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

SUGGESTED REVIEWERS:
Not Listed

REVIEWERS NOT TO INCLUDE:
Not Listed

Table A: List your Last Name, First Name, Middle Initial, and organizational affiliation (including considered affiliation) in the last 12 months.

A	Your Name:	Your Organizational Affiliation(s), last 12 r	Last Active Date
	Samtani, Sagar	University of South Florida	8/8/2018
		University of Arizona	7/7/2018

Table B: List names as Last Name, First Name, Middle Initial, and provide organizational affiliations, if known, for the following.

G: Your PhD Advisor(s)

T: All your PhD Thesis Advisees

P: Your Graduate Advisors

to disambiguate common names

B	Advisor/Advisee Name:	Organizational Affiliation	Optional (email, Department)
G:	Chen, Hsinchun	University of Arizona	Department of Management Information Systems

Table C: List names as Last Name, First Name, Middle Initial, and provide organizational affiliations, if known, for the following.

A: Co-authors on any book, article, report, abstract or paper (with collaboration in last 48 months; publication date may be later).

C: Collaborators on projects, such as funded grants, graduate research or others (in last 48 months).

to disambiguate common names

C	Name:	Organizational Affiliation	Optional (email, Department)	Last Active
A:	Chen, Hsinchun	University of Arizona	Department of Management Inform	2018
A:	Brown, Susan	University of Arizona	Department of Management Inform	2018
A:	Nunamaker, Jay F.	University of Arizona	Department of Management Inform	2018
A:	Patton, Mark	University of Arizona	Department of Management Inform	2018
A:	Matherly, John	Shodan		2018
A:	Yu, Shuo	University of Arizona	Department of Management Inform	2018
A:	Zhu, Hongyi	University of Arizona	Department of Management Inform	2018
A:	Chinn, Ryan	Ernst and Young		2017
A:	Williams, Ryan	University of Arizona	Department of Management Inform	2017
A:	McMahon, Emma	University of Arizona	Department of Management Inform	2017
A:	Grisham, John	University of Arizona	Department of Management Inform	2017
A:	El, Malaka	University of Arizona	Department of Management Inform	2017
A:	Benjamin, Victor	Arizona State University	Department of Management Inform	2016
A:	Larson, Cathy	University of Arizona	Department of Management Inform	2016

Table D: List editorial board, editor-in-chief and co-editors with whom you interact. An editor-in-chief should list the entire editorial board.

B: Editorial board: Name(s) of editor-in-chief and journal (in past 24 months).

E: Other Co-Editors of journals or collections with whom you directly interacted (in past 24 months).

to disambiguate common names

D	Name:	Organizational Affiliation	Journal/Collection	Last Active

Table E: List persons for whom a personal, family, or business relationship would otherwise preclude their service as a reviewer.

R: Additional names for whom some relationship would otherwise preclude their service as a reviewer.

to disambiguate common names

D	Name:	Organizational Affiliation	Optional (email, Department)	Last Active

COVER SHEET FOR PROPOSAL TO THE NATIONAL SCIENCE FOUNDATION

PROGRAM ANNOUNCEMENT/SOLICITATION NO./DUE DATE
NSF 18-554 **08/08/18**

☐ Special Exception to Deadline Date Policy

FOR NSF USE ONLY
NSF PROPOSAL NUMBER
1850362

FOR CONSIDERATION BY NSF ORGANIZATION UNIT(S) (Indicate the most specific unit known, i.e. program, division, etc.)
CNS - CRII CISE Research Initiation, (continued)

DATE RECEIVED	NUMBER OF COPIES	DIVISION ASSIGNED	FUND CODE	DUNS# (Data Universal Numbering System)	FILE LOCATION
08/08/2018	**2**	**05050000 CNS**	**026Y**	**069687242**	**08/08/2018 3:49pm**

EMPLOYER IDENTIFICATION NUMBER (EIN) OR TAXPAYER IDENTIFICATION NUMBER (TIN)
593102112

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 South Florida

AWARDEE ORGANIZATION CODE (IF KNOWN)
0015370000

ADDRESS OF AWARDEE ORGANIZATION, INCLUDING 9 DIGIT ZIP CODE
University of South Florida
3702 Spectrum Blvd.
Tampa, FL. 336129446

NAME OF PRIMARY PLACE OF PERF
Department of Information Systems and Decision Sciences

ADDRESS OF PRIMARY PLACE OF PERF, INCLUDING 9 DIGIT ZIP CODE
Department of Information Systems and Decision Sciences
4202 E. Fowler Ave., CIS 1040
Tampa ,FL ,336205500 ,US.

IS AWARDEE ORGANIZATION (Check All That Apply) ☐ SMALL BUSINESS ☐ MINORITY BUSINESS ☐ FOR-PROFIT ORGANIZATION ☐ WOMAN-OWNED BUSINESS

☐ IF THIS IS A PRELIMINARY PROPOSAL THEN CHECK HERE

TITLE OF PROPOSED PROJECT **CRII: SaTC: Identifying Emerging Threats in the Online Hacker Community for Proactive Cyber Threat Intelligence: A Diachronic Graph Convolutional Autoencoder Framework**

REQUESTED AMOUNT	PROPOSED DURATION (1-60 MONTHS)	REQUESTED STARTING DATE	SHOW RELATED PRELIMINARY PROPOSAL NO. IF APPLICABLE
$ **174,920**	**24** months	**07/01/19**	

THIS PROPOSAL INCLUDES ANY OF THE ITEMS LISTED BELOW
☐ BEGINNING INVESTIGATOR
☐ DISCLOSURE OF LOBBYING ACTIVITIES
☐ PROPRIETARY & PRIVILEGED INFORMATION
☐ HISTORIC PLACES
☐ VERTEBRATE ANIMALS IACUC App. Date _____
PHS Animal Welfare Assurance Number _____
☒ TYPE OF PROPOSAL **Research**

☐ HUMAN SUBJECTS Human Subjects Assurance Number _____
Exemption Subsection _____ or IRB App. Date _____
☐ INTERNATIONAL ACTIVITIES: COUNTRY/COUNTRIES INVOLVED
_____
☒ COLLABORATIVE STATUS
Not a collaborative proposal

PI/PD DEPARTMENT
Information Systems and Decision Science

PI/PD POSTAL ADDRESS
4202 E Fowler Ave
Tampa, FL 33620
United States

PI/PD FAX NUMBER

NAMES (TYPED)	High Degree	Yr of Degree	Telephone Number	Email Address
PI/PD NAME **Sagar Samtani**	**PhD**	**2018**	**520-971-4274**	**sagarsamtani96@gmail.com**
CO-PI/PD				
CO-PI/PD				
CO-PI/PD				
CO-PI/PD				

CERTIFICATION PAGE

Certification for Authorized Organizational Representative (or Equivalent) or Individual Applicant

By electronically signing and submitting this proposal, the Authorized Organizational Representative (AOR) 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 conflict of interest (when applicable), drug-free workplace, debarment and suspension, lobbying activities (see below), nondiscrimination, flood hazard insurance (when applicable), responsible conduct of research, organizational support, Federal tax obligations, unpaid Federal tax liability, and criminal convictions as set forth in the NSF Proposal & Award Policies & Procedures Guide (PAPPG). 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).

Certification Regarding Conflict of Interest

The AOR is required to complete certifications stating that the organization has implemented and is enforcing a written policy on conflicts of interest (COI), consistent with the provisions of PAPPG Chapter IX.A.; that, to the best of his/her knowledge, all financial disclosures required by the conflict of interest policy were made; and that conflicts of interest, if any, were, or prior to the organization's expenditure of any funds under the award, will be, satisfactorily managed, reduced or eliminated in accordance with the organization's conflict of interest policy. Conflicts that cannot be satisfactorily managed, reduced or eliminated and research that proceeds without the imposition of conditions or restrictions when a conflict of interest exists, must be disclosed to NSF via use of the Notifications and Requests Module in FastLane.

Drug Free Work Place Certification

By electronically signing the Certification Pages, the Authorized Organizational Representative (or equivalent), is providing the Drug Free Work Place Certification contained in Exhibit II-3 of the Proposal & Award Policies & Procedures 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 Certification Pages, the Authorized Organizational Representative (or equivalent) or Individual Applicant is providing the Debarment and Suspension Certification contained in Exhibit II-4 of the Proposal & Award Policies & Procedures Guide.

Certification Regarding Lobbying

This 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 Certification Pages, the Authorized Organizational Representative (or equivalent) is providing the Certification Regarding Nondiscrimination contained in Exhibit II-6 of the Proposal & Award Policies & Procedures 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 Certification Pages, the Authorized Organizational Representative (or equivalent) 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 Certification Pages, the Authorized Organizational Representative is certifying that, in accordance with the NSF Proposal & Award Policies & Procedures Guide, Chapter IX.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 AOR shall require that the language of this certification be included in any award documents for all subawards at all tiers.

CERTIFICATION PAGE - CONTINUED

Certification Regarding Organizational Support

By electronically signing the Certification Pages, the Authorized Organizational Representative (or equivalent) is certifying that there is organizational support for the proposal as required by Section 526 of the America COMPETES Reauthorization Act of 2010. This support extends to the portion of the proposal developed to satisfy the Broader Impacts Review Criterion as well as the Intellectual Merit Review Criterion, and any additional review criteria specified in the solicitation. Organizational support will be made available, as described in the proposal, in order to address the broader impacts and intellectual merit activities to be undertaken.

Certification Regarding Federal Tax Obligations

When the proposal exceeds $5,000,000, the Authorized Organizational Representative (or equivalent) is required to complete the following certification regarding Federal tax obligations. By electronically signing the Certification pages, the Authorized Organizational Representative is certifying that, to the best of their knowledge and belief, the proposing organization:
(1) has filed all Federal tax returns required during the three years preceding this certification;
(2) has not been convicted of a criminal offense under the Internal Revenue Code of 1986; and
(3) has not, more than 90 days prior to this certification, been notified of any unpaid Federal tax assessment for which the liability remains unsatisfied, unless the assessment is the subject of an installment agreement or offer in compromise that has been approved by the Internal Revenue Service and is not in default, or the assessment is the subject of a non-frivolous administrative or judicial proceeding.

Certification Regarding Unpaid Federal Tax Liability

When the proposing organization is a corporation, the Authorized Organizational Representative (or equivalent) is required to complete the following certification regarding Federal Tax Liability:

By electronically signing the Certification Pages, the Authorized Organizational Representative (or equivalent) is certifying that the corporation has no unpaid Federal tax liability that has been assessed, for which all judicial and administrative remedies have been exhausted or lapsed, and that is not being paid in a timely manner pursuant to an agreement with the authority responsible for collecting the tax liability.

Certification Regarding Criminal Convictions

When the proposing organization is a corporation, the Authorized Organizational Representative (or equivalent) is required to complete the following certification regarding Criminal Convictions:

By electronically signing the Certification Pages, the Authorized Organizational Representative (or equivalent) is certifying that the corporation has not been convicted of a felony criminal violation under any Federal law within the 24 months preceding the date on which the certification is signed.

Certification Dual Use Research of Concern

By electronically signing the certification pages, the Authorized Organizational Representative is certifying that the organization will be or is in compliance with all aspects of the United States Government Policy for Institutional Oversight of Life Sciences Dual Use Research of Concern.

AUTHORIZED ORGANIZATIONAL REPRESENTATIVE	SIGNATURE	DATE
NAME **Laura L Beagles**	**Electronic Signature**	**Aug 8 2018 3:37PM**

TELEPHONE NUMBER	EMAIL ADDRESS	FAX NUMBER
	lbeagles@usf.edu	

COVER SHEET FOR PROPOSAL TO THE NATIONAL SCIENCE FOUNDATION

FOR CONSIDERATION BY NSF ORGANIZATION UNIT(S) - continued from page 1
(Indicate the most specific unit known, i.e. program, division, etc.)

IIS - CRII CISE Research Initiation

PROJECT SUMMARY

Overview:

Computing technology has provided modern society with numerous benefits. Many private and public organizations employ complex information systems (IS) to execute financial transactions, maintain health records, and control critical infrastructure. Unfortunately, the rapid integration of IS has been met with an alarming rate of cyber-attacks conducted by malicious hackers using sophisticated exploits. Cybersecurity experts have appraised the total cost of hacking activities against major entities such as Equifax, Uber, and Yahoo! at $450B annually. To combat this societal issue, many organizations have aimed to develop timely, relevant, and actionable intelligence about emerging threats and key threat actors to enable effective cybersecurity decisions. This process, also referred to as Cyber Threat Intelligence (CTI) has quickly emerged as a key aspect of cybersecurity. Despite its value, existing CTI practices have been criticized as reactive to known exploits, rather than proactive to new and emerging threats from the hackers themselves. To combat these concerns, CTI experts have suggested proactively examining emerging exploits in the vast, international, and rapidly evolving online hacker community. The online hacker community is an attractive CTI data source as it motivates millions of hackers from the US, China, Russia, Middle East to share malicious tools and knowledge. Amongst various platforms, hacker forums offer particularly unique CTI value. Hackers have obtained exploits in forums to execute well-known breaches. Despite their untapped CTI value, hacker forums posts are unstructured, un-sanitized text. Existing CTI analytics are ill-equipped for these characteristics. Traditional text analytics require significant extensions to generate valuable CTI.

This CISE Research Initiation Initiative (CRII) proposal, carefully positioned under the Secure and Trustworthy Cyberspace (SaTC) program, proposes a novel CTI framework designed to collect and identify emerging threats from a multi-million record hacker forums. At the core of this framework, a novel computational algorithm called the Diachronic Graph Convolutional Autoencoder (D-GCAE) is proposed. The D-GCAE, rooted in emerging diachronic linguistics, network science, text mining, and deep learning methods, aims to deliver critically needed and innovative proactive CTI capabilities for identifying emerging threats in online hacker forums. This grant will enable five key activities to be pursued which would not possible otherwise: (1) collection and analysis a large test-bed of valuable hacker forum data, (2) training graduate students in advanced CTI analytics, (3) dissemination of research via high-impact research outlets, (4) integration selected results into highly-visible educational initiatives, and most importantly (5) results to become an independent researcher.

Intellectual Merit:

The proposed project shows promise in advancing knowledge not only in CTI, but for deep learning, network analysis, text mining, and social media analytics across numerous disciplines. Deep learning can be advanced by integrating graph convolutions into other popular architectures. The D-GCAE can help sociologists gain deeper insights from social networks to test relevant theories. Within text mining, the Association of Computational Linguistics (ACL) has run a text graphs workshop for over a decade. Deep learning has emerged as a novel approach to solve key issues. Finally, the proposed D-GCAE can be directly applied in other social media datasets.

Broader Impacts:

Innovative solutions for salient cybersecurity issues require inter-disciplinary efforts cutting across private and public sectors. From an academic perspective, this project will enable advanced research/education opportunities for graduate students, University of South Florida (USF) National Security Agency Center of Academic Excellence (NSA-CAE) certified courses, and the larger cybersecurity education community through the 12 universities in the Florida State University System (SUS). Beyond academia, collaborating with National Cyber Forensics Training Alliance (NCFTA) and The Society of Policing in Cyberspace (POLCYB) will allow collected data and knowledge of the proposed research will be widely disseminated to security practitioners in industry and governments.

TABLE OF CONTENTS

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

	Total No. of Pages	Page No.* (Optional)*
Cover Sheet for Proposal to the National Science Foundation		
Project Summary (not to exceed 1 page)	1	
Table of Contents	1	
Project Description (Including Results from Prior NSF Support) (not to exceed 15 pages) **(Exceed only if allowed by a specific program announcement/solicitation or if approved in advance by the appropriate NSF Assistant Director or designee)**	10	
References Cited	2	
Biographical Sketches (Not to exceed 2 pages each)	2	
Budget (Plus up to 3 pages of budget justification)	5	
Current and Pending Support	1	
Facilities, Equipment and Other Resources	1	
Special Information/Supplementary Documents (Data Management Plan, Mentoring Plan and Other Supplementary Documents)	2	
Appendix (List below.) **(Include only if allowed by a specific program announcement/ solicitation or if approved in advance by the appropriate NSF Assistant Director or designee)**		
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.

CRII: SaTC: Identifying Emerging Threats in the Online Hacker Community for Proactive Cyber Threat Intelligence: A Diachronic Graph Convolutional Autoencoder Framework

1. INTRODUCTION

Computing technology has provided modern society with numerous benefits. Many private and public organizations employ complex information systems (IS) to execute financial transactions, maintain health records, and control critical infrastructure. Unfortunately, the rapid integration of IS has been met with an alarming rate of cyber-attacks conducted by malicious hackers using sophisticated exploits. Cybersecurity experts have appraised the total cost of hacktivism, espionage, cyberwarfare, and other hacking activities against major entities such as Equifax, Uber, and Yahoo! at $450B annually (Graham 2017). To combat this societal issue, many organizations have aimed to develop timely, relevant, and actionable intelligence about emerging threats and key threat actors to enable effective cybersecurity decisions. This process, also referred to as Cyber Threat Intelligence (CTI) has quickly emerged as a key aspect of cybersecurity.

CTI is fundamentally a data-driven procedure. Many organizations collect data from Network Intrusion Detection/Prevention Systems (NIDS/NIPS), and log files generated from servers, workstations, firewalls, databases, and other internal network devices. Well-refined analytics such as event correlation, forensics, anomaly detection, and malware analysis are applied to collected data to generate intelligence about the exploits used against the networks. Selected intelligence is then disseminated to communities of interest. Despite the prevalence, maturity, and value of these methods, the data collected and analyzed are past network events. Thus, derived intelligence is inherently reactive to known exploits, rather than proactive to new and emerging threats from the hackers themselves. These shortcomings have led CTI professionals from the acclaimed SANS Institute to note that "most organizations are still reactive to alerts and incidents instead of proactively seeking out the threats" (Lee and Lee 2017). Consequently, the quantity, severity, and sophistication of exploits used in cyber-attacks increase annually.

To combat these concerns, CTI experts have suggested for organizations to proactively examining emerging exploits in the vast, international, and rapidly evolving online hacker community (Bromiley 2016; Shackleford 2018). The online hacker community is an attractive CTI data source as it motivates millions of hackers from major geo-political regions (e.g., US, China, Russia, Middle East) to share malicious tools and knowledge. Four major hacker community platforms exist: forums, DarkNet Marketplaces (DNMs), Internet-Relay-Chat (IRC) channels, and carding shops (Benjamin et al. 2015). Although each has its advantages, hacker forums offer particularly unique CTI value. Forums consistently provide metadata unavailable other hacker community platforms, namely post dates, and contain significantly richer post content. These characteristics have led to forums becoming the predominant platform for hackers to freely share and discuss malicious cyber-attack exploits (Samtani et al. 2017). Figure 1 provides an example of a hacker providing an exploit to hijack a computer's computing resources to mine Bitcoin, for other hackers to freely download and use. Such an exploit has become a significant CTI concern (Beek et al. 2018).

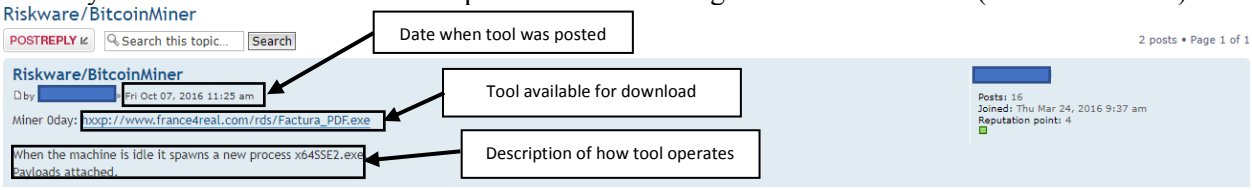

Figure 1. Example of a hacker providing a Bitcoin Miner 0-day exploit for free download

Overall, there are hundreds of forums containing tens of millions of posts made by hundreds of thousands of hackers. Hackers have obtained exploits in forums to execute well-known breaches. One notable example is Target, where hackers acquired the BlackPOS malware months before executing the attack (Kitten 2014). The severity of this event motivates the careful and systematic examination of hacker forum data to identify emerging exploits. However, hacker forums posts are unstructured, un-sanitized text. Hackers rapidly evolve their skills; thus, they develop new malware and augment existing exploits with novel functions. Compounding this issue is the unclear semantics of hacker terminology (e.g., injection can refer to memory, SQL, or process), and how they shift over time. Existing CTI analytics are ill-equipped for these unique characteristics. Moreover, text analytics approaches employed in hacker forum literature

require significant extensions to generate valuable CTI. Taken together, these challenges present numerous challenges for CTI professionals and motivate the development of innovative CTI analytics.

This Computer and Information Science and Engineering (CISE) Research Initiation Initiative (CRII) proposal, carefully positioned under the Secure and Trustworthy Cyberspace (SaTC) program, proposes a novel CTI framework designed to collect and identify emerging threats from multi-million record hacker forums. At the core of this framework, a novel computational algorithm called the Diachronic Graph Convolutional Autoencoder (D-GCAE) is proposed. The D-GCAE, rooted in emerging diachronic linguistics, network science, text mining, and deep learning methods, is positioned to deliver critically needed and innovative proactive CTI capabilities. The resources provided by this grant will enable five key activities to be pursued which would not possible otherwise: (1) collection and analysis a large test-bed of valuable hacker forum data, (2) training graduate students in advanced CTI analytics, a significant need in today's cybersecurity workforce, (3) dissemination of research via high-impact research outlets, (4) integration selected results into highly-visible educational initiatives, and most importantly (5) achievement promising results to become an independent researcher and pursue societally relevant and impactful research and funding opportunities in the future.

2. BACKGROUND OF HACKER FORUMS: PAST APPROACHES AND KEY CHALLENGES

As noted earlier, forums are the preferred hacker community platform to develop CTI due to their data richness, scale, and cybersecurity focus. DNMs contain significantly more pornography, weapon, and drugs than cybersecurity, while carding shops and IRC do not allow hackers to freely share exploits. Despite their value, collecting forums is a non-trivial task. Many forums employ anti-crawling mechanisms such as detecting bot-like collection, banning IPs, and CAPTCHA (i.e., human verification). Existing crawlers such as DARPA MEMEX and Scrapy are not designed to collect such data sources. Consequently, custom crawlers augmented with anti-crawling countermeasures are required (Benjamin et al. 2015).

Scholars have employed support vector machine (SVM), topic modeling, keyword approaches, and interviews with subject matter experts (SMEs) to identify exploits in forums (Hutchings and Holt 2014; Samtani et al. 2016; Zhao et al. 2016). Analysis reveals that hackers freely share exploits such as botnets, email hacks, exploit kits, keyloggers, web exploits, bank exploits, denial of service (DoS), and many others. However, only a few studies have gone beyond simply identifying exploits to detecting the overall and emerging exploit trends (the key goal of this proposal). The prevailing approach is monitoring the frequency of selected exploit keywords (e.g., "botnet," "crypter," "mobile malware" etc.) over a selected time period (Grisham et al. 2017; Samtani et al. 2016; Samtani et al. 2017; Sapienza et al. 2017). Major cybersecurity companies monitoring the online hacker community, such as OWL, SurfWatch Labs, or Recorded Future, also use similar summary statistics when reporting emerging threats.

Using term frequency and keyword-based approaches to identify exploit trends have several limitations. First, a term's context is ignored. For example, "*injection*" can refer to "*SQL*" or to "*memory*." Thus, results lack granularity. Second, hackers rapidly expand their exploit lexicon. As a result, new exploit terms can be overlooked. One example is Mirai, the botnet which executed the 2016 Internet DoS. Although appearing in the online hacker community months prior to the attack, it remained undetected as CTI experts were unaware of the new botnet terminology. Finally, term frequencies cannot capture the distance or relationships of a term with others. Thus, it is unclear how exploit terms evolve in usage over time (i.e., semantic shifts). These limitations necessitate an alternative approach to represent hacker forum text and highly customized novel algorithms to generate deep, relevant, timely, and thus actionable CTI.

3. PROPOSED RESEARCH

A novel research framework for identifying emerging threats from hacker forums is proposed. This framework has four major components: (1) Data Collection, (2) Time-Spell and Text-Graph Construction, (3) Emerging Threat Detection (D-GCAE), and (4) Evaluations. Each is described in detail in the following

sub-sections. Internationally-recognized cybersecurity experts National Cyber-Forensics Training Alliance (NCFTA), The Society for the Policing of Cyberspace (POLCYB), Cyber Florida have agreed to collaborate on selected activities.

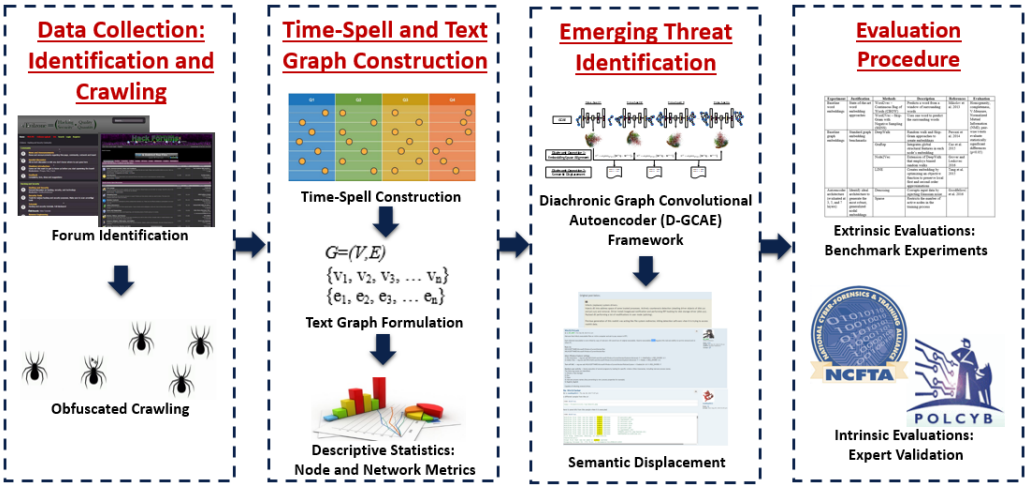

Figure 2. Research Framework for Identifying Emerging Hacker Exploits

a. Data Collection: Forum Identification and Obfuscated Crawling

The first stage in any analytics process (CTI or otherwise) is identifying the appropriate data sources for collection. Three approaches will be used to identify relevant hacker forums: suggestions from internationally-renowned cybersecurity experts, surface web and Tor search engines, and snowball identification. Using all three will help ensure a comprehensive, high-quality coverage. Only forums containing significant amounts of cybersecurity content will be collected. Forums with weapons, drugs, or pornography will be avoided. In the first strategy, PI Samtani will consult with NCFTA, a major non-profit organization focusing on the CTI sharing across 700 organizations in the private, public, and academic sectors, and POLCYB, a law enforcement entity operating at a global scale. Both have agreed in their letters of collaboration (attached) that they will recommend forums and provide keywords for input (second strategy) into surface web and Tor search engines. The identified platforms used as "seeds" for the final strategy: snowball identification. Hackers within these forums often post links to others. These links will be followed to identify more forums. All three strategies will continue until a comprehensive coverage of English, Russian, and Middle East forums are identified.

Forums will be collected with automated web spiders to ensure a comprehensive testbed. Recognizing that many forums employ anti-crawling measures, crawlers augmented with countermeasures will be developed. For example, crawling rates will be slowed to mimic human behaviors, HTTP requests and session cookies will collect more restricted forum pages. Throughout the project, additional countermeasures will be developed as needed. Although there is an upfront investment to set up the crawlers, they require minimal human intervention once they are collecting data. Customized parser programs will parse key data (e.g., post date, exploit post content) into a database. Regularity of collection to ensure fresh contents will be dictated by NCFTA and POLCYB feedback. Additional details of how this data will be managed during and after the project are provided in the data management plan. PI Samtani has nearly a half-decade in leading crawler development and periodic collections to acquire multi-million record testbeds during his time as lead research associate for Dr. Hsinchun Chen's highly-successful Scholarship-for-Service (SFS) and SaTC programs in University of Arizona's Artificial Intelligence Lab.

b. Time-Spell and Text Graph Construction: Formulation and Descriptive Statistics

The identification of emerging threats requires (1) a dataset to be split into time-spells, and (2) a representation of the hacker forum exploit texts to support in-depth analytics. Initially, exploit posts (i.e., posts with attached exploits) will be grouped into time-spells of three months. This length is consistent with

the time-frames used in many industry CTI reports. However, the appropriate time-spell length will be rigorously evaluated and validated based on expert feedback during project execution (see section 4b). With regards to representing hacker forum text, a novel Graph of Words (GoW) text graph is proposed to alleviate the issues associated with past keyword-based approaches of identifying emerging threats.

Based in graph theory, text graphs reveal relationships, patterns, and regularities within a corpora not captured in standard representations (e.g., bag of words) or word embeddings (Nastase et al. 2015). In a GoW, each word in the corpus is a node. Edges are the relationships between words. For the proposed project the hacker forum exploit text in each time-spell will be formally denoted as *G=(V,E)*. *G* will be an undirected graph, and *V* the node set, $\{v_1, v_2, v_3, \ldots v_n\}$ of all words appearing in exploit posts in that time-spell. *E* will be edge set, $\{e_1, e_2, e_3, \ldots e_n\}$. Nodes will have an edge if they appear in the same post. The GoW in each time-spell will build on the previous. To ensure a high-quality text graph, all posts will be pre-processed by lowercasing text, stripping punctuation, removing stop-words, and stemming. These steps and formulation will omit general, unrelated discussions while providing a high-quality, granular look at exploit terms and their relationships. More importantly, it provides access to an array of descriptive statistics (e.g., centrality measures) to understand overall network dynamics and pinpoint key nodes (i.e., words).

c. Emerging Threat Identification

Methodological Background and Foundation: GCNs, Autoencoders, and Diachronic Linguistics

While representing text as a GoW enables the calculation of numerous descriptive statistics not possible with traditional text models, detailed and comprehensive representations of each node are required to accurately identify how a word evolves and emerges over time. Deep learning has become the preeminent approach to learn embeddings (i.e., representations) of nodes within a network. Amongst various options, Graph Convolutional Networks (GCNs) have become the preferred method to learn embeddings from graphs for supervised learning tasks (Kipf and Welling 2017). The GCN simulates the convolutions and pooling operations in a Convolutional Neural Network (CNN) by learning a function f(X, A). X is an NxF input node feature matrix and A is the NxN adjacency matrix from graph G (N is # of nodes). The GCN outputs Z, an NxF feature matrix where F is the dimension of output embeddings. f(X, A) employs an Artificial Neural Network (ANN) layer-wise propagation function: $H^{l+1} = \sigma(AH^{(l)}W^{(l)})$. σ is an activation function, A is the adjacency matrix for G, $H^{(l)}$ is the row-wise embedding of the graph nodes in the lth layer, and $W^{(l)}$ is the a weight parameter matrix. Self-loops are captured (to prevent information loss) by adding the identity matrix I_N to A to form $\tilde{A}$. A symmetric normalization (to avoid scaling issues) is applied with $\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}$, where D is the node degree matrix. These adjustments create a generalized GCN propagation rule, $H^{l+1} = \sigma(\hat{A}H^{(l)}W^{(l)})$. More importantly, they capture a comprehensive set of information of each node (e.g., network position, connections, etc.). Scholars have used GCNs on text graphs to achieve state-of-the-art performances in link prediction (Kipf et al. 2017), node (Chen et al. 2018), document (Yasunaga et al. 2017), and part-of-speech (Demeril 2017) classification. Despite its success, GCN embeddings are task-specific (i.e., best features for mappings between the input data and pre-specified output labels in a gold-standard dataset). These embeddings do not generalize to other tasks. This drawback required adapting GCN operations into an architecture that can create generalized embeddings without a gold-standard dataset (a resource unavailable in hacker forums). One such architecture is the autoencoder.

The autoencoder is an unsupervised deep learning algorithm that learns a low-dimensional embedding for a data input (Goodfellow et al. 2016). The autoencoder has three components: the encoder, the compressed feature vector (i.e., embedding), and decoder. The encoder receives a data feature matrix *x* as input and learns a function $h(x) = \sigma(Wx + b)$. σ is the activation function, *W* is the weight matrix for layer *l* in the encoder, and *b* is the bias term. Each layer in the encoder takes the input from the previous layer and reduces its dimensionality. This process continues until a condensed representation (i.e., embedding) *h(x)* is reached. *h(x)* is used as input for the decoder, which learns a function $\hat{x} = \sigma(W^*h(x) + b)$ to reconstruct the original output. Each layer applies these operations while increasing previous layer's dimensionality. This process continues until the output dimensionality matches the original input. The Mean

Squared Error (MSE) between the input x and the reconstructed output $\hat{x}$ is backpropagated through all layers to update weights. The bottleneck in the autoencoder with minimal MSE serves as a general, condensed, latent embedding that can be used in subsequent tasks.

Despite autoencoders show promise in integrating GCN operations, they cannot capture embedding evolution, shifts, or changes in temporal datasets. While data can be split into time-spells and embeddings created in each, each time-spell's embedding space needs alignment to identify how embeddings change over time. A solution to align embedding spaces has been proposed in diachronic linguistics (i.e., historical linguistics; study of language over time) literature. A matrix of word embeddings at each time-spell, $\mathbf{W}^{(t)} \in \mathbb{R}^{d \, x \, |V|}$, where t is the time-spell is constructed (Hamilton et al. 2016). Embedding spaces are aligned across time-periods while preserving cosine similarities by optimizing the objective function $R^{(t)} = argmin_{Q^T Q=I} \left\| \mathbf{W}^{(t)} \mathbf{Q} - \mathbf{W}^{(t+1)} \right\|_F$. $\|\cdot\|_F$ denotes the Frobenius norm. Aligning spaces facilitates the measurement of semantic displacement (i.e., a word's usage evolution). Semantic displacement identifies a word's semantic shift across time-periods by measuring the cosine distance of a word at two time-periods (i.e., cosine-dist($w_t, w_{t+\Delta}$)). Such calculations offer significant value to detect emerging threats.

Proposed Approach: Diachronic Graph Convolutional Autoencoder (D-GCAE)

The aforementioned limitations of GCN's and autoencoders combined with the power of diachronic linguistics motivate the development of a novel deep learning approach to generate embeddings from GoW without a gold-standard dataset and map their shifts over multiple time-spells to identify emerging exploit trends: the D-GCAE. The D-GCAE has two major components: GCAE and diachronic operations. The GCAE incorporates custom graph convolutions into the autoencoder to create task-independent (i.e., general) embeddings for GoW's. Figure 3 illustrates GCAEs' architecture and operations.

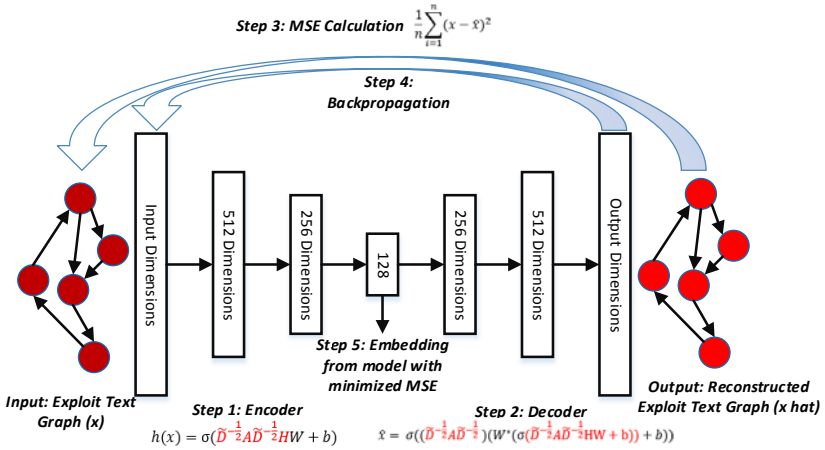

Figure 3. Architecture of the Graph Convolutional Autoencoder (GCAE)

The GCAE receives the *N*x*N* adjacency matrix *A* from the GoW *G*. While GCNs can also account for nodal attributes part of the input, nodes (i.e., words) in this formulation do not have any features. An *N*x*N* identity matrix *I* is used in lieu of a feature matrix. After taking *A* and *I* as input, the GCAE creates a low-dimensional embedding for each node by extending the standard autoencoder's encoder computation, $h(x) = \sigma(Wx + b)$, to include graph convolutions. The extended encoder is $h(x) = \sigma(\tilde{D}^{-\frac{1}{2}}A\tilde{D}^{-\frac{1}{2}}HW + b)$. σ is the activation function, *H* is the row-wise embedding of the graph nodes in a layer, *W* is the weight matrix, and *b* is the bias term. While Kipf and Welling (2017), used an alternate adjacency matrix, $\tilde{A}$, the standard adjacency matrix *A* will be used. $\tilde{A}$ includes I_N to capture self-loops and prevent information loss. In the case of GoW, however, self-loops naturally occur; a word may appear in the same post twice. Adding in *I* would result in redundant information that can skew the importance of words. Thus, *I* will be removed.

All other GCN operations will remain the same. Each encoder layer will use the extended layer-wise propagation rule and reduce the previous layer's dimensionality until the condensed embedding is reached.

The GCAE decoder will then reconstruct the original input by extending the standard autoencoder decoder function of $\sigma(W^*(h(x)) + b)$ (where h(x) is the encoder's output embedding) to $\hat{x} = \sigma((\widetilde{D}^{-\frac{1}{2}}A\widetilde{D}^{-\frac{1}{2}})(W^*(\sigma(\widetilde{D}^{-\frac{1}{2}}A\widetilde{D}^{-\frac{1}{2}}\mathrm{HW} + \mathrm{b})) + b))$ where σ is the activation function, $\widetilde{D}^{-\frac{1}{2}}A\widetilde{D}^{-\frac{1}{2}}$ is the symmetric normalization, W^* are the weights, and b is the bias. Each layer in the decoder will apply these operations and also increase the dimensionality of the previous layer. After reaching the input dimensionality, MSE between the input x and reconstructed output $\hat{x}$ will be calculated and backpropagation will adjust GCAE weights as needed. Encoding, decoding, MSE calculation, and backpropagation will continue until MSE is minimized. Embeddings will be generated and tabulated into matrix form in each time-spell (i.e., $\mathbf{W}^{(t)} \in \mathbb{R}^{\mathrm{d}\,x\,|V|}$, where t is the time-spell). The GCAE will be applied on each time-spell's GoW. Figure 4 illustrates the second D-GCAE component: diachronic operations.

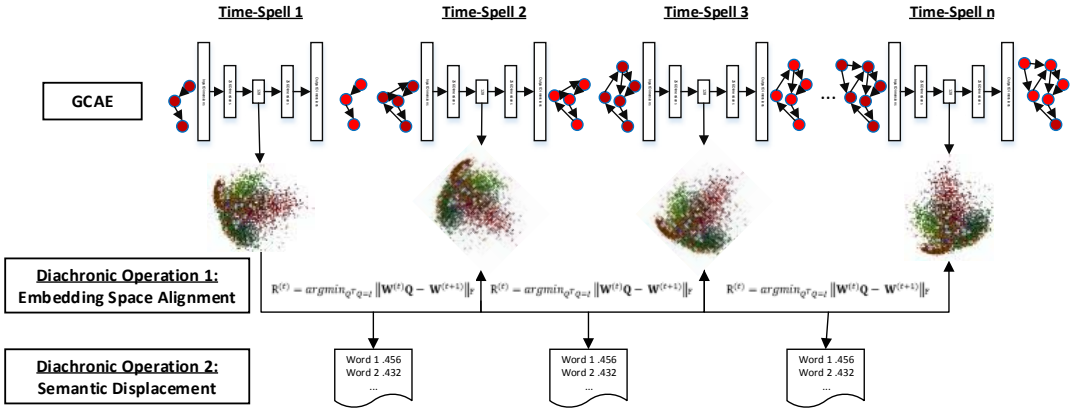

Figure 4. D-Graph Convolutional Autoencoder (D-GCAE) Procedure

The diachronic component of the D-GCAE will perform two tasks: aligns embedding spaces and computes semantic shifts for emerging threat detection. For the former, embedding spaces across time-spells will be aligned while retaining cosine similarities by optimizing the objective function $\mathrm{R}^{(t)} = argmin_{Q^TQ=I} \left\|\mathbf{W}^{(t)}\mathbf{Q} - \mathbf{W}^{(t+1)}\right\|_F$. Following alignment, emerging exploit terminology will be pinpointed by computing the cosine distance of a word's embedding at across time-periods (i.e., cosine-dist($w_t, w_{t+\Delta}$)). This computation identifies the magnitude and rate of semantic displacements.

a. Preliminary Results: Exploratory Ransomware Case Study

The proposed D-GCAE was applied on one English forum containing 24,933 posts (7,245 exploits) made by 1,418 hackers between 3/11/2010 and 10/20/2017 to conduct a preliminary exploration of emerging ransomware. Overall, 235 ransomware exploits were found, with 80-100 being posted quarterly. Text graphs at each time-spell were constructed, with each building upon the previous. Node and network level measures identified overall network dynamics. Selected results are summarized in (Table 1).

Category	Metric	2010	2011	2012	2013	2014	2015	2016	2017
Forum	# of ransomware	7	39	128	155	167	180	224	235
Network Level Metrics	# of nodes	334	569	935	1,040	1,109	1,225	1,463	1,511
	# of edges	27,742	36,918	48,017	50,663	53,070	60,058	69,129	70,117
	Graph density	0.499	0.228	0.110	0.094	0.086	0.080	0.065	0.061
	Avg. path length	1.501	1.818	1.980	2.008	2.027	2.037	2.060	2.076
Node Level Metrics	Minimum degree	10	4	2	2	2	2	2	1
	Maximum degree	220	397	606	681	700	804	1,005	1,038
	Average degree	166.12	129.764	102.710	97.429	95.708	98.054	94.503	92.809

Table 1. Topological and Node Level Descriptive Statistics Between 2010-2017

Table 1 reveals insights are not available in standard approaches to identify emerging threats. For example, the increases of ransomware in 2012 and 2016 coincided with a surge of new nodes (i.e., words) and decreases in graph density and average path length. This indicates that hackers developed ransomware with newer terminology and, in-turn, functionalities. Node level statistics reveals that each graph follows a

power law degree distribution. The top five words based on degree centrality (ransomware, name, ransom, malware, file) point to core ransomware features (e.g., deploying malware, demanding ransom). Deeper insight into specifically which features have been incorporated over the years to develop new ransomware is extracted by applying the semantic displacement computations to identify the average amount a word shifts between time-spells. The top 20 words with the highest average shift are summarized in Table 2.

Rank	Word	Amount Shifted*	Rank	Word	Amount Shifted*
1	Initiial	1.56378	11	Instal	1.53015
2	Variant	1.55563	12	Host	1.52990
3	Steal	1.55430	13	Vari	1.52902
4	Touch	1.55418	14	Strategi	1.52778
5	Organ	1.54652	15	Case	1.52593
6	Summer	1.53727	16	Financi	1.52273
7	Wolf	1.53707	17	August	1.52201
8	Mine	1.53594	18	Major	1.51968
9	Bitcoin	1.53217	19	Establish	1.51908
10	multicompon	1.53145	20	**infect**	**1.51857**

Table 2. Top Shifted Words Between 2010-2017 (* average shift per time-spell)

Words with the top 20 average shifts relate to specific ransomware functionalities (e.g., "bitcoin" for payment mechanisms, "steal," "variant," "organ" for exploitation). One top word providing actionable intelligence is "infect," which appears at rank 20 (shift of 1.51857). "Infect" appeared in various forms in a total of 19/235 (8.0551%) ransomware postings. Figure 5 provides a representative post from three time points, 2010, 2014, and 2017, to illustrate how infect's meaning has shifted.

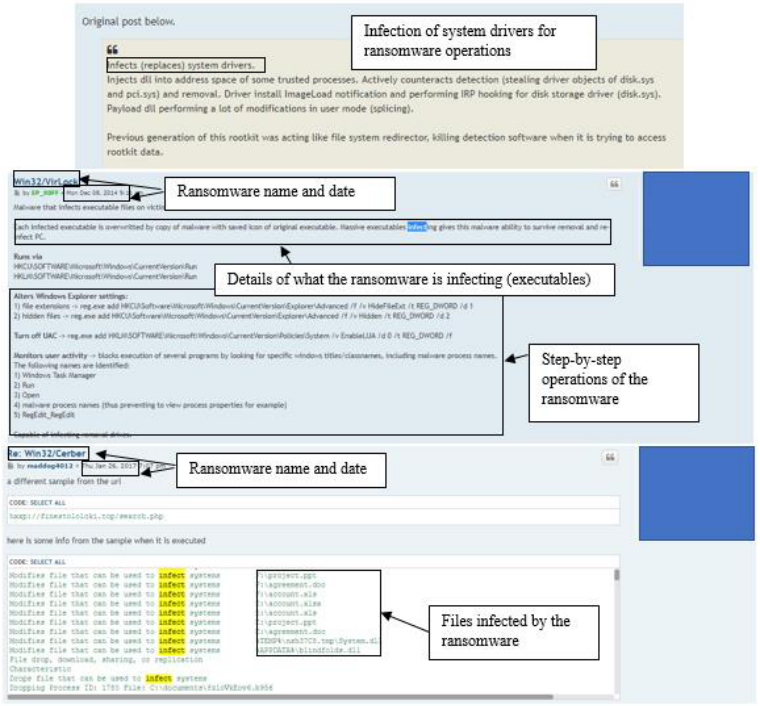

Figure 5. Three representative sample ransomware posts from 2010 (top), 2014 (middle), and 2017 (bottom) to illustrate how the term "infect" has changed in meaning.

The top post indicates that infect pertained to on injecting DLL's into the memory address space to steal driver objects. The VirLock ransomware in the middle post moves beyond infecting memory processes to causing damage to executables on a victim's machine and monitor user activities prior to encryption. Such capabilities can facilitate Advanced Persistent Threat (APT's). The final post provides Cerber, a strain which creates and/or modifies files (e.g., "project.ppt") to infect the victim's computer. Existing methods of identifying emerging threats presented cannot reveal these shifts. However, identifying them can provide

valuable tactical leads for CTI professionals (such as those monitoring hospitals, an industry widely afflicted with ransomware). One use case for the intelligence would be new rules for Security Information and Event Management (SIEM) systems. SIEM's are used by many CTI experts to monitor the status of machines on a network and detect Indicators of Compromise (IoC's). Using the intelligence provided by the third post, the SIEM would monitor and quarantine devices if the file names appear on them. While these preliminary results are promising, significant work is required to apply the D-GCAE in larger forums, identify multiple exploit trends (e.g., keyloggers), and more importantly, evaluate and validate the results.

4. EVALUATION PLAN: EXTRINSIC EXPERIMENTS AND INTRINSIC EXPERT INPUT

The development of any novel algorithm for a salient application (especially cybersecurity) requires rigorous evaluation to validate its technical superiority over benchmarks and ensure its practical utility. The proposed D-GCAE is fundamentally an unsupervised approach to (1) automatically generate word embeddings and (2) identify the semantic shifts of words. Although possessing significant descriptive capabilities, this type of framework faces numerous challenges as it pertains to the evaluation of generated outputs. Unlike other social media data, there is a lack of gold-standard datasets and well-established evaluation procedures (e.g., competing methods, metrics, etc.). Recognizing these issues, recent word embedding and diachronic linguistics literature has suggested two evaluation approaches: extrinsic and intrinsic (Bakarov 2018). Extrinsic evaluations input embeddings into a downstream task (e.g., clustering) and quantitatively compares performance of the proposed algorithm against benchmarks. Intrinsic evaluations rely on human judgements of semantic shifts. While extrinsic evaluations are favored within academia, the proposed research aims to set a solid foundation for attaining future funding which work closely with industry, law enforcement, and government to make a positive societal impact. As such, a comprehensive evaluation comprising of rigorous extrinsic experiments to evaluate embedding quality and intrinsic evaluations from NCFTA and POLCYB to identify accuracy of semantic shifts will be executed.

a. Extrinsic Evaluations: Selected Downstream Tasks and Quantitative Metrics

The definition of an appropriate downstream task is context specific. Within CTI, professionals often wish to comprehensively understand an exploit's functions and implementation by examining closely associated terms (Bromiley 2016; Shackleford 2016). These unique domain characteristics motivate the main downstream task D-GCAE's embeddings will be evaluated: k-means clustering. K-means' use of distance based metrics when creating groupings is ideal for this research. Identifying the semantic displacement of a word requires computing the distance a word travels across multiple embedding spaces. A high quality embedding is needed to ensure accurate distance calculations. This is also true for k-means. If embeddings are lower quality, then similar entities in the ground truth data will have a higher distance calculated between them, will be clustered apart, and the overall clustering performance suffers. This intuition has made k-means a popular approach to evaluating word embedding quality (Zhai et al. 2016).

In the proposed research, panels of cybersecurity experts will annotate data to develop gold-standard clusters. Inter-rater reliability will be calculated to ensure a high level of concordance between raters. From this, embeddings from various generated from the proposed method and competing state-of-the-art graph-based approaches such as LINE, DeepWalk, node2vec, and GraRep and word representation models (e.g., word2vec's skip-gram with negative sampling, continuous bag-of-words). Embeddings will be inputted into k-means and, when appropriate, other clustering algorithms (e.g., spectral). Clusters will be evaluated based on Normalized Mutual Information (NMI), homogeneity, completeness, and V-Measure. Statistical tests will identify significant differences between methods. Benchmarking in this fashion is commonly accepted practice in computer science and related disciplines. When appropriate, additional downstream tasks will be selected and used for evaluation, including classification and/or ranking/retrieving. Classification evaluation metrics will include accuracy, precision, recall, and F1. Information retrieval metrics will include Normalized Discounted Cumulative Gain and Mean Average Precision. Emerging literature within relevant conferences/journals will be monitored to identify new methods and metrics.

b. Intrinsic Evaluations: Subject Matter Expert (SME) Input and Feedback

Extrinsic testing is an excellent mechanism for quantitatively evaluating embedding quality. However, it cannot evaluate two critical aspects of the proposed D-GCAE: validation of emerging threats (e.g., semantic shifts) and selection of time-spell length. Both require expert feedback. As such, PI Samtani has requested the collaboration of NCFTA and POLCYB. Each has initiatives and/or partnerships relevant to this task. NCFTA runs the Malware and Cyber Threat program, an initiative designed to foster the sharing of malware across (1) governments: FBI, DHS, CIA, IRS; (2) finance and retail industry: Target, Best Buy, Citibank, Wells Fargo, and many others. This initiative has allowed them to generate 800+ CTI reports in the past three years. Selected data/results in the proposed project will be contributed to this initiative to validate the consistency of the emerging threats with those seen across organizations. For validating the time-spell lengths, POLCYB's extensive knowledge of the pace of cyber-crime will be leveraged. POLCYB's activities are conducted globally, with partners in India, Europe, North/South America, Africa, UAE, and Israel. As a result, they have extensive knowledge about the pace at which new exploits emerge across geo-political regions and can validate the selection of time-spell length.

5. INTEGRATION INTO EDUCATION

The proposed project will advance educational experiences for students through three mechanisms: (1) hands-on graduate research experience, (2) integration into PI Samtani's INFOSEC course, and (3) with the Cyber Florida (formerly known as FC^2). Each will provide students excellent hands-on training in advanced CTI knowledge for placement into government, industry, or academic cybersecurity positions. Graduate student(s) funded from this project will assist in data collection, analytics, and positioning/presenting results for selected outlets. Programming, database, and analytics skills valued by cybersecurity employers will be gained. From a course perspective, PI Samtani will integrate selected results into the threats and vulnerabilities sections of his INFOSEC class (~150 students per semester).

The final strategy will work with Cyber Florida (letter of collaboration attached). Founded in 2014 and headquartered at USF, Cyber Florida fosters collaboration across 12 universities in the Florida State University System (SUS) and provides resources (e.g., cyber range) to enhance cybersecurity education. Cyber Florida's efforts have gained significant traction at the national level and was designated a National Security Agency (NSA) Center for Academic Excellence (CAE) in 2014. Throughout the proposed project, selected results will be shared via their community forum, a platform which encourages the sharing of cybersecurity educational materials across the SUS. Similarly, results will be broadly disseminated through less formal educational channels such as cybersecurity blogs, newsletters, listservs, and workshops. For example, PI Samtani will use his SFS Alumni status to disseminate results to the 60+ SFS institutions.

6. DISSEMINATION OF RESEARCH: CONFERENCES, JOURNALS, AND DATA SHARING

Three strategies will be used to disseminate the proposed research to communities of interest. First, work will be presented at relevant cybersecurity and analytics conferences. Academic outlets include IEEE Intelligence and Security Informatics (ISI) (4,000+ member community), Workshop on Information Technologies and Systems (WITS) (120+ attendees annually), and Hawaiian International Conference on Systems Sciences (HICSS) (1,000+ attendees). PI Samtani has served on the program committee, volunteered, and/or presented at each. From an industry perspective, NCFTA's, POLCYB's, and Cyber Florida's Annual Conferences each draw 1,000+ attendees from law enforcement, government, and industry will be used. The second strategy will publish in premier CS/IS journals such as IEEE TKDE/IS/Computer, ACM TOIS/TMIS, and MISQ. Each has recently published or had special issues on cybersecurity, deep learning, and GCNs (Hui et al. 2016; Japkowicz and Elovici 2018; Luo et al. 2017). Thus, the generated research fit their current publishing interests. Finally data and code will be shared through the NSF-funded AZSecure Data Infrastructure Building Blocks-ISI repository designed to share dozens of multi-million record security datasets and tools (details in data management plan). Success will be assessed based on number of publications, citations, presentation attendees, and downloads.

7. PROPOSED PROJECT TIMELINE (TENTATIVE) FOR PROJECT MANAGEMENT

Activities	Year 1				Year 2			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Forum Collection	→	→	→					
D-GCAE Development			→	→	→			
SME Feedback	→	→	→	→	→	→	→	→
Educational Integration						→	→	→
CyberFlCon			→				→	→
Research Dissemination					→	→	→	→

The first half of the first year will be focused on gathering a comprehensive hacker forum testbed. NCFTA and POLCYB feedback will aide forum identification. In the second half of year one, D-GCAE will be fully developed. It is expected that results will be drawn during and after this time. The second year will refine and evaluate D-GCAE through the comprehensive plan listed above. Year 2 is also when selected results will be disseminated through journals, conferences, DIBBs-ISI and integrated into education.

8. SaTC RELEVANCE AND FOUNDATION FOR FUTURE TRANSFORMATIVE RESEARCH

Since its inception in 2012, the NSF SaTC program supports research that addresses cybersecurity and privacy. Cutting across multiple CISE divisions and drawing upon numerous technical perspectives, SaTC project PI's have made remarkable advances in adversarial data mining, anomaly detection, real-time log analytics, and many other areas. Apart from the SaTC Hacker Web project (2014 – 2017), however, there is a lack of projects focusing on CTI from the online hacker community. The proposed project will help provide a timely solutions for this societally relevant issue. Moreover, it would improve PI Samtani's capacity to write successful proposals to support promising future transformative research. Several promising research findings are anticipated from the proposed research. Each can help set the foundation for future proposals. First, this research can form the basis of multi-lingual cross-forum diachronic linguistics to identify the velocity at which new exploits emerge across the globe. Second, D-GCAE's can be applied in social network analysis to identify key emerging hackers. Finally, the detection of emerging exploits can become predictive and forecast future threats. Each has significant promise for the development of advanced and transformative CTI academic research/education and industry practices. Consequently, programs such as SaTC TTP, CORE, and EDU can be viable funding opportunities.

9. INTELLECTUAL MERIT

The proposed project shows promise in advancing knowledge not only in CTI, but for deep learning, network analysis, text mining, and social media analytics across numerous disciplines. Deep learning can be advanced by integrating graph convolutions into other popular architectures, such as Generative Adversarial Networks (GANs). The D-GCAE can help sociologists gain deeper insights from social networks to test relevant theories. Within text mining and natural language processing, the Association of Computational Linguistics (ACL) has run a text graphs workshop for over a decade. Deep learning has emerged as a novel representation learning approach to solve key issues. Finally, the proposed D-GCAE can be directly applied in other social media datasets (e.g., Twitter, Facebook, etc.) to extract new knowledge and pursue novel inquiries which would have otherwise been unavailable.

10. BROADER IMPACTS

Cybersecurity is a grand societal challenge. Innovative solutions for salient issues require inter-disciplinary efforts cutting across private and public sectors. This project will have broader impacts as it aims to involve multiple stakeholders across this breadth. From an academic perspective, this project will enable advanced research/education opportunities for graduate students, USF NSA CAE certified courses, and the larger cybersecurity education community (e.g., Florida SUS). PI Samtani will actively work to ensure that under-represented minorities and women are integrated. Beyond academia, collaborating with NCFTA and POLCYB will allow collected data and knowledge of the proposed research will be widely disseminated to security practitioners in industry and governments.

References

Bakarov, A. 2018. *A Survey of Word Embeddings Evaluation Methods*. (http://arxiv.org/abs/1801.09536).

Beek, C., Dunton, T., Grobman, S., Karlton, M., Minihane, N., Palm, C., Peterson, E., Samani, R., Schmugar, C., Sims, R., Sommer, D., Sun, B. 2018, "McAfee Labs Threat Report" (https://www.mcafee.com/enterprise/en-us/assets/reports/rp-quarterly-threats-jun-2018.pdf, accessed August 1, 2018).

Benjamin, V., Li, W., Holt, T., and Chen, H. 2015. "Exploring Threats and Vulnerabilities in Hacker Web: Forums, IRC and Carding Shops," in *2015 IEEE International Conference on Intelligence and Security Informatics (ISI)*, IEEE, May, pp. 85–90. (https://doi.org/10.1109/ISI.2015.7165944).

Bromiley, M. 2016. "Threat Intelligence: What It Is, and How to Use It Effectively," *SANS Institute*. (https://www.sans.org/reading-room/whitepapers/analyst/threat-intelligence-is-effectively-37282, accessed June 5, 2017).

Chen, J., Ma, T., & Xiao, C. 2018. "FastGCN: Fast Learning with Graph Convolutional Networks via Importance Sampling," In *2018 International Conference on Learning Representations*. Forthcoming. *arXiv preprint arXiv:1801.10247*.

Demirel, S. 2017. *Spectral Graph Convolutional Networks for Part-of-Speech Tagging* (Doctoral dissertation, Universität Koblenz-Landau).

Goodfellow, I., Bengio, Y., and Courville. 2016. *Deep Learning*. The MIT Press.

Graham, L. 2017. "Cybercrime costs the global economy $450 billion" (https://www.cnbc.com/2017/02/07/cybercrime-costs-the-global-economy-450-billion-ceo.html)

Grisham, J., Samtani, S., Patton, M., and Chen, H. 2017. "Identifying mobile malware and key threat actors in online hacker forums for proactive cyber threat intelligence," *2017 IEEE International Conference on Intelligence and Security Informatics: Security and Big Data, ISI 2017*, pp. 13–18 (https://doi.org/10.1109/ISI.2017.8004867).

Hui, K.L., Vance, A., Zhdanov, D. "Securing Digital Assets," in MIS Quarterly Research Curations, Ashley Bush, ed., http://misq.org/research-curations, May 27, 2016.

Hutchings, A., and Holt, T. J. 2015. "A Crime Script Analysis of the Online Stolen Data Market," *British Journal of Criminology* (55:3), pp. 596–614. (https://doi.org/10.1093/bjc/azu106).

Japkowicz, N and Elovici, Y. 2018. "Introduction to the Special Issue on Data Mining for Cybersecurity," *IEEE Intelligent Systems,* (33:2), 3-4.

Kipf, T. N., & Welling, M. 2017. "Semi-supervised classification with graph convolutional networks," In *International Conference on Learning Representations, arXiv preprint.* (http://arxiv.org/abs/1609.02907).

Kitten, T. 2014. "Target Malware: Exploring the Origins." (http://www.bankinfosecurity.com/interviews/intelcrawler-i-2161, accessed June 5, 2017).

Luo, Z., Liu, L., Yin, J., Li, Y., and Wu, Z. 2018. "Deep Learning of Graphs with Ngram Convolutional Neural Networks," *IEEE Transactions on Knowledge and Data Engineering,* (29:10), 2125-2139

Nastase, V., Mihalcea, R., and Radev, D. R. 2015. "A survey of graphs in natural language processing," *Natural Language Engineering*, (*21*:5), 665-698.

Samtani, S. and Chen, H. 2016. "AZSecure Hacker Assets Portal: Cyber Threat Intelligence and Malware Analysis," in *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, IEEE, September, pp. 319–321. (https://doi.org/10.1109/ISI.2016.7745500).

Samtani, S., Chinn, R., Chen, H., and Nunamaker, J. F. 2017. "Exploring Emerging Hacker Assets and Key Hackers for Proactive Cyber Threat Intelligence," *Journal of Management Information Systems* (34:4), pp. 1023–1053. (https://doi.org/10.1080/07421222.2017.1394049).

Sapienza, A., Bessi, A., Damodaran, S., Shakarian, P., Lerman, K., & Ferrara, E. 2018. "Early Warnings of Cyber Threats in Online Discussions," in *IEEE International Conference on Data Mining Workshops (ICDM),* IEEE, November, pp. 667 – 674. (https://doi.org/10.1109/ICDMW.2017.94)

Shackleford, D. 2018. "CTI in Security Operations: SANS 2018 Cyber Threat Intelligence Survey. (https://www.sans.org/reading-room/whitepapers/threats/cti-security-operations-2018-cyber-threat-intelligence-survey-38285, accessed August 1, 2018).

Yasunaga, M., Zhang, R., Meelu, K., Pareek, A., Srinivasan, K., and Radev, D. 2017. *Graph-Based Neural Multi-Document Summarization*. (http://arxiv.org/abs/1706.06681).

Zhao, Z., Sankaran, M., Ahn, G.-J., Holt, T. J., Jing, Y., and Hu, H. 2016. "Mules, Seals, and Attacking Tools: Analyzing 12 Online Marketplaces," *IEEE Security & Privacy* (14:3), pp. 32–43. (https://doi.org/10.1109/MSP.2016.46).

Zhai, M., Tan, J., and Choi, J. D. 2016. "Intrinsic and Extrinsic Evaluations of Word Embeddings," *The 30th AAAI Conference on Artificial Intelligence*, pp. 4282–4283. (http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12454/12257).

SAGAR SAMTANI

Assistant Professor of Information Systems and Decision Sciences
University of South Florida
ssamtani@usf.edu

A. PROFESSIONAL PREPARATION

University of Arizona	Tucson, AZ	Management Information Systems	Ph.D., 2018
University of Arizona	Tucson, AZ	Management Information Systems	M.S., 2014
University of Arizona	Tucson, AZ	Management Information Systems	B.S.B.A., 2013

B. APPOINTMENTS

Assistant Professor of Information Systems and Decision Sciences:
2018 – Present, Muma College of Business, University of South Florida

Adjunct Lecturer:
May 2018 – July 2018, Department of Management Information Systems, University of Arizona

Research Associate and Project Lead:
2014 – 2018, Artificial Intelligence Lab, University of Arizona, Tucson, AZ

C. SELECTED PUBLICATIONS

Five Most Relevant to Proposed Work:

1) **S. Samtani**, S. Yu, H. Zhu, M. Patton, J. Matherly, and H. Chen. Identifying Supervisory Control and Data Acquisition (SCADA) "Devices and their Vulnerabilities on the Internet of Things (IoT): A Text Mining Approach" *IEEE Intelligent Systems*. Volume 33, Issue 2, Pages 63-73, 2018.
2) **S. Samtani,** R. Chinn, H. Chen, and J. Nunamaker, "Exploring Emerging Hacker Assets and Key Hackers for Proactive Cyber Threat Intelligence" *Journal of Management Information Systems*, Volume 34, Number 3, Pages 1023-1053, 2017.
3) M. El, **S.Samtani**, H. Chen, M. Patton, and E. McMahon, "Benchmarking Vulnerability Scanners: An Experiment on SCADA Devices and Scientific Instruments," ISI 2017, Proceedings of 2017 IEEE Conference on Intelligence and Security Informatics. Beijing, China. July, 2017.
4) **S. Samtani**, S. Yu, H. Zhu, M. Patton, and H. Chen, "Identifying SCADA Vulnerabilities Using Passive and Active Vulnerability Assessment Techniques," ISI 2016, Proceedings of 2016 IEEE Conference on Intelligence and Security Informatics, Tucson, Arizona, September 2016.
5) **S. Samtani,** R. Chinn, and H. Chen., "Exploring Hacker Assets in Underground Forums," ISI 2015, Proceedings of 2015 IEEE Conference on Intelligence and Security Informatics. Baltimore, Maryland. May, 2015.

Five Other Works:

1) R. Williams, E. McMahon, **S. Samtani**, M. Patton, and H. Chen, "Identifying Vulnerabilities of Consumer Internet of Things (IoT) Devices: A Scalable Approach," ISI 2017, Proceedings of 2017 IEEE Conference on Intelligence and Security Informatics. Beijing, China. July, 2017.
2) E. McMahon, R. Williams, M. El, **S. Samtani**, M. Patton, and H. Chen, "Assessing Medical Device Vulnerabilities on the Internet of Things," ISI 2017, Proceedings of 2017 IEEE Conference on Intelligence and Security Informatics. Beijing, China. July, 2017.
3) J. Grisham, **S. Samtani,** M. Patton, and H. Chen, "Identifying Mobile Malware and Key Threat

Actors in Online Hacker Forums for Proactive Cyber Threat Intelligence," ISI 2017, Proceedings of 2017 IEEE Conference on Intelligence and Security Informatics. Beijing, China. July, 2017.

4) **S. Samtani,** K. Chinn, C. Larson, and H. Chen, "AZSecure Hacker Assets Portal: Cyber Threat Intelligence and Malware Analysis," ISI 2016, Proceedings of 2016 IEEE Conference on Intelligence and Security Informatics. Tucson, Arizona. September, 2016.

5) **S. Samtani** and H. Chen, "Using Social Network Analysis to Identify Key Hackers for Keylogging Tools in Hacker Forums," ISI 2016, Proceedings of 2016 IEEE Conference on Intelligence and Security Informatics. Tucson, Arizona. September, 2016.

D. SYNERGISTIC ACTIVITIES

1) Program Committee, Workshop on Information Technology and Systems (WITS), 2018
2) Session Chair, 2017 INFORMS Annual Meeting. Session Title: "Artificial Intelligence for Social Media Applications"
3) Designed, developed, and delivered University of Arizona's first online Cyber Threat Intelligence course for MS in Cybersecurity program
4) Led teams in hacker community collection and analytics and large-scale vulnerability assessment for proactive Cyber Threat Intelligence research as Lead Research Associate in University of Arizona's Artificial Intelligence Lab
5) Reviewer for IEEE Intelligence and Security Informatics (ISI), International Conference on Information Systems (ICIS), Springer Security Informatics, Hawaii International Conference on System Sciences (HICSS)

SUMMARY PROPOSAL BUDGET

YEAR 1

FOR NSF USE ONLY		
PROPOSAL NO.	DURATION (months) Proposed	Granted
AWARD NO.		

ORGANIZATION
University of South Florida

PRINCIPAL INVESTIGATOR / PROJECT DIRECTOR
Sagar Samtani

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. **Sagar Samtani - PI**	0.00	0.50	0.00	**8,333**	
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	0.50	0.00	**8,333**	
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. (**2**) GRADUATE STUDENTS				**34,242**	
4. (**0**) UNDERGRADUATE STUDENTS				**0**	
5. (**0**) SECRETARIAL - CLERICAL (IF CHARGED DIRECTLY)				**0**	
6. (**0**) OTHER				**0**	
TOTAL SALARIES AND WAGES (A + B)				**42,575**	
C. FRINGE BENEFITS (IF CHARGED AS DIRECT COSTS)				**2,110**	
TOTAL SALARIES, WAGES AND FRINGE BENEFITS (A + B + C)				**44,685**	
D. EQUIPMENT (LIST ITEM AND DOLLAR AMOUNT FOR EACH ITEM EXCEEDING $5,000.)					
TOTAL EQUIPMENT				**0**	
E. TRAVEL 1. DOMESTIC (INCL. U.S. POSSESSIONS)				**1,800**	
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				**6,400**	
2. PUBLICATION COSTS/DOCUMENTATION/DISSEMINATION				**0**	
3. CONSULTANT SERVICES				**0**	
4. COMPUTER SERVICES				**0**	
5. SUBAWARDS				**0**	
6. OTHER				**10,354**	
TOTAL OTHER DIRECT COSTS				**16,754**	
H. TOTAL DIRECT COSTS (A THROUGH G)				**63,239**	
I. INDIRECT COSTS (F&A)(SPECIFY RATE AND BASE) **MTDC (Rate: 49.5000, Base: 52885)**					
TOTAL INDIRECT COSTS (F&A)				**26,178**	
J. TOTAL DIRECT AND INDIRECT COSTS (H + I)				**89,417**	
K. SMALL BUSINESS FEE				**0**	
L. AMOUNT OF THIS REQUEST (J) OR (J MINUS K)				**89,417**	

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

PI/PD NAME
Sagar Samtani

ORG. REP. NAME*
Laura Beagles

FOR NSF USE ONLY		
INDIRECT COST RATE VERIFICATION		
Date Checked	Date Of Rate Sheet	Initials - ORG

 ***ELECTRONIC SIGNATURES REQUIRED FOR REVISED BUDGET**

SUMMARY PROPOSAL BUDGET

YEAR 2

FOR NSF USE ONLY		
PROPOSAL NO.	DURATION (months) Proposed	Granted
AWARD NO.		

ORGANIZATION
University of South Florida

PRINCIPAL INVESTIGATOR / PROJECT DIRECTOR
Sagar Samtani

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. **Sagar Samtani - PI**	0.00	0.50	0.00	**8,583**	
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	0.50	0.00	**8,583**	
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. (**2**) GRADUATE STUDENTS				**35,270**	
4. (**0**) UNDERGRADUATE STUDENTS				**0**	
5. (**0**) SECRETARIAL - CLERICAL (IF CHARGED DIRECTLY)				**0**	
6. (**0**) OTHER				**0**	
TOTAL SALARIES AND WAGES (A + B)				**43,853**	
C. FRINGE BENEFITS (IF CHARGED AS DIRECT COSTS)				**2,175**	
TOTAL SALARIES, WAGES AND FRINGE BENEFITS (A + B + C)				**46,028**	
D. EQUIPMENT (LIST ITEM AND DOLLAR AMOUNT FOR EACH ITEM EXCEEDING $5,000.)					
TOTAL EQUIPMENT				**0**	
E. TRAVEL 1. DOMESTIC (INCL. U.S. POSSESSIONS)				**3,200**	
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				**11,907**	
TOTAL OTHER DIRECT COSTS				**11,907**	
H. TOTAL DIRECT COSTS (A THROUGH G)				**61,135**	
I. INDIRECT COSTS (F&A)(SPECIFY RATE AND BASE) **MTDC (Rate: 49.5000, Base: 49228)**					
TOTAL INDIRECT COSTS (F&A)				**24,368**	
J. TOTAL DIRECT AND INDIRECT COSTS (H + I)				**85,503**	
K. SMALL BUSINESS FEE				**0**	
L. AMOUNT OF THIS REQUEST (J) OR (J MINUS K)				**85,503**	

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

PI/PD NAME
Sagar Samtani

ORG. REP. NAME*
Laura Beagles

FOR NSF USE ONLY		
INDIRECT COST RATE VERIFICATION		
Date Checked	Date Of Rate Sheet	Initials - ORG

 ***ELECTRONIC SIGNATURES REQUIRED FOR REVISED BUDGET**

SUMMARY PROPOSAL BUDGET

Cumulative

FOR NSF USE ONLY		
PROPOSAL NO.	DURATION (months) Proposed	Granted
AWARD NO.		

ORGANIZATION: **University of South Florida**

PRINCIPAL INVESTIGATOR / PROJECT DIRECTOR: **Sagar Samtani**

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. **Sagar Samtani - PI**	0.00	1.00	0.00	**16,916**	
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	1.00	0.00	**16,916**	
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. (**4**) GRADUATE STUDENTS				**69,512**	
4. (**0**) UNDERGRADUATE STUDENTS				**0**	
5. (**0**) SECRETARIAL - CLERICAL (IF CHARGED DIRECTLY)				**0**	
6. (**0**) OTHER				**0**	
TOTAL SALARIES AND WAGES (A + B)				**86,428**	
C. FRINGE BENEFITS (IF CHARGED AS DIRECT COSTS)				**4,285**	
TOTAL SALARIES, WAGES AND FRINGE BENEFITS (A + B + C)				**90,713**	
D. EQUIPMENT (LIST ITEM AND DOLLAR AMOUNT FOR EACH ITEM EXCEEDING $5,000.)					
TOTAL EQUIPMENT				**0**	
E. TRAVEL 1. DOMESTIC (INCL. U.S. POSSESSIONS)				**5,000**	
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				**6,400**	
2. PUBLICATION COSTS/DOCUMENTATION/DISSEMINATION				**0**	
3. CONSULTANT SERVICES				**0**	
4. COMPUTER SERVICES				**0**	
5. SUBAWARDS				**0**	
6. OTHER				**22,261**	
TOTAL OTHER DIRECT COSTS				**28,661**	
H. TOTAL DIRECT COSTS (A THROUGH G)				**124,374**	
I. INDIRECT COSTS (F&A)(SPECIFY RATE AND BASE)					
TOTAL INDIRECT COSTS (F&A)				**50,546**	
J. TOTAL DIRECT AND INDIRECT COSTS (H + I)				**174,920**	
K. SMALL BUSINESS FEE				**0**	
L. AMOUNT OF THIS REQUEST (J) OR (J MINUS K)				**174,920**	

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

PI/PD NAME: **Sagar Samtani**

ORG. REP. NAME*: **Laura Beagles**

FOR NSF USE ONLY		
INDIRECT COST RATE VERIFICATION		
Date Checked	Date Of Rate Sheet	Initials - ORG

C ***ELECTRONIC SIGNATURES REQUIRED FOR REVISED BUDGET**

BUDGET JUSTIFICATION - SAMTANI, USF

Senior Personnel	Year 1	Year 2
	$8,333	$8,583

Assistant Professor Sagar Samtani, in the Information Systems and Decision Sciences Department at University of South Florida (USF), will serve as the Principal Investigator on this project. He will manage the entire project, which includes supervising two students. He will commit .5 academic months to this project. His reequested salary for year one is $8,333 and year 2 is $8,583.

Other Personnel	Year 1	Year 2
	$34,242	$35,270

One doctoral student and one masters degree student will assist with the data analysis. The doctoral student will commit 12 calendar months to the project (with an FTE of .5) and the requested salary is $22,152 for year 1 and $22,817 for year 2, which is the agreed-upon rate for University of South Florida doctoral students. The masters degree student will commit 9 calendar months to the project (with an FTE of .5) and the requested salary is $12,090 for year 1 and $12,453 for year 2. When hiring, we recruit from 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 classes.

Fringe Benefits	Year 1	Year 2
	$2,110	$2,175

Dr. Samtani's fringe benefits for year 1 is $2,042 ($8,333 * .19 regular fringe percentage + $688.58 insurance rate * 12 months * (.5 academic months of efforts/9 academic months). Year 2 is $2,104 ($8,583 * .19 regular fringe percentage + $709.13 insurance rate * 12 months * (.5 academic months of efforts/9 academic months). The doctoral student's regular fringes total is $44 for year 1, which is .20% of the requested salary ($22,152 * .0020 regular fringe percentage rate); years 2 is $46 ($22,817 * .0020). The masters student's regular fringes total for year 1 is $24 ($12,090 requested salary * .0020 regular percentage rate) and year 2 is $25 ($12,453 request salary * .0020).

Travel	Year 1	Year 2
	$1,800	$3,200

Dr. Samtani and the doctoral student will travel on this project. During year 1 only Dr. Samtani will attend one conference and year 2 both will attend a conference. The total travel cost of $5,000 includes: $2,600 registration fee ($2,000 for the PI + $600 student rate for the doctoral student); $1,110 for lodging (2 nights at an average rate of $185 for 3 trips); $42 for milage (.445/mile for 96 miles); $900 for round trip airfare ($300/trip * 3 trips); $102 ground transportation ($34/roundtrip * 3 trips); $216 for meals [breakfast $36 ($6 * 2 days * 3 trips); lunch $66 ($11 * 2 days * 3 trips), and dinner $114 ($19 * 2 days * 3 trips)]; and $30 for incidentals ($10/conference * 3 trips).

BUDGET JUSTIFICATION - SAMTANI, USF

Other Direct Costs	Year 1	Year 2
	$16,754	**$11,907**

Tuition remission totaled $22,261 for two years. Tuition is budgeted at a rate of $431.44 per semester hour for the doctoral student during year one - covering 24 hours. Year 1 totaled $10,354 ($431.44 * 24). A 15% tuition inflation rate was added to subsequent year 2 totaling $11,907. Also, $6,400 is budgeted during year 1 to purchase graphical processing units (GPU) required for the proposed deep learning-based analytics and a dedicated server to manage the terabytes of hacker forum data which will be collected throughout the research. A GPU average cost is $1,200 * 2 = $2,400 and the dedicated server averaged cost is $4,000.

Indirect Costs	Year 1	Year 2
	$26,178	**$24,368**

The Modified Total Direct Costs (MTDC) calculation is used to determine the indirect cost. USF's negotiated indirect cost rate is 49.5%. The cognizant federal agency is the Department of Health and Human Services. The indirect cost base for year 1 is $52,885 * .495 = $26,178; year 2 $49,228 * .495 = $24,368.

Total Budget	Year 1	Year 2
$174,920	**$89,417**	**$85,503**

Current and Pending Support

(See PAPPG Section II.C.2.h for guidance on information to include on this form.)

The following information should be provided for each investigator and other senior personnel. Failure to provide this information may delay consideration of this proposal.

Investigator: Sagar Samtani

Other agencies (including NSF) to which this proposal has been/will be submitted.

Support: ☐ Current ☒ Pending ☐ Submission Planned in Near Future ☐ *Transfer of Support

Project/Proposal Title: CRII: SaTC: Identifying Emerging Threats in the Online Hacker Community for Proactive Cyber Threat Intelligence: A Diachronic Graph Convolutional Autoencoder Framework

Source of Support: Federal

Total Award Amount: $ 174,940 Total Award Period Covered: 07/01/19 - 08/30/21

Location of Project: USF Tampa

Person-Months Per Year Committed to the Project. Cal: 0.00 Acad: 0.50 Sumr: 0.00

Support: ☐ Current ☐ Pending ☐ Submission Planned in Near Future ☐ *Transfer of Support

Project/Proposal Title:

Source of Support:

Total Award Amount: $ Total Award Period Covered:

Location of Project:

Person-Months Per Year Committed to the Project. Cal: Acad: Sumr:

Support: ☐ Current ☐ Pending ☐ Submission Planned in Near Future ☐ *Transfer of Support

Project/Proposal Title:

Source of Support:

Total Award Amount: $ Total Award Period Covered:

Location of Project:

Person-Months Per Year Committed to the Project. Cal: Acad: Sumr:

Support: ☐ Current ☐ Pending ☐ Submission Planned in Near Future ☐ *Transfer of Support

Project/Proposal Title:

Source of Support:

Total Award Amount: $ Total Award Period Covered:

Location of Project:

Person-Months Per Year Committed to the Project. Cal: Acad: Sumr:

Support: ☐ Current ☐ Pending ☐ Submission Planned in Near Future ☐ *Transfer of Support

Project/Proposal Title:

Source of Support:

Total Award Amount: $ Total Award Period Covered:

Location of Project:

Person-Months Per Year Committed to the Project. Cal: Acad: Summ:

---

*If this project has previously been funded by another agency, please list and furnish information for immediately preceding funding period.

Facilities, Equipment and Other Resources

No laboratory work is necessary in order to accomplish the goals set forth in this project. Dr. Samtani and the graduate students will need a computer and office space in order to conduct the research.

Laboratory
Not Applicable

Clinical
Not Applicable

Animal
Not Applicable

Computer

University of South Florida provides Dr. Samtani and the graduate students with a desktop computer. They will use their assigned computing devices for communication purpose. The computers have a set of licensed software programs.

Office

The University of South Florida also provides Dr. Samtani and the graduate students with office spaces—where they will conduct the research.

Other Resources

The staff at Florida Center for Cybersecurity, at the University of South Florida, will help disseminate selected information related to the research via a community forum to other schools in the Florida State University System. When it is possible, they will also share research information with industry, academic, and government partners within the Community Outreach and Engagement channels.

The staff at the National Cyber-Forensics and Training Alliance will serve as subject matter experts and provide feedback/advice throughout the duration of this project.

Personnel at the Society for the Policing of Cyberspace will participate in discussions via telephone consultation and provide advice on identification of relevant hacker community platforms for data collection and evaluate the usefulness of output generated during the research.

Data Management Plan

1. Types of Data, Samples, and Other Materials to Be Produced In the Course Of the Project

Data

Data will be obtained from online hacker forums by constructing customized web crawlers. The principal investigator will use his nearly a half-decade in leading crawler development and periodic collections to acquire multi-million record testbeds during his time as lead research associate for Dr. Hsinchun Chen's highly-successful Scholarship-for-Service (SFS) and Secure and Trustworthy Cyberspace (SaTC) programs in University of Arizona's Artificial Intelligence (AI) Lab.

Publications and presentations

It is anticipated that the proposed project will result in published papers and presentations at conferences, workshops, and other events. Additionally, it is anticipated that educational materials generated for and CAE courses. PI Samtani is committed to disseminating work, mentoring students to publish in top-tier journals, and upholding the NSF's commitment to open access. PI Samtani will monitor open access journals at the national level. This includes, but is not limited to, NSF's open access portal initiative which provides papers' abstracts and other metadata, linking to the full-text articles on publishers' websites. Accepted and published papers and presentations will be available through the NSF funded Data Infrastructure Building Blocks for Intelligence and Security Informatics (DIBBs-ISI) portal (http://azsecure-data.org/). This system is headed by Dr. Hsinchun Chen at the AI Lab and is designed to share multi-million record datasets and papers for the larger cybersecurity community to download and access for their research. Accepted and published papers and presentations will be available through the DIBBs portal, this project's website, and presentation at conferences and workshops. Regarding re-use and distribution, publication, publications may be subject to copyright terms and conditions. As such, archiving will be made online from journal websites. Links will be provided on the PI's project website. PI Samtani will also publicize educational materials through Cyber Florida's community forum.

2. Standards to Be Used For Data and Metadata Format and Content

The data will be stored in the following formats: HTML and SQL formats. The types of metadata that will be included are postid, post date, author screen name, post content which conform to the standards in this field. Institutional data is defined as all data elements created, maintained, received, or transmitted as a result of business, educational or research activities of a USF System unit. Institutional Data is the Property of the USF System and is not "Owned" by Any Individual, Academic, or Business Unit - Data and the meta-data about that data are business and technical resources owned by the USF System. All employees must recognize that the proper management of strategic Institutional Data is critical to the success of the organization. Institutional data in all electronic formats shall be safeguarded and secured based on recorded and approved requirements and compliance guidelines. These requirements are determined by the Stewardship Committee Data in conjunction with the Office of Information Security.

3. Methods and policies for providing access and enabling sharing

Institutional Data and information about that data (meta-data) shall be readily accessible to all, except where determined to be restricted. When restrictions are made, the appropriate Stewardship Committee is accountable for defining specific individuals and levels of access privileges that are to be enabled. The Office of Information Security is responsible for ensuring the proper implementation of the security controls. Data and code in the DIBBs-ISI portal may be freely downloaded.

4. Policies and provisions for re-use, re-distribution, and the production of derivatives

All strategic Institutional Data shall be modeled, named, and defined consistently (according to standards) across the business and academic divisions of the USF System. Every effort must be made by management to share data across divisions, and not to maintain redundant data without justification. Stewardship Committees responsible for the data must recognize the informational needs of downstream processes and academic units that may require said data.

5. Plans for archiving and Preservation of access

University of South Florida's data retention rules will be applied Any tools and documentation produced in the project will be maintained in a concurrent versions system (CVS), that support bug tracking. Any code developed from this project will be distributed from one or more open source software (OSS) portals. One such distribution outlet will include the DIBBs portal. PI Samtani will also aim to resach the broader research community by including selected code into other OSS portals. These include GitHub (http://github.com/), SourceForge (http://sourceforge.net/), and Open Source Initiative (http://opensource.org/). PI Samtani will continuously monitor current NSF guidelines for best practices and advice about the best portals for ensuring that government agencies and employees are aware of software and tools.

July 30, 2018

National Science Foundation CRII Program
2415 Eisenhower Ave.
Alexandria, VA 22314

Dear Dr. Gilinert and Review Team,

As chair of the Information Systems and Decision Sciences (ISDS) Department in the Muma College of Business at the University of South Florida (USF), it is my pleasure to certify Dr. Sagar Samtani's eligibility for the National Science Foundation (NSF) Computer and Information Science and Engineering (CISE) Research Initiation Initiative (CRII) program. As per the requirements of the CRII solicitation 18-554, this letter has two sections. The first section certifies Dr. Samtani's eligibility for NSF CRII, while the second summarizes how his proposal meets the goals of the CRII program.

CRII Eligibility:

Dr. Samtani joined the ISDS department in August 2018 as a first year tenure-track assistant professor (currently untenured). He graduated with his Ph.D. under the supervision of Dr. Hsinchun Chen from the Artificial Intelligence (AI) Lab in the Management Information Systems (MIS) Department at the University of Arizona (UA) in May 2018 (final defense in March 2018). This is within the five year limit specified by the solicitation. He has not and will not apply for NSF CAREER within the calendar year of 2018. He also has not yet attained any independent NSF support. Taken together, these characteristics make him eligible for CRII funding.

Alignment of Proposal with CRII Program Goals:

Dr. Samtani's proposal, entitled "CRII: SaTC: Identifying Emerging Threats from the Online Hacker Community for Proactive Cyber Threat Intelligence: A Diachronic Graph Convolutional Autoencoder Approach," is well-positioned to make interesting and critically-needed contributions to cybersecurity. Specifically, this proposal aims to develop advanced, proactive Cyber Threat Intelligence (CTI) capabilities by (1) identifying and automatically collecting a multi-million record testbed of hacker forum posts and (2) analyzing the rich textual nature of these posts to identify emerging threats, specifically malicious hacker exploits (malware), via innovative text graphs, novel deep learning algorithms, and diachronic linguistics.

Dr. Samtani is an ideal candidate for receiving CRII funds. While new faculty members in our department are outfitted with state-of-the-art computers and allocated start-up funds (in his case, $17.5K over three years), his proposed research necessitates additional, specialized resources which can be attained through the CRII program. Collecting and analyzing (via novel deep learning approaches) a large testbed of hacker community data requires dedicated, customized servers designed to continuously crawl, store, and back up terabytes of unique and valuable hacker social media data and Graphical Processing Units (GPUs) to analyze collected data. More important than these supplies, however, is the time commitment and allocated graduate student. Both will support and significantly enhance his ability to effectively

gather/clean a large corpus of data, implement the proposed algorithms, and disseminate relevant selected results in a timely fashion. Most importantly, it will enable him to attain promising research results such that it improves his capacity to write successful proposals in the future.

The lack of CRII funding would prevent him from fully executing the breadth and depth of his proposed research activities. The lack of these funds would also prevent him from achieving research independence and pursuing further high-impact and relevant CTI research inquiries of interest to the NSF Secure and Trustworthy Cyberspace (SaTC) and larger cybersecurity communities. In a domain where hackers advance their skill-sets at staggering rates, the lack of comprehensive collection capabilities and delays in analytical processing can have significant societal ramifications.

To summarize, Dr. Samtani meets the CRII eligibility criteria with his graduation date and his current employment (start date and position). His proposal aims to develop timely and relevant cybersecurity capabilities and help him establish research independence (a key goal of the CRII program). More importantly, however, the proposal aligns with NSF's larger vision of funding transformative research which advance knowledge and contribute to the achievement of specific societal outcomes.

If you have any further questions or would like additional information, please feel free to contact me.

With best regards,

Manish Agrawal, Ph.D.
Department Head, Information Systems and Decision Sciences (ISDS)
Professor
Email: magrawal@usf.edu
Phone: 813-947-6716

Florida Center for
Cybersecurity
at the University of South Florida

August 3, 2018

Dr. Sagar Samtani
Assistant Professor, ISDS Department
4202 E. Fowler Ave., CIS 2071
Tampa, FL 33620

Dear Dr. Samtani:

This letter is to state Cyber Florida's willingness to be an unpaid collaborator in your proposal to the National Science Foundation (NSF) grant titled, "CRII: SaTC: Identifying Emerging Threats in the Online Hacker Community for Proactive Cyber Threat Intelligence: A Diachronic Graph Convolutional Autoencoder Framework." Should the proposal be funded, Cyber Florida will:

- Help disseminate selected results via our community forum to other schools in the Florida State University Systems for possible integration into cybersecurity curriculums
- Share, when possible, selected research results to industry, academic, and government partners through our Community Outreach and Engagement channels.

As you know, Cyber Florida's mission is to make Florida the leading Cyber State within the US. The goals of your proposal and the foundation this funding will provide for your future research agenda are in line with Cyber Florida's mission. Thus, it is our pleasure in helping you achieve your goal of attaining research independence. We look forward to your future success!

Please feel free to reach out to me or my team if you need any additional information.

Sincerely,

Sri Sridharan
Director, Florida Center for Cybersecurity (Cyber Florida)
sris@usf.edu

NATIONAL CYBER-FORENSICS AND TRAINING ALLIANCE

NCFTA | PITTSBURGH
2000 Technology Drive
Suite 450
Pittsburgh PA 15219
P | 412-802-8000
F | 412-802-8510
E | info@ncfta.net

NCFTA | NEW YORK
One Penn Plaza
250 West 34th Street
Suite 2124
New York NY 10119

NCFTA | LOS ANGELES
Wilshire Building
601 S Figueroa Street
Suite 1460
Los Angeles CA 90017

July 30, 2018

Dr. Sagar Samtani
Assistant Professor, ISDS Department
4202 E. Fowler Ave., CIS 2071
Tampa, FL 33620

Dear Dr. Samtani:

We are very pleased to continue our collaboration with you as you make the transition from a Ph.D. student at the University of Arizona to an Assistant Professor at the University of South Florida. We acknowledge that we are included as an unpaid collaborator in your proposal to the National Science Foundation, "CRII: SaTC: Identifying Emerging Threats in the Online Hacker Community for Proactive Cyber Threat Intelligence: A Diachronic Graph Convolutional Autoencoder Framework."

The National Cyber-Forensics & Training Alliance (NCFTA) is committed to serving as an SME and providing feedback and advice throughout the duration of this project. As described in the proposal's project description, our role as will include the following:

- Having you work with our malware team to integrate selected hacker community tools and malware into our Malware Analysis Portal (MAP)
- Helping to identify relevant hacker community platforms of interest
- Soliciting feedback and disseminate selected results through our partner network

As you are very well familiar, NCFTA is a 501c3 nonprofit organization which specializes in real time information sharing on cyber threats between industry and law enforcement partners. NCFTA's industry partner base represents over 140 of the most impactful global brands across financial, manufacturing, retail, technology, and pharmaceutical industries. The proposed project will provide immediate benefit back to NCFTA's global partners both in industry and law enforcement.

Ultimately, NCFTA is willing to work with you on this project because we believe it has the potential to help transform our analytical work, enabling us to gain key insights of emerging threats in cyberspace faster, more readily, and more accurately. We also assume that, as you achieve research independence, you will continue to pursue grant funding in the future with NCFTA involvement.

Let me know if you need additional information.

Sincerely,

Matt LaVigna
CEO and President, NCFTA
mlavigna@ncfta.net

THE SOCIETY FOR THE POLICING OF CYBERSPACE

Mailing address: Suite 212, 185 – 9040 Blundell Road, Richmond, B.C. V6Y 1K3, Canada
Email: polcyb@polcyb.org Website: http://www.polcyb.org

July 30, 2018

Dr. Sagar Samtani
Assistant Professor, ISDS Department
4202 E. Fowler Ave., CIS 2071
Tampa, FL 33620

Dear Dr. Samtani:

This letter is to state my willingness to be an unpaid collaborator in your proposal to the National Science Foundation (NSF), "CRII: SaTC: Identifying Emerging Threats in the Online Hacker Community for Proactive Cyber Threat Intelligence: A Diachronic Graph Convolutional Autoencoder Framework."

Please note that, due to the limited resources of our not-for-profit organization, contribution from The Society for the Policing of Cyberspace (POLCYB) will be limited to the confines of POLCYB's available resources at the time of consultation.

Should the proposal be funded, we, The Society for the Policing of Cyberspace (POLCYB), will participate in discussions via **telephone consultation with you** on the following:

- Provide advice on identification of relevant hacker community platforms for data collection
- Evaluate the usefulness of generated research outputs (i.e., validation of emerging threat detection).

POLCYB would also consider disseminating selected research results to our partners as our Board of Directors see fit.

It is our pleasure to assist you in your goal of attaining research independence. As you are well aware, POLCYB is a not-for-profit Canadian organization with a large, international network of professional partners in the public and private sectors. POLCYB's goal focuses on sharing information and best practices on cybercrime prevention, detection, and/or response. One of the key issues we focus on is identifying emerging trends and threats in cybercrime. Your overall proposal goal and future research agenda align well with POLCYB's mission.

Should you need further information, please feel free to contact me via email at bessie-pang@polcyb.org .

Sincerely,

Bessie Pang
Executive Director, (POLCYB)