4. (**1**) UNDERGRADUATE STUDENTS				**1,768**	
5. (**0**) SECRETARIAL - CLERICAL (IF CHARGED DIRECTLY)				**0**	
6. (**0**) OTHER				**0**	
TOTAL SALARIES AND WAGES (A + B)				**35,404**	
C. FRINGE BENEFITS (IF CHARGED AS DIRECT COSTS)				**15,716**	
TOTAL SALARIES, WAGES AND FRINGE BENEFITS (A + B + C)				**51,120**	
D. EQUIPMENT (LIST ITEM AND DOLLAR AMOUNT FOR EACH ITEM EXCEEDING $5,000.)					
TOTAL EQUIPMENT				**0**	
E. TRAVEL 1. DOMESTIC (INCL. CANADA, MEXICO AND U.S. POSSESSIONS)				**1,379**	
2. INTERNATIONAL				**0**	
F. PARTICIPANT SUPPORT COSTS 1. STIPENDS $ **0** 2. TRAVEL **0** 3. SUBSISTENCE **0** 4. OTHER **0**					
TOTAL NUMBER OF PARTICIPANTS (**0**) TOTAL PARTICIPANT COSTS				**0**	
G. OTHER DIRECT COSTS					
1. MATERIALS AND SUPPLIES				**0**	
2. PUBLICATION COSTS/DOCUMENTATION/DISSEMINATION				**0**	
3. CONSULTANT SERVICES				**0**	
4. COMPUTER SERVICES				**0**	
5. SUBAWARDS				**0**	
6. OTHER				**0**	
TOTAL OTHER DIRECT COSTS				**0**	
H. TOTAL DIRECT COSTS (A THROUGH G)				**52,499**	
I. INDIRECT COSTS (F&A)(SPECIFY RATE AND BASE) **MTDC (Rate: 53.5000, Base: 43241)**					
TOTAL INDIRECT COSTS (F&A)				**23,134**	
J. TOTAL DIRECT AND INDIRECT COSTS (H + I)				**75,633**	
K. RESIDUAL FUNDS				**0**	
L. AMOUNT OF THIS REQUEST (J) OR (J MINUS K)				**75,633**	

M. COST SHARING PROPOSED LEVEL $ **0**	AGREED LEVEL IF DIFFERENT $

PI/PD NAME **Thomas Holt**	**FOR NSF USE ONLY** INDIRECT COST RATE VERIFICATION		
ORG. REP. NAME* **Mary Gerrow**	Date Checked	Date Of Rate Sheet	Initials - ORG

***ELECTRONIC SIGNATURES REQUIRED FOR REVISED BUDGET**

SUMMARY PROPOSAL BUDGET

YEAR 3

ORGANIZATION Michigan State University	**FOR NSF USE ONLY** PROPOSAL NO.	DURATION (months) Proposed	 Granted
PRINCIPAL INVESTIGATOR / PROJECT DIRECTOR Thomas Holt	AWARD NO.		

A. SENIOR PERSONNEL: PI/PD, Co-PI's, Faculty and Other Senior Associates (List each separately with title, A.7. show number in brackets)	NSF Funded Person-months CAL	 ACAD	 SUMR	Funds Requested By proposer	Funds granted by NSF (if different)
1. Thomas Holt - PI	0.00	0.90	0.99	19,353	
2.					
3.					
4.					
5.					
6. (0) OTHERS (LIST INDIVIDUALLY ON BUDGET JUSTIFICATION PAGE)	0.00	0.00	0.00	0	
7. (1) TOTAL SENIOR PERSONNEL (1 - 6)	0.00	0.90	0.99	19,353	
B. OTHER PERSONNEL (SHOW NUMBERS IN BRACKETS)					
1. (0) POST DOCTORAL SCHOLARS	0.00	0.00	0.00	0	
2. (0) OTHER PROFESSIONALS (TECHNICIAN, PROGRAMMER, ETC.)	0.00	0.00	0.00	0	
3. (1) GRADUATE STUDENTS				15,101	
4. (1) UNDERGRADUATE STUDENTS				1,839	
5. (0) SECRETARIAL - CLERICAL (IF CHARGED DIRECTLY)				0	
6. (0) OTHER				0	
TOTAL SALARIES AND WAGES (A + B)				36,293	
C. FRINGE BENEFITS (IF CHARGED AS DIRECT COSTS)				16,333	
TOTAL SALARIES, WAGES AND FRINGE BENEFITS (A + B + C)				52,626	
D. EQUIPMENT (LIST ITEM AND DOLLAR AMOUNT FOR EACH ITEM EXCEEDING $5,000.)					
TOTAL EQUIPMENT				0	
E. TRAVEL 1. DOMESTIC (INCL. CANADA, MEXICO AND U.S. POSSESSIONS)				1,379	
2. INTERNATIONAL				0	
F. PARTICIPANT SUPPORT COSTS 1. STIPENDS $ 0 2. TRAVEL 0 3. SUBSISTENCE 0 4. OTHER 0					
TOTAL NUMBER OF PARTICIPANTS (0) TOTAL PARTICIPANT COSTS				0	
G. OTHER DIRECT COSTS					
1. MATERIALS AND SUPPLIES				0	
2. PUBLICATION COSTS/DOCUMENTATION/DISSEMINATION				0	
3. CONSULTANT SERVICES				0	
4. COMPUTER SERVICES				0	
5. SUBAWARDS				0	
6. OTHER				0	
TOTAL OTHER DIRECT COSTS				0	
H. TOTAL DIRECT COSTS (A THROUGH G)				54,005	
I. INDIRECT COSTS (F&A)(SPECIFY RATE AND BASE) MTDC (Rate: 53.5000, Base: 44377)					
TOTAL INDIRECT COSTS (F&A)				23,742	
J. TOTAL DIRECT AND INDIRECT COSTS (H + I)				77,747	
K. RESIDUAL FUNDS				0	
L. AMOUNT OF THIS REQUEST (J) OR (J MINUS K)				77,747	

M. COST SHARING PROPOSED LEVEL $ 0	AGREED LEVEL IF DIFFERENT $		
PI/PD NAME Thomas Holt	**FOR NSF USE ONLY** INDIRECT COST RATE VERIFICATION		
ORG. REP. NAME* Mary Gerrow	Date Checked	Date Of Rate Sheet	Initials - ORG

***ELECTRONIC SIGNATURES REQUIRED FOR REVISED BUDGET**

SUMMARY PROPOSAL BUDGET

Cumulative

	FOR NSF USE ONLY		
ORGANIZATION Michigan State University	PROPOSAL NO.	DURATION (months)	
		Proposed	Granted
PRINCIPAL INVESTIGATOR / PROJECT DIRECTOR Thomas Holt	AWARD NO.		

A. SENIOR PERSONNEL: PI/PD, Co-PI's, Faculty and Other Senior Associates (List each separately with title, A.7. show number in brackets)	NSF Funded Person-months			Funds Requested By proposer	Funds granted by NSF (if different)
	CAL	ACAD	SUMR		
1. Thomas Holt - PI	0.00	2.70	2.97	56,929	
2.					
3.					
4.					
5.					
6. ( ) OTHERS (LIST INDIVIDUALLY ON BUDGET JUSTIFICATION PAGE)	0.00	0.00	0.00	0	
7. (1) TOTAL SENIOR PERSONNEL (1 - 6)	0.00	2.70	2.97	56,929	
B. OTHER PERSONNEL (SHOW NUMBERS IN BRACKETS)					
1. (0) POST DOCTORAL SCHOLARS	0.00	0.00	0.00	0	
2. (0) OTHER PROFESSIONALS (TECHNICIAN, PROGRAMMER, ETC.)	0.00	0.00	0.00	0	
3. (3) GRADUATE STUDENTS				43,998	
4. (3) UNDERGRADUATE STUDENTS				5,307	
5. (0) SECRETARIAL - CLERICAL (IF CHARGED DIRECTLY)				0	
6. (0) OTHER				0	
TOTAL SALARIES AND WAGES (A + B)				106,234	
C. FRINGE BENEFITS (IF CHARGED AS DIRECT COSTS)				47,174	
TOTAL SALARIES, WAGES AND FRINGE BENEFITS (A + B + C)				153,408	
D. EQUIPMENT (LIST ITEM AND DOLLAR AMOUNT FOR EACH ITEM EXCEEDING $5,000.)					
TOTAL EQUIPMENT				0	
E. TRAVEL 1. DOMESTIC (INCL. CANADA, MEXICO AND U.S. POSSESSIONS)				2,758	
2. INTERNATIONAL				0	
F. PARTICIPANT SUPPORT COSTS					
1. STIPENDS $ 0					
2. TRAVEL 0					
3. SUBSISTENCE 0					
4. OTHER 0					
TOTAL NUMBER OF PARTICIPANTS (0) TOTAL PARTICIPANT COSTS				0	
G. OTHER DIRECT COSTS					
1. MATERIALS AND SUPPLIES				0	
2. PUBLICATION COSTS/DOCUMENTATION/DISSEMINATION				0	
3. CONSULTANT SERVICES				0	
4. COMPUTER SERVICES				0	
5. SUBAWARDS				0	
6. OTHER				0	
TOTAL OTHER DIRECT COSTS				0	
H. TOTAL DIRECT COSTS (A THROUGH G)				156,166	
I. INDIRECT COSTS (F&A)(SPECIFY RATE AND BASE)					
TOTAL INDIRECT COSTS (F&A)				68,683	
J. TOTAL DIRECT AND INDIRECT COSTS (H + I)				224,849	
K. RESIDUAL FUNDS				0	
L. AMOUNT OF THIS REQUEST (J) OR (J MINUS K)				224,849	

M. COST SHARING PROPOSED LEVEL $ 0	AGREED LEVEL IF DIFFERENT $		
PI/PD NAME Thomas Holt	FOR NSF USE ONLY INDIRECT COST RATE VERIFICATION		
ORG. REP. NAME* Mary Gerrow	Date Checked	Date Of Rate Sheet	Initials - ORG

C *ELECTRONIC SIGNATURES REQUIRED FOR REVISED BUDGET

Michigan State University - Budget Narrative

Personnel

Dr. Thomas J. Holt, an Associate Professor in the School of Criminal Justice at Michigan State University, will serve as the Co-Principal Investigator for this project. Dr. Holt will assist in the identification of various websites and CMCs of interest where individuals are discussing computer hacking, malware, and cybercrime generally, as well as aid in the development of Honeynet infrastructure utilized in the course of the project. He will also develop metrics to assess and classify the participants, and then assist in the development of algorithms and statistical techniques to automate these analysis processes. In addition, Dr. Holt will assist in the development of publications and presentations generated from this research. He is budgeted for 10% in the academic year and one month of summer salary in each year of the project. The first funding year $18,602 is requested. The year two estimate is $18,974 and year three estimate is $19,353. Annual increases for Dr. Holt are included at 2% per academic year.

In addition, a graduate student will be hired to work on this project from the Ph. D. program in the School of Criminal Justice at Michigan State University. This student will be identified based on their background and experience with social science research and statistical analyses and is budgeted for 9 months of funding during the each year of the project. A 3% increase is included each of the following two years.

Fringe

MSU utilizes the Specific Identification (SI) method to calculate fringe benefit rates. This system relies on an actual fringe rate for all personnel based on the actual rate for that employee. It is thus variable across employees. For the faculty personnel, fringe includes retirement, FICA, health, prescription, and dental.

Holt's summer fringe rate of 7.65% includes only the cost of FICA and Medicare.

Graduate assistant health care costs are included at $2,310 for year one, $2,426 for year two and $2,548 for year three of the project.

Travel

Travel is included in the budget to allow for the principal investigator to present the findings from this study at academic conferences and professional venues during years two and three of the project. This includes $485 for airfare, $170 per night for lodging, as well as $71 per day for per diem, and $50 for ground travel for two trips for a total of $2,758. (Lodging and per diem are based on travel to a major US city.)

Equipment

None.

Supplies

None.

Construction

None.

Consultant/Subcontract

None.

Other

Tuition is included for the graduate assistant as specified by the MSU Controller's Office rate schedule. This provides a total of tuition in year one is $8,902, in year two is $9,258 and $9,628 in year three of the project.

Facilities and Administration

Michigan State University On Campus F&A rate is calculated at 53.5% of Direct Costs, excluding tuition and fees. The total F&A costs are $21,807 in year one, $23,134 in year two and $23,742 in year three of the project.

The total Project Direct Costs are $156,166.

The total Project F&A Costs are $68,683.

The total Project Costs are $224,849.

Hsinchun Chen McClelland Professor of Management Information Systems; Director, Artificial Intelligence Lab, University of Arizona, Tucson, AZ 85721; http://ai.arizona.edu

Current Support

EAGER: Two Decades of Nanotechnology Development: Global Competitive Landscape and Knowledge Diffusion via ERGM and SIR Analysi**s**. PI.

NSF	$298,358	10/1/12 – 09/30/14	Sumr: 1.0 mo.

EXP-LA: Explosives and IEDs in the Dark Web: Discovery, Categorization, and Analysis. PI.

NSF	$ 797,447	12/1/07 – 11/30/12	Acad: 0.2 mo.

WMD Intent Identification and Interaction Analysis Using the Dark Web

DTRA	$1,047,652	7/14/09 – 07/19/13	Acad: 2.0 mo.

Pending Support

SBE TTP: Medium: Securing Cyber Space: Understanding the Cyber Attackers and Attacks via Social Media Analytics. PI.

NSF	$1,195,721	09/01/2013 – 08/31/16	Acad. .36 mo; Sumr, 1 mo.

Cybersecurity Scholarship-for-Service at The University of Arizona. PI.

NSF	$2,708,338	08/01/13-07/31/18	Acad. .6 mo

CI-ADDO-NEW: The GeoPolitical Web: A Social Media Collection and Infrastructure for the Computational and Social Sciences. PI.

NSF	$998,739	7/1/13 – 6/30/16	Acad. .6 mo.

Personalized Healthcare and Community Mapping. PI.

NIH/NLM	$1,582,571	7/1/12 – 6/30/15	Acad: 1.0 mo

Current and Pending Support for Salim Hariri

Current:

Investigator: Salim Hariri
Project/Proposal Title: **NSF Center for Autonomic Computing**
Source of Support: NSF
Total Award Amount: **280,000**
Total Award Period Covered: **01/15/08 -12/31/12**
Location of Project: UA
Person-Months Per Year Committed to the Project: Cal: Acad: Sumr: 1.0
Support: Current

Investigator: Salim Hariri (PI)
Project/Proposal Title: **Collaborative Research: II-NEW: An Instrumented Data Center Infrastructure for Research on Cross-Layer Autonomics**
Source of Support: NSF
Total Award Amount: $210,000
Total Award Period Covered: **10/01/09 – 09/30/12**
Location of Project: UA
Person-Months Per Year Committed to the Project: Cal: Acad: Sumr: 0.10
Support: current

Investigator: Salim Hariri (Co-PI)
Project/Proposal Title: Collaborative Research: Towards Unified Cloud Computing and Management
Source of Support: NSF
Total Award Amount: $50,000
Total Award Period Covered: **08/15/11 – 07/31/12**
Location of Project: UA
Person-Months Per Year Committed to the Project: Cal: Acad: Sumr: 0.16
Support: Current

Pending:

Role: Co-PI Investigator: Hsinchun Chen (PI)
Project/Proposal Title: SBE TTP: Medium: Securing Cyber Space: Understanding the Cyber Attackers and Attacks via Social Media Analytics.
Source of Support: NSF
Total Award Amount: $1,195,721
Total Award Period Covered: 09/01/2013 – 08/31/16
Location of Project: UA

Person-Months Per Year Committed to the Project: Cal: Acad: .33 Sumr: .95
Support: Pending

Role: Co-PI Investigator: Hsinchun Chen (PI)
Project/Proposal Title: Cybersecurity Scholarship-for-Service at The University of Arizona
Source of Support: NSF
Total Award Amount: $ 2,708,338
Total Award Period Covered: 08/01/13-07/31/18
Location of Project: UA
Person-Months Per Year Committed to the Project: Cal: Acad: .36 Sumr:
Support: Pending

Investigator: Salim Hariri (PI)
Project/Proposal Title: Co-Design for Orders of Magnitude Improvement of Navier-Stokes based Virtual Wind Tunnel Funding Agency: AFOSR Proposed Period of Performance
Source of Support: AFOSR
Total Award Amount: $5,997,234
Total Award Period Covered: 10/01/12 - 9/30/17
Location of Project: UA
Person-Months Per Year Committed to the Project: Cal: Acad: Sumr: 1.0
Support: Pending

Investigator: Salim Hariri (PI)
Project/Proposal Title: CPS: Synergy: Collaborative Research: Adaptive Control Environment in Plant Production Systems (ACEPPS)
Source of Support: NSF
Total Award Amount: $1,073,626
Total Award Period Covered: 09/15/2012 – 09/14/2016
Location of Project: UA
Person-Months Per Year Committed to the Project: Cal: Acad: Sumr: 1.0
Support: Pending

Investigator: Salim Hariri (PI)
Project/Proposal Title: CPS: Synergy: Collaborative Research: Autonomic Cyber Physical Systems for Accelerated Earth Systems Science Discovery
Source of Support: NSF
Total Award Amount: $1,360,000
Total Award Period Covered: 01/01/2013-12/31/2016
Location of Project: UA
Person-Months Per Year Committed to the Project: Cal: Acad: Sumr: 1.0
Support: Pending

Investigator: Salim Hariri (PI)
Project/Proposal Title: BioARC: Biologically-inspired Autonomic Cloud Resilient
Source of Support: NSF
Total Award Amount: $1,200,000
Total Award Period Covered: 08/10/12 – 07/31/16
Location of Project: UA
Person-Months Per Year Committed to the Project: Cal: Acad: Sumr: 1.0
Support: Pending

CURRENT AND PENDING SUPPORT – THOMAS J. HOLT

CURRENT SUPPORT

2011-2013
Holt, Thomas J. (PI). "Examining the Structure, Organization, and Processes of the International Market Stolen Data Online."
Funding: $272,891
National Institute of Justice.
YR 1 Summer 2 mo pay = .66 x 3 = 1.98 mo
YR 1 AY 10% = .10 x 9 = .9 mo
YR 2 Summer 1.5 mo pay = .5 x 3 = 1.5 mo
YR 2 AY 10% = .10 x 9 = .9 mo

2012-2013
Holt, Thomas J. (PI), Kristie R. Blevins, David Foran and Ruth W. Smith. "An Examination of the Conditions Affecting Forensic Scientists: Workplace Productivity and Occupational Stress."
Funding: $129,376
National Institute of Justice
YR 1 Summer 2 mo pay = .66 x 3 = 1.98 mo
YR 1 AY 10% = .10 x 9 = .9 mo

PENDING SUPPORT

2013-2016
Chen, Hsinchun; Hariri, Salim; Holt, Thomas J. (PI for MSU). "SBE TTP: Medium: Securing Cyber Space: Understanding the Cyber Attackers and Attacks via Social Media Analytics." (MSU Subcontract from University of Arizona.)
Funding: $268,349 for MSU subcontract (total project funding requested by UA is $1,195,721)
National Science Foundation
YR 1 Summer 1 mo pay = .66 x 2 = .99 mo
YR 1 AY 10% = .10 x 9 = .9 mo

Facilities

1. MIS Research Facilities

A. Artificial Intelligence Lab at The University of Arizona

The Artificial Intelligence Lab (AI Lab) is an internationally known research group in the areas of digital libraries, intelligent retrieval, collaborative computing, and knowledge management. The group is distinguished for its adaptation and development of scalable and practical artificial intelligence, neural networks, genetic algorithms, statistical analysis, computational linguistics, and visualization techniques. The AI Lab and the Hoffman E-Commerce Lab collaborate on projects and provide access to each other's facilities. Both also have access to facilities provided by the University of Arizona. Dr. Chen is Founding Director of the Hoffman Ecommerce Lab.

- Servers:
  - Dell Poweredge R900, Dell Poweredge 2650, Supermicro SuperBlade, and IntelSC5200 ESX/ESXI Servers as numerous virtual machines for fast data collection
  - Dell Poweredge 1650, Dell Poweredge 2950, Intel SC5600, Lian Li, Supermicro CSE-825TQ, and 3 Supermicro SuperBlade for Project Servers
  - 16-core HPC host with 64 GB Ram; Windows Server 2008
  - Dell 4 cpu Intel Xeon 1.9 GHz processors with 12 GB RAM, Hard Disk storage: 770 GB disk
  - SGI Origin2000 supercomputer, 8 processors and 1 GB RAM, running IRIX (Legacy Platform). 520 GB total storage capacity.
  - 14 other assorted single and dual cpu servers for development, staging, production, and management. 1.8 to 3+ GHz, 1 to 8 GB RAM.
- Server Hosting
  - Dedicated environmentally controlled server room with secured electronic access
- Online Storage
  - 45 TB (30 TB usable) iSCSI storage area network (SAN)
  - 4 TB NAS Box
  - Two 2 TB Direct Attached Storage Units
  - 2 direct attached storage 2 Terabyte storage arrays
  - 1 legacy storage server: Windows 2003 Enterprise Server, 2.6 Terabytes
- Backup and Archival
  - 1 Dell Tape Library
- Workstations
  - Every Lab staff member is provided with one or more workstation(s) fully equipped for their individual needs; specifications vary depending on need. Individuals working on computing-intensive applications are allocated additional workstations and resources depending on need.
  - For special applications, 6 Quad core spidering machines /workstations with quad monitors and 2 TB disk storage are available to Lab members.

B. Facilities available through the MIS Commons & MicroAge Lab

The MIS Commons & MicroAge Lab is a teaching and research facility originally founded by Dr. Hsinchun Chen. This state-of-the-art computer lab is combined with a large multimedia classroom containing an art visual/multimedia presentation system and 26 high-end workstations. Labs and other units part of Eller College have access rights to the resources, which include:

- Servers:
  - 1 Enterprise-class VMware ESX based Xeon virtual server host, 32 GB RAM
  - 2 HP enterprise-class 4 way Windows 2000 Servers, each with 4GB memory
  - 1 Enterprise-class IBM/AIX server P660
  - 4 AMD 1700+ Secondary servers

  - 1 Pentium IV secondary server
  - 2 Pentium III secondary servers
- Server Hosting
  - Dedicated environmentally controlled server room with secured electronic access
- Online Storage
  - 1 IBM enterprise-level storage server (2 Terabytes) connected to all Enterprise class servers through a Fiber Channel Storage Area Network (SAN)
- Backup and Archival
  - 1 IBM automated tape storage and retrieval server (126 Terabytes)

2. Facilities available through the College of Engineering and Autonomic Computing Laboratory

The Autonomic Computing Laboratory (ACL), previously known as High Performance Distributed Computing (HPDC) Lab, provides a controlled environment to better understand the operations, behavior and limitations of the current technologies used to implement power efficient self-optimized, self-protect, and self-managed systems. Below are brief descriptions of current facilities.

- **Distributed Cloud/Data Center Testbed**

  The ACL bought an IBM Blade System with 168 cores with two 40 Giga Bit per second (Gbps) Infiniband switches and two Gbps Ethernet switches, 14 Xeon Processors with 12 cores, each, 24 Giga Byte Ram. The lab consists of the following testbeds:
  - Autonomia Testbed
  - Network security testebeds for both wired and wireless networks
  - HPDC Testbed.
  - Network Teaching Testbed.
  - Sensor Network Testbed.
  - AOL Connectivity Testbed.
- **4 Tera Flops Personal Supercomputer**

  High performance Personal Supercomputing (PSC) cluster capable of delivering 4 teraflops peak performance with 960 cores of streaming processors, using four nVidia Tesla C1060 GPUs (each capable of 1 Tera Flops and with 4 GB of memory, two 2.4 GHz (8 cores), and a SSD hard drive.
- **High Productivity Computing (HPC) Testbed (Microsoft/HP testbed)**

  ACL has an HP Cluster Setup that consists of 6 nodes, in a form of 12 sockets and 48 cores. The cluster is divided into 1 cluster head which has 4GB of memory node and 5 computational nodes with 8GB memory each. A high speed Gigabit network integrates the nodes together. The node has a state of the art operating system, Windows HPC server 2008. This test bed is a perfect environment for developing research in Security in Computer Clusters and Cloud computing, Virtualization, Autonomic Computing, Acceleration of Scientific Research, etc.
- **Industrial Control System Testbed**

  As energy critical infrastructures (power, water, gas and oil) are starting to modernize their Industrial Control Systems (ICS) and build what is referred to as a “Smart Grid”: a networks that use advanced computing and communications technologies to increase operational efficiency. These networks have become a prime target for cyber attacks owing to built-in vulnerabilities. To address this massive security challenge, we are building a state-of-the-art ICS Cybersecurity Testbed at the UA site of the NSF CAC. In this testbed, our goal is to integrate the Autonomic Management tools developed at Center, AVIRTEK and Raytheon to experiment with and evaluate innovative cybersecurity technologies to secure and protect ICS services. The testbed will also be a valuable tool for teaching, and training the next generation of ICS workforce to become expert in securing and protecting critical infrastructure, resources, and services.

- **Test bed for Online Monitoring and Abnormality Analysis of Cyber Attacks**

  We have set up an instrumented test bed environment using the resources in ACL to develop, implement, experiment with, and demonstrate our approach in achieving proactive detection of network attacks and self-protection of network systems and their services from these attacks. We will use the Cisco routers available in ACL as the core network's backbone routers. In addition, several Linux routers will be used as the access routers and are managed using Autonomia online monitoring and analysis engines. This testbed consists of 4 Cisco 2800 series routers, 1 Cisco 2600 series router, and 3 10/100M switches, 1 10/100/1000 Regular Switch, and 1 10/100/100 and 34 PCs, with all computers configured into 4 sub networks.

3. Facilities available through the University of Arizona

- **CCIT High Performance Computing Facility**

  The AI Lab and ACL both have access to The University of Arizona SGI Altix ICE cluster system which was ranked in June 2008 as the 237th most powerful computer in the world and as the 50th greenest in the world in electricity consumption. The SGI Altix ICE cluster has1392-core. Each node has an Intel Xeon quad-core processors and a 2GB memory per core. It uses the DDR Infiniband interconnect to form a huge shared memory cluster. This cluster has a computational power of 19.4 TeraFLOPS.

Data Management Plan

1. The Hacker Web Research Portal

We plan to create a sustainable, web-based Hacker Web Research Portal (HWRP) testbed for use by computer/information science and social/political science researchers, where data will be updated periodically (an important characteristics of CMC digital contents) and the database and portal are developed and maintained via a multi-tier web architecture.

1.1. Data Collection

All data collected for and served through the portal will be open source data of the type typically found in social media, such as forum postings, tweets, and the like, in English, Chinese, Russian, and other languages useful to the scope of research. Semi-automated spiders, tuned to the specific environment in which they are used, will run on our collection personnel's desktops and read the files to our centralized data storage stack, a high-performance blade server which is backed up daily and easy to configure for growth. A parser will place each thread and reply into a MS SQL database; database procedures will check all integrity rules. The data will also be manually spot-checked for quality to help prevent errors or incorrectly parsed data from reaching the portal. Although it is difficult to estimate the amount of data that will be collected, in previous projects (Dark Web and GeoPolitical Web) we have collected more than 10K websites, 60M+ messages from over 100 forums, and more than 1M tweets, in English, Arabic, French, German, Indonesian, Pashto, Russian, and Urdu.

Our proposed research framework (Figure 1) identifies three other categories of information we will collect: honeypot, malware analysis, and P2P network information. The collection methods for these data are extensively described in Section 4.2.2 and not repeated here. These additional sources of information will be used in our research to help in developing the tools and algorithms.

1.2. Portal and System Development

Architecturally, the proposed system technology stack will be modeled on our previous work and will be built using Java and run on Apache Tomcat Server. We will adopt popular enterprise-level open-source web frameworks for better scalability, flexibility, compatibility, and extendibility. Front-end view will use JSP, Javascript, Jquery, HTML, CSS, and Google Chart. Struts 2 MVC Framework or similar will serve as the web component tier. Translation functions will be handled through the Google Translate API. The use of MS SQL Server and JDBC JDO will allow the system to handle complex queries through the search engine Apache Lucene, a high-performance, full-text search engine written in Java. It supports ranked searching, fielded searching, searching by date and date range, and multiple languages.Work in refining multilingual searching and improved relevance ranking started in previous projects will continue here to provide the best search experience for users (precise, comprehensive, and ordered in an expected manner).

1.3. Tool Development for Analytics and User Interface Design

The analysis and translation tools (see Section 5 for examples) will allow social science researchers and security researchers and practitioners to detect and track the spread of ideas, identify important and influential cyber criminals, and recognize hacker identity signals. The flexible searching coupled with tightly integrated analysis and visualization will make the portal easier to use and results easier to interpret.

The testbed portal will serve as the main user interface of the HWRP testbed. This HTML-based web application will allow users to easily search, browse, download, translate, analyze, and discuss the Hacker Web collections. Our behind-the-scenes search translator will allow the user to type in the search term in any supported language, and retrieve matching results from all forums in supported languages. Google Translate or other translation tools (e.g., Bing) can be called up by the user to translate results into the language of choice. The interface design will be of paramount importance. Although our previous portals have been successfully accepted and adopted by users, we will be prepared to take a fresh look at the needs of this particular audience and involve our potential users in the design. Guidelines such as Nielsens' "Ten Usability Heuristics" and the ISO 9241 standard will be consulted to inform the design and ensure, for example, that portal searching is efficient and effective; presentation is clear, concise, and consistent; etc.

1.4. Documentation

Documentation will be produced which will record descriptive metadata on the data sources (forum names, known URLs, etc.) as well as administrative data (archiving data, collection personnel, etc.). The structural metadata is an organic part of each set of data collected and is what allows the reply-network to be reconstructed for each search. Documentation will be for both internal audiences (shared through our wiki/other online mechanism and our bug tracker) and for external users (shared directly on the portal).

1.5. Security and Archiving

All data collected will be stored in multiple locations and backed up nightly to our archival tape drive and archived externally (using University resources). The data itself is not sensitive, but backing it up securely will prevent the need to recollect or reparse if the original data is accidentally destroyed or lost. The tools and algorithms will be shared via an open source repository (such as SourceForge or other reputable, stable site). The testbed and portal will be replicated on an external mirror (location TBD) as an added precaution.

1.6. Access

Researchers and others will access the portal, testbed, and tools through accounts set up in response to their requests, following general website terms of use. Information which individually identifies registered users will not be shared with anyone outside of the immediate project team.

2. Dissemination

Data and research dissemination activities will include active presentation of the research and Hacker Web collection at computing, social/political, and cybercrime conferences, including the DOD's Cyber Crime Conference (we are eligible as we run a National Center of Academic Excellence in Information Assurance). Results of our research will also be shared via scholarly publications for a variety of audiences. We will make every effort to publish in top-tier journals and to reach all disciplines represented in the research community. For example, a paper in *Decision Support Systems* will be read by the information systems community; the *International Journal of Cyber Criminology* will reach criminologists and others involved in justice and law; *Social Science Computer Review* will reach a broad social sciences audience. We will also broadly disseminate the findings and availability of the data through conference presentations and through less formal channels such as newsletters, listservs, and announcements. For additional details, see Section 7.

3. Human Subjects Related Data

Our processes for interacting with and protecting human subjects follow our University requirements which in turn adhere to federal regulations established by the Department of Health and Human Services (and other agencies when applicable). All investigators are required to complete human subjects training prior to engaging with human subjects. In the case of our proposed work, our user studies will be considered human subjects research and cannot be started without an IRB-approved plan. Plans include a description and justification of the research goals and methods; key personnel; recruitment activities and materials; the intended target audience and their relevance; procedures for obtaining informed consent; all instruments, surveys, questions, tools, methods, etc. that will be used in collecting participant data; a description of how data will be stored and protected. In addition, special reviews may be needed if participants are from certain vulnerable or other populations. Our user studies are generally found to be exempt from needing full committee review, but are reviewed and approved first at the departmental level, then by designated staff in the Human Subjects Protection Office, under the Office of the VP of Research.

Data collected in our user studies will fall into these categories: 1) Participants names and consent records; 2) Demographic information (for example, age, gender, education level); 3) Other relevant personal information (e.g., comfort level with technology, use of the Internet); 4) Data collected from the study (e.g., participants' responses on questionnaires, their performance on tasks, etc.) In all cases, only the consent records (1) contain personally identifiable information; all other data (2-4) are de-identified related only through codes which cannot be traced back to individual participants. If stored online, consent records will be encrypted; if on paper, they will be securely stored in a locked file, initially in the PI's office. All data collected from human subjects research is stored for a total of six years: three years within the managing office, and an additional three years at RMA. Results of the studies will be submitted for publication.

Supplementary Document

Project Personnel and Partner Institutions

1. Hsinchun Chen; University of Arizona; PI
2. Salim Hariri; University of Arizona; Co-PI
3. Thomas Holt; Michigan State University; Subawardee; Co-PI
4. Catherine A. Larson; University of Arizona; Senior Personnel