Building the UA/Eller/MIS AZSecure Cybersecurity Analytics Program: My Journey

Hsinchun Chen, Ph. D.

Regents' Professor, Thomas R. Brown Chair

Director, AI Lab, Azsecure Cybersecurity Program

Fellow, ACM, IEEE, AAAS

University of Arizona

Funding Acknowledgements: NSF (SaTC, SFS, ACI)

Outline

- **Security Informatics & Analytics**: COPLINK, BorderSafe, Dark Web

- **Azsecure Cybersecurity Analytics:**

(1) ***Dark Web Analytics*** for studying international hacker community, forums, and markets;

(2) ***Privacy and PII (Personally Identifiable Information) Analytics*** for identifying and alleviating privacy risks for vulnerable populations;

(3) ***Adversarial Malware Generation and Evasion*** for adversarial AI in cybersecurity; and

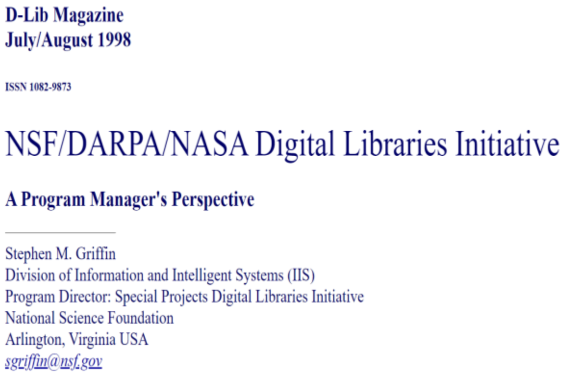
(4) ***Smart Vulnerability Assessment*** for scientific workflows and OSS (Open Source Software) vulnerability analytics and mitigation.

Computational Design Science Research at UA/Eller/MIS AI Lab

- Applications/problems: digital libraries, search engines, biomedical informatics, healthcare data mining, security informatics, business intelligence, cybersecurity analytics
- Approaches: web collection/spidering, databases, data warehousing, data mining, text mining, web mining, statistical NLP, machine learning, deep learning, ontologies, social media analytics, interface design, information visualization, economic modeling, assessment
- Structure: federal funding (NSF/DOD/NIH), director, affiliated faculty, post-docs, Ph.D./MS/BS students ➔ tech transfer, commercialization

- Major phases: DLI ➔ COPLINK ➔ Dark Web ➔ AZSecure

Security Informatics & Analytics: COPLINK & Dark Web



D-Lib Magazine
July/August 1998

ISSN 1082-9873

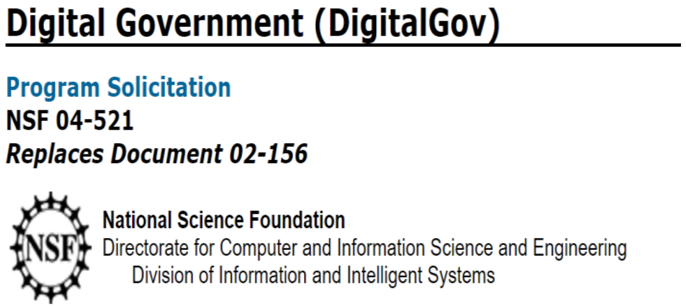
NSF/DARPA/NASA Digital Libraries Initiative

A Program Manager's Perspective

Stephen M. Griffin
Division of Information and Intelligent Systems (IIS)
Program Director: Special Projects Digital Libraries Initiative
National Science Foundation
Arlington, Virginia USA
sgriffin@nsf.gov

NIJ National Institute of Justice

STRENGTHEN SCIENCE. ADVANCE JUSTICE.



Digital Government (DigitalGov)

Program Solicitation
NSF 04-521
Replaces Document 02-156

National Science Foundation
Directorate for Computer and Information Science and Engineering
Division of Information and Intelligent Systems

Global Security Impacts

- “War on terror” (Iraq and Afghanistan) surpassed cost of Second World War, $5 trillion…Time Magazine

- Hacker costing $1 trillion globally… President Obama

From the Surface Web to the Dark Web

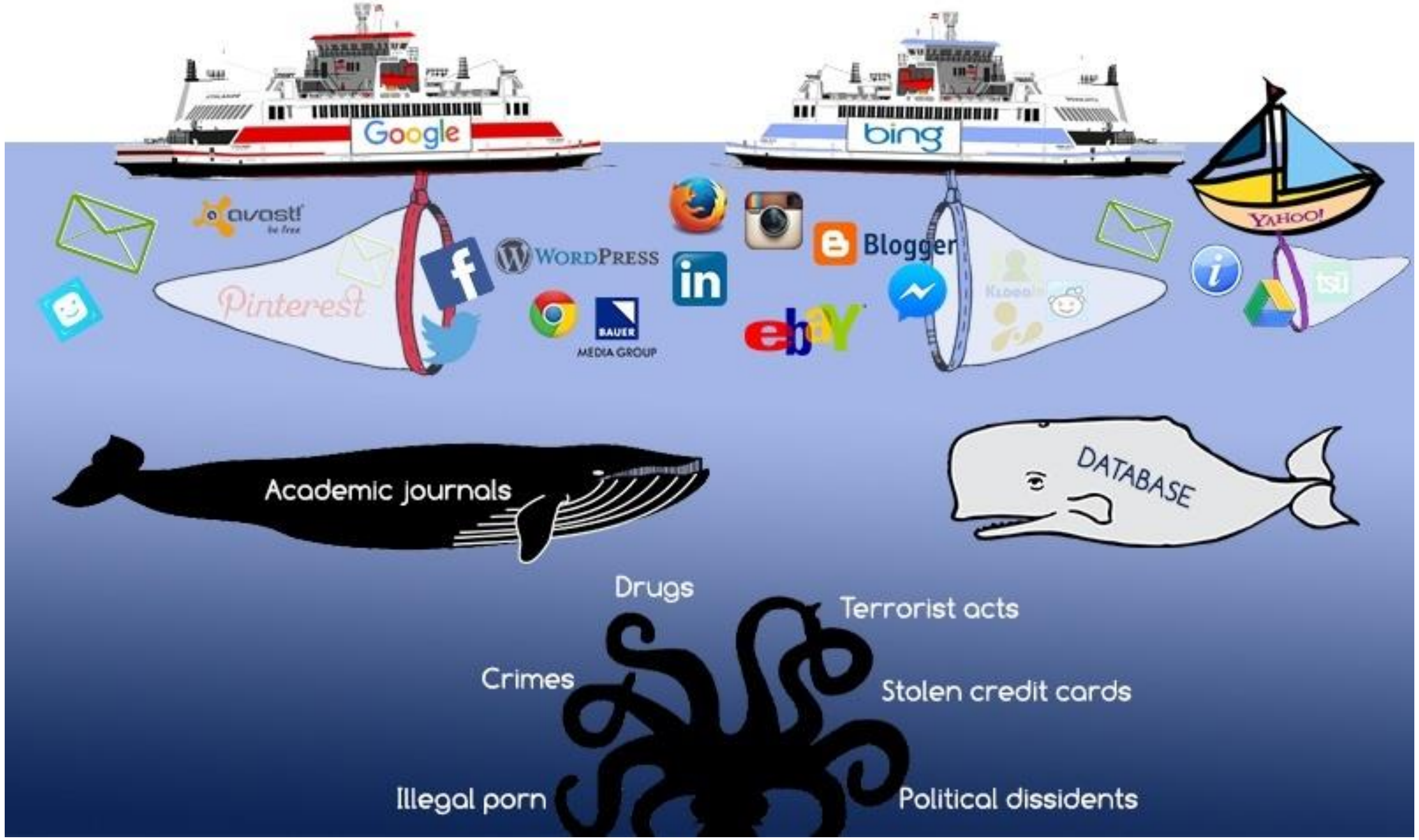
Surface Web

Deep Web

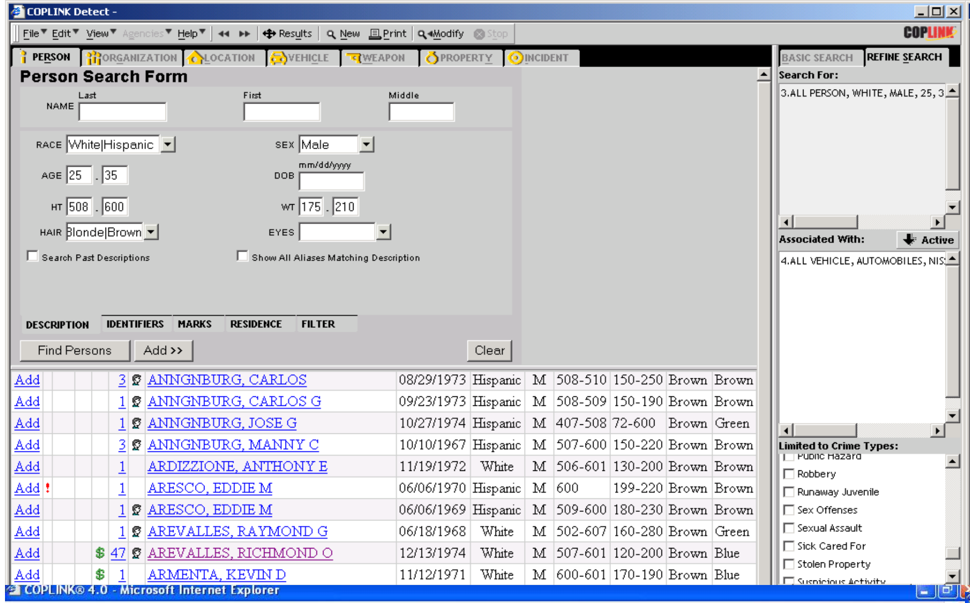
Dark Web

DarkNet

Hacker Web



COPLINK: Crime Data Mining (1997-2009)



COPLINK Detect -

Person Search Form

NAME Last		First	Middle
RACE	White|Hispanic	SEX	Male
AGE	25 - 35	DOB (mm/dd/yyyy)	
HT	508 - 600	WT	175 - 210
HAIR	Blonde|Brown	EYES	

□ Search Past Descriptions   □ Show All Aliases Matching Description

DESCRIPTION | IDENTIFIERS | MARKS | RESIDENCE | FILTER

Find Persons | Add >> | Clear

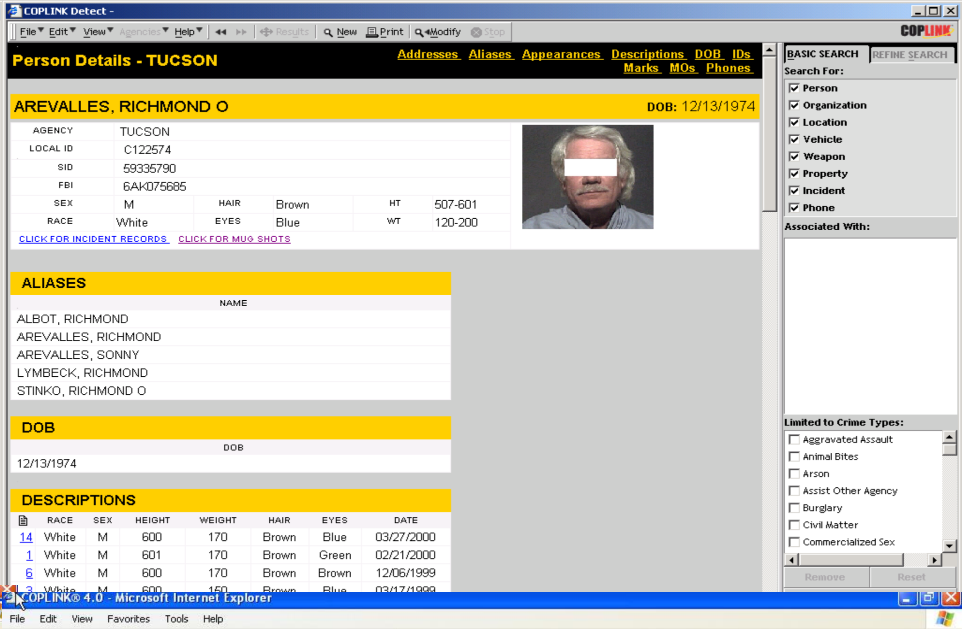
Add			3	ANNGNBURG, CARLOS	08/29/1973	Hispanic	M	508-510	150-250	Brown	Brown
Add			1	ANNGNBURG, CARLOS G	09/23/1973	Hispanic	M	508-509	150-190	Brown	Brown
Add			1	ANNGNBURG, JOSE G	10/27/1974	Hispanic	M	407-508	72-600	Brown	Green
Add			3	ANNGNBURG, MANNY C	10/10/1967	Hispanic	M	507-600	150-220	Brown	Brown
Add			1	ARDIZZIONE, ANTHONY E	11/19/1972	White	M	506-601	130-200	Brown	Brown
Add	!		1	ARESCO, EDDIE M	06/06/1970	Hispanic	M	600	199-220	Brown	Brown
Add			1	ARESCO, EDDIE M	06/06/1969	Hispanic	M	509-600	180-230	Brown	Brown
Add			1	AREVALLES, RAYMOND G	06/18/1968	White	M	502-607	160-280	Brown	Green
Add		$	47	AREVALLES, RICHMOND O	12/13/1974	White	M	507-601	120-200	Brown	Blue
Add		$	1	ARMENTA, KEVIN D	11/12/1971	White	M	600-601	170-190	Brown	Blue

BASIC SEARCH | REFINE SEARCH

Search For:
3.ALL PERSON, WHITE, MALE, 25, 3

Associated With: Active
4.ALL VEHICLE, AUTOMOBILES, NIS

Limited to Crime Types:
□ Public Hazard
□ Robbery
□ Runaway Juvenile
□ Sex Offenses
□ Sexual Assault
□ Sick Cared For
□ Stolen Property
□ Suspicious Activity



COPLINK Detect -

Person Details - TUCSON

AREVALLES, RICHMOND O DOB: 12/13/1974

AGENCY	TUCSON				
LOCAL ID	C122574				
SID	59335790				
FBI	6AK075685				
SEX	M	HAIR	Brown	HT	507-601
RACE	White	EYES	Blue	WT	120-200

CLICK FOR INCIDENT RECORDS   CLICK FOR MUG SHOTS

ALIASES

NAME
ALBOT, RICHMOND
AREVALLES, RICHMOND
AREVALLES, SONNY
LYMBECK, RICHMOND
STINKO, RICHMOND O

DOB

DOB
12/13/1974

DESCRIPTIONS

	RACE	SEX	HEIGHT	WEIGHT	HAIR	EYES	DATE
14	White	M	600	170	Brown	Blue	03/27/2000
1	White	M	601	170	Brown	Green	02/21/2000
6	White	M	600	170	Brown	Brown	12/06/1999
3	White	M	600	150	Brown	Blue	03/17/1999

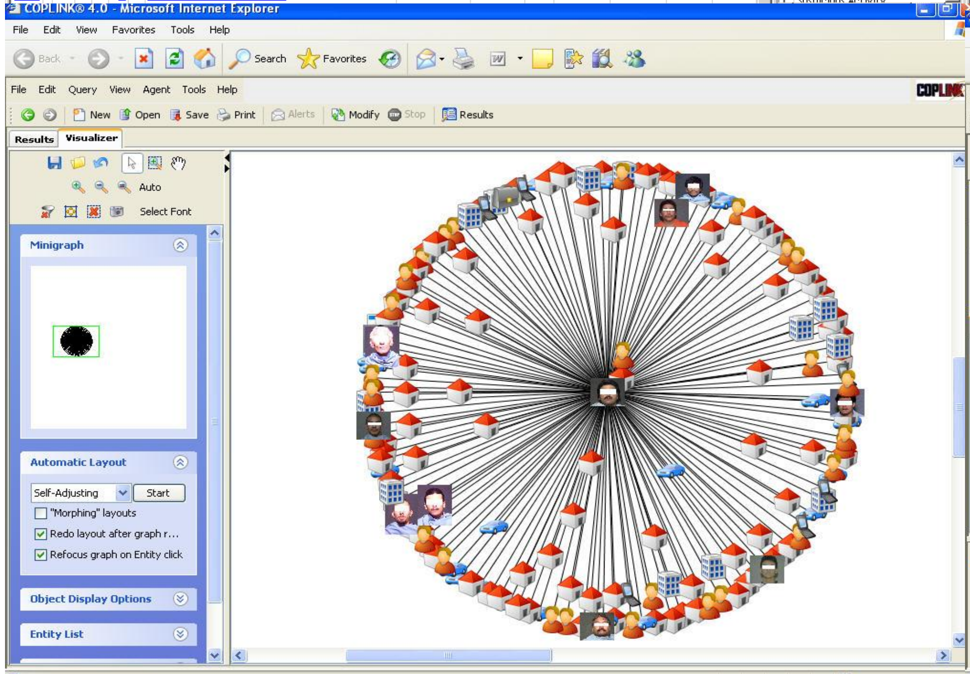
BASIC SEARCH | REFINE SEARCH

Search For:
☑ Person
☑ Organization
☑ Location
☑ Vehicle
☑ Weapon
☑ Property
☑ Incident
☑ Phone

Associated With:

Limited to Crime Types:
□ Aggravated Assault
□ Animal Bites
□ Arson
□ Assist Other Agency
□ Burglary
□ Civil Matter
□ Commercialized Sex

Remove | Reset



COPLINK® 4.0 - Microsoft Internet Explorer

Minigraph

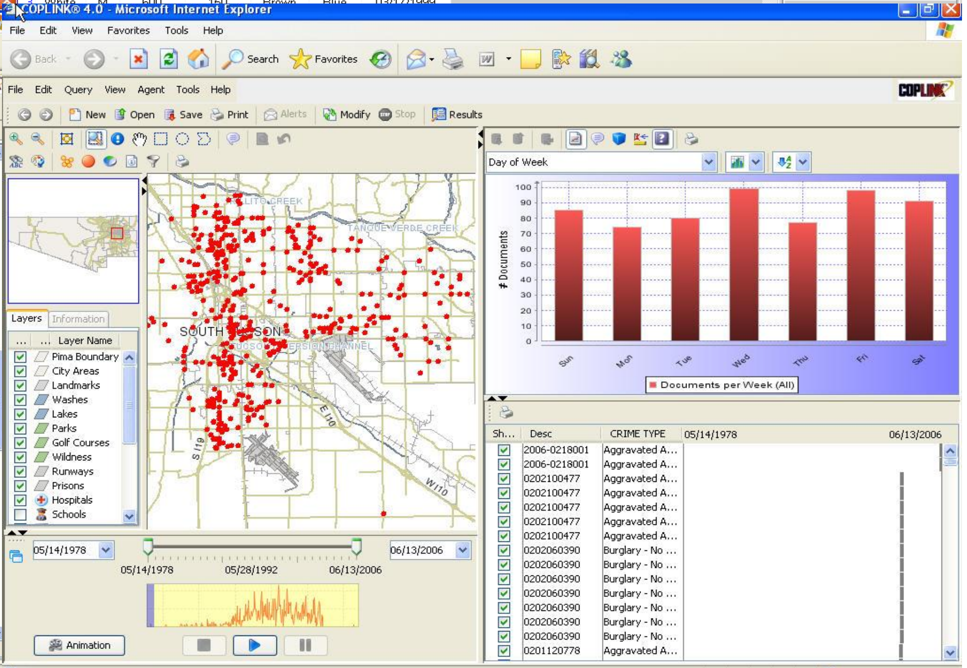
Automatic Layout
Self-Adjusting | Start
□ "Morphing" layouts
☑ Redo layout after graph r...
☑ Refocus graph on Entity click

Object Display Options

Entity List

Applet coplinkApplet started

My Computer



COPLINK® 4.0 - Microsoft Internet Explorer

...	...	Layer Name
☑		Pima Boundary
☑		City Areas
☑		Landmarks
☑		Washes
☑		Lakes
☑		Parks
☑		Golf Courses
☑		Wildness
☑		Runways
☑		Prisons
☑	✚	Hospitals
□		Schools

05/14/1978 – 06/13/2006

05/14/1978   05/28/1992   06/13/2006

Sh...	Desc	CRIME TYPE	05/14/1978 06/13/2006
☑	2006-0218001	Aggravated A...	
☑	2006-0218001	Aggravated A...	
☑	0202100477	Aggravated A...	
☑	0202100477	Aggravated A...	
☑	0202100477	Aggravated A...	
☑	0202100477	Aggravated A...	
☑	0202100477	Aggravated A...	
☑	0202060390	Burglary - No ...	
☑	0202060390	Burglary - No ...	
☑	0202060390	Burglary - No ...	
☑	0202060390	Burglary - No ...	
☑	0202060390	Burglary - No ...	
☑	0202060390	Burglary - No ...	
☑	0202060390	Burglary - No ...	
☑	0201120778	Aggravated A...	

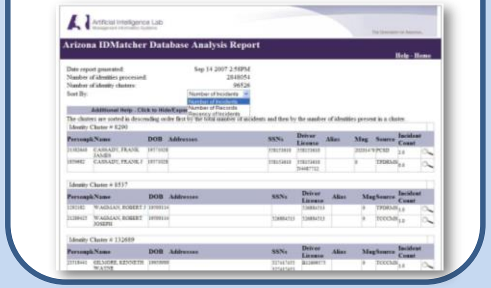
Applet coplinkApplet started

My Computer

COPLINK Identity Resolution and Criminal Network Analysis

Cross-jurisdictional Information Sharing/Collaboration

Arizona IDMatcher



Artificial Intelligence Lab

Arizona IDMatcher Database Analysis Report

Help - Home

Date report generated: Sep 14 2007 2:58PM
Number of identities processed: 2848054
Number of identity clusters: 96526
Sort By: Number of Incidents

Additional Help : Click to Hide/Expand

The clusters are sorted in descending order first by the total number of incidents and then by the number of identities present in a cluster.

Identity Cluster # 8290

PersonphiName		DOB	Addresses	SSNs	Driver License	Alias	Mug	Source	Incident Count
[illegible]	CASSADY, FRANK JAMES	[illegible]		[illegible]	[illegible]		[illegible]	PCSD	2.0
[illegible]	CASSADY, FRANK J	[illegible]		[illegible]	[illegible]		0	TPDRMS	0.0

Identity Cluster # 8537

PersonphiName		DOB	Addresses	SSNs	Driver License	Alias	Mug	Source	Incident Count
[illegible]	WAGMAN, ROBERT J	[illegible]			[illegible]		0	TPDRMS	1.0
[illegible]	WAGMAN, ROBERT JOSEPH	[illegible]		[illegible]	[illegible]		0	TCCCMS	1.0

Identity Cluster # 132689

PersonphiName		DOB	Addresses	SSNs	Driver License	Alias	Mug	Source	Incident Count
[illegible]	GILMORE, KENNETH WAYNE	[illegible]		[illegible]	[illegible]		0	TCCCMS	1.0

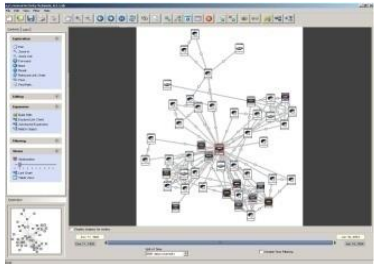
Law-enforcement Data

AZ | CA | TX

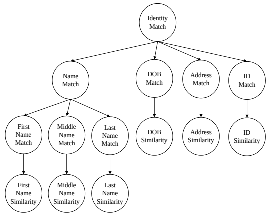
Border Crossing Data (AZ, CA, TX)

Vehicles | People

CAN Visualizer



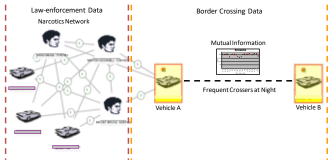
Identity Resolution



Detect false and deceptive identities across jurisdictions using a probabilistic naïve-Bayes based resolution system.

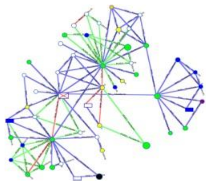
Criminal Network Analysis

High-risk Vehicle Identification



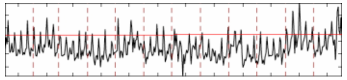
Identify high-risk vehicles using association techniques like mutual information using border crossing and law enforcement data.

Criminal Link Prediction



Predict interaction between individuals and vehicles using link prediction techniques to identify high-risk border crossers.

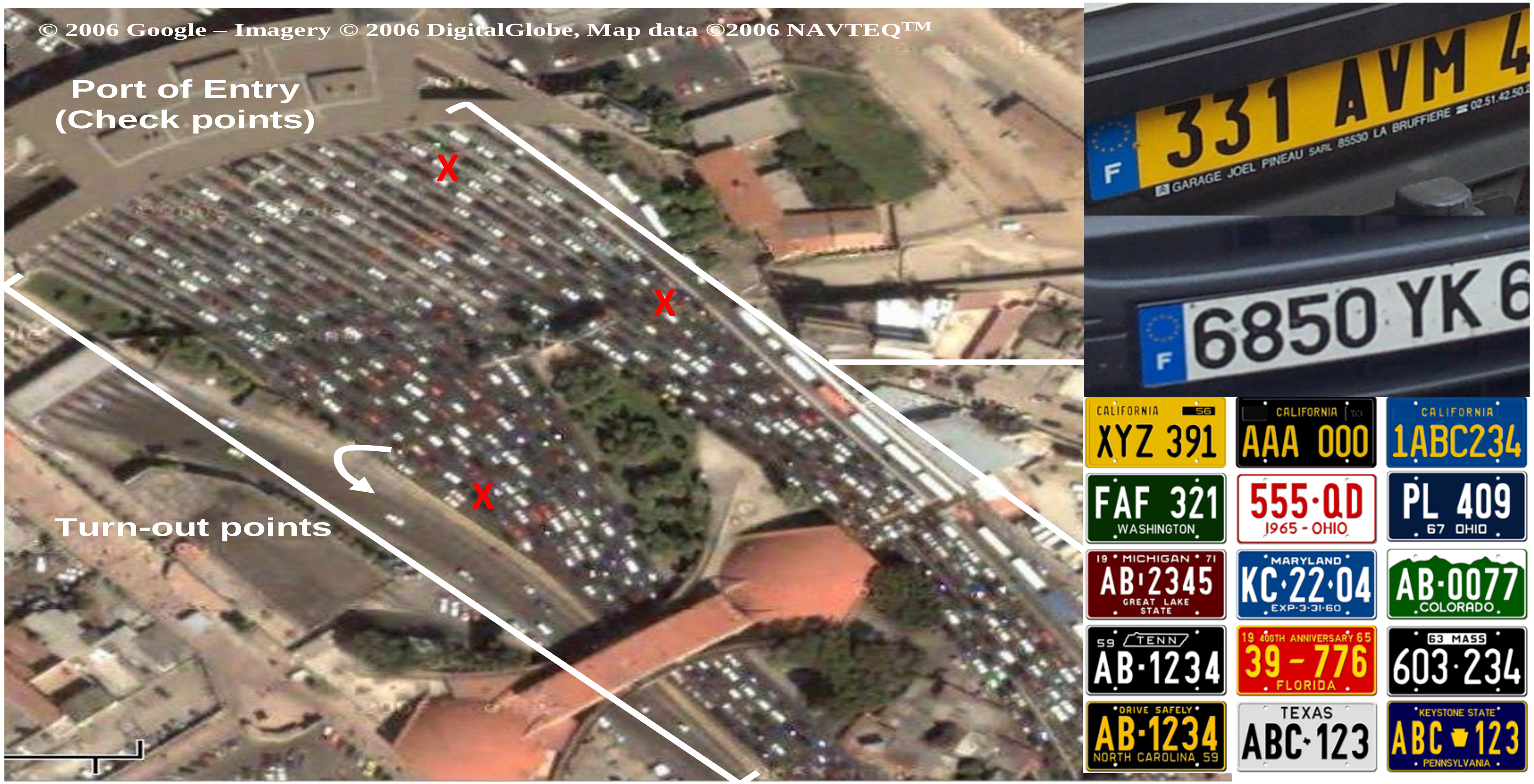
Suspect Traffic Burst Detection



Detect real-time anomalies and threats in border traffic using Markov switching and other models.

* Only the grayed datasets are available to the AI Lab

Border Security: High-risk Vehicle Identification (LPR + DM/SNA)



© 2006 Google – Imagery © 2006 DigitalGlobe, Map data ©2006 NAVTEQ™

Port of Entry
(Check points)
X
X
X
Turn-out points
331 AVM 4
F
GARAGE JOEL PINEAU SARL 85530 LA BRUFFIERE 02.51.42.50
6850 YK 6
F
CALIFORNIA XYZ 391
CALIFORNIA AAA 000
CALIFORNIA 1ABC234
FAF 321 WASHINGTON
555·QD 1965 - OHIO
PL 409 67 OHIO
19 MICHIGAN 71 AB 2345 GREAT LAKE STATE
MARYLAND KC·22·04 EXP·3·31·60
AB·0077 COLORADO
59 TENN AB·1234
19 400TH ANNIVERSARY 65 39 – 776 FLORIDA
63 MASS 603·234
DRIVE SAFELY AB·1234 NORTH CAROLINA 59
TEXAS ABC·123
KEYSTONE STATE ABC 123 PENNSYLVANIA

COPLINK: Crime Data Mining

ABC News April 15, 2003

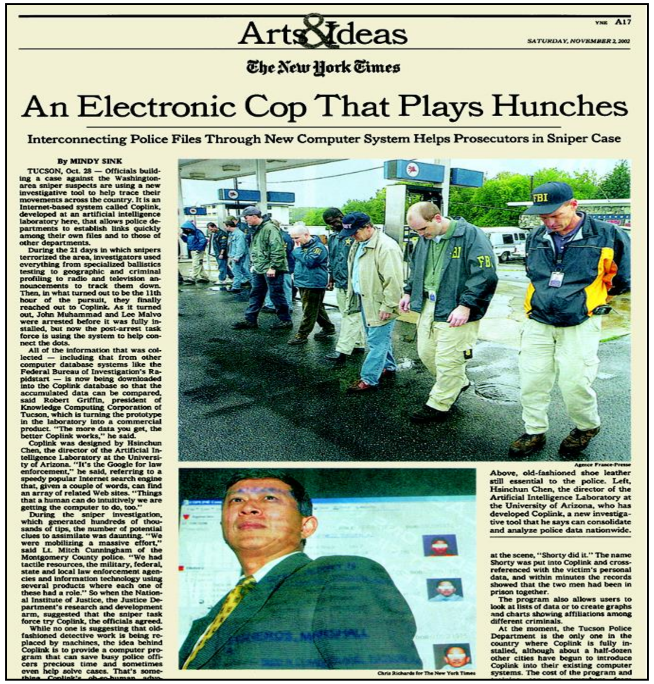
Google for Cops: Coplink software helps police search for cyber clues to bust criminals

IBM i2 COPLINK

Accelerating law enforcement investigations

Palantir

($54B, IPO 2020)



YNE A17

Arts&Ideas

SATURDAY, NOVEMBER 2, 2002

The New York Times

An Electronic Cop That Plays Hunches

Interconnecting Police Files Through New Computer System Helps Prosecutors in Sniper Case

By MINDY SINK

TUCSON, Oct. 28 — Officials building a case against the Washington-area sniper suspects are using a new investigative tool to help trace their movements across the country. It is an Internet-based system called Coplink, developed at an artificial intelligence laboratory here, that allows police departments to establish links quickly among their own files and to those of other departments.

During the 21 days in which snipers terrorized the area, investigators used everything from specialized ballistics testing to geographic and criminal profiling to radio and television announcements to track them down. Then, in what turned out to be the 11th hour of the pursuit, they finally reached out to Coplink. As it turned out, John Muhammad and Lee Malvo were arrested before it was fully installed, but now the post-arrest task force is using the system to help connect the dots.

All of the information that was collected — including that from other computer database systems like the Federal Bureau of Investigation's Rapidstart — is now being downloaded into the Coplink database so that the accumulated data can be compared, said Robert Griffin, president of Knowledge Computing Corporation of Tucson, which is turning the prototype in the laboratory into a commercial product. "The more data you get, the better Coplink works," he said.

Coplink was designed by Hsinchun Chen, the director of the Artificial Intelligence Laboratory at the University of Arizona. "It's the Google for law enforcement," he said, referring to a speedy popular Internet search engine that, given a couple of words, can find an array of related Web sites. "Things that a human can do intuitively we are getting the computer to do, too."

During the sniper investigation, which generated hundreds of thousands of tips, the number of potential clues to assimilate was daunting. "We were mobilizing a massive effort," said Lt. Mitch Cunningham of the Montgomery County police. "We had tactile resources, the military, federal, state and local law enforcement agencies and information technology using several products where each one of these had a role." So when the National Institute of Justice, the Justice Department's research and development arm, suggested that the sniper task force try Coplink, the officials agreed.

While no one is suggesting that old-fashioned detective work is being replaced by machines, the idea behind Coplink is to provide a computer program that can save busy police officers precious time and sometimes even help solve cases. That's something Coplink's oh-so-human advo-

Agence France-Presse

Chris Richards for The New York Times

Above, old-fashioned shoe leather still essential to the police. Left, Hsinchun Chen, the director of the Artificial Intelligence Laboratory at the University of Arizona, who has developed Coplink, a new investigative tool that he says can consolidate and analyze police data nationwide.

at the scene, "Shorty did it." The name Shorty was put into Coplink and cross-referenced with the victim's personal data, and within minutes the records showed that the two men had been in prison together.

The program also allows users to look at lists of data or to create graphs and charts showing affiliations among different criminals.

At the moment, the Tucson Police Department is the only one in the country where Coplink is fully installed, although about a half-dozen other cities have begun to introduce Coplink into their existing computer systems. The cost of the program and

abc NEWS

Newsweek

COPLINBK Commercialization Timeline

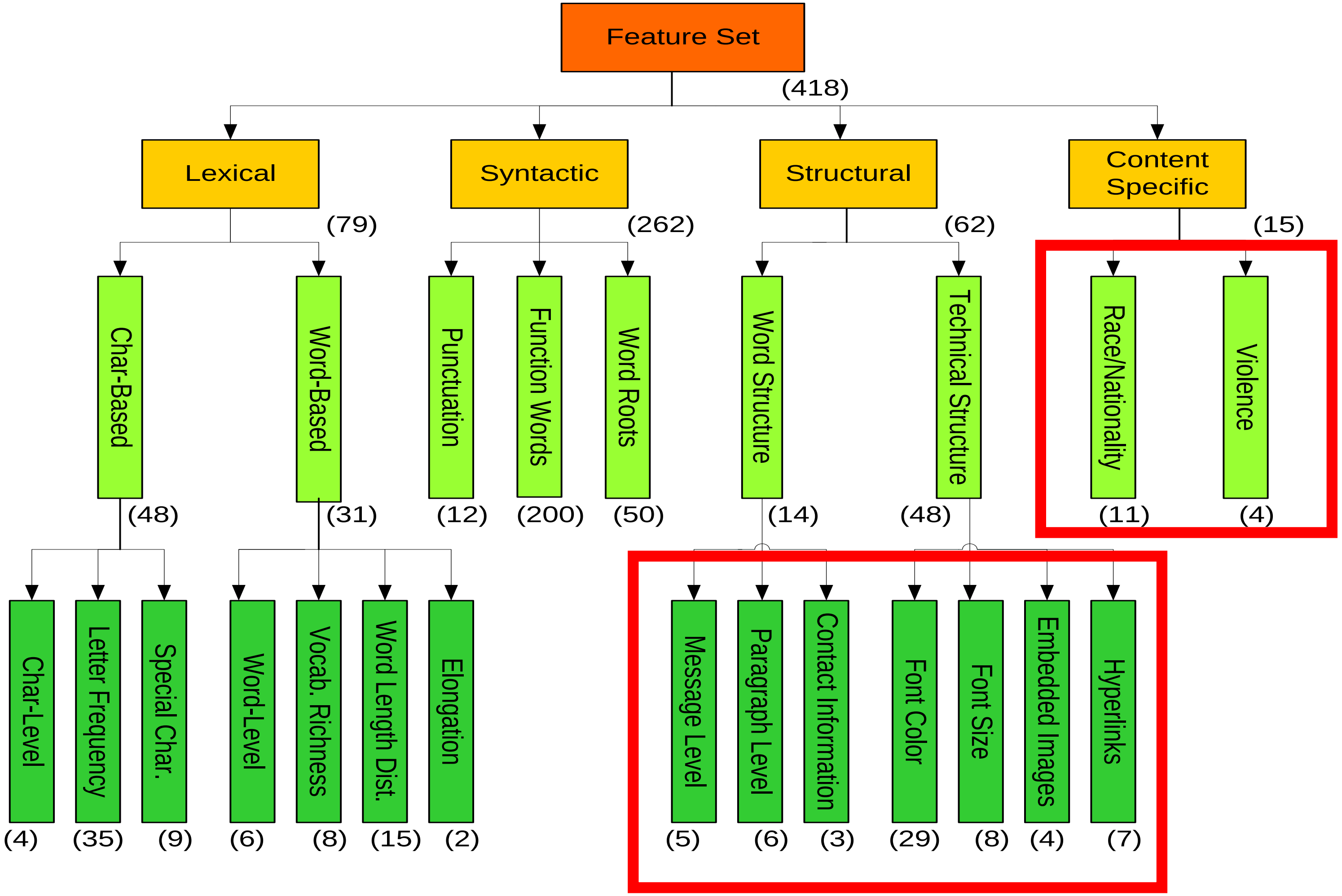
- 1994-1997, NSF DLI projects, DL, SE
- 1997, NIJ $1.2M project, UA/TPD
- 2000, NSF DG $1.6M, UA/TPD/PPD
- 2000, KCC founding, UA tech transfer; $2.6M VC funding
- 2001, Tucson, Phoenix, San Diego
- 2002, bubble burst, $2M additional funding (anti-dilution clause)
- 2003, DC snipper investigation use, NYT cover article; AZ, CA, NJ, IL
- 2009, SilverLake PE fund; COPLINK + i2
- 2011, sold to IBM ($500M); Chen exit
- 2017, IBM sold COPLINK to Forensic Logic

➔ COPLINK is in use in 5,000+ law enforcement jurisdictions and intelligence agencies in the U.S. and Europe, making significant contribution to public safety worldwide.

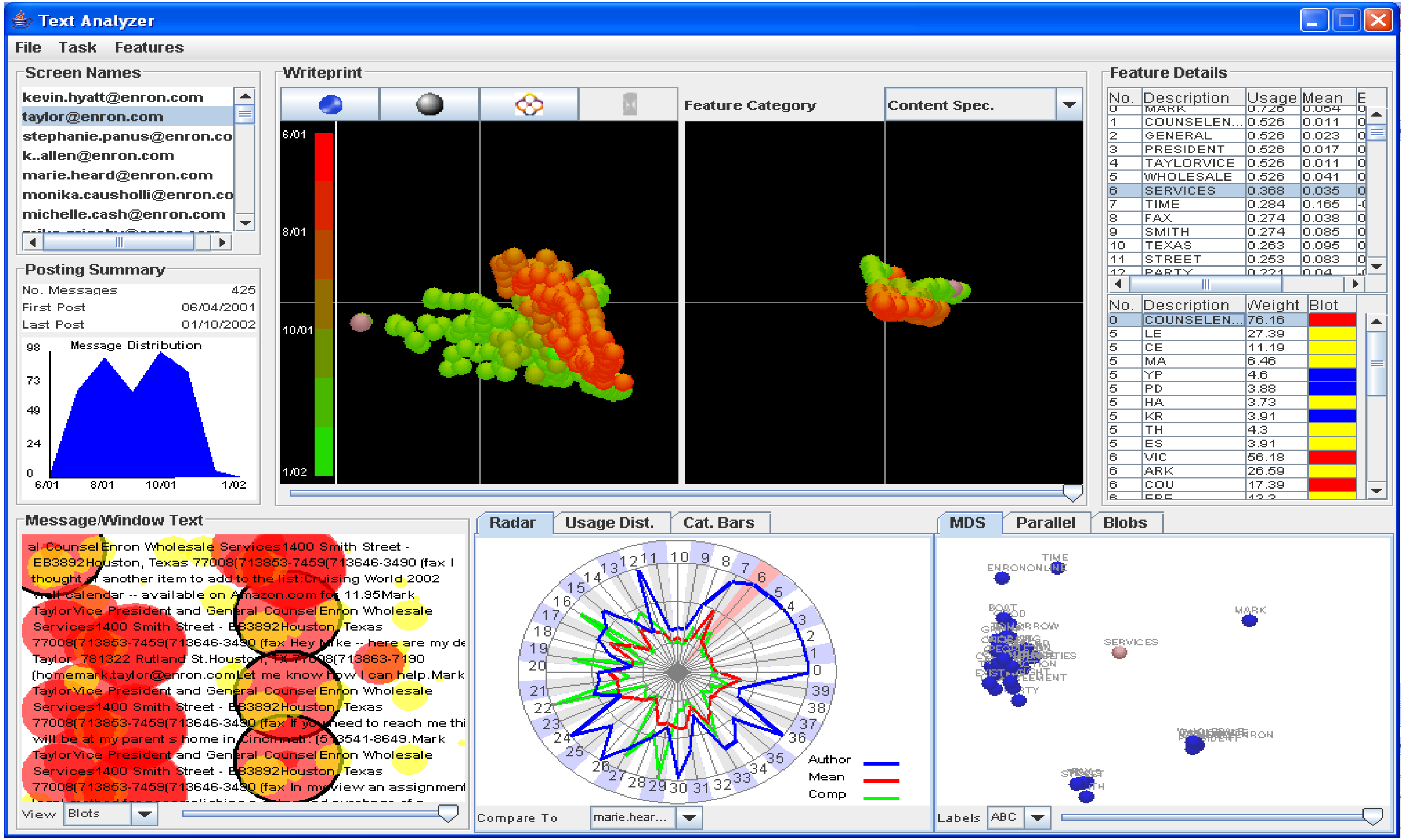
Dark Web: Countering Terrorism (2003-2014)

- Dark Web: Terrorists' and cyber criminals' use of the Internet
- Collection: Web sites, forums, blogs, YouTube, etc.
- 20 TBs in size, with close to 10B pages/files/messages (the entire LOC collection: 15 TBs)

Arabic Writeprint Feature for Authorship Analysis



CyberGate (Abbasi, et al., MISQ, 2008)



Screen Names

kevin.hyatt@enron.com
taylor@enron.com
stephanie.panus@enron.co
k..allen@enron.com
marie.heard@enron.com
monika.causholli@enron.co
michelle.cash@enron.com

Posting Summary

No. Messages	425
First Post	06/04/2001
Last Post	01/10/2002

Message Distribution

Feature Details

No.	Description	Usage	Mean	E
0	MARK	0.726	0.054	0
1	COUNSELEN...	0.526	0.011	0
2	GENERAL	0.526	0.023	0
3	PRESIDENT	0.526	0.017	0
4	TAYLORVICE	0.526	0.011	0
5	WHOLESALE	0.526	0.041	0
6	SERVICES	0.368	0.035	0
7	TIME	0.284	0.165	-0
8	FAX	0.274	0.038	0
9	SMITH	0.274	0.085	0
10	TEXAS	0.263	0.095	0
11	STREET	0.253	0.083	0
12	PARTY	0.221	0.04	-0

No.	Description	Weight	Blot
0	COUNSELEN...	76.16	🟥
5	LE	27.39	🟨
5	CE	11.19	🟨
5	MA	6.46	🟨
5	YP	4.6	🟦
5	PD	3.88	🟦
5	HA	3.73	🟨
5	KR	3.91	🟦
5	TH	4.3	🟨
5	ES	3.91	🟨
6	VIC	56.18	🟥
6	ARK	26.59	🟨
6	COU	17.39	🟥
6	EPE	13.3	🟨

Message/Window Text

al CounselEnron Wholesale Services1400 Smith Street - EB3892Houston, Texas 77008(713853-7459(713646-3490 (fax I thought of another item to add to the list:Cruising World 2002 wall calendar -- available on Amazon.com for 11.95Mark TaylorVice President and General CounselEnron Wholesale Services1400 Smith Street - EB3892Houston, Texas 77008(713853-7459(713646-3490 (fax Hey Mike -- here are my de Taylor 781322 Rutland St.Houston, TX 77008(713863-7190 (homemark.taylor@enron.comLet me know how I can help.Mark TaylorVice President and General CounselEnron Wholesale Services1400 Smith Street - EB3892Houston, Texas 77008(713853-7459(713646-3490 (fax If you need to reach me thi will be at my parent s home in Cincinnati: (513541-8649.Mark TaylorVice President and General CounselEnron Wholesale Services1400 Smith Street - EB3892Houston, Texas 77008(713853-7459(713646-3490 (fax In my view an assignment

View Blots

The Dark Web project in the Press

AP Associated Press

Project Seeks to Track Terror Web Posts, 11/11/2007

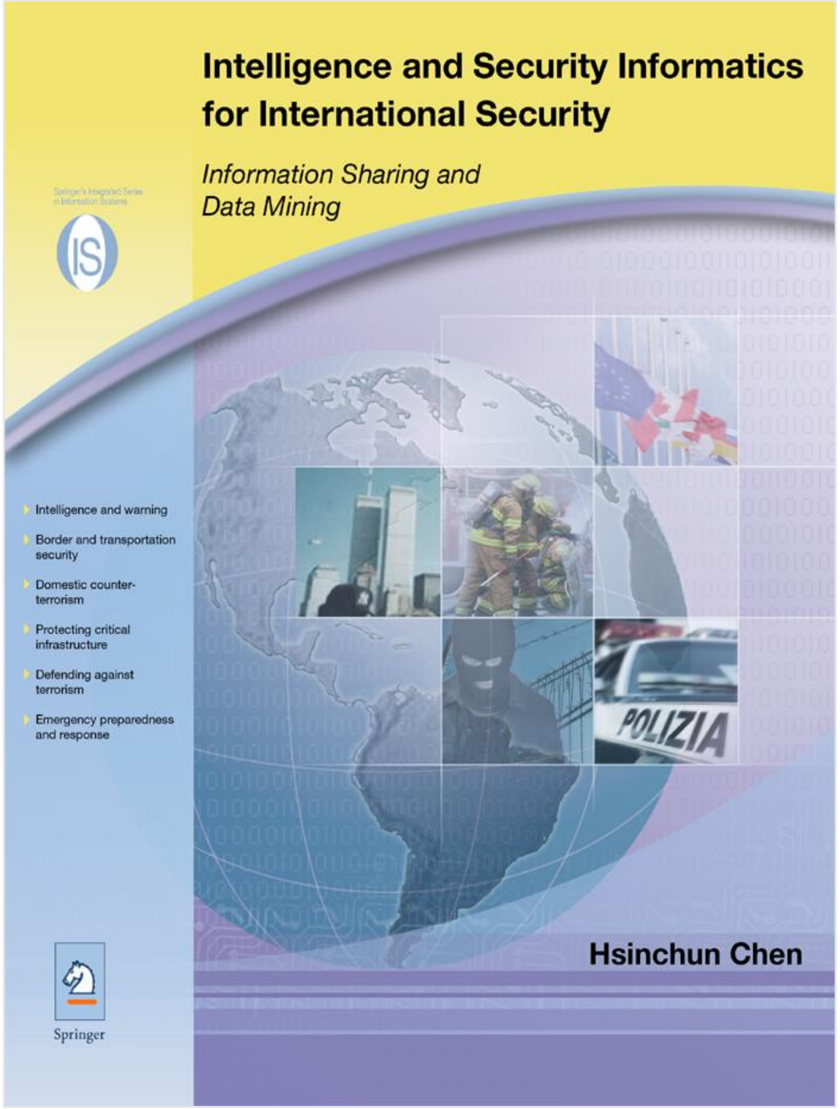
USA TODAY

Researchers say tool could trace online posts to terrorists, 11/11/2007

npr

Mathematicians Work to Help Track Terrorist Activity, 9/14/2007

ISI, Springer, 2006



Intelligence and Security Informatics for International Security

Information Sharing and Data Mining

Springer's Integrated Series in Information Systems

IS

- Intelligence and warning
- Border and transportation security
- Domestic counter-terrorism
- Protecting critical infrastructure
- Defending against terrorism
- Emergency preparedness and response

Hsinchun Chen

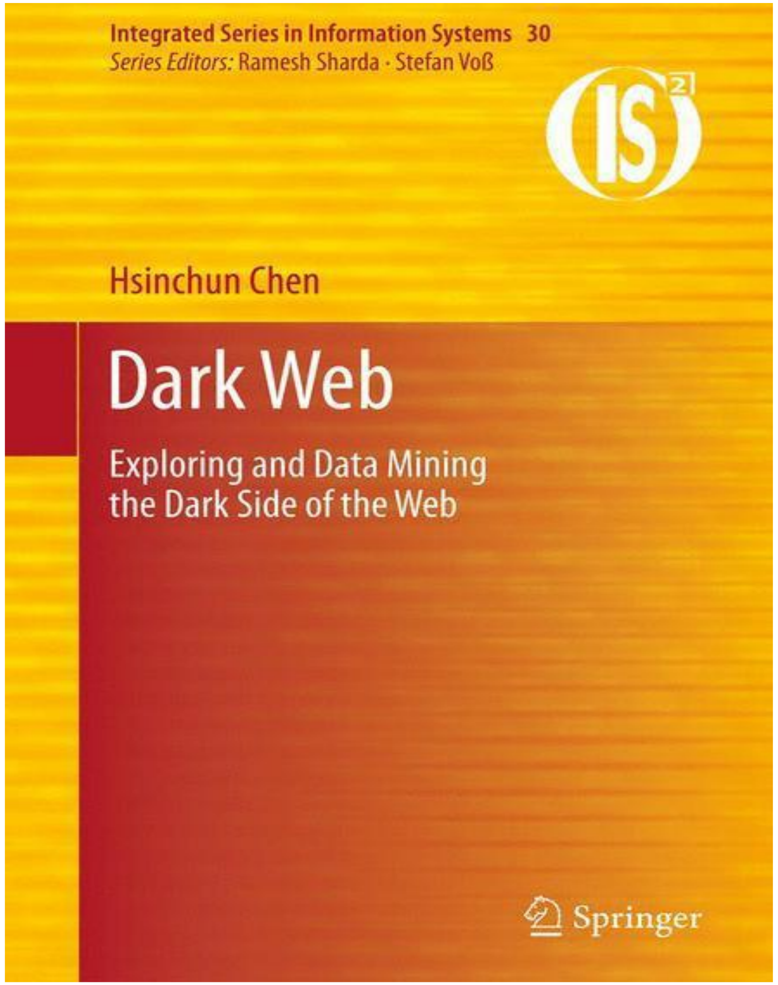
Springer

- Intelligence and Security Informatics (ISI) (Chen, *2006*)

- Data, text, and web mining
- From COPLINK to Dark Web

- **IEEE ISI, EISIC, PAISI ➔ 4000+ scholars, since 2003**

Dark Web, Springer, 2012



Integrated Series in Information Systems 30

Series Editors: Ramesh Sharda · Stefan Voß

Hsinchun Chen

Dark Web

Exploring and Data Mining the Dark Side of the Web

Springer

22 chapters, 451 pages, 150 illustrations (81 in color); Springer Integrated Series in Information Systems, 2012.

Selected TOC:

- Forum Spidering
- Link and Content Analysis
- Dark Network Analysis
- Interactional Coherence Analysis
- Dark Web Attribution System
- Authorship Analysis
- Sentiment Analysis
- Affect Analysis
- CyberGate Visualization
- Dark Web Forum Portal
- Case Studies: Jihadi Video Analysis, Extremist YouTube Videos, IEDs, WMDs, Women’s Forums

Pivoting to Cyber Security

AZProtector (Abbasi, Chen, et al., 2010; MISQ best paper)

MIS Quarterly

SPECIAL ISSUE

DETECTING FAKE WEBSITES: THE CONTRIBUTION OF STATISTICAL LEARNING THEORY[1]

Fraud Cues

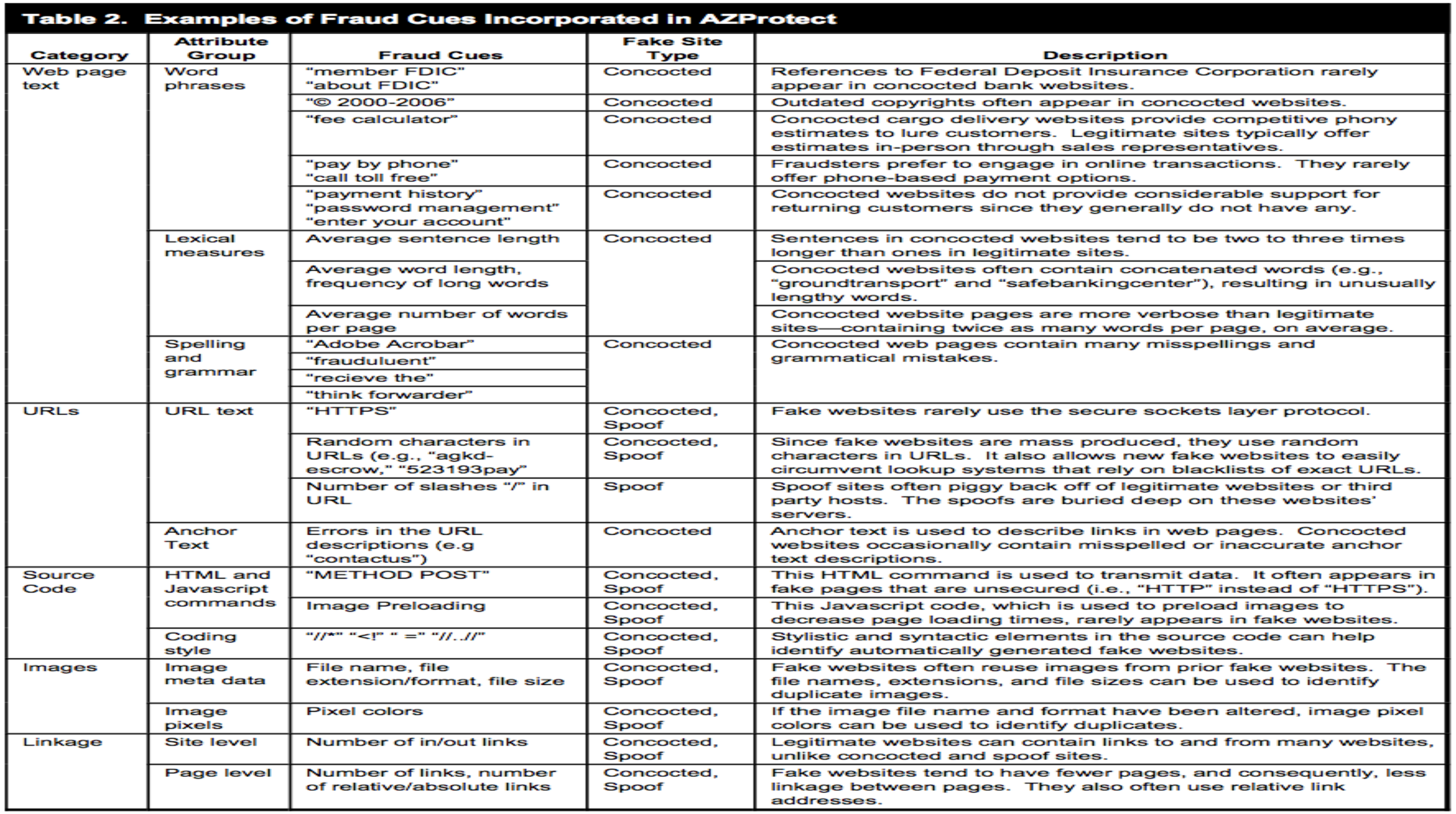
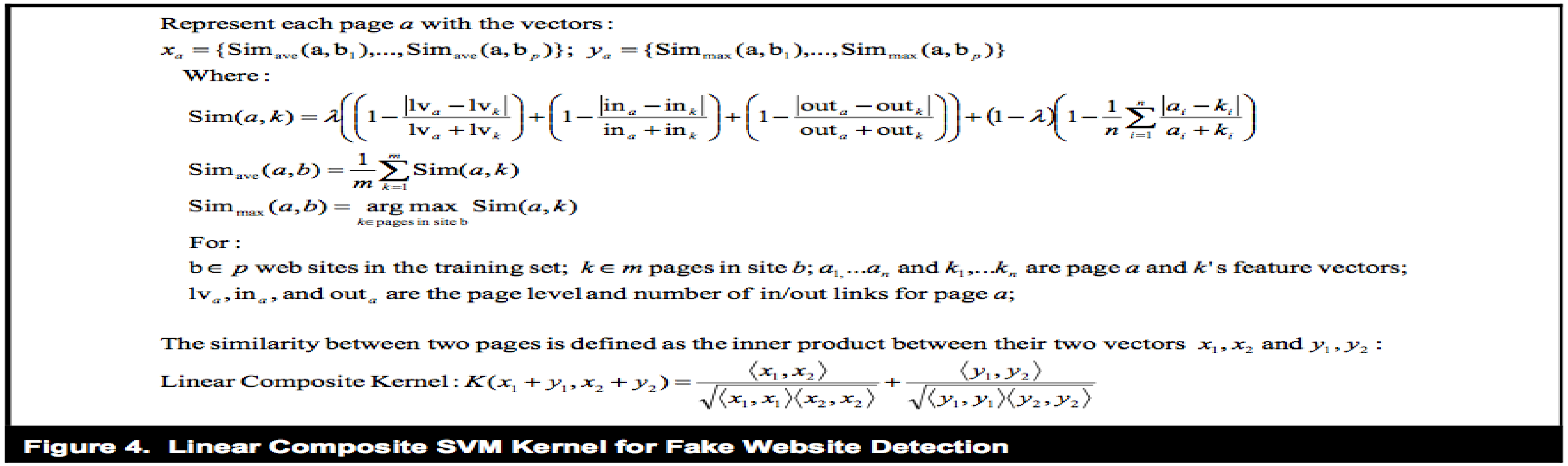


Table 2. Examples of Fraud Cues Incorporated in AZProtect

Category	Attribute Group	Fraud Cues	Fake Site Type	Description
Web page text	Word phrases	"member FDIC" "about FDIC"	Concocted	References to Federal Deposit Insurance Corporation rarely appear in concocted bank websites.
		"© 2000-2006"	Concocted	Outdated copyrights often appear in concocted websites.
		"fee calculator"	Concocted	Concocted cargo delivery websites provide competitive phony estimates to lure customers. Legitimate sites typically offer estimates in-person through sales representatives.
		"pay by phone" "call toll free"	Concocted	Fraudsters prefer to engage in online transactions. They rarely offer phone-based payment options.
		"payment history" "password management" "enter your account"	Concocted	Concocted websites do not provide considerable support for returning customers since they generally do not have any.
	Lexical measures	Average sentence length	Concocted	Sentences in concocted websites tend to be two to three times longer than ones in legitimate sites.
		Average word length, frequency of long words		Concocted websites often contain concatenated words (e.g., "groundtransport" and "safebankingcenter"), resulting in unusually lengthy words.
		Average number of words per page		Concocted website pages are more verbose than legitimate sites—containing twice as many words per page, on average.
	Spelling and grammar	"Adobe Acrobar"	Concocted	Concocted web pages contain many misspellings and grammatical mistakes.
		"frauduluent"		
		"recieve the"		
		"think forwarder"		
URLs	URL text	"HTTPS"	Concocted, Spoof	Fake websites rarely use the secure sockets layer protocol.
		Random characters in URLs (e.g., "agkd-escrow," "523193pay"	Concocted, Spoof	Since fake websites are mass produced, they use random characters in URLs. It also allows new fake websites to easily circumvent lookup systems that rely on blacklists of exact URLs.
		Number of slashes "/" in URL	Spoof	Spoof sites often piggy back off of legitimate websites or third party hosts. The spoofs are buried deep on these websites' servers.
	Anchor Text	Errors in the URL descriptions (e.g "contactus")	Concocted	Anchor text is used to describe links in web pages. Concocted websites occasionally contain misspelled or inaccurate anchor text descriptions.
Source Code	HTML and Javascript commands	"METHOD POST"	Concocted, Spoof	This HTML command is used to transmit data. It often appears in fake pages that are unsecured (i.e., "HTTP" instead of "HTTPS").
		Image Preloading	Concocted, Spoof	This Javascript code, which is used to preload images to decrease page loading times, rarely appears in fake websites.
	Coding style	"//*" "<!" " =" "//../ /"	Concocted, Spoof	Stylistic and syntactic elements in the source code can help identify automatically generated fake websites.
Images	Image meta data	File name, file extension/format, file size	Concocted, Spoof	Fake websites often reuse images from prior fake websites. The file names, extensions, and file sizes can be used to identify duplicate images.
	Image pixels	Pixel colors	Concocted, Spoof	If the image file name and format have been altered, image pixel colors can be used to identify duplicates.
Linkage	Site level	Number of in/out links	Concocted, Spoof	Legitimate websites can contain links to and from many websites, unlike concocted and spoof sites.
	Page level	Number of links, number of relative/absolute links	Concocted, Spoof	Fake websites tend to have fewer pages, and consequently, less linkage between pages. They also often use relative link addresses.

Escrow Kernnel for Detecting Fake Web Sites



Represent each page a with the vectors:

$$x_a = \{\text{Sim}_{\text{ave}}(\text{a}, \text{b}_1), \ldots, \text{Sim}_{\text{ave}}(\text{a}, \text{b}_p)\}; \; y_a = \{\text{Sim}_{\text{max}}(\text{a}, \text{b}_1), \ldots, \text{Sim}_{\text{max}}(\text{a}, \text{b}_p)\}$$

Where:

$$\text{Sim}(a,k) = \lambda\left(\left(1-\frac{|\text{lv}_a - \text{lv}_k|}{\text{lv}_a + \text{lv}_k}\right) + \left(1-\frac{|\text{in}_a - \text{in}_k|}{\text{in}_a + \text{in}_k}\right) + \left(1-\frac{|\text{out}_a - \text{out}_k|}{\text{out}_a + \text{out}_k}\right)\right) + (1-\lambda)\left(1-\frac{1}{n}\sum_{i=1}^{n}\frac{|a_i - k_i|}{a_i + k_i}\right)$$

$$\text{Sim}_{\text{ave}}(a,b) = \frac{1}{m}\sum_{k=1}^{m}\text{Sim}(a,k)$$

$$\text{Sim}_{\text{max}}(a,b) = \underset{k \in \text{pages in site b}}{\arg\max}\ \text{Sim}(a,k)$$

For:

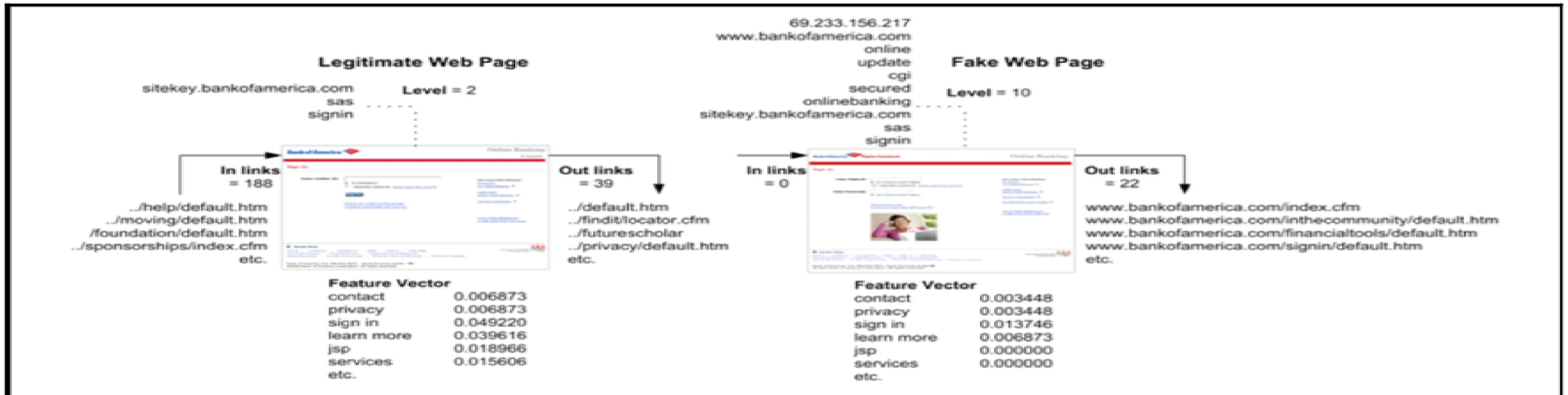
b $\in p$ web sites in the training set; $k \in m$ pages in site b; $a_{1,} \ldots a_n$ and $k_1, \ldots k_n$ are page a and k's feature vectors;

lv_a, in_a, and out_a are the page level and number of in/out links for page a;

The similarity between two pages is defined as the inner product between their two vectors x_1, x_2 and y_1, y_2:

$$\text{Linear Composite Kernel}: K(x_1 + y_1, x_2 + y_2) = \frac{\langle x_1, x_2\rangle}{\sqrt{\langle x_1, x_1\rangle\langle x_2, x_2\rangle}} + \frac{\langle y_1, y_2\rangle}{\sqrt{\langle y_1, y_1\rangle\langle y_2, y_2\rangle}}$$

Figure 4. Linear Composite SVM Kernel for Fake Website Detection



Category	Comparsion Websites	Legitimate Bank Web Page		Spoof Bank Web Page	
		$\text{Sim}_{\text{ave}}(a, b)$	$\text{Sim}_{\text{max}}(a, b)$	$\text{Sim}_{\text{ave}}(a, b)$	$\text{Sim}_{\text{max}}(a, b)$
Legitimate	www.bankofamerica.com	0.495098	1.000000	0.314208	0.392708
	www.citibank.com	0.387834	0.577704	0.247735	0.386828
Fake	beliebo.nl\media\2007\Paypal	0.317838	0.383798	0.522842	0.555874
	65.85.238.84\ebay.monitors\store.ebay	0.257576	0.297611	0.510915	0.577157

Figure 5. Comparing Two Web Pages Against Legitimate and Fake Websites

Performance vs. Classifier and Lookup Systems

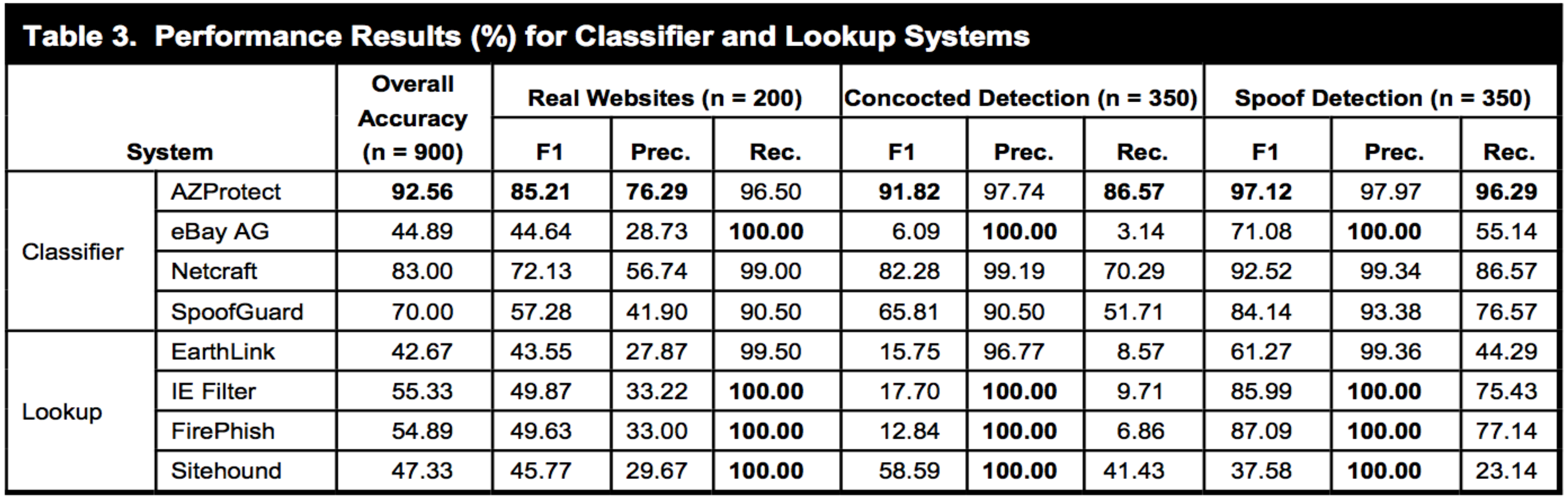


Table 3. Performance Results (%) for Classifier and Lookup Systems

System		Overall Accuracy (n = 900)	Real Websites (n = 200)			Concocted Detection (n = 350)			Spoof Detection (n = 350)		
			F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.
Classifier	AZProtect	**92.56**	**85.21**	**76.29**	96.50	**91.82**	97.74	**86.57**	**97.12**	97.97	**96.29**
	eBay AG	44.89	44.64	28.73	**100.00**	6.09	**100.00**	3.14	71.08	**100.00**	55.14
	Netcraft	83.00	72.13	56.74	99.00	82.28	99.19	70.29	92.52	99.34	86.57
	SpoofGuard	70.00	57.28	41.90	90.50	65.81	90.50	51.71	84.14	93.38	76.57
Lookup	EarthLink	42.67	43.55	27.87	99.50	15.75	96.77	8.57	61.27	99.36	44.29
	IE Filter	55.33	49.87	33.22	**100.00**	17.70	**100.00**	9.71	85.99	**100.00**	75.43
	FirePhish	54.89	49.63	33.00	**100.00**	12.84	**100.00**	6.86	87.09	**100.00**	77.14
	Sitehound	47.33	45.77	29.67	**100.00**	58.59	**100.00**	41.43	37.58	**100.00**	23.14

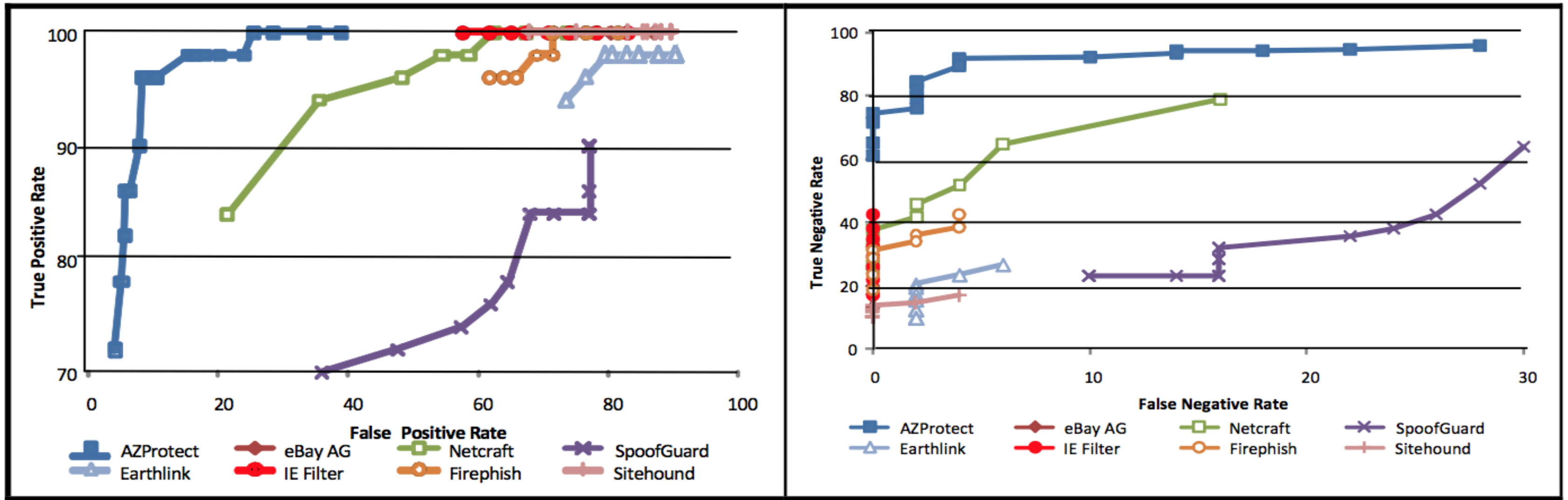


Figure 8. ROC Curves for Classifier and Lookup Systems

Performance vs. Other ML Techniques

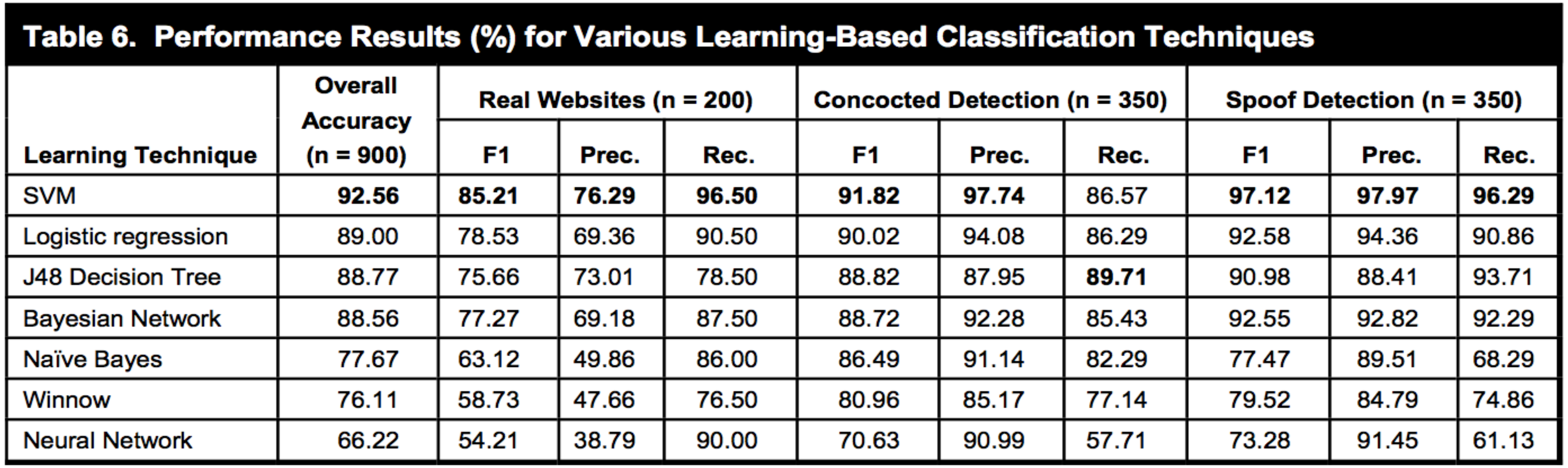


Table 6. Performance Results (%) for Various Learning-Based Classification Techniques

Learning Technique	Overall Accuracy (n = 900)	Real Websites (n = 200)			Concocted Detection (n = 350)			Spoof Detection (n = 350)		
		F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.
SVM	**92.56**	**85.21**	**76.29**	**96.50**	**91.82**	**97.74**	86.57	**97.12**	**97.97**	**96.29**
Logistic regression	89.00	78.53	69.36	90.50	90.02	94.08	86.29	92.58	94.36	90.86
J48 Decision Tree	88.77	75.66	73.01	78.50	88.82	87.95	**89.71**	90.98	88.41	93.71
Bayesian Network	88.56	77.27	69.18	87.50	88.72	92.28	85.43	92.55	92.82	92.29
Naïve Bayes	77.67	63.12	49.86	86.00	86.49	91.14	82.29	77.47	89.51	68.29
Winnow	76.11	58.73	47.66	76.50	80.96	85.17	77.14	79.52	84.79	74.86
Neural Network	66.22	54.21	38.79	90.00	70.63	90.99	57.71	73.28	91.45	61.13

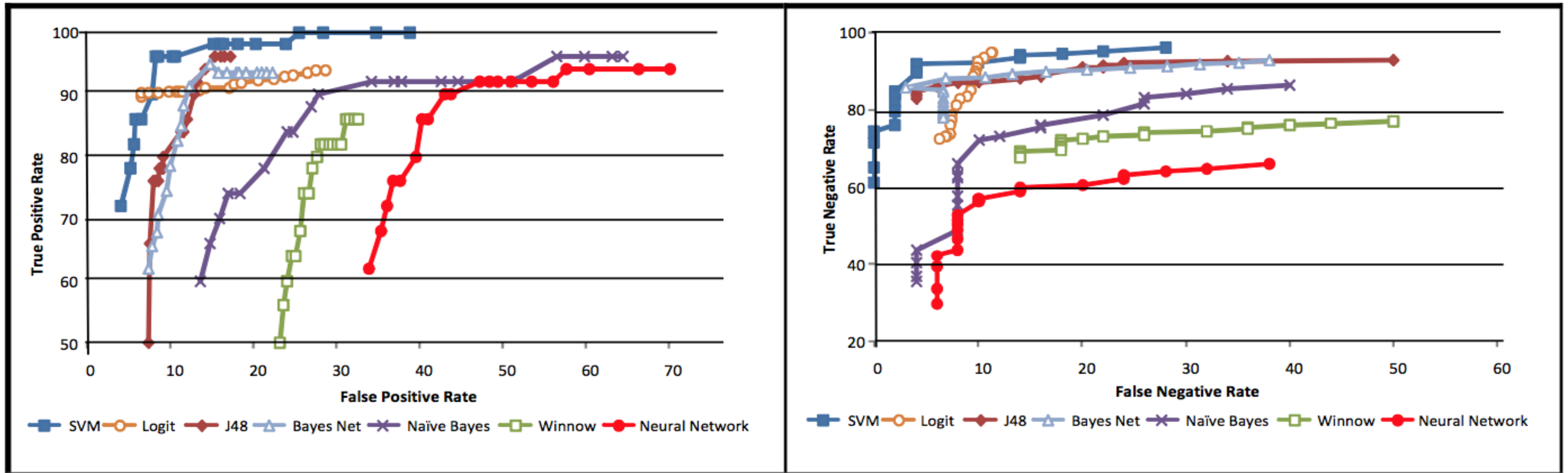


Figure 9. ROC Curves for Various Learning Classifiers

Azsecure Cybersecurity Analytics Program:

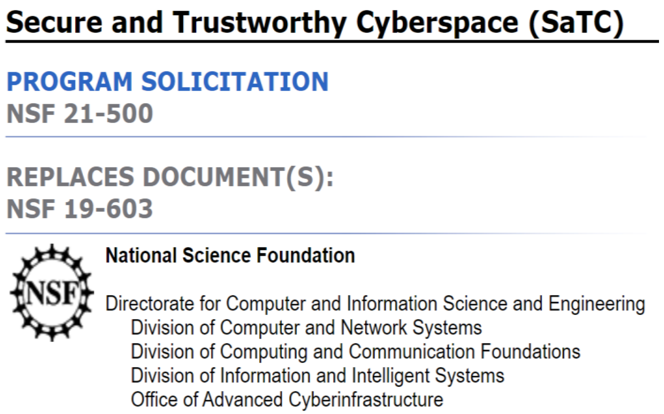
(1) Dark Web Analytics for studying international hacker community, forums, and markets;

(2) Privacy and PII (Personally Identifiable Information) Analytics for identifying and alleviating privacy risks for vulnerable populations;

(3) Adversarial Malware Generation and Evasion for adversarial AI in cybersecurity; and

(4) Smart Vulnerability Assessment for scientific workflows and OSS (Open Source Software) vulnerability analytics and mitigation.

AZSecure Cybersecurity Analytics Program (2010-present): SaTC, SFS, ACI



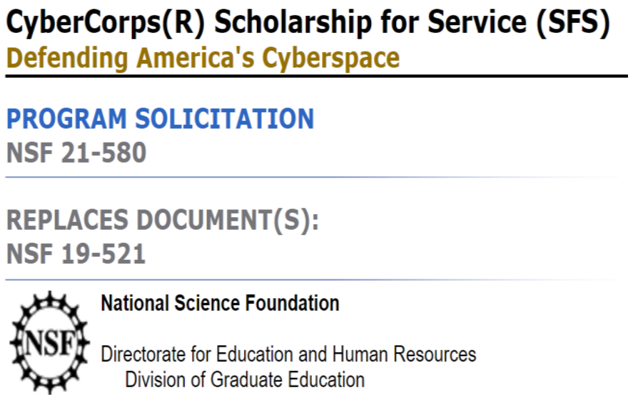
Secure and Trustworthy Cyberspace (SaTC)

PROGRAM SOLICITATION
NSF 21-500

REPLACES DOCUMENT(S):
NSF 19-603

National Science Foundation

Directorate for Computer and Information Science and Engineering
Division of Computer and Network Systems
Division of Computing and Communication Foundations
Division of Information and Intelligent Systems
Office of Advanced Cyberinfrastructure



CyberCorps(R) Scholarship for Service (SFS)

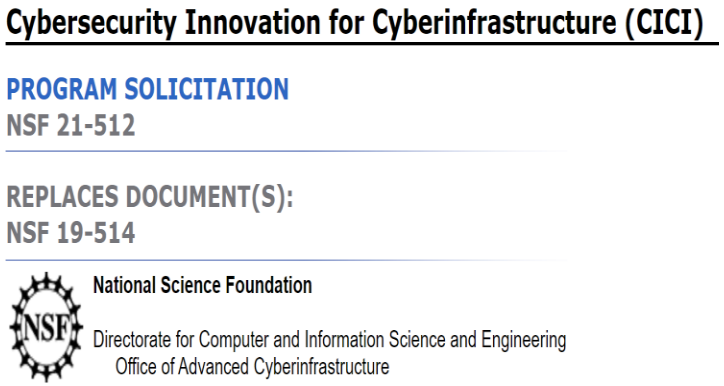
Defending America's Cyberspace

PROGRAM SOLICITATION
NSF 21-580

REPLACES DOCUMENT(S):
NSF 19-521

National Science Foundation

Directorate for Education and Human Resources
Division of Graduate Education



Cybersecurity Innovation for Cyberinfrastructure (CICI)

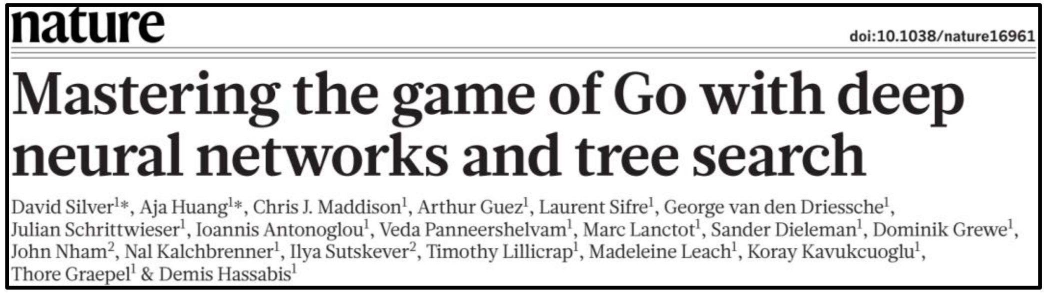
PROGRAM SOLICITATION
NSF 21-512

REPLACES DOCUMENT(S):
NSF 19-514

National Science Foundation

Directorate for Computer and Information Science and Engineering
Office of Advanced Cyberinfrastructure

nature
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
At last — a computer program that can beat a champion Go player PAGE 484
ALL SYSTEMS GO
CONSERVATION
SONGBIRDS À LA CARTE
Illegal harvest of millions of Mediterranean birds
PAGE 452
RESEARCH ETHICS
SAFEGUARD TRANSPARENCY
Don't let openness backfire on individuals
PAGE 459
POPULAR SCIENCE
WHEN GENES GOT 'SELFISH'
Dawkins's calling card forty years on
PAGE 462
NATURE.COM/NATURE
28 January 2016 £10
Vol. 529, No. 7587

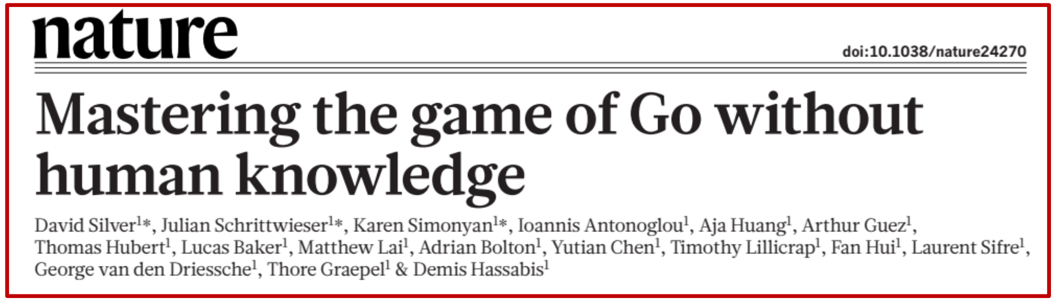


nature

doi:10.1038/nature16961

Mastering the game of Go with deep neural networks and tree search

David Silver[1]*, Aja Huang[1]*, Chris J. Maddison[1], Arthur Guez[1], Laurent Sifre[1], George van den Driessche[1], Julian Schrittwieser[1], Ioannis Antonoglou[1], Veda Panneershelvam[1], Marc Lanctot[1], Sander Dieleman[1], Dominik Grewe[1], John Nham[2], Nal Kalchbrenner[1], Ilya Sutskever[2], Timothy Lillicrap[1], Madeleine Leach[1], Koray Kavukcuoglu[1], Thore Graepel[1] & Demis Hassabis[1]



nature

doi:10.1038/nature24270

Mastering the game of Go without human knowledge

David Silver[1]*, Julian Schrittwieser[1]*, Karen Simonyan[1]*, Ioannis Antonoglou[1], Aja Huang[1], Arthur Guez[1], Thomas Hubert[1], Lucas Baker[1], Matthew Lai[1], Adrian Bolton[1], Yutian Chen[1], Timothy Lillicrap[1], Fan Hui[1], Laurent Sifre[1], George van den Driessche[1], Thore Graepel[1] & Demis Hassabis[1]

AI & Deep Learning: From AlphaGo to Autonomous Vehicles (2012-)

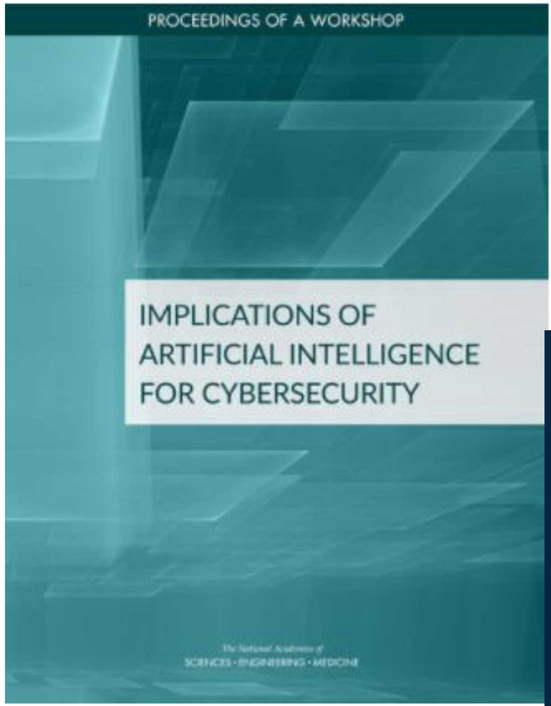
➔

Hacker Web, AZSecure projects at UA/MIS AI Lab (2010-present)

AI and Cybersecurity

- AI and Cybersecurity → not just buzzwords!
  - Noted as a national security priority by NSF, NSTC, and NAS.

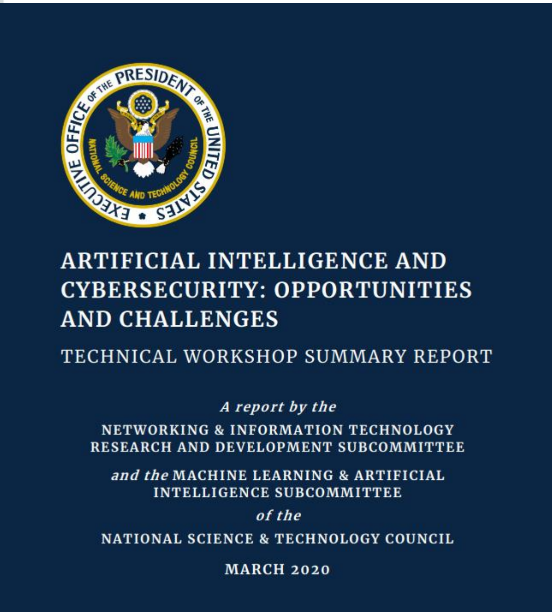
- Role of AI for Cybersecurity :
  1. Automate common cybersecurity tasks
  2. Identify patterns in large datasets



PROCEEDINGS OF A WORKSHOP

IMPLICATIONS OF ARTIFICIAL INTELLIGENCE FOR CYBERSECURITY

The National Academies of SCIENCES · ENGINEERING · MEDICINE



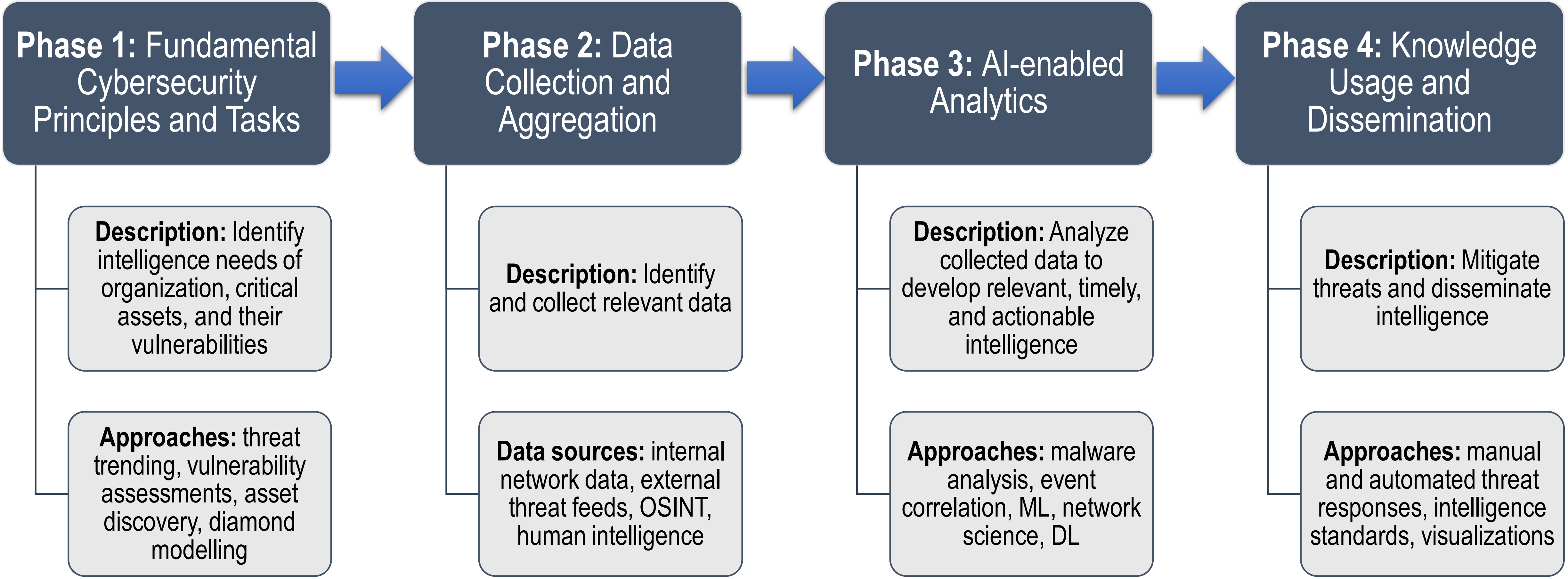
ARTIFICIAL INTELLIGENCE AND CYBERSECURITY: OPPORTUNITIES AND CHALLENGES

TECHNICAL WORKSHOP SUMMARY REPORT

A report by the
NETWORKING & INFORMATION TECHNOLOGY RESEARCH AND DEVELOPMENT SUBCOMMITTEE
and the MACHINE LEARNING & ARTIFICIAL INTELLIGENCE SUBCOMMITTEE
of the
NATIONAL SCIENCE & TECHNOLOGY COUNCIL

MARCH 2020

AI for Cybersecurity – An Analytics Approach



Phase 1: Fundamental Cybersecurity Principles and Tasks

Description: Identify intelligence needs of organization, critical assets, and their vulnerabilities

Approaches: threat trending, vulnerability assessments, asset discovery, diamond modelling

Phase 2: Data Collection and Aggregation

Description: Identify and collect relevant data

Data sources: internal network data, external threat feeds, OSINT, human intelligence

Phase 3: AI-enabled Analytics

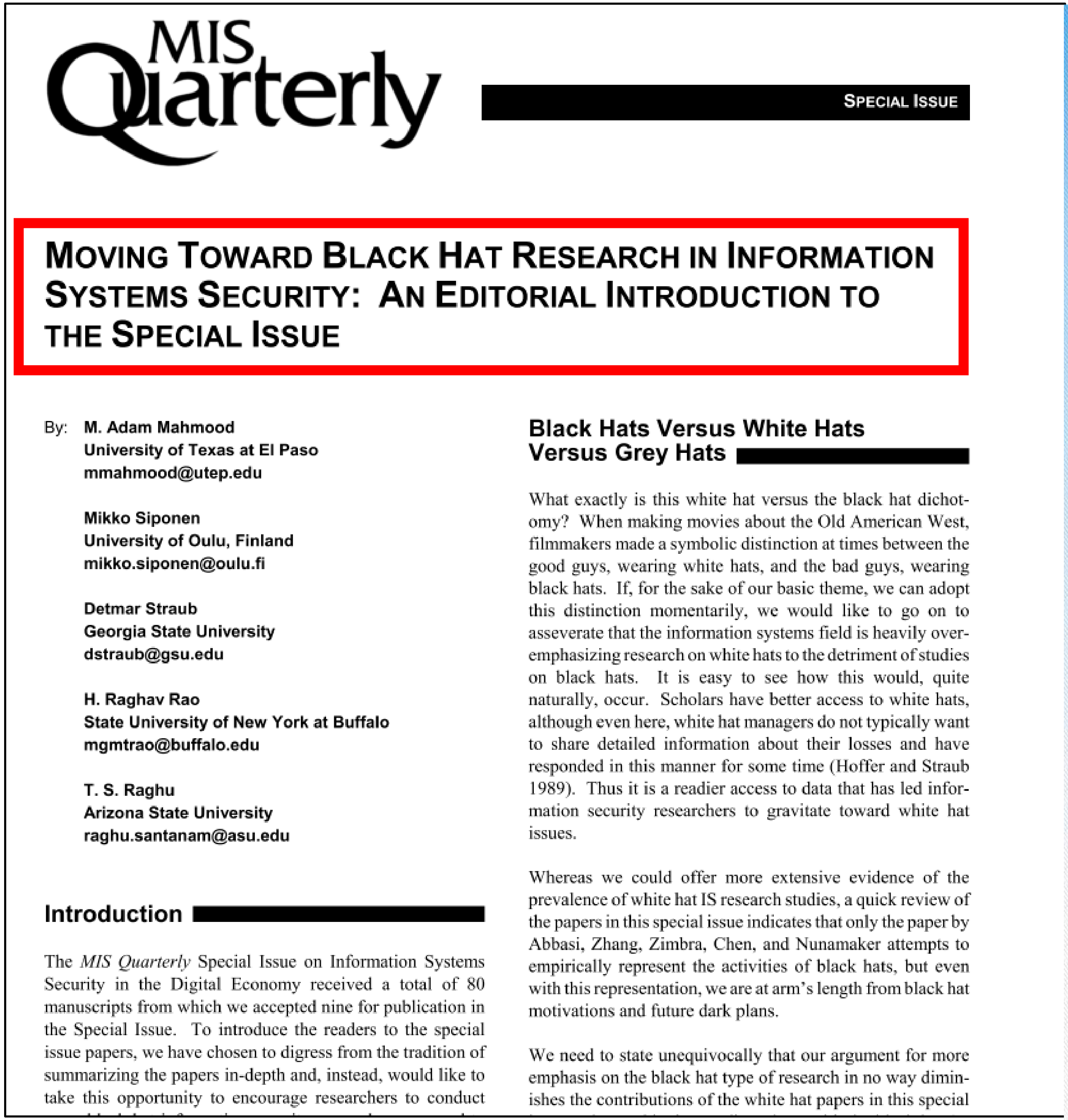
Description: Analyze collected data to develop relevant, timely, and actionable intelligence

Approaches: malware analysis, event correlation, ML, network science, DL

Phase 4: Knowledge Usage and Dissemination

Description: Mitigate threats and disseminate intelligence

Approaches: manual and automated threat responses, intelligence standards, visualizations



MIS Quarterly

Special Issue

Moving Toward Black Hat Research in Information Systems Security: An Editorial Introduction to the Special Issue

By: M. Adam Mahmood
University of Texas at El Paso
mmahmood@utep.edu

Mikko Siponen
University of Oulu, Finland
mikko.siponen@oulu.fi

Detmar Straub
Georgia State University
dstraub@gsu.edu

H. Raghav Rao
State University of New York at Buffalo
mgmtrao@buffalo.edu

T. S. Raghu
Arizona State University
raghu.santanam@asu.edu

Introduction

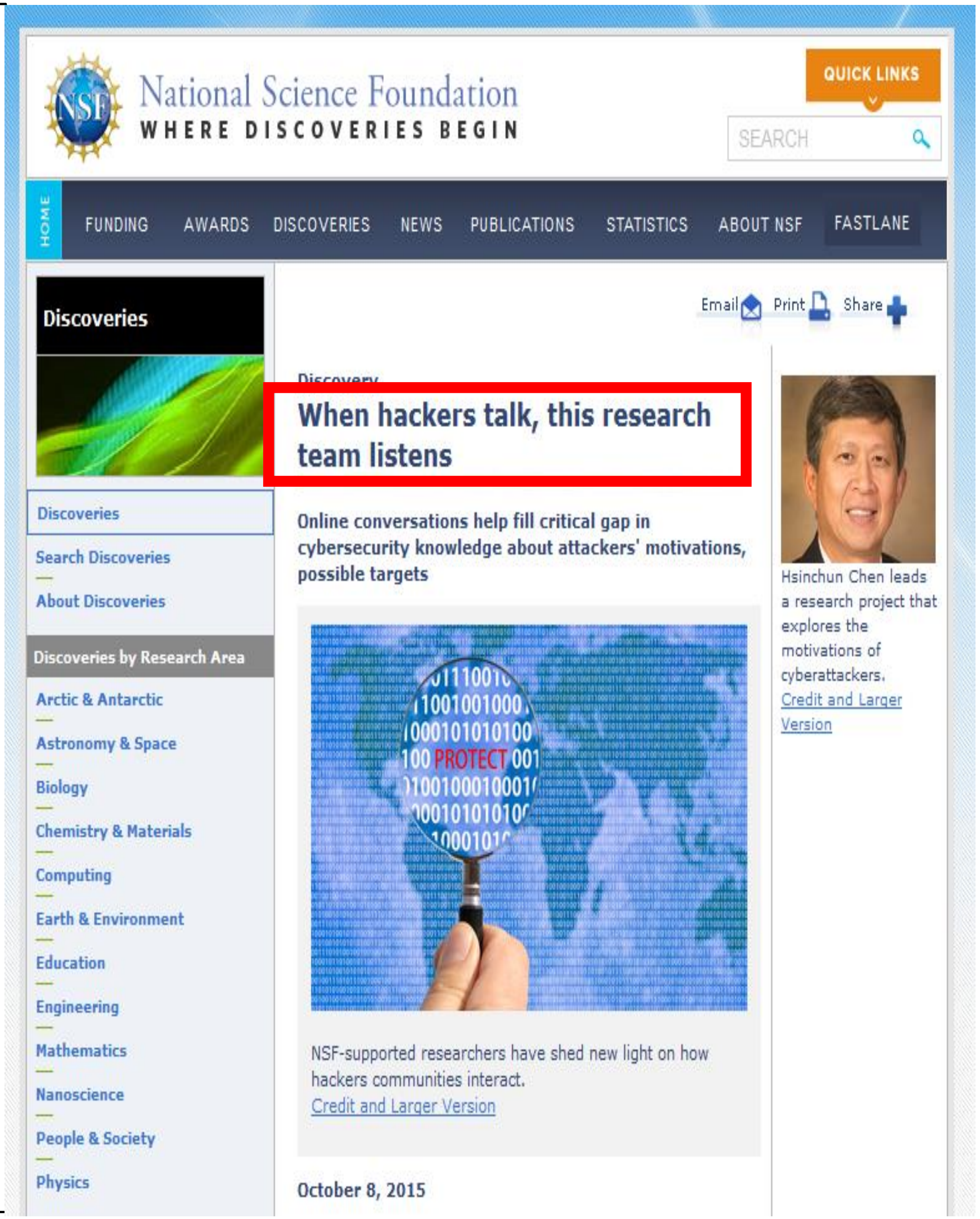
The *MIS Quarterly* Special Issue on Information Systems Security in the Digital Economy received a total of 80 manuscripts from which we accepted nine for publication in the Special Issue. To introduce the readers to the special issue papers, we have chosen to digress from the tradition of summarizing the papers in-depth and, instead, would like to take this opportunity to encourage researchers to conduct

Black Hats Versus White Hats Versus Grey Hats

What exactly is this white hat versus the black hat dichotomy? When making movies about the Old American West, filmmakers made a symbolic distinction at times between the good guys, wearing white hats, and the bad guys, wearing black hats. If, for the sake of our basic theme, we can adopt this distinction momentarily, we would like to go on to asseverate that the information systems field is heavily overemphasizing research on white hats to the detriment of studies on black hats. It is easy to see how this would, quite naturally, occur. Scholars have better access to white hats, although even here, white hat managers do not typically want to share detailed information about their losses and have responded in this manner for some time (Hoffer and Straub 1989). Thus it is a readier access to data that has led information security researchers to gravitate toward white hat issues.

Whereas we could offer more extensive evidence of the prevalence of white hat IS research studies, a quick review of the papers in this special issue indicates that only the paper by Abbasi, Zhang, Zimbra, Chen, and Nunamaker attempts to empirically represent the activities of black hats, but even with this representation, we are at arm's length from black hat motivations and future dark plans.

We need to state unequivocally that our argument for more emphasis on the black hat type of research in no way diminishes the contributions of the white hat papers in this special



Discovery

When hackers talk, this research team listens

Online conversations help fill critical gap in cybersecurity knowledge about attackers' motivations, possible targets

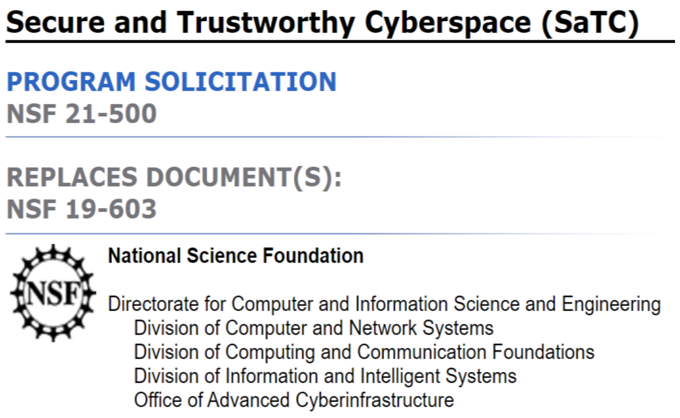
NSF-supported researchers have shed new light on how hackers communities interact.
Credit and Larger Version

October 8, 2015

Hsinchun Chen leads a research project that explores the motivations of cyberattackers.
Credit and Larger Version

Dark Web Analytics: studying international hacker community, forums, and markets

(ACI, 2012-2017; SaTC 2013-2018; SFS-1, 2012-2018; SaTC 2019-; SFS-2, 2019-)



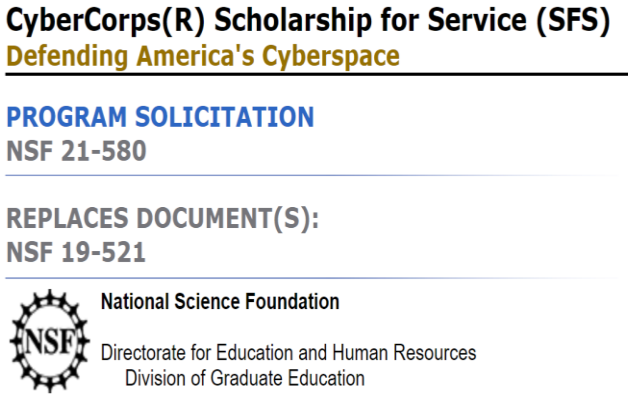
Secure and Trustworthy Cyberspace (SaTC)

PROGRAM SOLICITATION
NSF 21-500

REPLACES DOCUMENT(S):
NSF 19-603

National Science Foundation

Directorate for Computer and Information Science and Engineering
Division of Computer and Network Systems
Division of Computing and Communication Foundations
Division of Information and Intelligent Systems
Office of Advanced Cyberinfrastructure



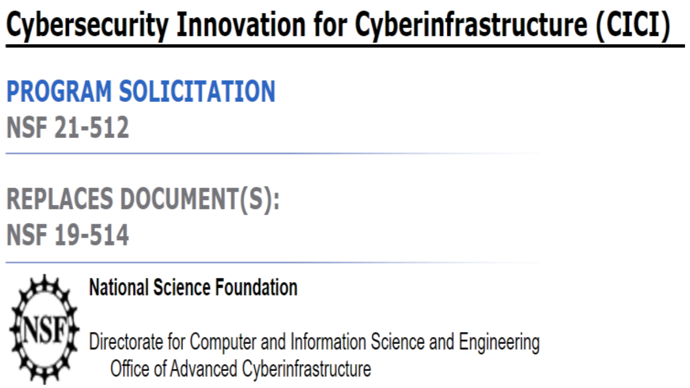
CyberCorps(R) Scholarship for Service (SFS)
Defending America's Cyberspace

PROGRAM SOLICITATION
NSF 21-580

REPLACES DOCUMENT(S):
NSF 19-521

National Science Foundation

Directorate for Education and Human Resources
Division of Graduate Education



Cybersecurity Innovation for Cyberinfrastructure (CICI)

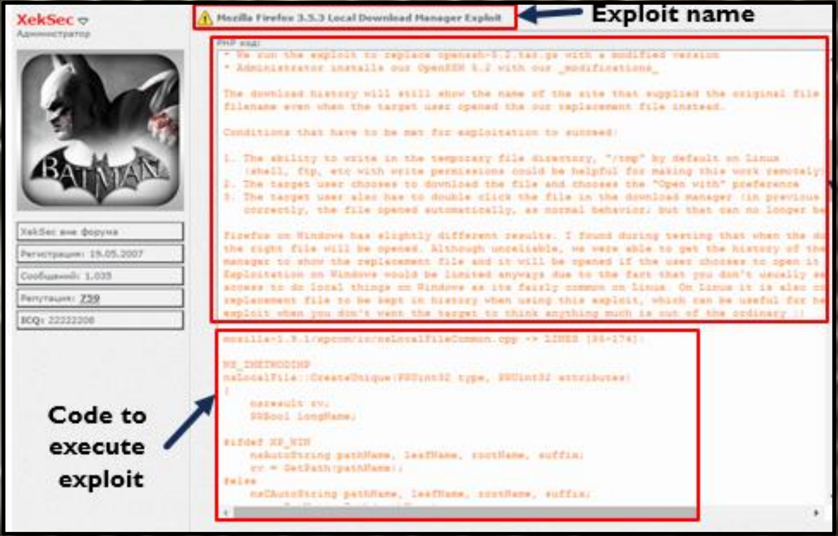
PROGRAM SOLICITATION
NSF 21-512

REPLACES DOCUMENT(S):
NSF 19-514

National Science Foundation

Directorate for Computer and Information Science and Engineering
Office of Advanced Cyberinfrastructure

Hacker Web



Exploit name

Mozilla Firefox 3.5.3 Local Download Manager Exploit

XekSec
Администратор

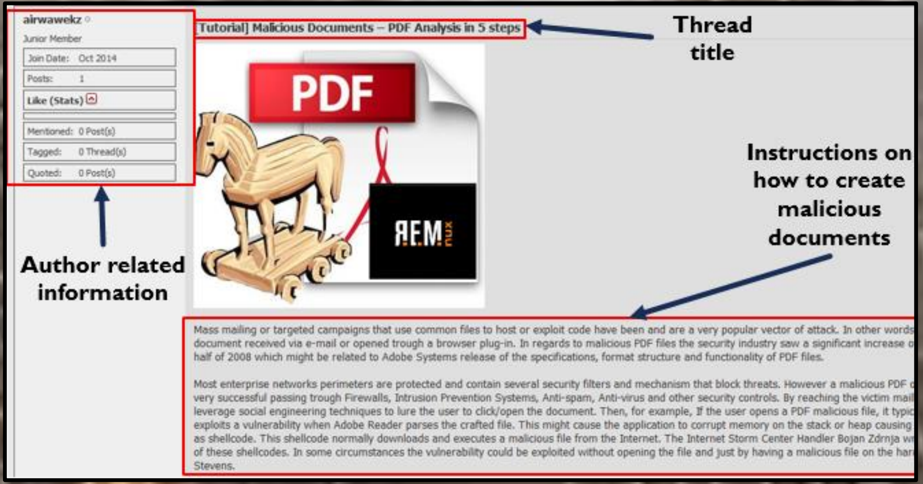
XekSec вне форума
Регистрация: 19.05.2007
Сообщений: 1,035
Репутация: 739
ICQ: 22222208

```
PHP код:
* We run the exploit to replace openssh-5.2.tar.gz with a modified version
* Administrator installs our OpenSSH 5.2 with our _modifications_

The download history will still show the name of the site that supplied the original file
filename even when the target user opened the our replacement file instead.

Conditions that have to be met for exploitation to succeed:

1. The ability to write in the temporary file directory, "/tmp" by default on Linux
   (shell, ftp, etc with write permissions could be helpful for making this work remotely
2. The target user chooses to download the file and chooses the "Open with" preference
3. The target user also has to double click the file in the download manager (in previous
   correctly, the file opened automatically, as normal behavior) but that can no longer be

Firefox on Windows has slightly different results. I found during testing that when the do
the right file will be opened. Although unreliable, we were able to get the history of the
manager to show the replacement file and it will be opened if the user chooses to open it
Exploitation on Windows would be limited anyways due to the fact that you don't usually se
access to do local things on Windows as its fairly common on Linux. On Linux it is also co
replacement file to be kept in history when using this exploit, which can be useful for he
exploit when you don't want the target to think anything much is out of the ordinary :)
```

Code to execute exploit

```
mozilla-1.9.1/xpcom/io/nsLocalFileCommon.cpp -> LINES [86-174]:

NS_IMETHODIMP
nsLocalFile::CreateUnique(PRUint32 type, PRUint32 attributes)
{
    nsresult rv;
    PRBool longName;

#ifdef XP_WIN
    nsAutoString pathName, leafName, rootName, suffix;
    rv = GetPath(pathName);
#else
    nsCAutoString pathName, leafName, rootName, suffix;
```

Forum post with source code to exploit Mozilla Firefox 3.5.3



[Tutorial] Malicious Documents – PDF Analysis in 5 steps

Thread title

airwawekz
Junior Member
Join Date: Oct 2014
Posts: 1
Like (Stats)
Mentioned: 0 Post(s)
Tagged: 0 Thread(s)
Quoted: 0 Post(s)

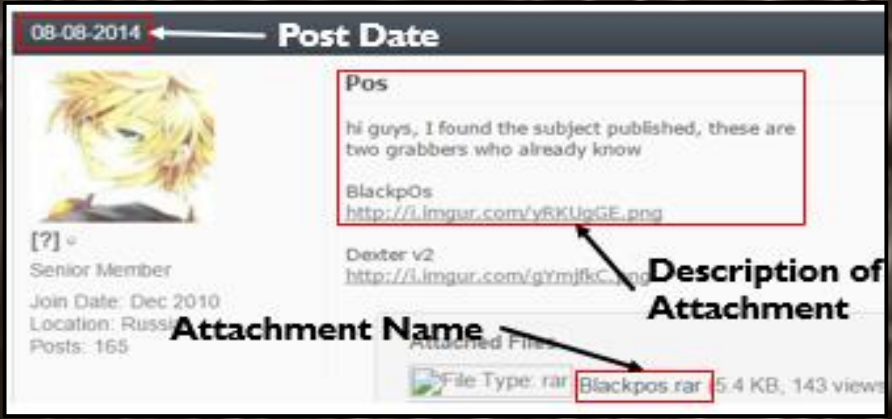
Author related information

Instructions on how to create malicious documents

Mass mailing or targeted campaigns that use common files to host or exploit code have been and are a very popular vector of attack. In other words
document received via e-mail or opened trough a browser plug-in. In regards to malicious PDF files the security industry saw a significant increase o
half of 2008 which might be related to Adobe Systems release of the specifications, format structure and functionality of PDF files.

Most enterprise networks perimeters are protected and contain several security filters and mechanism that block threats. However a malicious PDF c
very successful passing trough Firewalls, Intrusion Prevention Systems, Anti-spam, Anti-virus and other security controls. By reaching the victim mail
leverage social engineering techniques to lure the user to click/open the document. Then, for example, If the user opens a PDF malicious file, it typic
exploits a vulnerability when Adobe Reader parses the crafted file. This might cause the application to corrupt memory on the stack or heap causing
as shellcode. This shellcode normally downloads and executes a malicious file from the Internet. The Internet Storm Center Handler Bojan Zdrnja wr
of these shellcodes. In some circumstances the vulnerability could be exploited without opening the file and just by having a malicious file on the har
Stevens.

Tutorial on how to create malicious documents



08-08-2014

Post Date

[?]
Senior Member
Join Date: Dec 2010
Location: Russi
Posts: 165

Pos

hi guys, I found the subject published, these are two grabbers who already know

BlackpOs
http://i.imgur.com/yRKUgGE.png

Dexter v2
http://i.imgur.com/gYmjfkC.png

Description of Attachment

Attachment Name

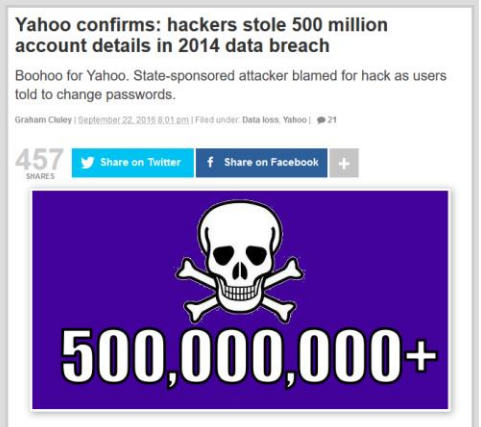
Attached Files

File Type: rar Blackpos.rar (5.4 KB, 143 views

Forum post with
BlackPOS malware attachment.

Selected data breaches in 2014

Victim	Date	Ramification
Target	2013.12	**40M** credit/debit cards; **70M** customer records; 46% drop in annual profits (**seller: Rescator**)
Neiman Marcus	2014.3	282K credit/debit cards
Sally Beauty	2014.3	25K credit/debit cards
P.F. Chang	2014.6	8 month of customer data from 33 stores
J.P. Morgan Chase	2014.8	**83M** accounts
UPS	2014.8	51 stores customers
Dairy Queen	2014.9	395 store systems
Home Depot	2014.9	**56M** credit/debit cards
Jimmy Jones	2014.9	216 store systems
Staples	2014.10	51 store systems



Yahoo confirms: hackers stole 500 million account details in 2014 data breach

Boohoo for Yahoo. State-sponsored attacker blamed for hack as users told to change passwords.

Graham Cluley | September 22, 2016 8:01 pm | Filed under: **Data loss**, **Yahoo** | 💬 21

Are your data breached? Do you even know?

Hacker Community Platforms – *“Know your enemy”*

Hacker Forums

TUXEDO CREW
CONTACT ADMIN :
YIM: TXCREW@YAHOO.COM
ICQ: 661108350

before you can post: click the register link above to proceed. To start viewing

Discussion board allowing hackers to freely share malicious tools and knowledge

DarkNet Markets

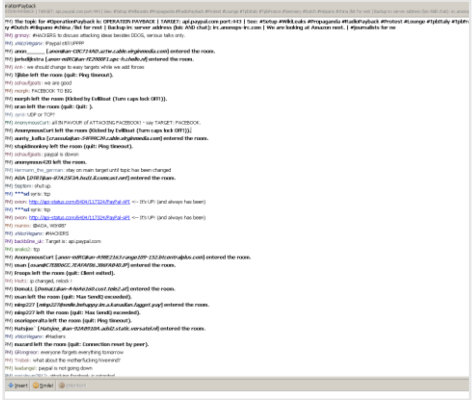
Dream Market
Established 201

Markets facilitating the sale of illicit goods (e.g., new exploits, drugs, weapons)

Carding Shops

Shops selling sensitive information (e.g., credit cards, SSN’s)

IRC Channels



```
rationPayback
PM) The topic for #OperationPayback is: OPERATION PAYBACK | TARGET: api.paypal.com port:443 | See: #Setup #WikiLeaks #Propaganda #RadioPayback #Protest #Lounge #TpbItaly #Tpbfr
y #Dutch #Hispano #china /list for rest | Backup irc server address (loic AND chat): irc.anonops-irc.com | We are looking at Amazon next. | #journalists for me
PM) grimzy: #HACKERS to discuss attacking ideas besides DDOS, serious talks only.
PM) xNicoVeganx: Paypal still UPPPP
PM) anon______ [anon@an-C0C714AD.aztw.cable.virginmedia.com] entered the room.
PM) jorisdijkstra [anon-mIRC@an-fE290DF1.upc-h.chello.nl] entered the room.
PM) Ant-: we should change to easy targets while we add forces
PM) Tjibbe left the room (quit: Ping timeout).
PM) cohaufgoats: we are good
PM) morph: FACEBOOK TO BIG
PM) morph left the room (Kicked by EvilBoat (Turn caps lock OFF!)).
PM) oran left the room (quit: Quit: ).
PM) syris: UDP or TCP?
PM) AnonymousCurt: all IN FAVOUR of ATTACKING FACEBOOK! - say TARGET: FACEBOOK.
PM) AnonymousCurt left the room (Kicked by EvilBoat (Turn caps lock OFF!)).
PM) aunty_kafka [crassula@an-54F8C20.cable.virginmedia.com] entered the room.
PM) stupidmonkey left the room (quit: Ping timeout).
PM) cohaufgoats: paypal is down
PM) anonymous420 left the room.
PM) Hermann_the_german: stay on main target until topic has been changed
PM) ADA [D1R7@an-07A25F3A.hsd1.il.comcast.net] entered the room.
PM) Septem: shut up.
PM) ***sd syris: tcp
PM) ovion: http://api-status.com/6404/117324/PayPal-API <-- It's UP! (and always has been)
PM) ***sd syris: tcp
PM) ovion: http://api-status.com/6404/117324/PayPal-API <-- It's UP! (and always has been)
PM) rsanier: @ADA, WINS?
PM) xNicoVeganx: #HACKERS
PM) backb0ne_uk: Target is: api.paypal.com
PM) anako2: tcp
PM) AnonymousCurt [anon-mIRC@an-A98E2163.range109-152.btcentralplus.com] entered the room.
PM) osan [osan@C7E8D6CC.7EAFAF06.386FAD40.IP] entered the room.
PM) Froops left the room (quit: Client exited).
PM) Matt1: ip changed, relock !
PM) DonatL [DonatL@an-A4e4e160.cust.tele2.at] entered the room.
PM) osan left the room (quit: Max SendQ exceeded).
PM) minp227 [minp227@smile.behappy.im.a.kanadian.faggot.yay] entered the room.
PM) minp227 left the room (quit: Max SendQ exceeded).
PM) osoriaperalta left the room (quit: Ping timeout).
PM) Hatsjoe` [Hatsjoe_@an-92A0918A.adsl2.static.versatel.nl] entered the room.
PM) xNicoVeganx: #Hackers
PM) mazard left the room (quit: Connection reset by peer).
PM) GRingreor: everyone forgets everything tomorrow
PM) Trobek: what about the motherfucking hivemind?
PM) leadangel: paypal is not going down
```

Plain-text IM service commonly used by hacktivist groups (e.g., Anonymous)

US → cybercrime and general hacking
Russia → underground economy, financial fraud
China → cyberwarfare content

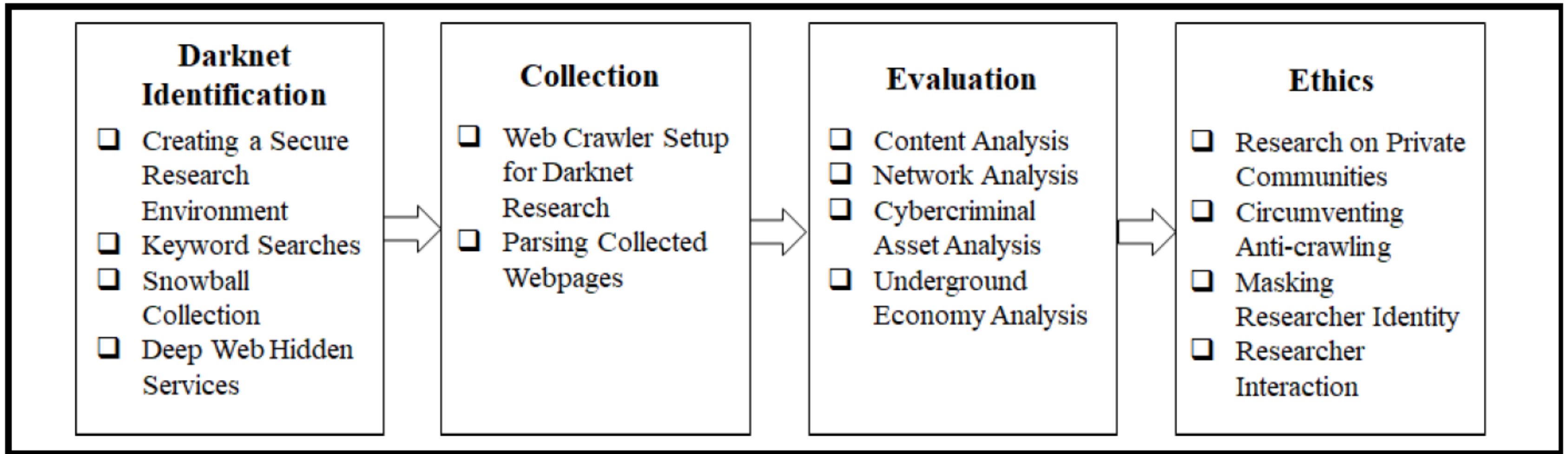
MIS Quarterly

MIS Quarterly Vol. 43 No. 1, pp. 1-22/March 2019

METHODS ARTICLE

DICE-E: A FRAMEWORK FOR CONDUCTING DARKNET IDENTIFICATION, COLLECTION, EVALUATION WITH ETHICS[1]

Victor Benjamin



Darknet Identification

- ❑ Creating a Secure Research Environment
- ❑ Keyword Searches
- ❑ Snowball Collection
- ❑ Deep Web Hidden Services

Collection

- ❑ Web Crawler Setup for Darknet Research
- ❑ Parsing Collected Webpages

Evaluation

- ❑ Content Analysis
- ❑ Network Analysis
- ❑ Cybercriminal Asset Analysis
- ❑ Underground Economy Analysis

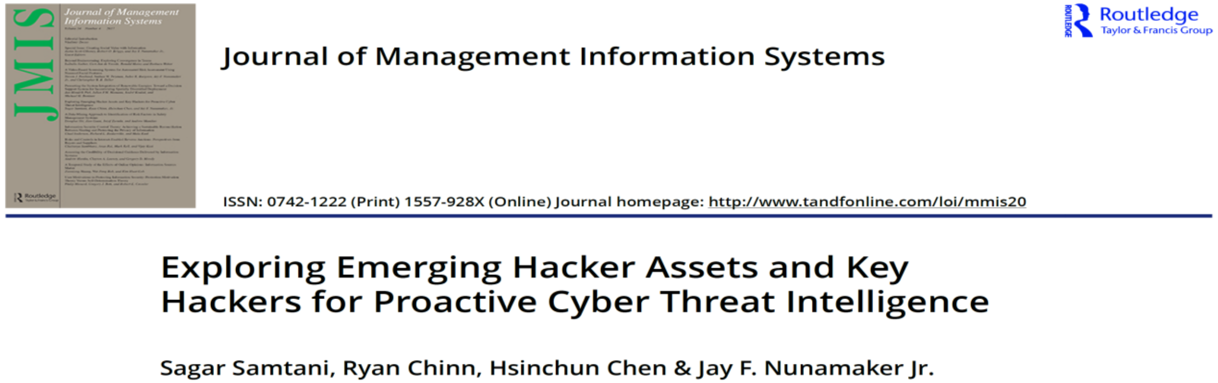
Ethics

- ❑ Research on Private Communities
- ❑ Circumventing Anti-crawling
- ❑ Masking Researcher Identity
- ❑ Researcher Interaction

Figure 2. The DICE-E Framework

Identify Hacker Assets/Tools

Sagar Samtani (JMIS, January 2018)



Routledge
Taylor & Francis Group

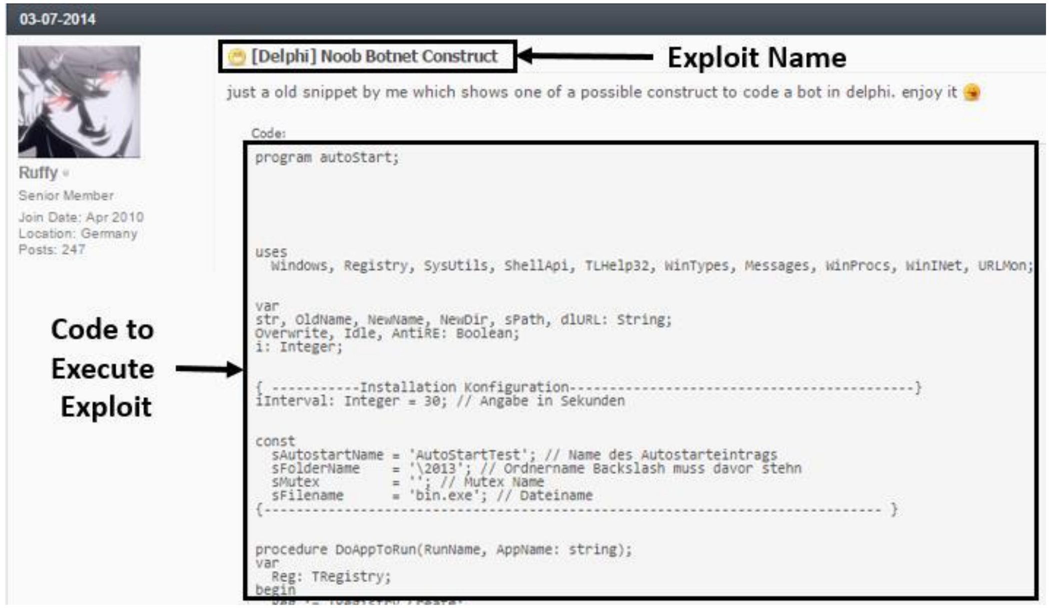
Journal of Management Information Systems

ISSN: 0742-1222 (Print) 1557-928X (Online) Journal homepage: http://www.tandfonline.com/loi/mmis20

Exploring Emerging Hacker Assets and Key Hackers for Proactive Cyber Threat Intelligence

Sagar Samtani, Ryan Chinn, Hsinchun Chen & Jay F. Nunamaker Jr.

Hacker Asset/Tool Examples



03-07-2014

Ruffy

Senior Member

Join Date: Apr 2010
Location: Germany
Posts: 247

[Delphi] Noob Botnet Construct

Exploit Name

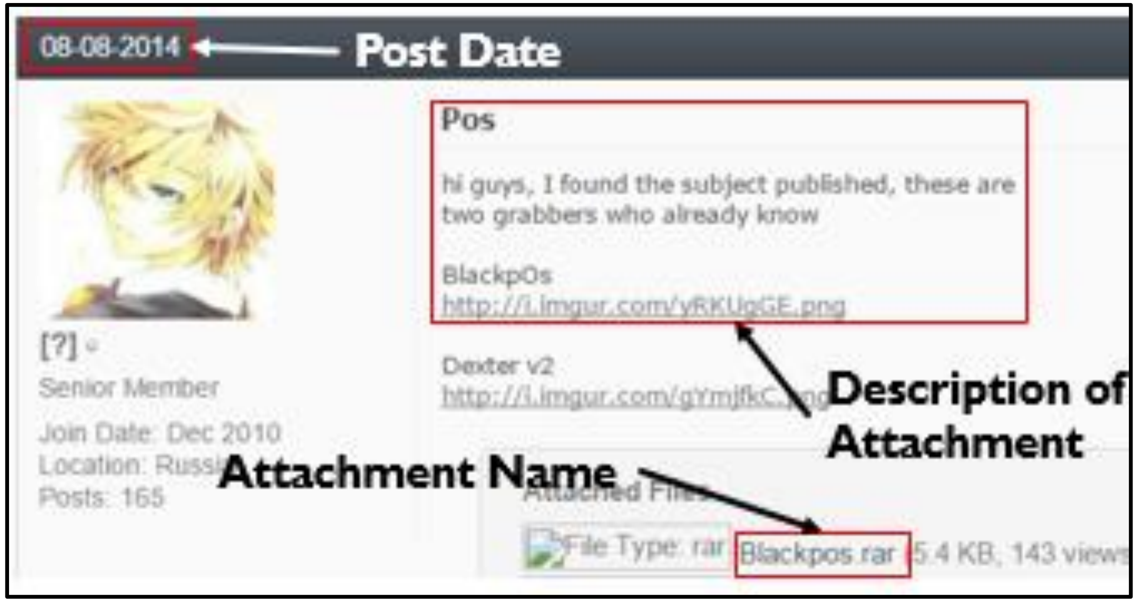
just a old snippet by me which shows one of a possible construct to code a bot in delphi. enjoy it

Code to Execute Exploit

Code:

```
program autoStart;



uses
  Windows, Registry, SysUtils, ShellApi, TLHelp32, WinTypes, Messages, WinProcs, WinINet, URLMon;

var
str, OldName, NewName, NewDir, sPath, dlURL: String;
Overwrite, Idle, AntiRE: Boolean;
i: Integer;

{ ----------Installation Konfiguration---------------------------------------}
iInterval: Integer = 30; // Angabe in Sekunden

const
  sAutostartName = 'AutoStartTest'; // Name des Autostarteintrags
  sFolderName    = '\2013'; // Ordnername Backslash muss davor stehn
  sMutex         = ''; // Mutex Name
  sFilename      = 'bin.exe'; // Dateiname
{------------------------------------------------------------------------ }

procedure DoAppToRun(RunName, AppName: string);
var
  Reg: TRegistry;
begin
```

Figure 1. Forum post with source code to create botnets



08-08-2014

Post Date

[?]

Senior Member

Join Date: Dec 2010
Location: Russi
Posts: 165

Pos

hi guys, I found the subject published, these are two grabbers who already know

BlackpOs
http://i.imgur.com/yRKUgGE.png

Dexter v2
http://i.imgur.com/gYmjfkC ng

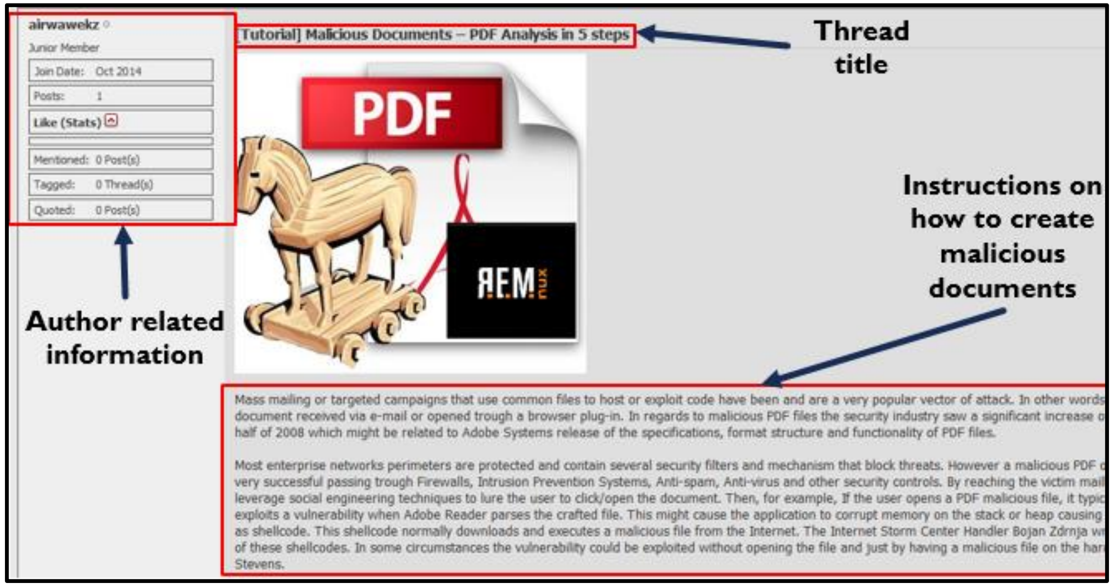
Description of Attachment

Attachment Name

Attached Files

File Type: rar Blackpos.rar 5.4 KB, 143 views

Figure 2. Forum post with BlackPOS malware attachment



airwawekz

Junior Member

Join Date:	Oct 2014
Posts:	1
Like (Stats)	
Mentioned:	0 Post(s)
Tagged:	0 Thread(s)
Quoted:	0 Post(s)

Author related information

[Tutorial] Malicious Documents – PDF Analysis in 5 steps

Thread title

Instructions on how to create malicious documents

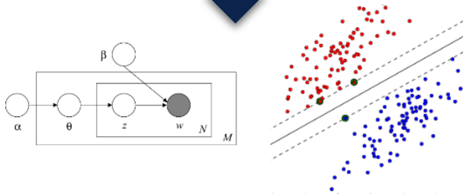
Mass mailing or targeted campaigns that use common files to host or exploit code have been and are a very popular vector of attack. In other words document received via e-mail or opened trough a browser plug-in. In regards to malicious PDF files the security industry saw a significant increase o half of 2008 which might be related to Adobe Systems release of the specifications, format structure and functionality of PDF files.

Most enterprise networks perimeters are protected and contain several security filters and mechanism that block threats. However a malicious PDF c very successful passing trough Firewalls, Intrusion Prevention Systems, Anti-spam, Anti-virus and other security controls. By reaching the victim mail leverage social engineering techniques to lure the user to click/open the document. Then, for example, If the user opens a PDF malicious file, it typic exploits a vulnerability when Adobe Reader parses the crafted file. This might cause the application to corrupt memory on the stack or heap causing as shellcode. This shellcode normally downloads and executes a malicious file from the Internet. The Internet Storm Center Handler Bojan Zdrnja w of these shellcodes. In some circumstances the vulnerability could be exploited without opening the file and just by having a malicious file on the har Stevens.

Figure 3. Tutorial on how to create malicious documents

AZSecure Hacker Assets Portal System

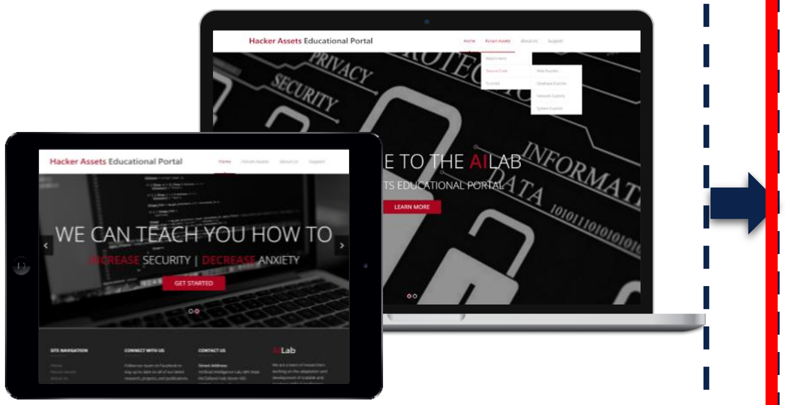
Data Collection and Analytics



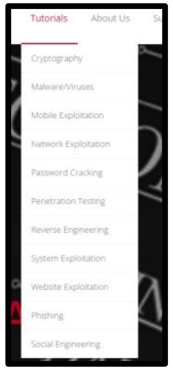
Latent Dirichlet Allocation (LDA) and Support Vector Machine (SVM) Analytics

987 tutorials, 15,576 source code, and 14,851 attachments

Web Hosting and Access



System Functionalities

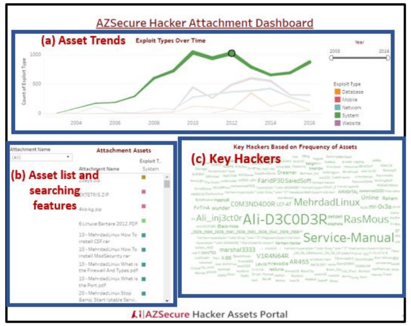


Browsing

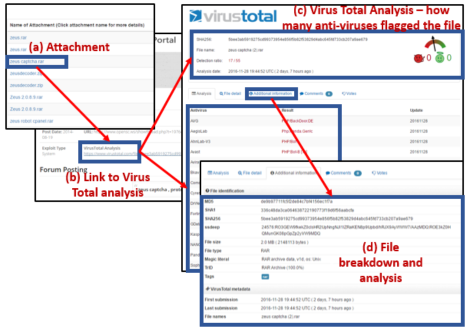
Searching

Downloading

System Analytics


AZSecure Hacker Attachment Dashboard
(a) Asset Trends
Exploit Types Over Time
(b) Asset list and searching features
(c) Key Hackers
Key Hackers Based on Frequency of Assets
AZSecure Hacker Assets Portal

Cyber Threat Intelligence Dashboard

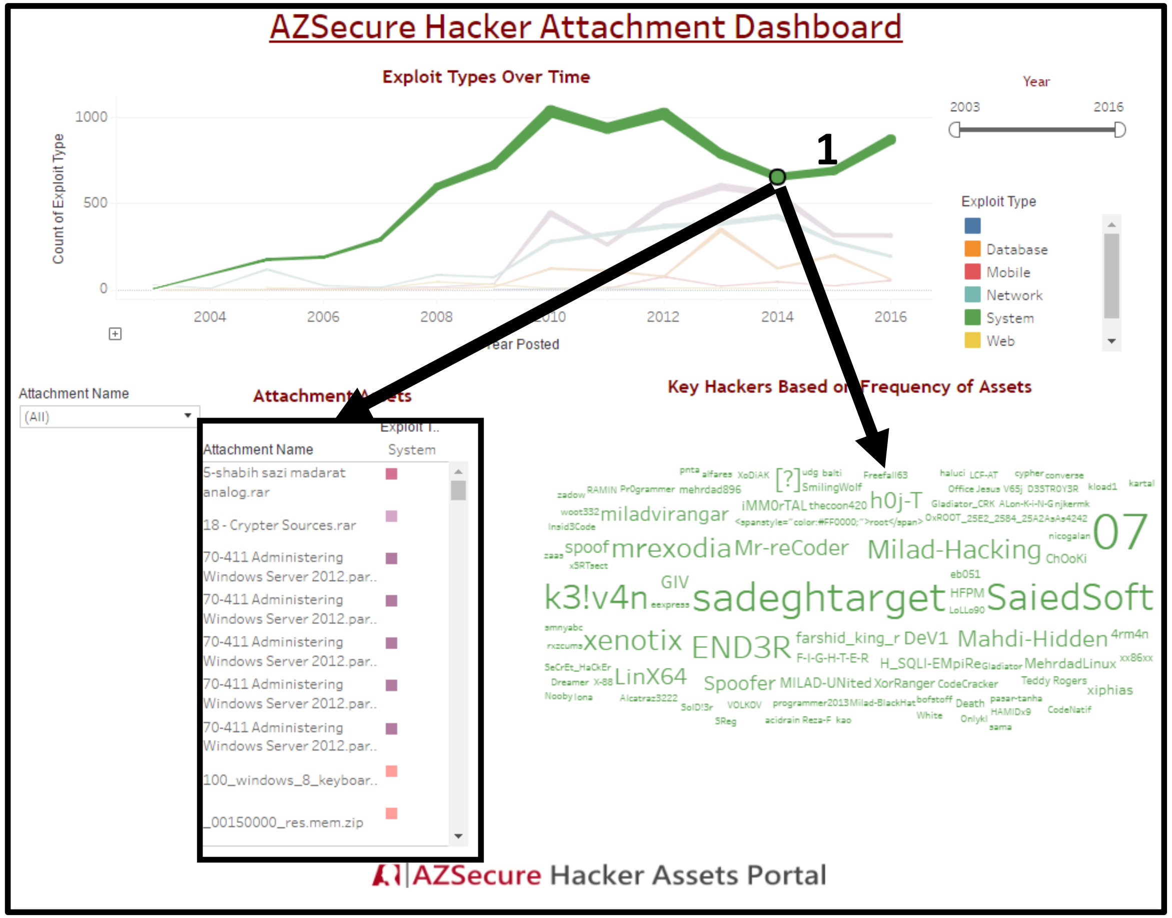

virustotal
(c) Virus Total Analysis – how many anti-viruses flagged the file
(a) Attachment
(b) Link to Virus Total analysis
(d) File breakdown and analysis

VirusTotal Malware Analysis

AZSecure Hacker Assets Portal (English, Russian, Arabic)

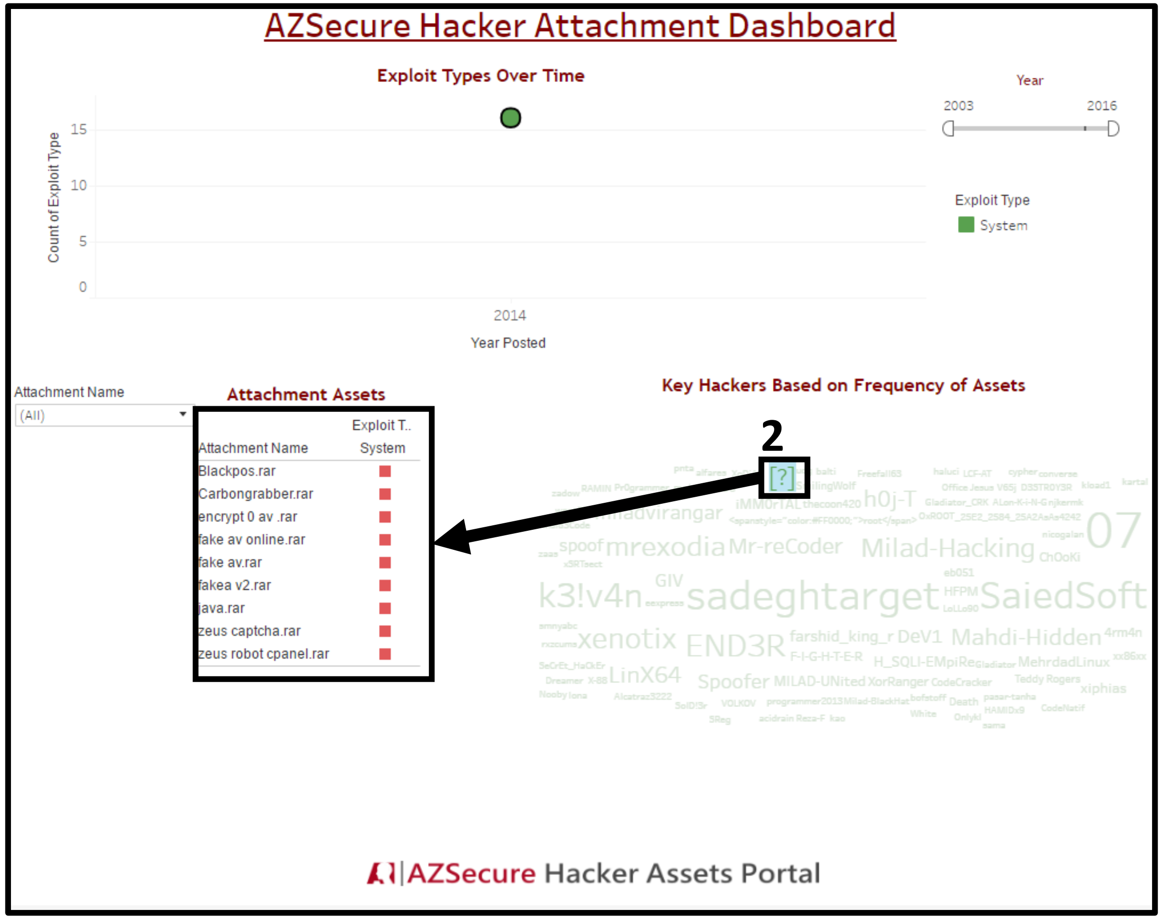
Forum	Language	Date Range	# of Posts	# of Members	# of source code	# of attachments	# of tutorials
OpenSC	English	02/07/2005-02/21/2016	124,993	6,796	2,590	2,349	628
Xeksec	Russian	07/07/2007- 9/15/2015	62,316	18,462	2,456	-	40
Ashiyane	Arabic	5/30/2003 – 9/24/2016	34,247	6,406	5,958	10,086	80
tuts4you	English	6/10/2006 – 10/31/2016	40,666	2,539	-	2,206	38
exelab	Russian	8/25/2008 – 10/27/2016	328,477	13,289	4,572	-	628
Total:	**-**	**02/07/2005- 10/31/2016**	**590,699**	**47,492**	**15,576**	**14,851**	**987**

Cyber Threat Intelligence (CTI) Example – Bank Exploits (e.g., BlackPOS)



Attachment Assets

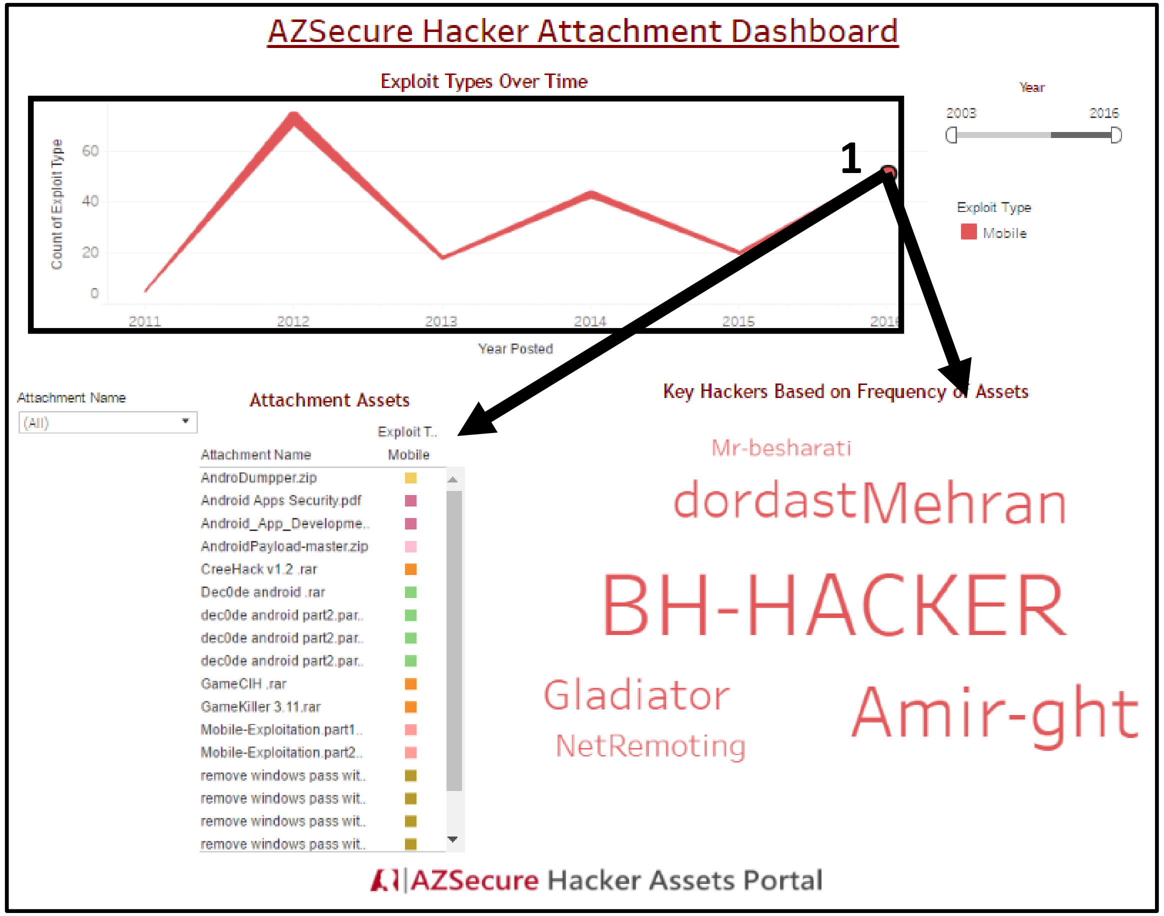
Attachment Name	Exploit T.. System
5-shabih sazi madarat analog.rar	🟥
18 - Crypter Sources.rar	🟥
70-411 Administering Windows Server 2012.par..	🟥
70-411 Administering Windows Server 2012.par..	🟥
70-411 Administering Windows Server 2012.par..	🟥
70-411 Administering Windows Server 2012.par..	🟥
70-411 Administering Windows Server 2012.par..	🟥
100_windows_8_keyboar..	🟥
_00150000_res.mem.zip	🟥



Attachment Assets

Attachment Name	Exploit T.. System
Blackpos.rar	🟥
Carbongrabber.rar	🟥
encrypt 0 av .rar	🟥
fake av online.rar	🟥
fake av.rar	🟥
fakea v2.rar	🟥
java.rar	🟥
zeus captcha.rar	🟥
zeus robot cpanel.rar	🟥

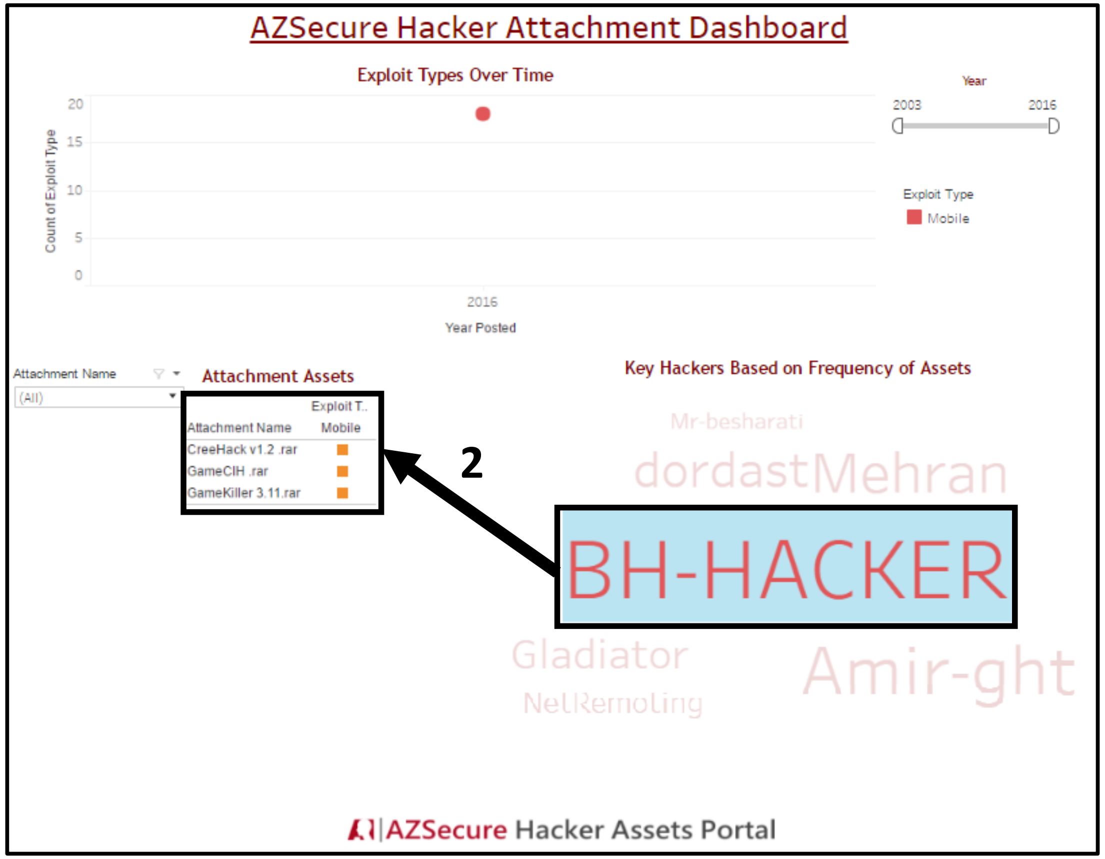
Cyber Threat Intelligence (CTI) Example – Mobile Malware



Attachment Name
(All)

Attachment Assets

Attachment Name	Exploit T.. Mobile
AndroDumpper.zip	🟨
Android Apps Security.pdf	🟥
Android_App_Developme..	🟥
AndroidPayload-master.zip	🟥
CreeHack v1.2 .rar	🟧
Dec0de android .rar	🟩
dec0de android part2.par..	🟩
dec0de android part2.par..	🟩
dec0de android part2.par..	🟩
GameCIH .rar	🟧
GameKiller 3.11.rar	🟧
Mobile-Exploitation.part1..	🟥
Mobile-Exploitation.part2..	🟥
remove windows pass wit..	🟫
remove windows pass wit..	🟫
remove windows pass wit..	🟫
remove windows pass wit..	🟫



Attachment Name
(All)

Attachment Assets

Attachment Name	Exploit T.. Mobile
CreeHack v1.2 .rar	🟧
GameCIH .rar	🟧
GameKiller 3.11.rar	🟧

Labeling Hacker Exploits for Proactive Cyber Threat Intelligence: A Deep Transfer Learning Approach

Benjamin Ampel (MISQ, 2nd Round))

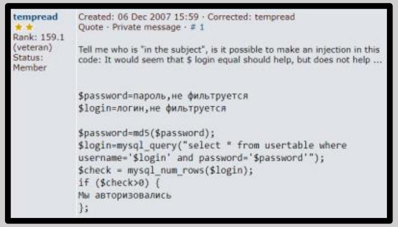
Literature Review: Hacker Forum Exploit Analysis

Year	Author	1. Data Source	2. Data Type Used	Analytics	Identified Exploits	3. Purpose
2019	Schafer et al.	General purpose forums	Forum titles, users, message, topic, keywords	SNA, LDA	Leaks, botnets, DDoS	Trend identification
2019	Benjamin et al.	General purpose forums	Post content, attachments, source code, keywords, reputation	OLS Regression	Rootkit, XSS, SQLi, DDoS, shellcode, drive-by	Darknet identification, collection, evaluation
2018	Williams et al.	General purpose forums	Sub-forum name, author, post content, attachment metadata	LSTM	Crypters, keyloggers, RATs, DDoS, SQLi	Exploit categorization
2018	Goyal et al.	Forums, Twitter, Blogs	Post content, Tweet content, blog content	LSTM, RNN	Trojan, Windows, Apple OSX, phishing	Cyber attack prediction
2018	Deliu et al.	Nulled.IO leak	Post content	SVM, CNN	Botnet, crypter, keylogger, malware, rootkit	Exploit categorization
2017	Samtani et al.	General purpose forums	Post content, assets, thread, author, source code	LDA, SVM	Crypters, keyloggers, RATs, botnets	Exploit categorization
2017	Grisham et al.	General purpose forums	Post content, date, author, role, attachments	RNN	Mobile malware	Malware identification/ Proactive CTI
2017	Deliu et al.	Nulled.IO leak	Post content	SVM, LDA	Backdoor, botnet, crypter, DDoS, exploit, malware, password, rootkit	Exploit categorization

- **Key Observations:**
  1. Studies focus on general forums, but not exploit DNMs or public repositories.
  2. Although source code contains valuable information, many studies omit them from analysis.
  3. The most common task is to categorize post content by exploit category.

Proposed Research Design

Data Collection



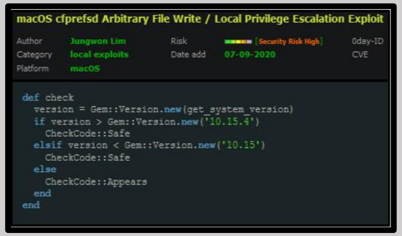
tempread
★★
Rank: 159.1
(veteran)
Status:
Member

Created: 06 Dec 2007 15:59 · Corrected: tempread
Quote · Private message · # 1

Tell me who is "in the subject", is it possible to make an injection in this code: It would seem that $ login equal should help, but does not help ...

```
$password=пароль,не фильтруется
$login=логин,не фильтруется

$password=md5($password);
$login=mysql_query("select * from usertable where
username='$login' and password='$password'");
$check = mysql_num_rows($login);
if ($check>0) {
Мы авторизовались
};
```

Traditional Hacker Forums

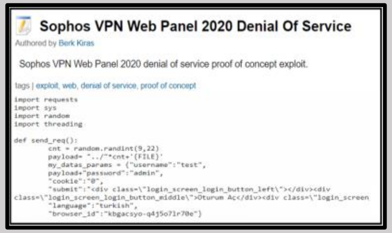


macOS cfprefsd Arbitrary File Write / Local Privilege Escalation Exploit

Author	Jungwon Lim	Risk	[Security Risk High]	0day-ID
Category	local exploits	Date add	07-09-2020	CVE
Platform	macOS			

```
def check
  version = Gem::Version.new(get_system_version)
  if version > Gem::Version.new('10.15.4')
    CheckCode::Safe
  elsif version < Gem::Version.new('10.15')
    CheckCode::Safe
  else
    CheckCode::Appears
  end
end
```

Exploit DNMs



Sophos VPN Web Panel 2020 Denial Of Service

Authored by Berk Kiras

Sophos VPN Web Panel 2020 denial of service proof of concept exploit.

tags | exploit, web, denial of service, proof of concept

```
import requests
import sys
import random
import threading

def send_req():
        cnt = random.randint(9,22)
        payload= "../"*cnt+'{FILE}'
        my_datas_params = {"username":"test",
        payload+"password":"admin",
        "cookie":"0",
        "submit":"<div class=\"login_screen_login_button_left\"></div><div
class=\"login_screen_login_button_middle\">Oturum Aç</div><div class=\"login_screen
        "language":"turkish",
        "browser_id":"kbgacsyo-q4j5o7lr70e"}
```

Public Exploit Repositories

Pre-Processing and Dataset Construction

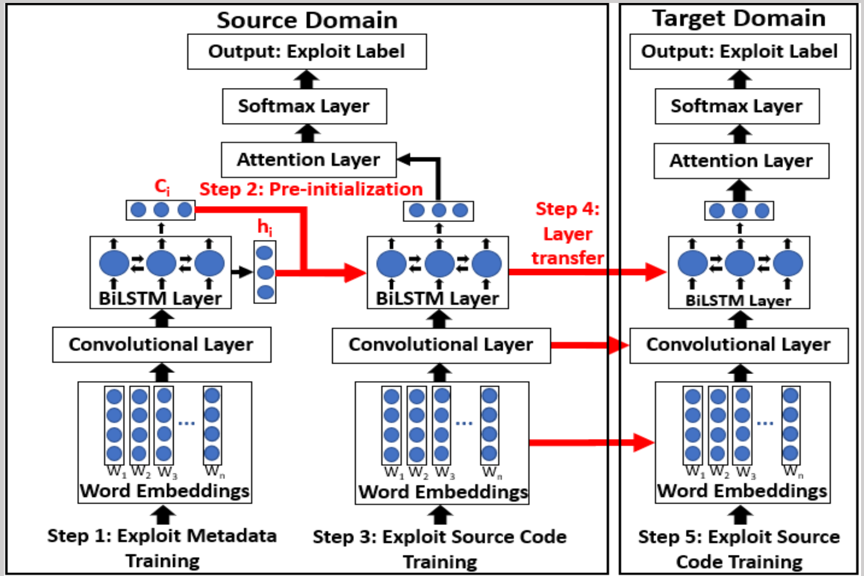
Programming Language Classifier

Remove stop words, low frequency words, and lemmatize

Tokenizer and Sequence Padder

Gold-Standard Dataset Construction

Deep Transfer Learning Exploit Labeler (DTL-EL) Model



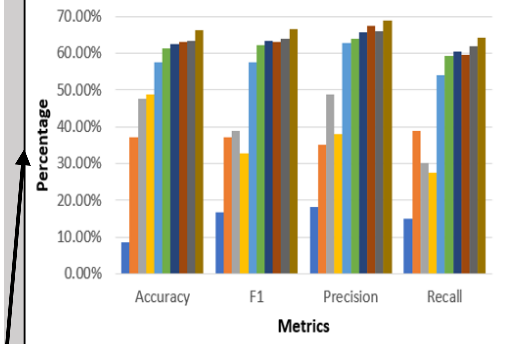
Experiment 1: DTL-EL against prevailing classification methods on source domain

Experiment 2: DTL-EL against non-transfer learning approaches on target domain

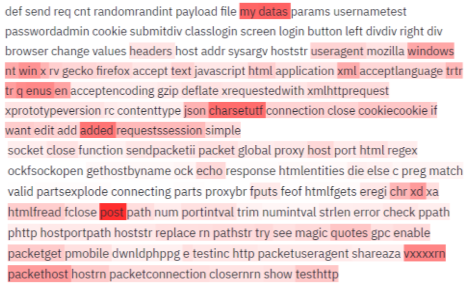
Experiment 3: DTL-EL against transfer learning layer selection on target domain

Evaluations and Visualization

Benchmark Comparisons



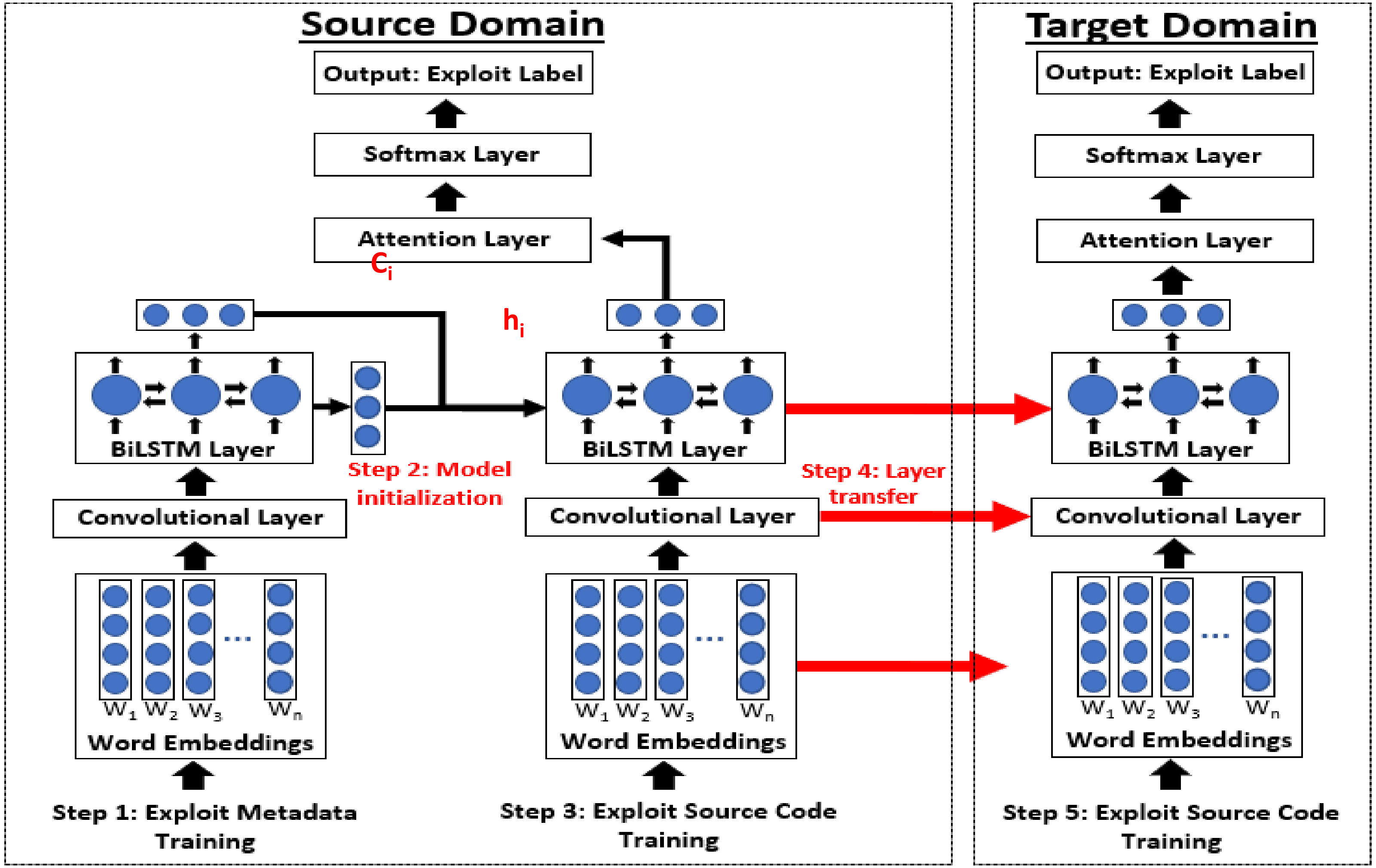
Visualized Attention Comparison



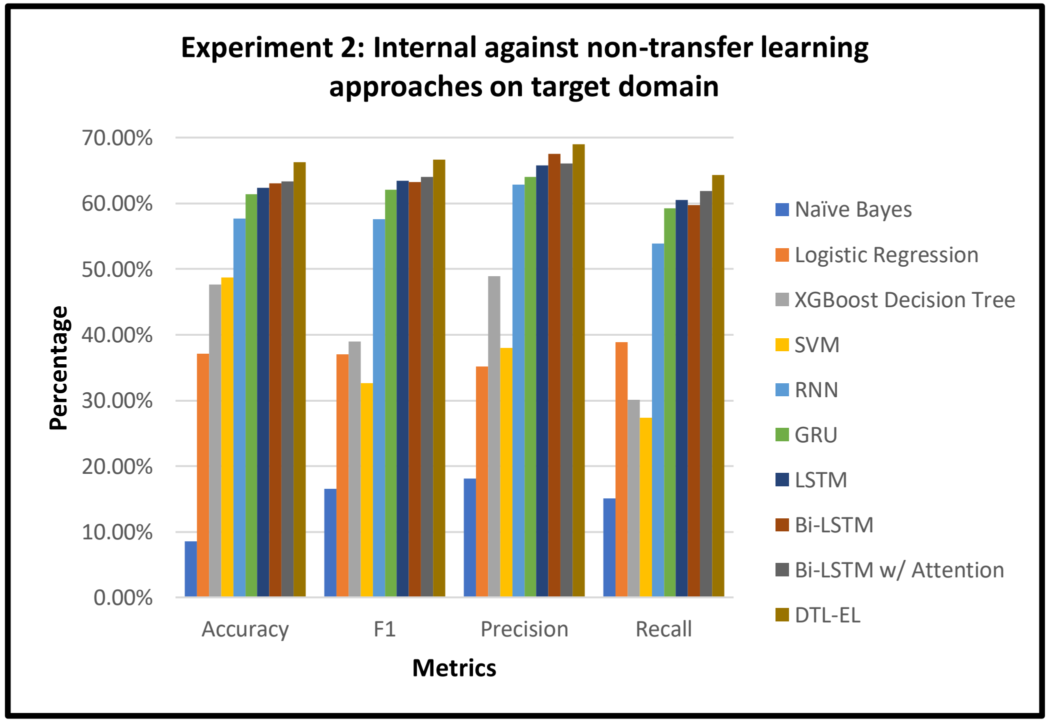
def send req cnt randomrandint payload file my datas params usernametest passwordadmin cookie submitdiv classlogin screen login button left divdiv right div browser change values headers host addr sysargv hoststr useragent mozilla windows nt win x rv gecko firefox accept text javascript html application xml acceptlanguage trtr tr q enus en acceptencoding gzip deflate xrequestedwith xmlhttprequest xprototypeversion rc contenttype json charsetutf connection close cookiecookie if want edit add added requestssession simple

socket close function sendpacketii packet global proxy host port html regex ockfsockopen gethostbyname ock echo response htmlentities die else c preg match valid partsexplode connecting parts proxybr fputs feof htmlfgets eregi chr xd xa htmlfread fclose post path num portintval trim numintval strlen error check ppath phttp hostportpath hoststr replace rn pathstr try see magic quotes gpc enable packetget pmobile dwnldphppg e testinc http packetuseragent shareaza vxxxxrn packethost hostrn packetconnection closernrn show testhttp

Research Design: DTL-EL



Results and Discussion: DTL-EL Model

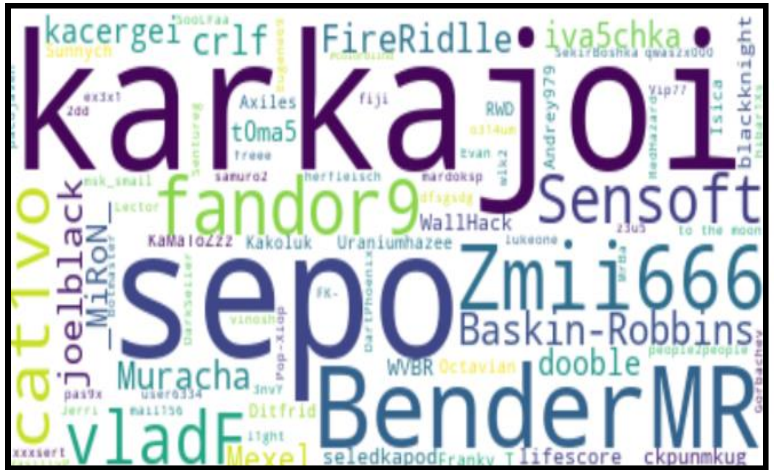


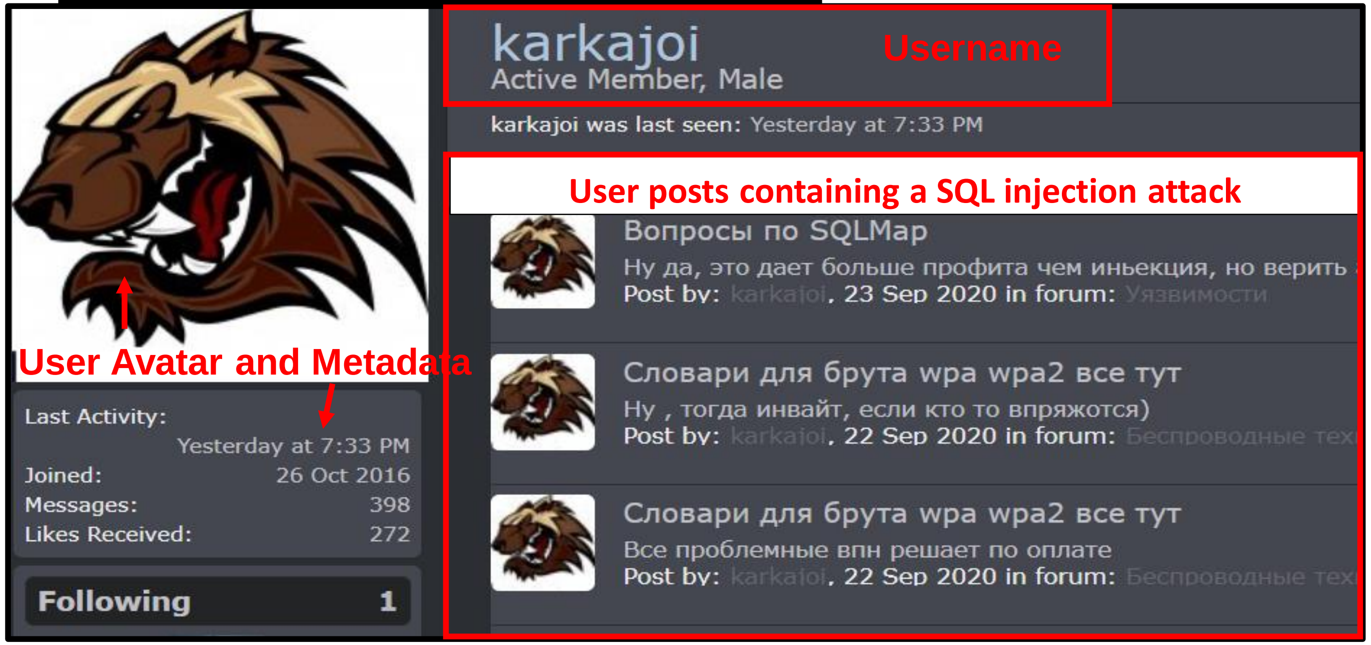
Experiment 2: Internal against non-transfer learning models		**Results**			
Model	**Layer Weights**	**Accuracy**	**Precision**	**Recall**	**F1**
Naïve Bayes	Random	8.59% ***	18.09% ***	15.08% ***	16.45% ***
Logistic Regression	Random	37.16% ***	35.13% ***	38.85% ***	36.9% ***
XGBoost Decision Tree	Random	47.65% ***	48.87% ***	30.06% ***	37.22% ***
SVM	Random	48.72% ***	37.98% ***	27.38% ***	31.82% ***
RNN	Random	57.64% ***	62.89% ***	53.93% ***	57.62% ***
GRU	Random	61.34% ***	64.06% ***	59.27% ***	62.09% ***
LSTM	Random	62.39% ***	65.77% ***	60.49% ***	63.42% ***
BiLSTM	Random	63.05% ***	67.56% ***	59.71% ***	63.21% ***
BiLSTM w/ Attention	Random	63.38% ***	66.04% ***	61.88% ***	64.02% ***
DTL-EL (Our model)	Transferred	**66.17%**	**68.25%**	**64.99%**	**66.61%**

Case Study: Identifying Key Hackers - SQLi

- Since 2017, SQL injections are the most prevalent exploit in Russian forums.

- The five hackers with the most SQL injections posted on Russian forums are:
  1. **karkajoi (13 exploits)**
  2. sepo (12 exploits)
  3. BenderMR (12 exploits)
  4. Zmii666 (6 exploits)
  5. fandor9 (6 exploits)





karkajoi
Active Member, Male

karkajoi was last seen: Yesterday at 7:33 PM

Last Activity: Yesterday at 7:33 PM
Joined: 26 Oct 2016
Messages: 398
Likes Received: 272

Following 1

Вопросы по SQLMap
Ну да, это дает больше профита чем иньекция, но верить
Post by: karkajoi, 23 Sep 2020 in forum: Уязвимости

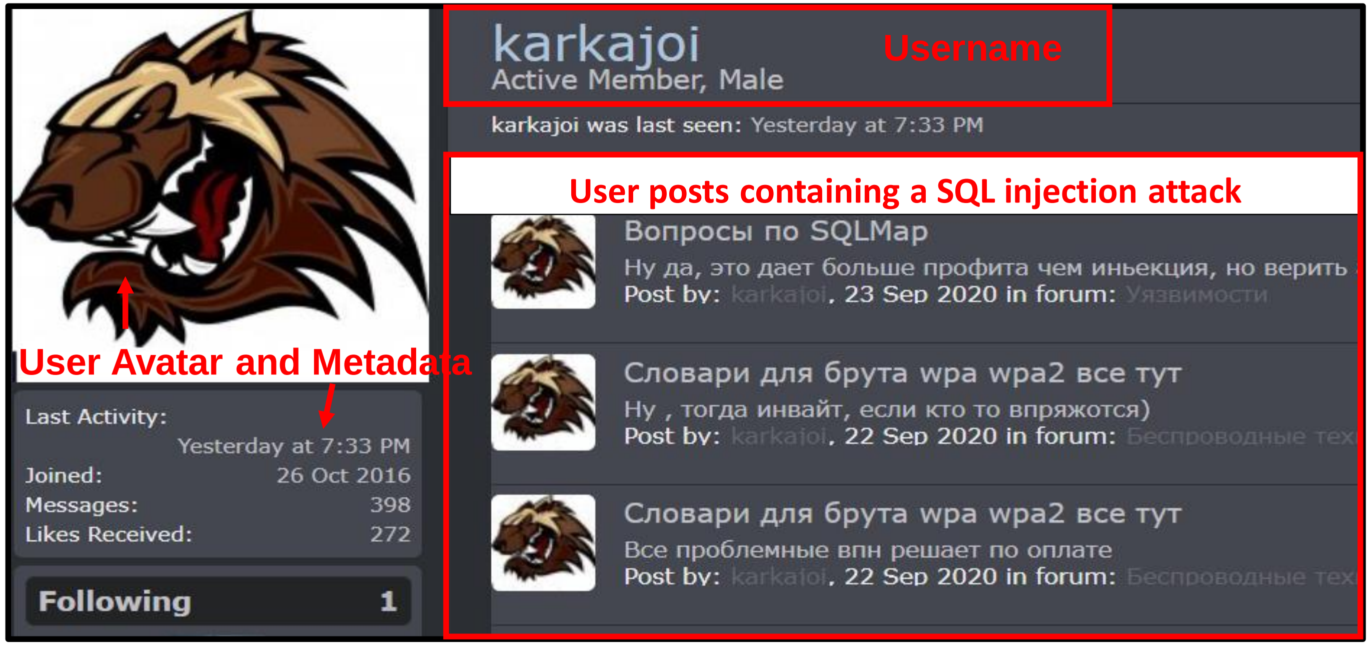
Словари для брута wpa wpa2 все тут
Ну , тогда инвайт, если кто то впряжотся)
Post by: karkajoi, 22 Sep 2020 in forum: Беспроводные тех

Словари для брута wpa wpa2 все тут
Все проблемные впн решает по оплате
Post by: karkajoi, 22 Sep 2020 in forum: Беспроводные тех

Figure 15. Hacker Profile Page on Antichat

Case Study: Hacker Profile - Karkajoi

- “Karkajoi” is a unique username and can be found on separate Russian hacker forums (e.g. root-me, raidforums), suggesting he is an active contributor in the larger hacker community.
- Along with SQL injections, he also cracks various hashes and posts web application exploits.



karkajoi
Active Member, Male

karkajoi was last seen: Yesterday at 7:33 PM

Last Activity: Yesterday at 7:33 PM
Joined: 26 Oct 2016
Messages: 398
Likes Received: 272

Following 1

Вопросы по SQLMap
Ну да, это дает больше профита чем иньекция, но верить
Post by: karkajoi, 23 Sep 2020 in forum: Уязвимости

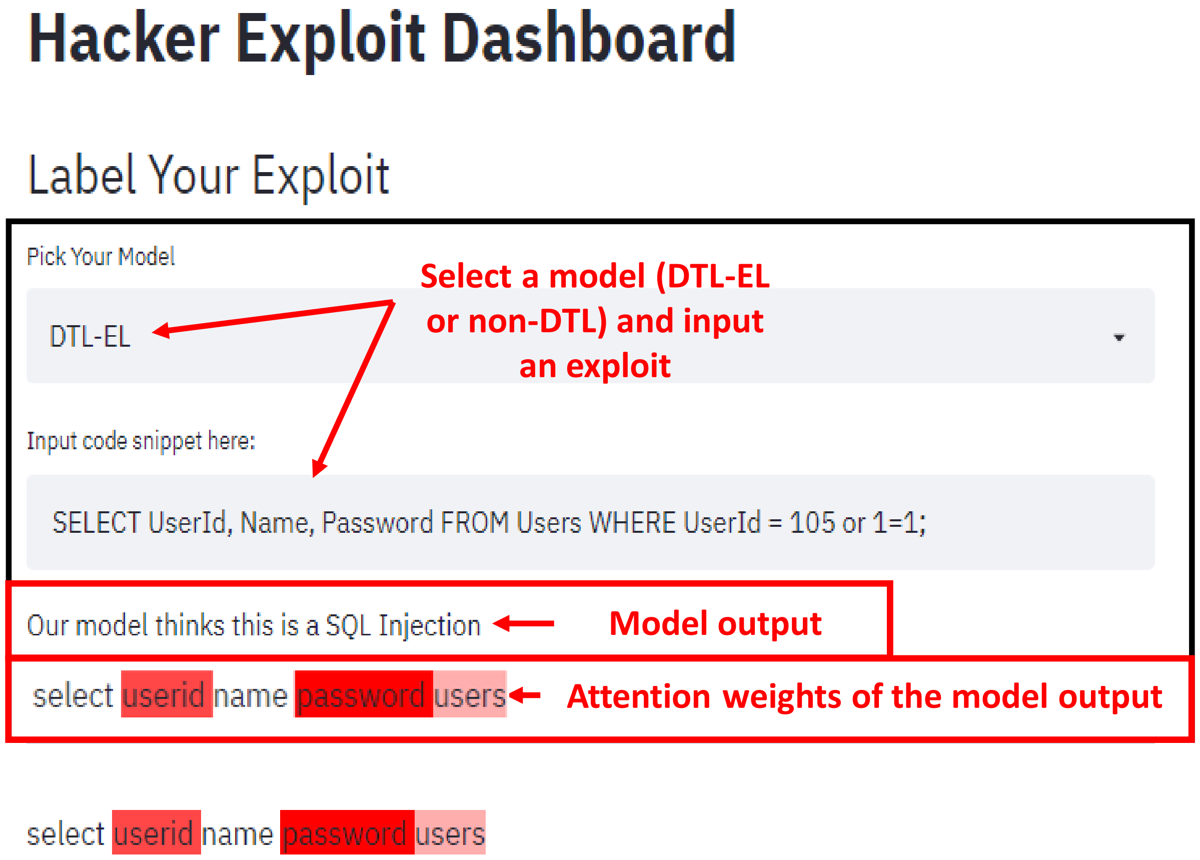
Словари для брута wpa wpa2 все тут
Ну , тогда инвайт, если кто то впряжотся)
Post by: karkajoi, 22 Sep 2020 in forum: Беспроводные тех

Словари для брута wpa wpa2 все тут
Все проблемные впн решает по оплате
Post by: karkajoi, 22 Sep 2020 in forum: Беспроводные тех

Figure 15. Hacker Profile Page on Antichat

Case Study: System Integration

- Hacker exploit source code can be input for classification with attention weights.
- The system applies a DTL-EL label upon the collection of new hacker forum text, providing real-time information to researchers.
  - APIs allow for forums to be downloaded in their entirety with related programming languages and exploit labels for source code.



Hacker Exploit Dashboard

Label Your Exploit

Pick Your Model

DTL-EL

Input code snippet here:

```
SELECT UserId, Name, Password FROM Users WHERE UserId = 105 or 1=1;
```

Our model thinks this is a SQL Injection

select userid name password users

select userid name password users

Select a model (DTL-EL or non-DTL) and input an exploit

Model output

Attention weights of the model output

Figure 16. Hacker Exploit Portal For Further Analysis

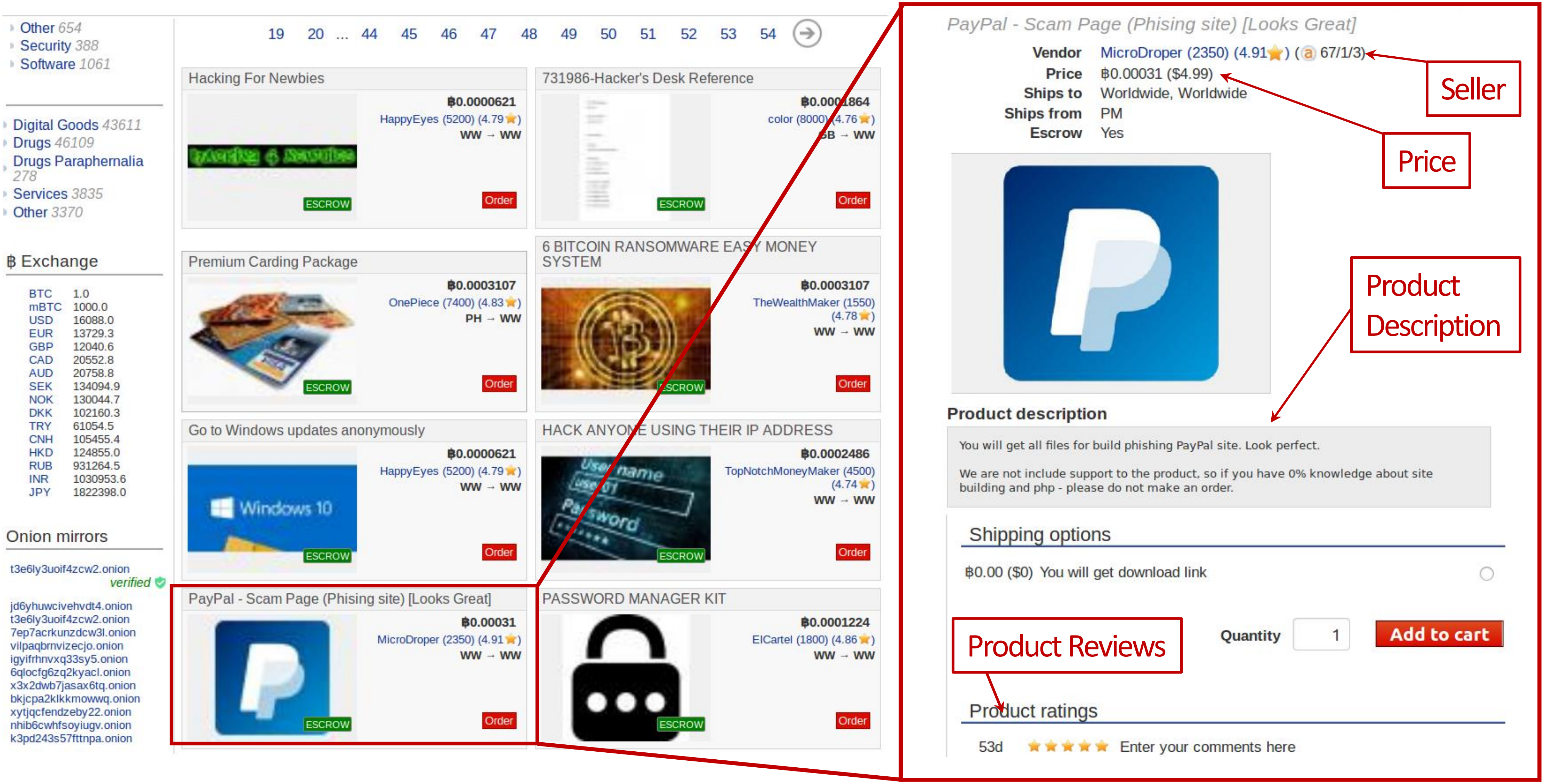
Detecting Cyber Threats with AI Agents: Multilingual, Multimedia DNM Content

Reza Ebrahimi (JMIS, MIS, IEEE PAMI)

Detecting Cyber Threats with AI Agents

- **Intelligence Source:** Dark web
  - A large conglomerate of platforms that facilitate illegal transactions among hackers
- **DarkNet Market Places** (Amazon for illegal products; hidden from search engines) → Attract cybercriminals
  - **Hacker Assets:** Hacking tools (Remote Access Trojan); malicious executables; hacking tutorials
  - **Non-Hacker Assets:** Digital goods (credit card information); copyrighted software; pirated e-books; counterfeits; drugs; forged documents

Dark Net Marketplaces (DNMs)



- Other 654
- Security 388
- Software 1061

- Digital Goods 43611
- Drugs 46109
- Drugs Paraphernalia 278
- Services 3835
- Other 3370

฿ Exchange

BTC	1.0
mBTC	1000.0
USD	16088.0
EUR	13729.3
GBP	12040.6
CAD	20552.8
AUD	20758.8
SEK	134094.9
NOK	130044.7
DKK	102160.3
TRY	61054.5
CNH	105455.4
HKD	124855.0
RUB	931264.5
INR	1030953.6
JPY	1822398.0

Onion mirrors

t3e6ly3uoif4zcw2.onion
verified

jd6yhuwcivehvdt4.onion
t3e6ly3uoif4zcw2.onion
7ep7acrkunzdcw3l.onion
vilpaqbrnvizecjo.onion
igyifrhnvxq33sy5.onion
6qlocfg6zq2kyacl.onion
x3x2dwb7jasax6tq.onion
bkjcpa2klkkmowwq.onion
xytjqcfendzeby22.onion
nhib6cwhfsoyiugv.onion
k3pd243s57fttnpa.onion

19 20 ... 44 45 46 47 48 49 50 51 52 53 54

Hacking For Newbies
฿0.0000621
HappyEyes (5200) (4.79)
WW → WW
ESCROW Order

731986-Hacker's Desk Reference
฿0.0001864
color (8000) (4.76)
GB → WW
ESCROW Order

Premium Carding Package
฿0.0003107
OnePiece (7400) (4.83)
PH → WW
ESCROW Order

6 BITCOIN RANSOMWARE EASY MONEY SYSTEM
฿0.0003107
TheWealthMaker (1550) (4.78)
WW → WW
ESCROW Order

Go to Windows updates anonymously
฿0.0000621
HappyEyes (5200) (4.79)
WW → WW
ESCROW Order

HACK ANYONE USING THEIR IP ADDRESS
฿0.0002486
TopNotchMoneyMaker (4500) (4.74)
WW → WW
ESCROW Order

PayPal - Scam Page (Phising site) [Looks Great]
฿0.00031
MicroDroper (2350) (4.91)
WW → WW
ESCROW Order

PASSWORD MANAGER KIT
฿0.0001224
ElCartel (1800) (4.86)
WW → WW
ESCROW Order

PayPal - Scam Page (Phising site) [Looks Great]

Vendor	MicroDroper (2350) (4.91) (a 67/1/3)
Price	฿0.00031 ($4.99)
Ships to	Worldwide, Worldwide
Ships from	PM
Escrow	Yes

Product description

You will get all files for build phishing PayPal site. Look perfect.

We are not include support to the product, so if you have 0% knowledge about site building and php - please do not make an order.

Shipping options

฿0.00 ($0) You will get download link ○

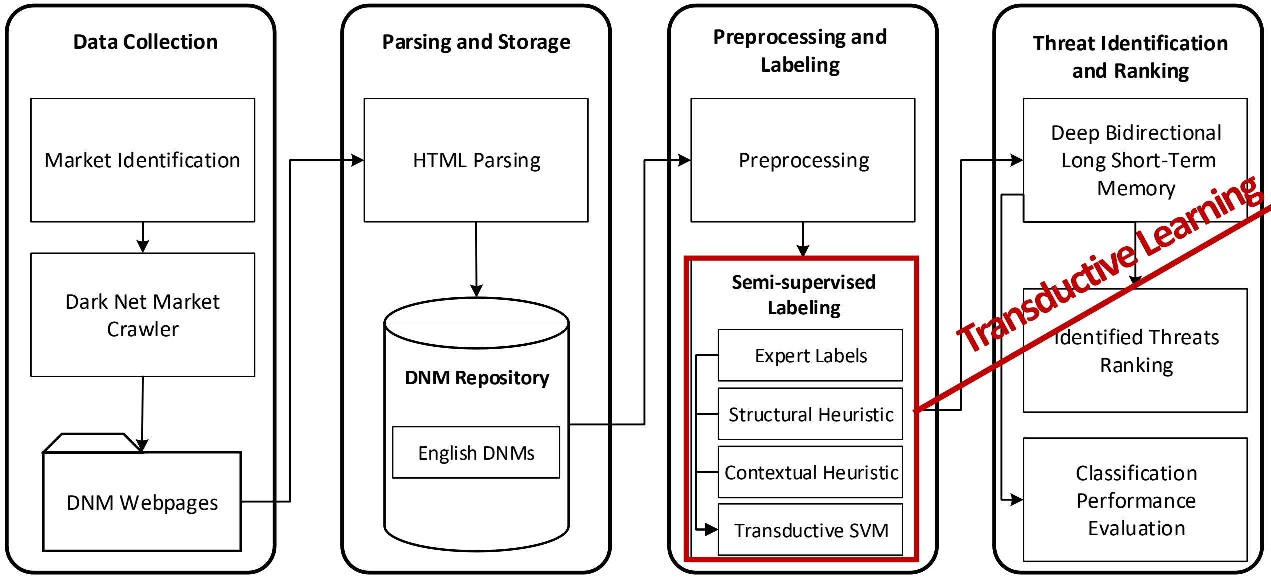
Quantity 1 Add to cart

Product ratings

53d ★★★★★ Enter your comments here

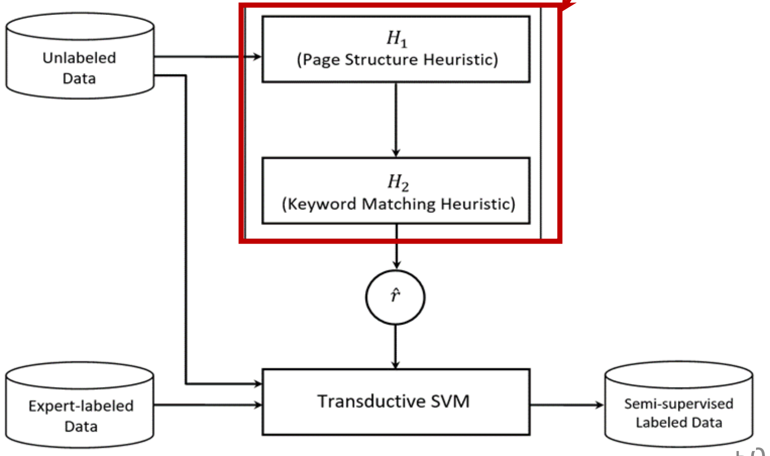
Essay I: Learning From Unlabeled Cybersecurity Content (JMIS, March 2020)

- Learning from examples → supervised by human-labeled data → Expensive!
- Unlabeled data improves cyber threat detection with **transductive learning theory**



$$\min_{w}\left(\frac{\lambda}{2}\|w\|^2+\frac{1}{2L}\sum_{i=1}^{L} l(y_i\,,w^T x_i)+\frac{\lambda'}{2U}\sum_{j=1}^{U} l(y_i'\,,w^T x_j')\right)$$

$$\text{subject to:}\quad \frac{1}{U}\sum_{j=1}^{u}\max\left(0,\operatorname{sign}\left(w^T x_j'\right)\right)=\boxed{r}$$

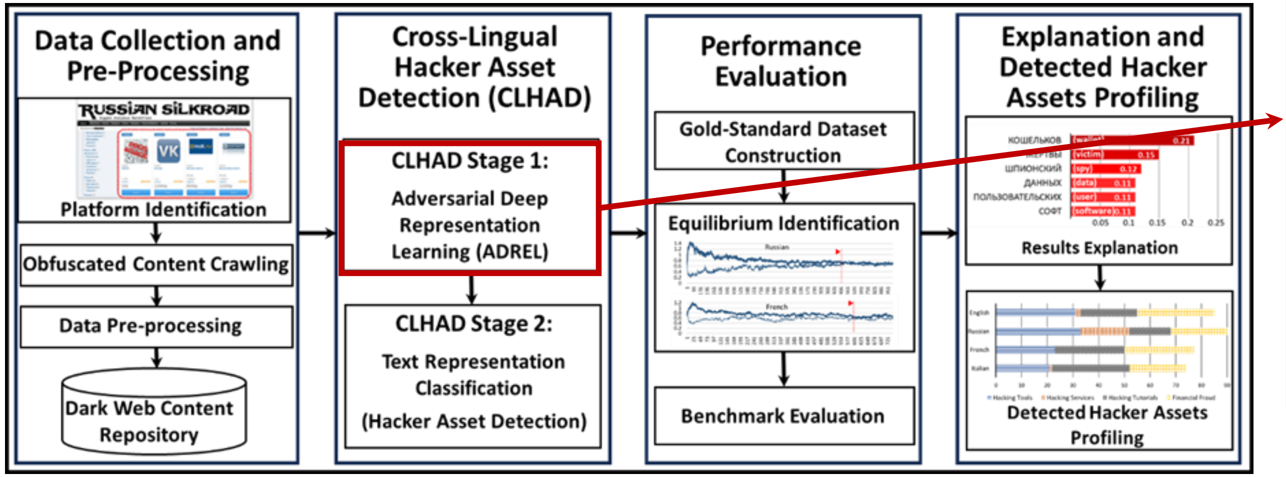


Unlabeled Data
H_1 (Page Structure Heuristic)
H_2 (Keyword Matching Heuristic)
$\hat{r}$
Expert-labeled Data
Transductive SVM
Semi-supervised Labeled Data

- Significantly decreased reliance on human supervision for cyber threat detection.

Essay II: Learning from Heterogeneous Cybersecurity Content (MISQ, Forthcoming)

- Cyber threat detection in non-English content → lack of non-English training data
- Transfer cyber threat knowledge from high-resource English platforms to non-English ones with **transfer learning theory**



Data Collection and Pre-Processing

Platform Identification

Obfuscated Content Crawling

Data Pre-processing

Dark Web Content Repository

Cross-Lingual Hacker Asset Detection (CLHAD)

CLHAD Stage 1: Adversarial Deep Representation Learning (ADREL)

CLHAD Stage 2: Text Representation Classification (Hacker Asset Detection)

Performance Evaluation

Gold-Standard Dataset Construction

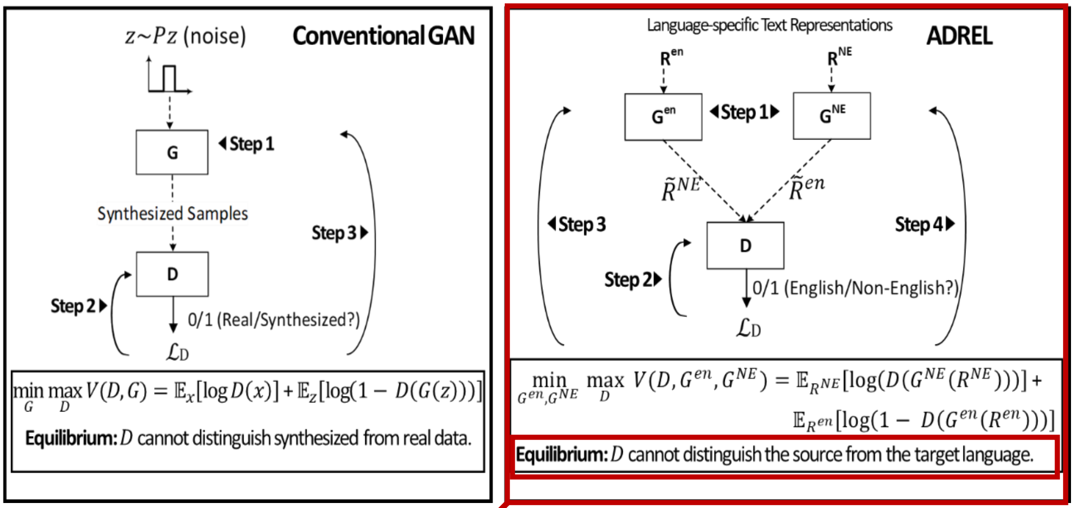
Equilibrium Identification

Benchmark Evaluation

Explanation and Detected Hacker Assets Profiling

Results Explanation

Detected Hacker Assets Profiling



Conventional GAN

$z \sim Pz$ (noise)

G ◂Step 1

Synthesized Samples

Step 3▸

Step 2▸

D

0/1 (Real/Synthesized?)

$\mathcal{L}_D$

$$\min_G \max_D V(D,G) = \mathbb{E}_x[\log D(x)] + \mathbb{E}_z[\log(1 - D(G(z)))]$$

Equilibrium: *D* cannot distinguish synthesized from real data.

ADREL

Language-specific Text Representations

R^{en} R^{NE}

G^{en} ◂Step 1▸ G^{NE}

$\tilde{R}^{NE}$ $\tilde{R}^{en}$

◂Step 3

Step 4▸

Step 2▸

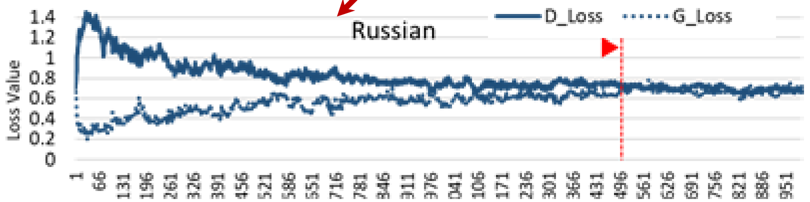
D

0/1 (English/Non-English?)

$\mathcal{L}_D$

$$\min_{G^{en},G^{NE}} \max_D V(D, G^{en}, G^{NE}) = \mathbb{E}_{R^{NE}}[\log(D(G^{NE}(R^{NE})))] + \mathbb{E}_{R^{en}}[\log(1 - D(G^{en}(R^{en})))]$$

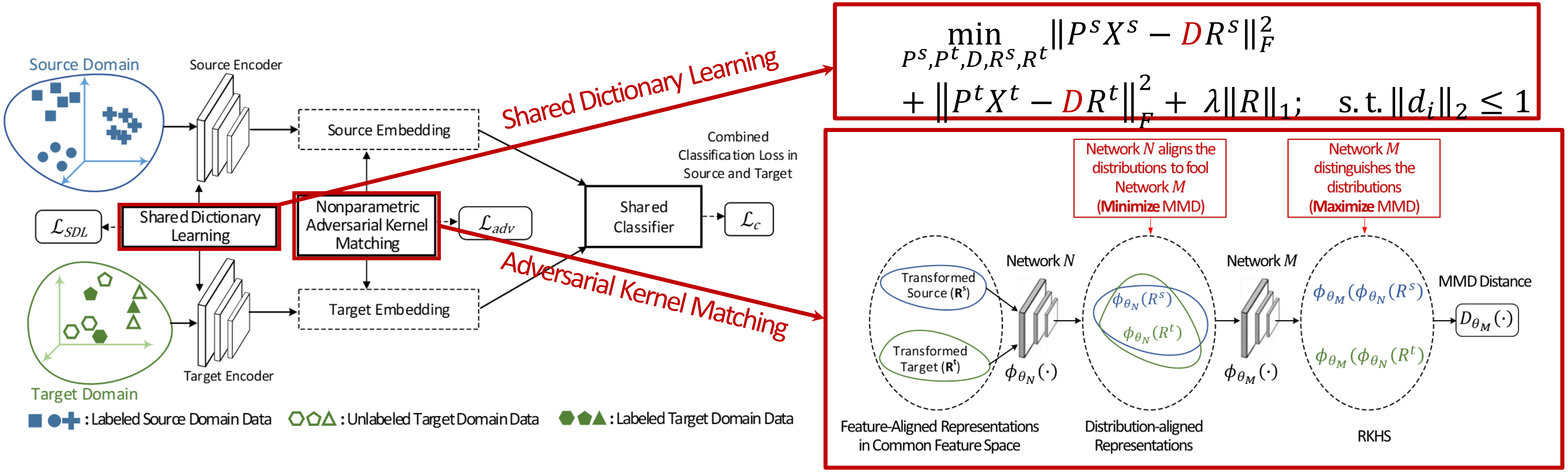
Equilibrium: *D* cannot distinguish the source from the target language.



- Significantly decreased reliance on human supervision and outperformed machine translation.

Essay III: Learning from Heterogeneous Cybersecurity Content (IEEE TPAMI, 2nd Round)

- Learning from two domains (multilingual text, source code, image representations)
- Align different data distributions & feature spaces with **domain adaptation theory**



$$\min_{P^s,P^t,D,R^s,R^t} \|P^sX^s - DR^s\|_F^2 + \|P^tX^t - DR^t\|_F^2 + \lambda\|R\|_1; \quad \text{s.t.}\ \|d_i\|_2 \le 1$$

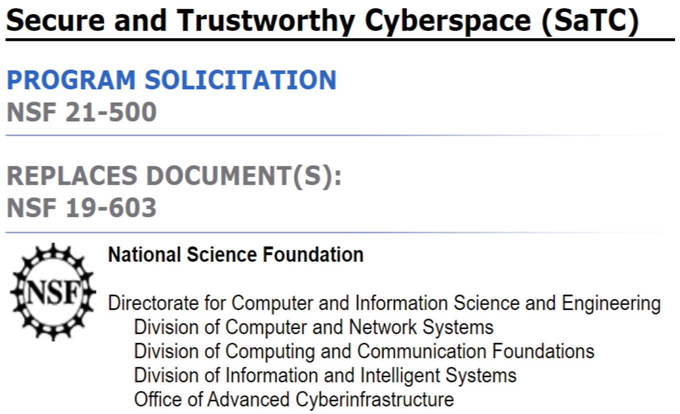
Source Domain
Source Encoder
Source Embedding
Shared Dictionary Learning
Combined Classification Loss in Source and Target
$\mathcal{L}_{SDL}$
Shared Dictionary Learning
Nonparametric Adversarial Kernel Matching
$\mathcal{L}_{adv}$
Shared Classifier
$\mathcal{L}_c$
Adversarial Kernel Matching
Target Embedding
Target Encoder
Target Domain
■●✚: Labeled Source Domain Data
⬡⬠△: Unlabeled Target Domain Data
⬢⬟▲: Labeled Target Domain Data
Network N aligns the distributions to fool Network M (**Minimize** MMD)
Network M distinguishes the distributions (**Maximize** MMD)
Transformed Source ($\mathbf{R}^s$)
Transformed Target ($\mathbf{R}^t$)
Network N
$\phi_{\theta_N}(\cdot)$
$\phi_{\theta_N}(R^s)$
$\phi_{\theta_N}(R^t)$
Network M
$\phi_{\theta_M}(\cdot)$
$\phi_{\theta_M}(\phi_{\theta_N}(R^s))$
$\phi_{\theta_M}(\phi_{\theta_N}(R^t))$
MMD Distance
$D_{\theta_M}(\cdot)$
Feature-Aligned Representations in Common Feature Space
Distribution-aligned Representations
RKHS

- Enables heterogeneous data analytics (multilingual text, images) in any online market.

Privacy and PII (Personally Identifiable Information) Analytics:

identifying and alleviating privacy risks for vulnerable populations

(SaTC 2019-; SFS-2, 2019-)



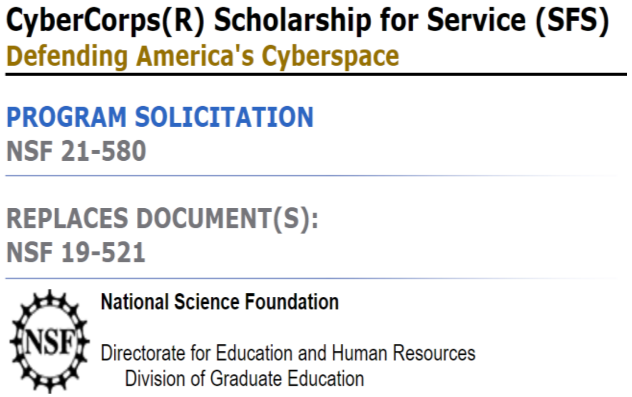
Secure and Trustworthy Cyberspace (SaTC)

PROGRAM SOLICITATION
NSF 21-500

REPLACES DOCUMENT(S):
NSF 19-603

National Science Foundation

Directorate for Computer and Information Science and Engineering
Division of Computer and Network Systems
Division of Computing and Communication Foundations
Division of Information and Intelligent Systems
Office of Advanced Cyberinfrastructure



CyberCorps(R) Scholarship for Service (SFS)
Defending America's Cyberspace

PROGRAM SOLICITATION
NSF 21-580

REPLACES DOCUMENT(S):
NSF 19-521

National Science Foundation

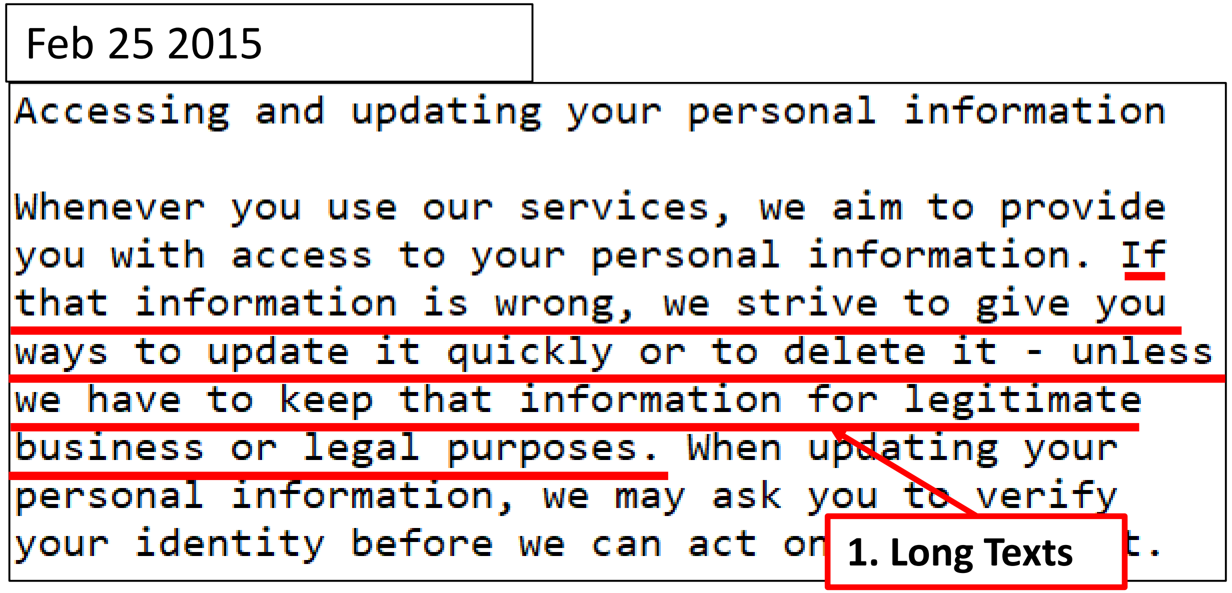
Directorate for Education and Human Resources
Division of Graduate Education

Automated Analysis of Changes in Privacy Policies: A Structured Self-Attentive Sentence Embedding Approach

Fangyu Lin (MIS, 2nd Round)

Privacy Policy, Before and After GDPR:

- Privacy policies contain lengthy texts.
- Require a college reading level to decode legalistic, confusing, or jargon-laden phrases (Gluck et al. 2016; Jain et al. 2016)



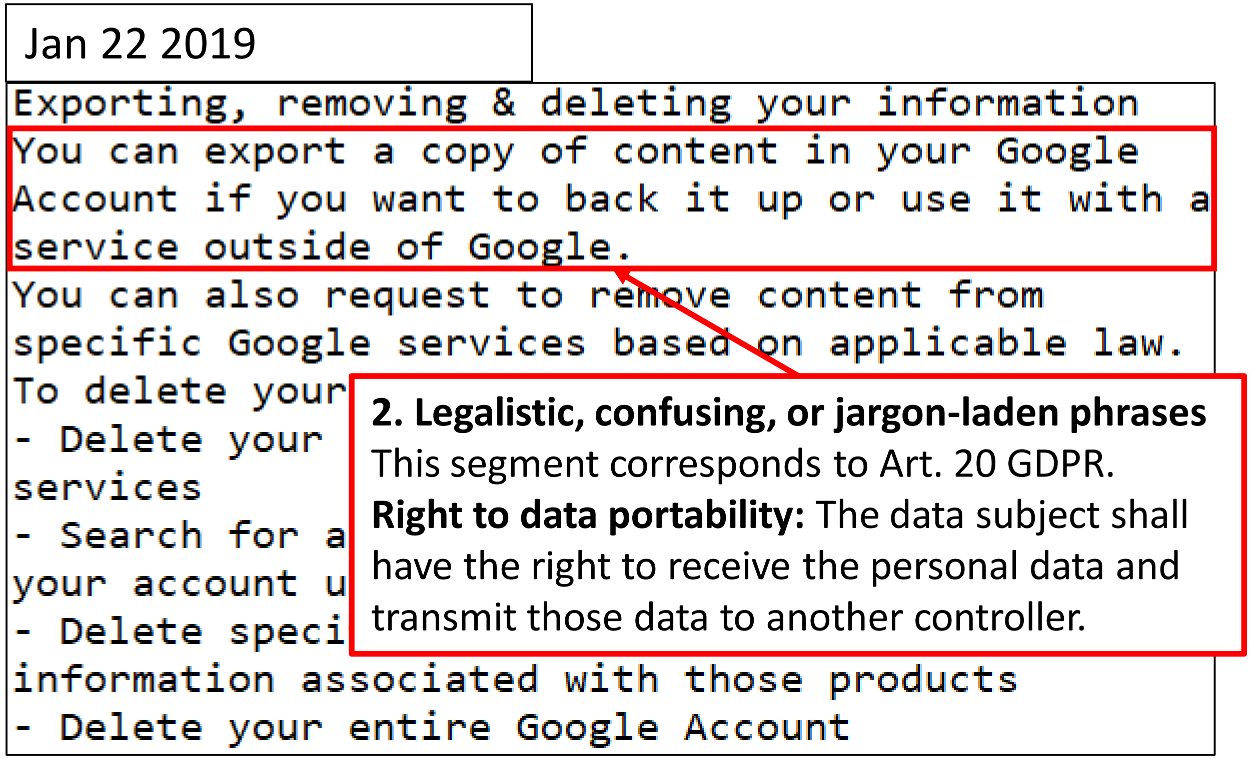
Feb 25 2015

Accessing and updating your personal information

Whenever you use our services, we aim to provide you with access to your personal information. If that information is wrong, we strive to give you ways to update it quickly or to delete it - unless we have to keep that information for legitimate business or legal purposes. When updating your personal information, we may ask you to verify your identity before we can act on t.

1. Long Texts

Figure 1. A "User Access, Edit, & Deletion" Segment in Google Privacy Policy Before and After GDPR: "Right to data portability" was added to the new version.



Jan 22 2019

Exporting, removing & deleting your information

You can export a copy of content in your Google Account if you want to back it up or use it with a service outside of Google.

You can also request to remove content from specific Google services based on applicable law.

To delete your

- Delete your

services

- Search for a

your account u

- Delete speci

information associated with those products

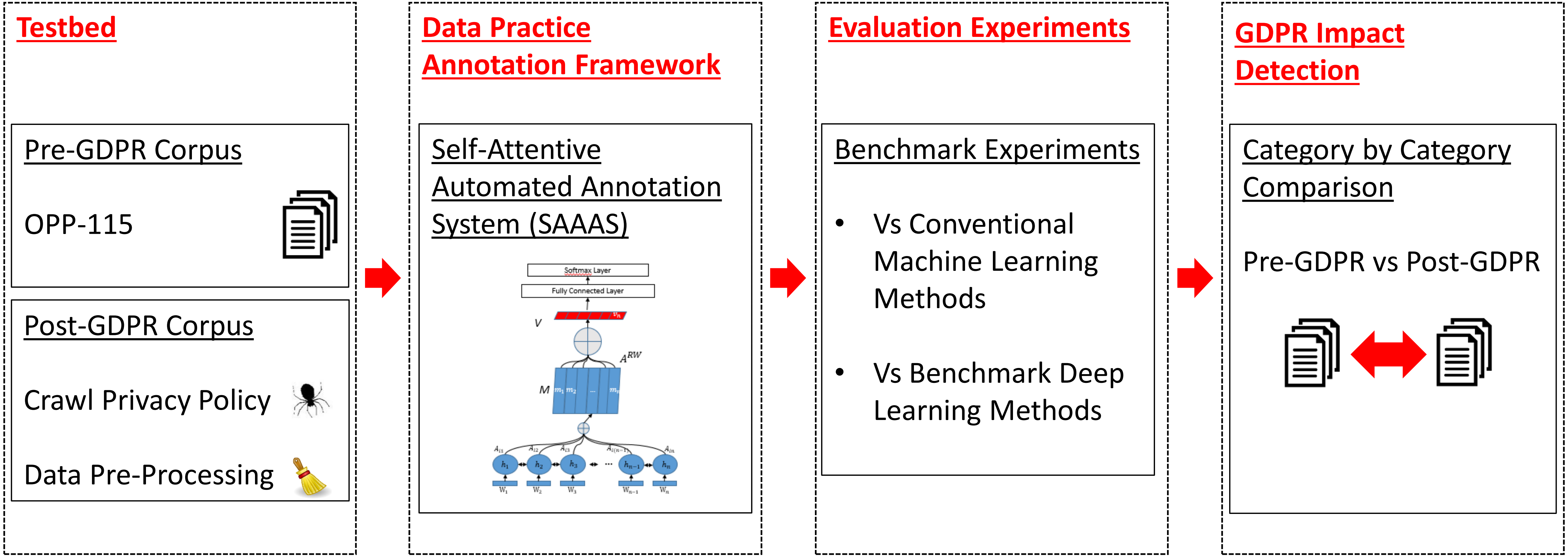
- Delete your entire Google Account

2. Legalistic, confusing, or jargon-laden phrases

This segment corresponds to Art. 20 GDPR.

Right to data portability: The data subject shall have the right to receive the personal data and transmit those data to another controller.

Research Design and Testbed



Testbed

Pre-GDPR Corpus

OPP-115

Post-GDPR Corpus

Crawl Privacy Policy

Data Pre-Processing

Data Practice Annotation Framework

Self-Attentive Automated Annotation System (SAAAS)

Softmax Layer
Fully Connected Layer
v
v_n
A^{RW}
M
m_1 m_2 ... m_r
A_{i1} A_{i2} A_{i3} $A_{i(n-1)}$ A_{in}
h_1 h_2 h_3 ... h_{n-1} h_n
W_1 W_2 W_3 W_{n-1} W_n

Evaluation Experiments

Benchmark Experiments

- Vs Conventional Machine Learning Methods
- Vs Benchmark Deep Learning Methods

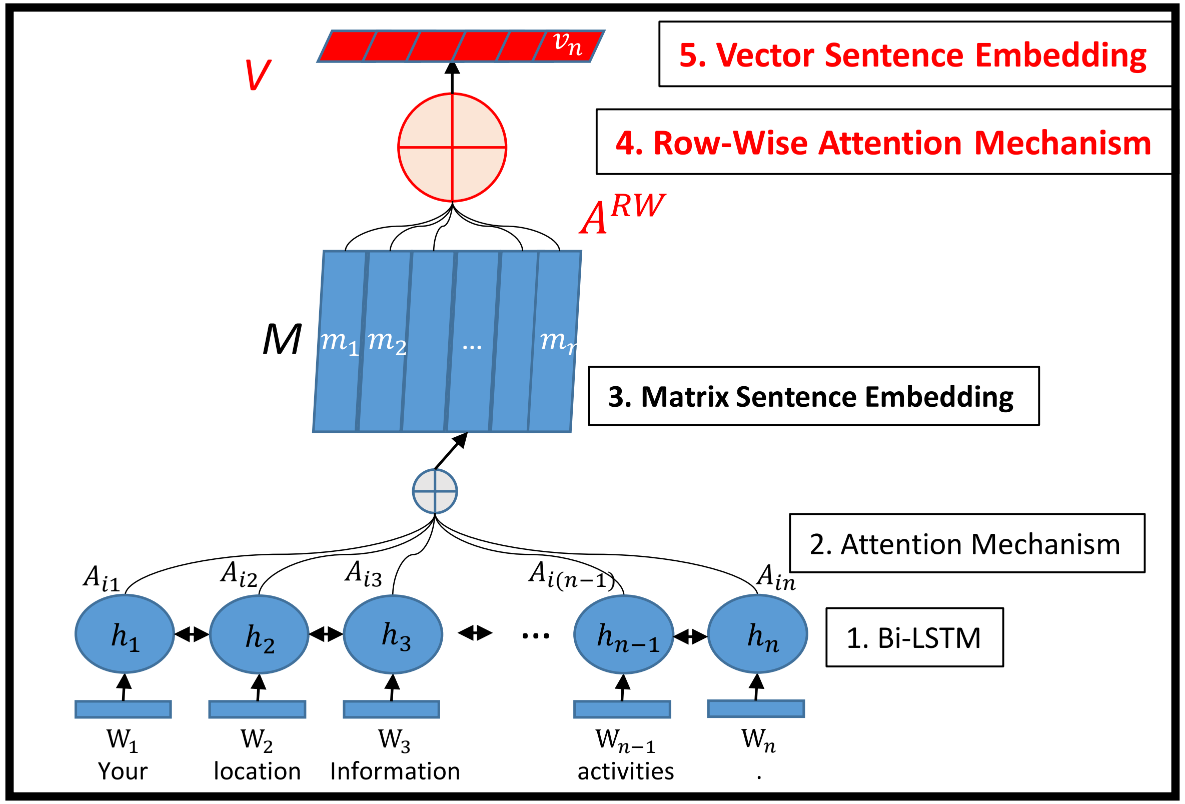
GDPR Impact Detection

Category by Category Comparison

Pre-GDPR vs Post-GDPR

Figure 3. Research Design for the Proposed Privacy Policy Evolution Analysis Framework

Data Practice Annotation Framework – SAAAS



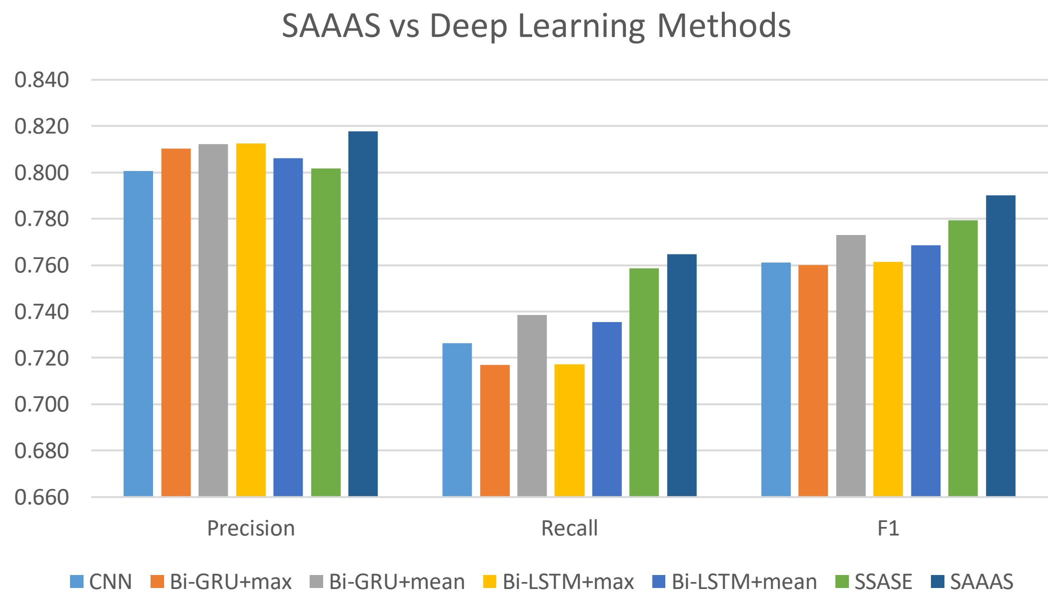
V
v_n
5. Vector Sentence Embedding
4. Row-Wise Attention Mechanism
A^{RW}
M
m_1 m_2 … m_r
3. Matrix Sentence Embedding
2. Attention Mechanism
A_{i1} A_{i2} A_{i3} $A_{i(n-1)}$ A_{in}
h_1 h_2 h_3 … h_{n-1} h_n
1. Bi-LSTM
W_1 Your
W_2 location
W_3 Information
W_{n-1} activities
W_n .

Figure 4. Row-Wise Self-Attentive Sentence Embedding: Key contribution is in red.

Row-Wise Self-Attentive Sentence Embedding

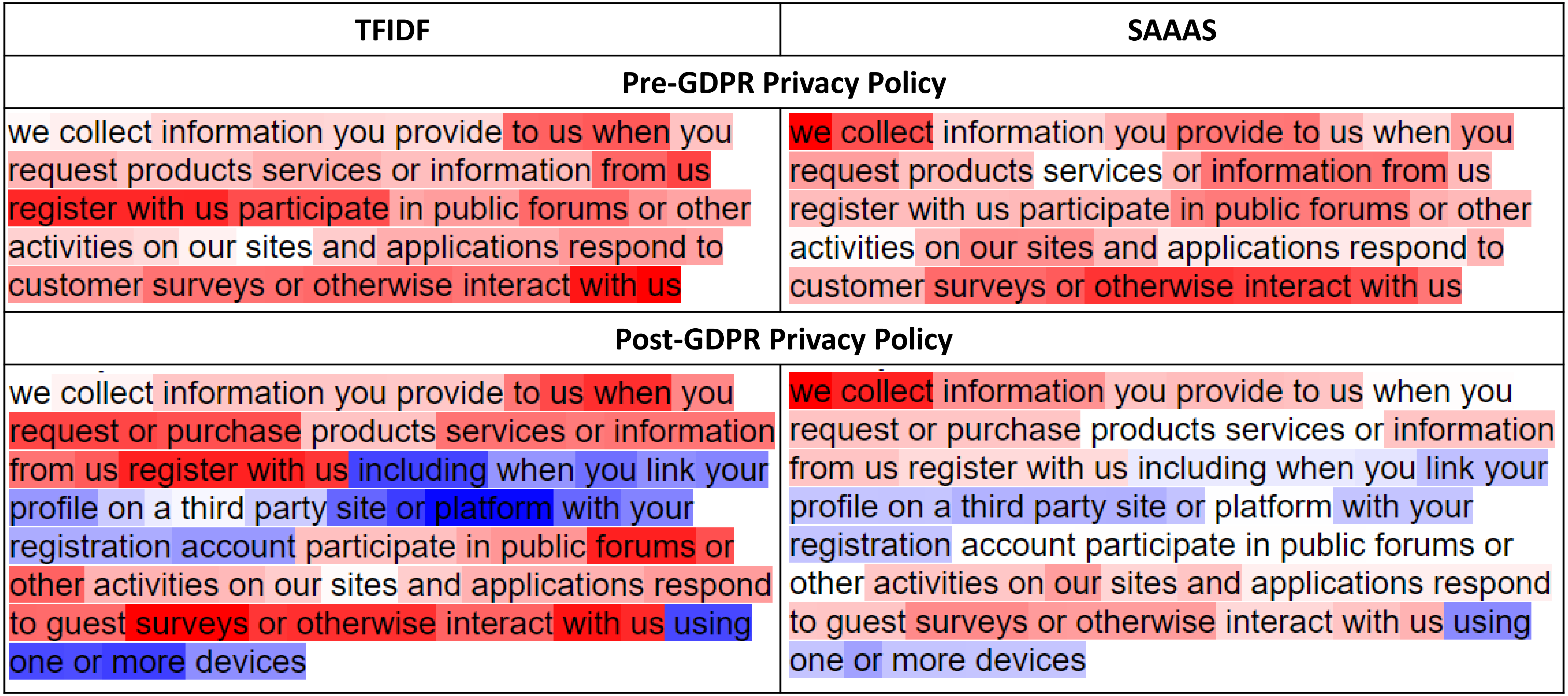
1. Bi-LSTM
2. Attention Mechanism
3. Matrix Sentence Embedding *M*
4. Row-Wise Attention Mechanism
   - **Input:** Matrix sentence embedding ***M***
   - **Output:** Row-wise attention weight matrix $\boldsymbol{A^{RW}}$
5. Vector Sentence Embedding
   - **Input:** Matrix sentence embedding ***M*** and $\boldsymbol{A^{RW}}$
   - **Output:** Vector sentence embedding ***V***
   - ***V*** is the dot product of ***M*** and $\boldsymbol{A^{RW}}$

Results and Discussion – SAAAS vs Benchmark Deep Learning Methods



	Precision	Recall	F1
CNN	0.801*	0.726*	0.761*
Bi-GRU+max	0.810*	0.717*	0.760*
Bi-GRU+mean	0.812*	0.738*	0.773*
Bi-LSTM+max	0.812*	0.717*	0.761*
Bi-LSTM+mean	0.806*	0.736*	0.769*
SSASE	0.802*	0.759*	0.779*
SAAAS	0.818	0.765	0.790

Case Study: GDPR Impact Detection (An Example)



TFIDF	SAAAS
Pre-GDPR Privacy Policy	
we collect information you provide to us when you request products services or information from us register with us participate in public forums or other activities on our sites and applications respond to customer surveys or otherwise interact with us	we collect information you provide to us when you request products services or information from us register with us participate in public forums or other activities on our sites and applications respond to customer surveys or otherwise interact with us
Post-GDPR Privacy Policy	
we collect information you provide to us when you request or purchase products services or information from us register with us including when you link your profile on a third party site or platform with your registration account participate in public forums or other activities on our sites and applications respond to guest surveys or otherwise interact with us using one or more devices	we collect information you provide to us when you request or purchase products services or information from us register with us including when you link your profile on a third party site or platform with your registration account participate in public forums or other activities on our sites and applications respond to guest surveys or otherwise interact with us using one or more devices

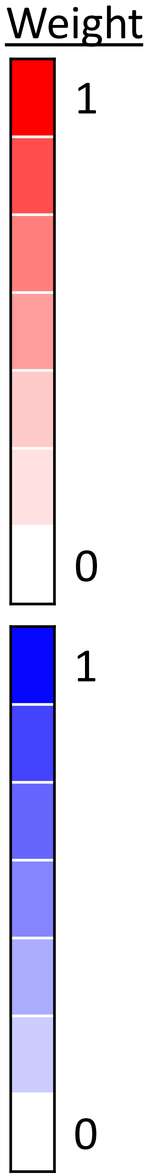
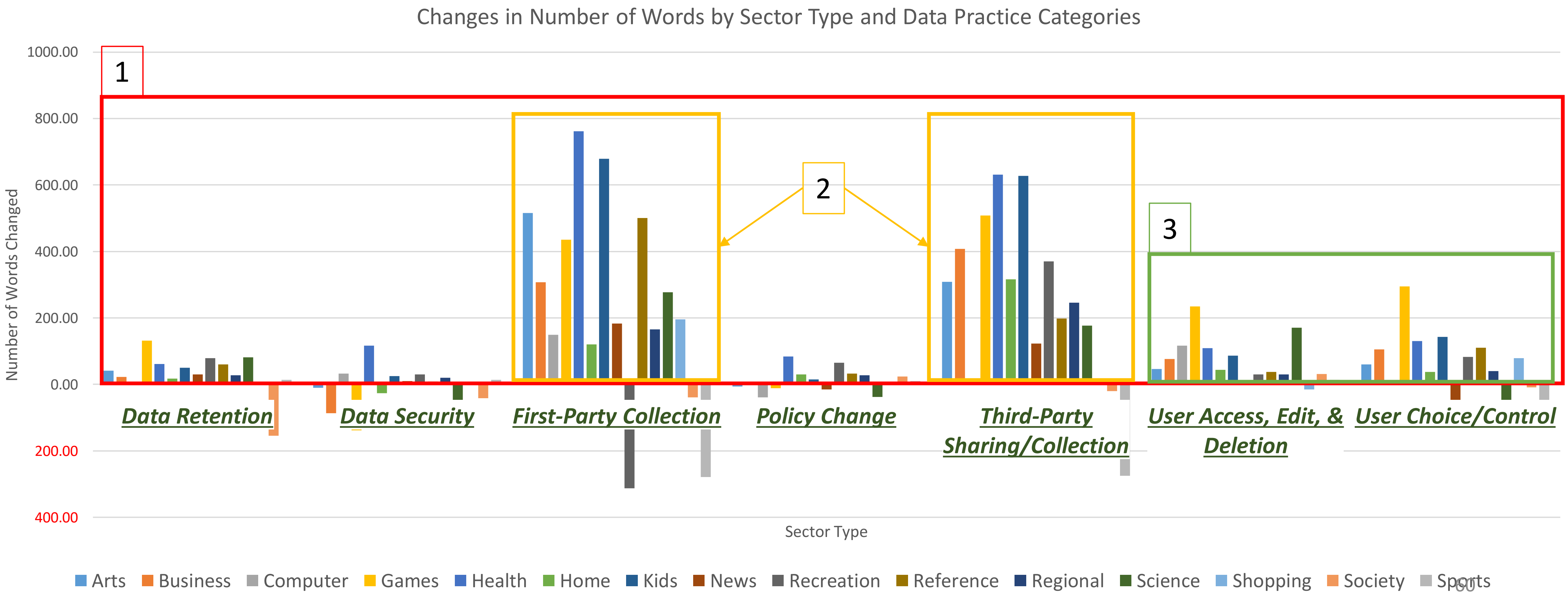


Table 9. An example of corresponding segment in pre- and post-GDPR "Disney" privacy policy. The red part is unmodified, and the blue part is new content. In the heatmap, the shade of red and blue corresponds to the weight, ranging from 0 to 1.

Case Study: GDPR Impact Detection

- The number of words increased in most of the categories. Complies with GDPR and CCPA requirements to provide comprehensive information related to data processing
- First Party Collection and Third-Party Collection categories changed the most.

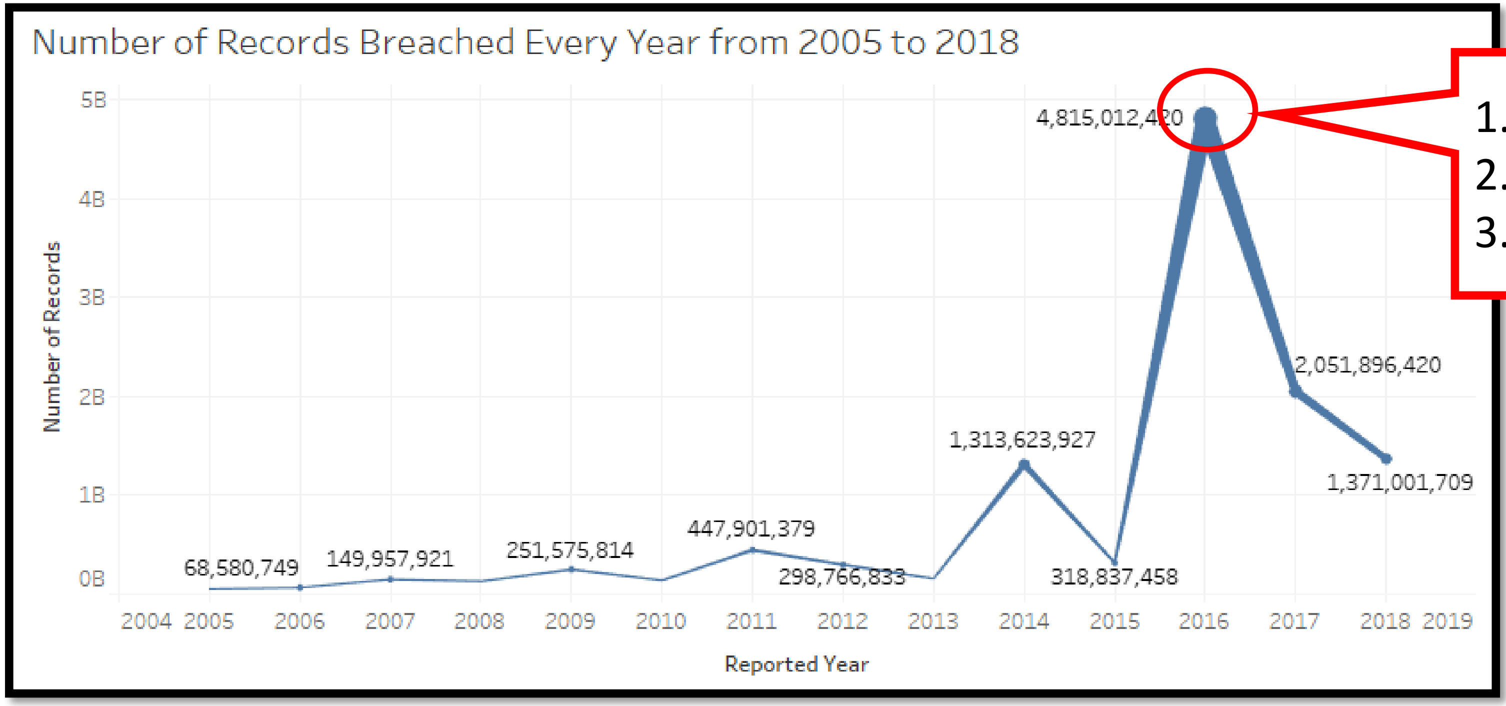


Exploring Privacy Risk of Exposed Digital Personally Identifiable Information (PII): A Neighbor Attention-Based Approach

Fangyu Lin and Hsinchun Chen

Data Breaches since 2005 (FTC, Clearinghouse, 2019)

- # of records breached: 11,582,808,013
- # of data breaches: 9,071



2016 Data Breach

1. Yahoo! : 3.5B user accounts
2. FriendFinder : 412M user accounts
3. MySpace : 360M passwords

Revealing and Protecting PII:
From Dark Web to Surface Web

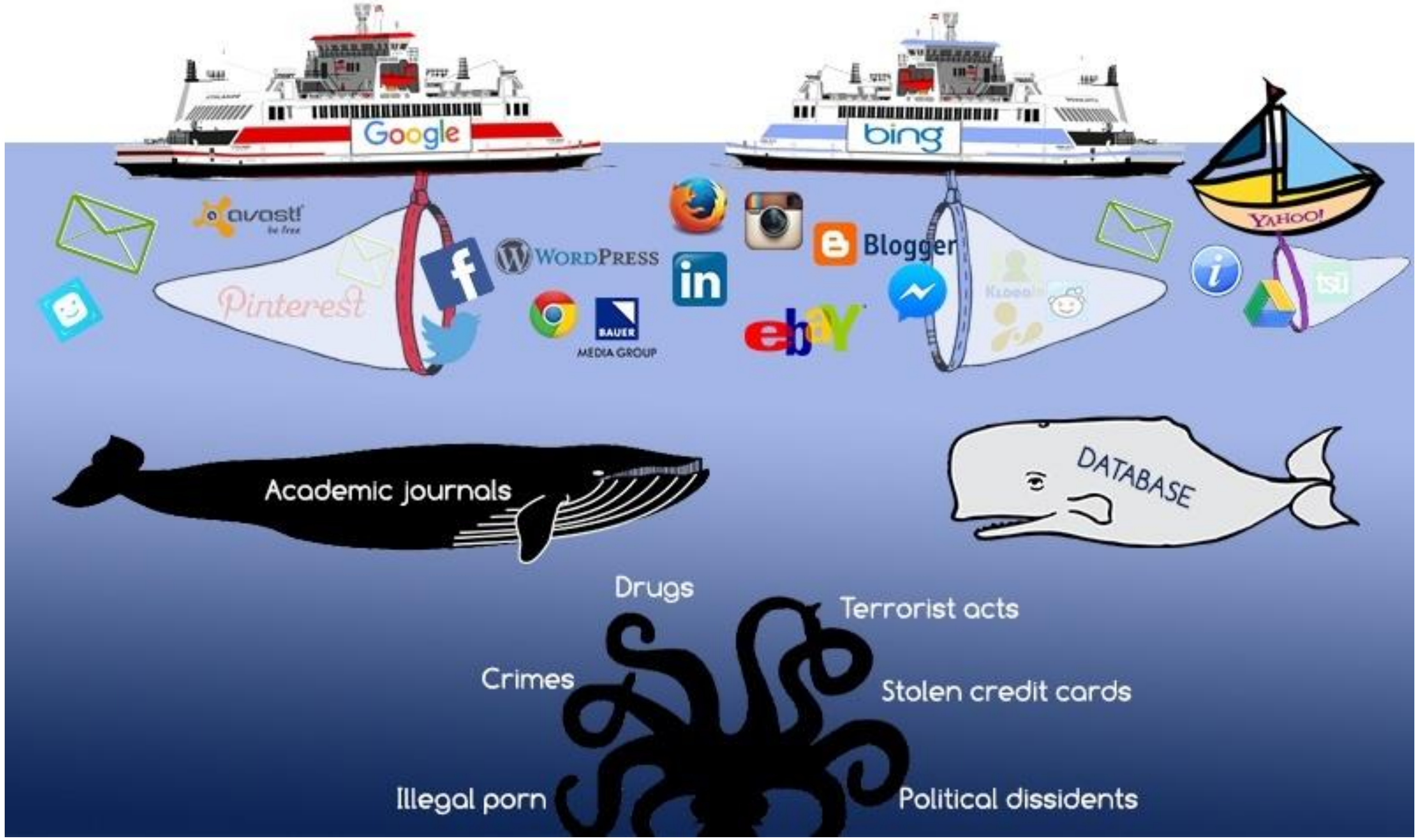
Surface Web

Deep Web

Dark Web

DarkNet

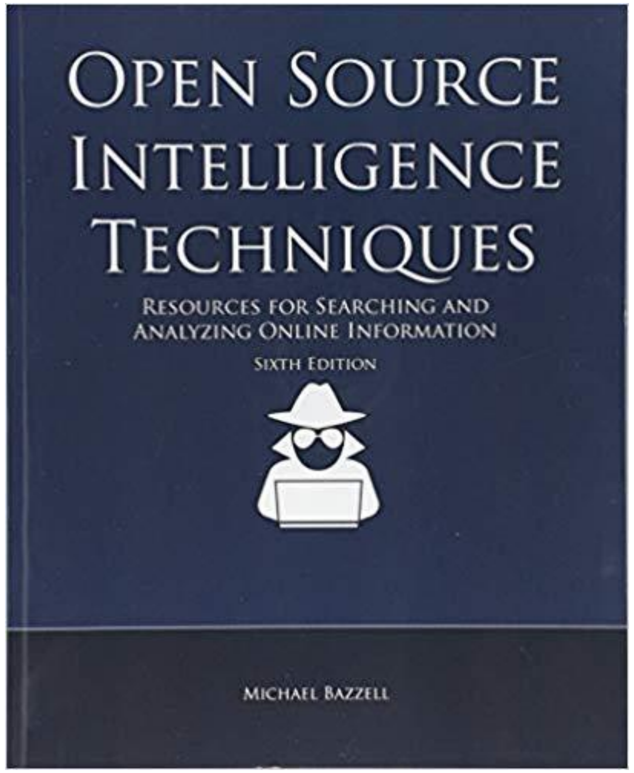
Hacker Web

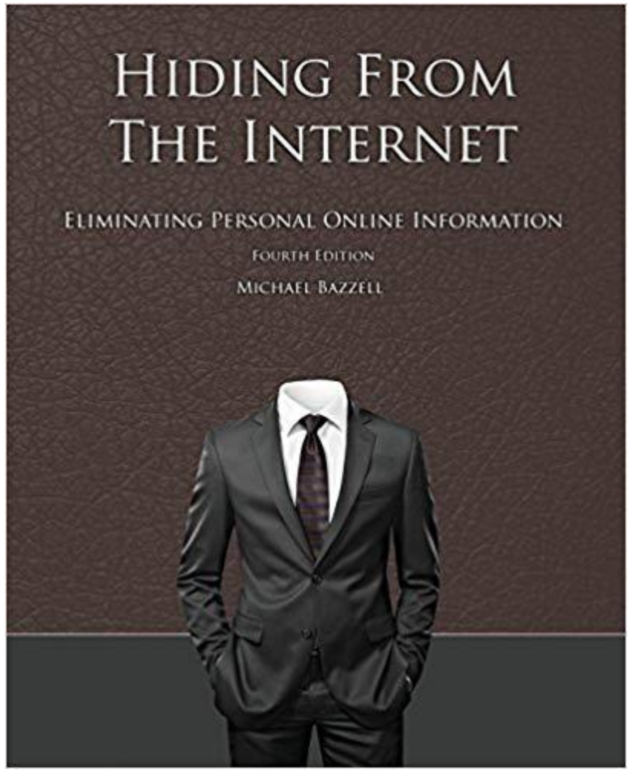


IRB, HIPAA, GDPR, PII

➔ Cybersecurity to Privacy

➔ **Michael Bazzell** + From Dark Web to Surface Web


Open Source Intelligence Techniques
Resources for Searching and Analyzing Online Information
Sixth Edition
Michael Bazzell



Dark Web Intelligence Sources (May, 2019)

Source	**Description**	**Size***	**Promising Attributes**
Stolen Account Collection	Stolen social media and e-mail accounts	**25 billions**	**Username**
			Password
Stolen Credit Card - Tormarket	Stolen credit and debit card owner information * No card number	**832 thousands**	**Full name**
			Country
			State
			City
			Zip
Stolen SSN - Buyssn	Personal information of SSN owners *No SSN	**5.75 millions**	**Full name**
			YOB
			City
			State
			Zip
			Country

“Passwords are like underwear…
change often, don’t share…”

Stolen Accounts

Rank	E-mail Domains	Numbers	Percentage
1	yahoo.com	244,769,117	20.41%
2	hotmail.com	182,564,724	15.22%
3	gmail.com	103,435,791	8.62%
4	mail.ru	90,371,699	7.53%
5	aol.com	44,830,568	3.74%
6	yandex.ru	36,336,003	3.03%
7	rambler.ru	23,521,080	1.96%
8	hotmail.fr	16,571,495	1.38%
9	web.de	12,918,595	1.08%
10	live.com	11,661,375	0.97%
11	msn.com	11,248,354	0.94%
12	gmx.de	10,800,404	0.90%
13	163.com	10,492,032	0.87%
14	bk.ru	9,416,062	0.78%
15	yahoo.fr	8,886,223	0.74%
Total	-	**817,823,522**	**68.18%**

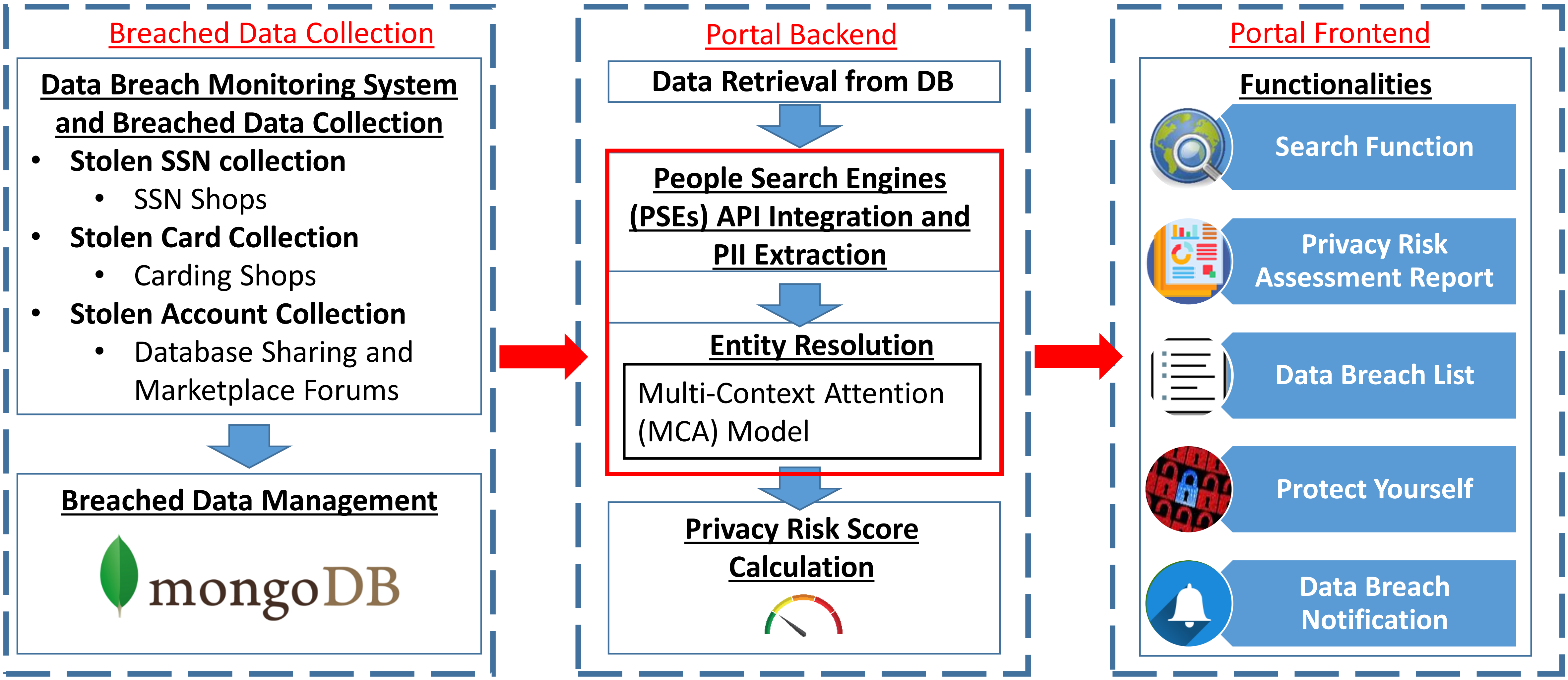
Popular Passwords

Rank	Passwords	Numbers
1	**123456**	3,370,644
2	**123456789**	1,187,812
3	**Homelesspa***	546,648
4	**password**	522,529
5	**abc123**	516,091
6	**password1**	435,753
7	**12345**	382,970
8	**qwerty**	376,099
9	**12345678**	357,654
10	**1234567**	287,453
11	**1234567890**	252,929
12	**111111**	236,852
13	**iloveyou**	211,593
14	**123456a**	205,807
15	**123123**	191,450
Total	**-**	**9,082,284**

AZSecure Privacy Portal Design

LifeLock®

Relentlessly Protecting Your Identity™



Breached Data Collection

Data Breach Monitoring System and Breached Data Collection

- **Stolen SSN collection**
  - SSN Shops
- **Stolen Card Collection**
  - Carding Shops
- **Stolen Account Collection**
  - Database Sharing and Marketplace Forums

Breached Data Management

mongoDB

Portal Backend

Data Retrieval from DB

People Search Engines (PSEs) API Integration and PII Extraction

Entity Resolution

Multi-Context Attention (MCA) Model

Privacy Risk Score Calculation

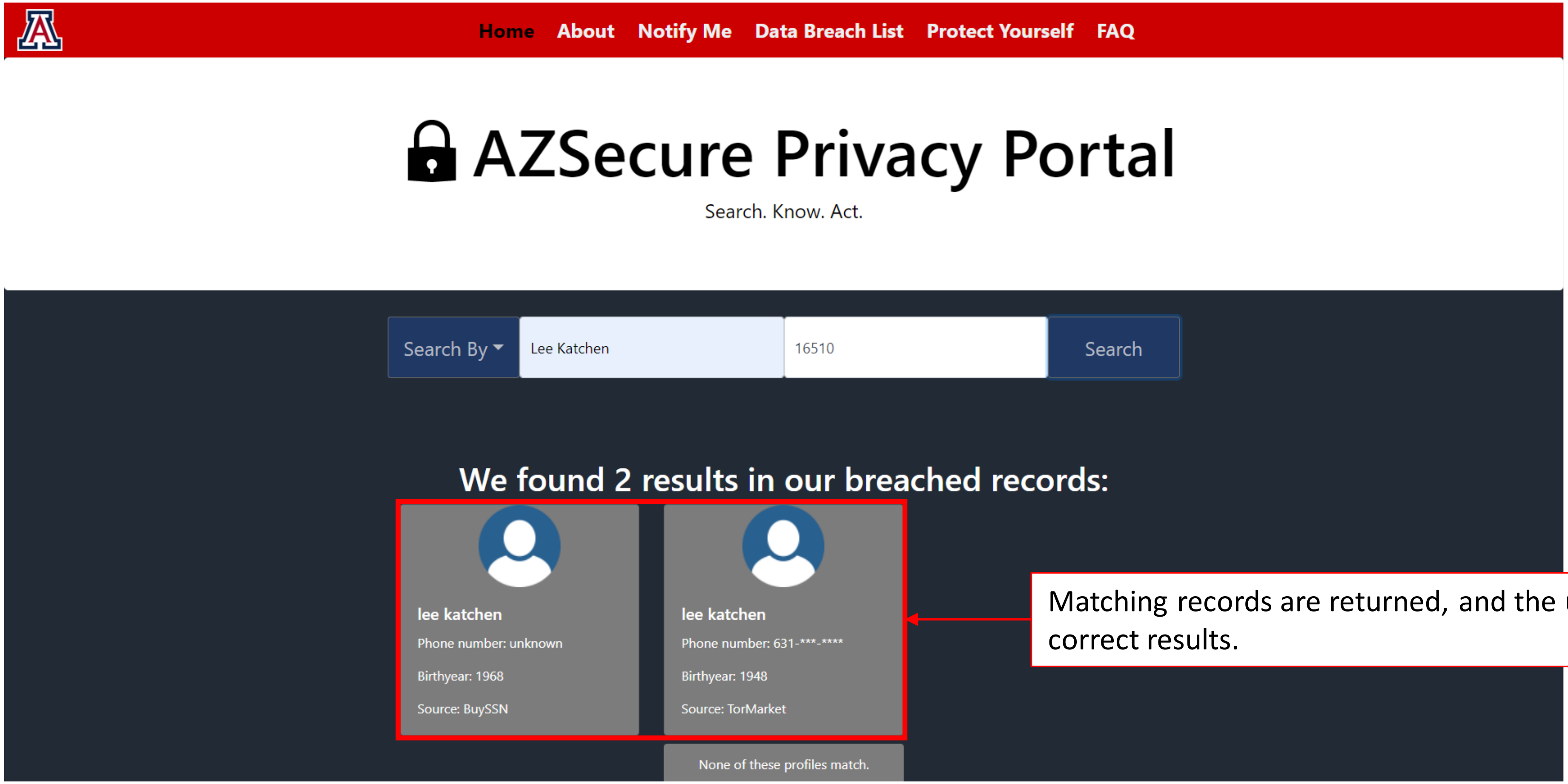
Portal Frontend

Functionalities

- Search Function
- Privacy Risk Assessment Report
- Data Breach List
- Protect Yourself
- Data Breach Notification

Figure 1. AZSecure Privacy Portal Project Overview

Search in AZSecure Privacy Portal



AZSecure Privacy Portal

Search. Know. Act.

Lee Katchen

16510

We found 2 results in our breached records:

lee katchen

Phone number: unknown

Birthyear: 1968

Source: BuySSN

lee katchen

Phone number: 631-***-****

Birthyear: 1948

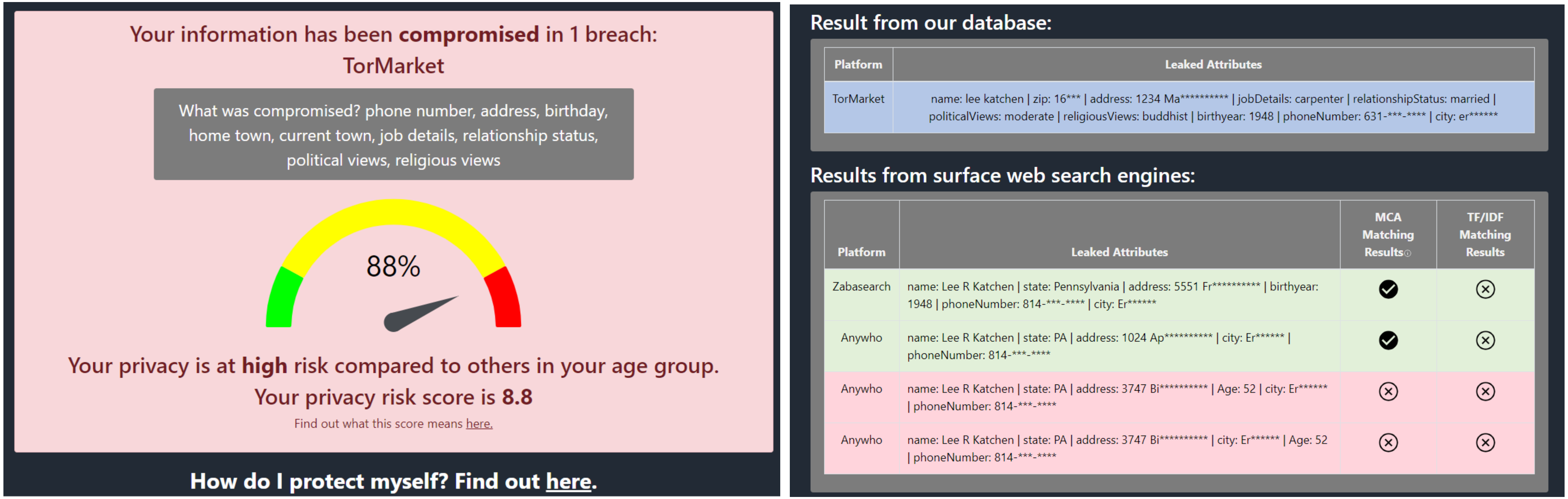
Source: TorMarket

None of these profiles match.

Matching records are returned, and the user can select the correct results.

Figure 5. A mock-up response when records are found

Return Exposed PII



Your information has been **compromised** in 1 breach: TorMarket

What was compromised? phone number, address, birthday, home town, current town, job details, relationship status, political views, religious views

88%

Your privacy is at **high** risk compared to others in your age group.

Your privacy risk score is **8.8**

Find out what this score means here.

How do I protect myself? Find out here.

Result from our database:

<table>
<tr><th>Platform</th><th>Leaked Attributes</th></tr>
<tr><td>TorMarket</td><td>name: lee katchen | zip: 16*** | address: 1234 Ma********** | jobDetails: carpenter | relationshipStatus: married | politicalViews: moderate | religiousViews: buddhist | birthyear: 1948 | phoneNumber: 631-***-**** | city: er******</td></tr>
</table>

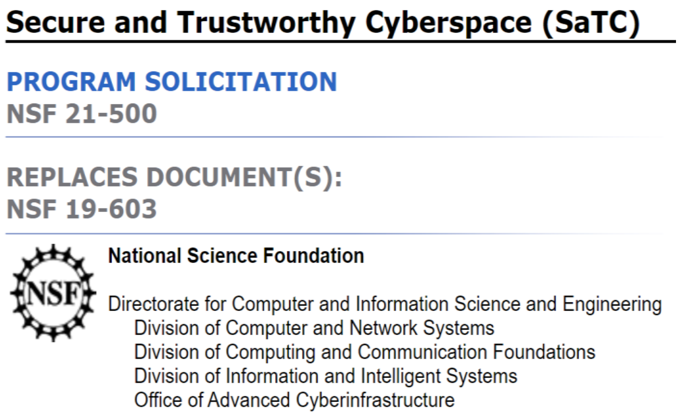
Results from surface web search engines:

<table>
<tr><th>Platform</th><th>Leaked Attributes</th><th>MCA Matching Results</th><th>TF/IDF Matching Results</th></tr>
<tr><td>Zabasearch</td><td>name: Lee R Katchen | state: Pennsylvania | address: 5551 Fr********** | birthyear: 1948 | phoneNumber: 814-***-**** | city: Er******</td><td>✓</td><td>⊗</td></tr>
<tr><td>Anywho</td><td>name: Lee R Katchen | state: PA | address: 1024 Ap********** | city: Er****** | phoneNumber: 814-***-****</td><td>✓</td><td>⊗</td></tr>
<tr><td>Anywho</td><td>name: Lee R Katchen | state: PA | address: 3747 Bi********** | Age: 52 | city: Er****** | phoneNumber: 814-***-****</td><td>⊗</td><td>⊗</td></tr>
<tr><td>Anywho</td><td>name: Lee R Katchen | state: PA | address: 3747 Bi********** | city: Er****** | Age: 52 | phoneNumber: 814-***-****</td><td>⊗</td><td>⊗</td></tr>
</table>

Figure 9. Mock-ups of a comprehensive exposed PII profile

Adversarial Malware Generation and Evasion:
adversarial AI in cybersecurity

(SaTC 2019-; SFS-2, 2019-)



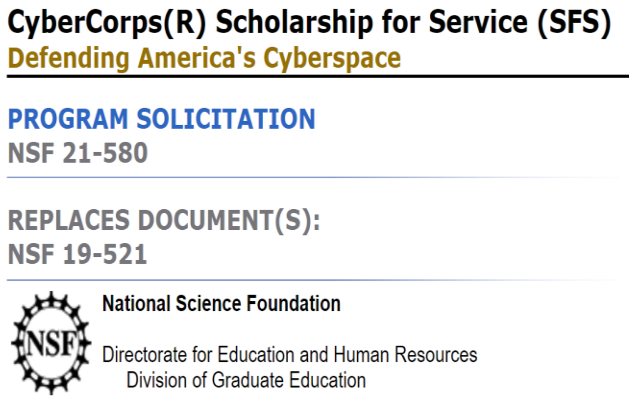
Secure and Trustworthy Cyberspace (SaTC)

PROGRAM SOLICITATION
NSF 21-500

REPLACES DOCUMENT(S):
NSF 19-603

National Science Foundation

Directorate for Computer and Information Science and Engineering
Division of Computer and Network Systems
Division of Computing and Communication Foundations
Division of Information and Intelligent Systems
Office of Advanced Cyberinfrastructure



CyberCorps(R) Scholarship for Service (SFS)
Defending America's Cyberspace

PROGRAM SOLICITATION
NSF 21-580

REPLACES DOCUMENT(S):
NSF 19-521

National Science Foundation

Directorate for Education and Human Resources
Division of Graduate Education

Defending Cybersecurity AI Agents

Reza Ebrahimi (JMIS, MISQ)

- **Essay I:** Learning to Protect Malware Detectors
- **Essay 2:** Learning to Protect any Defense AI agent

Defending Cybersecurity AI Agents

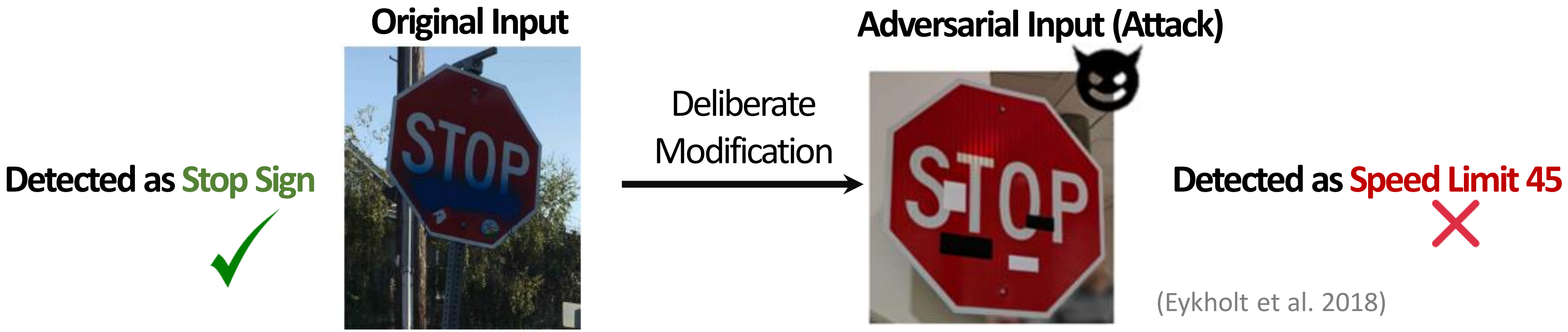
Symantec unveils artificial intelligence powered endpoint security

(expresscomputer.in)

- Cybersecurity firms are adopting AI agents for autonomous cyber defense (Rai et al. 2019).
  - Automate threat detection and remediation at a large scale (Tolido et al. 2019).

- However, AI agents have shown to be vulnerable to adversarial attacks.

- Inputs meticulously modified to mislead them (Yuan et al. 2019). → Known as adversarial attacks (Apruzzese et al. 2019).



Original Input

Adversarial Input (Attack)

Deliberate Modification

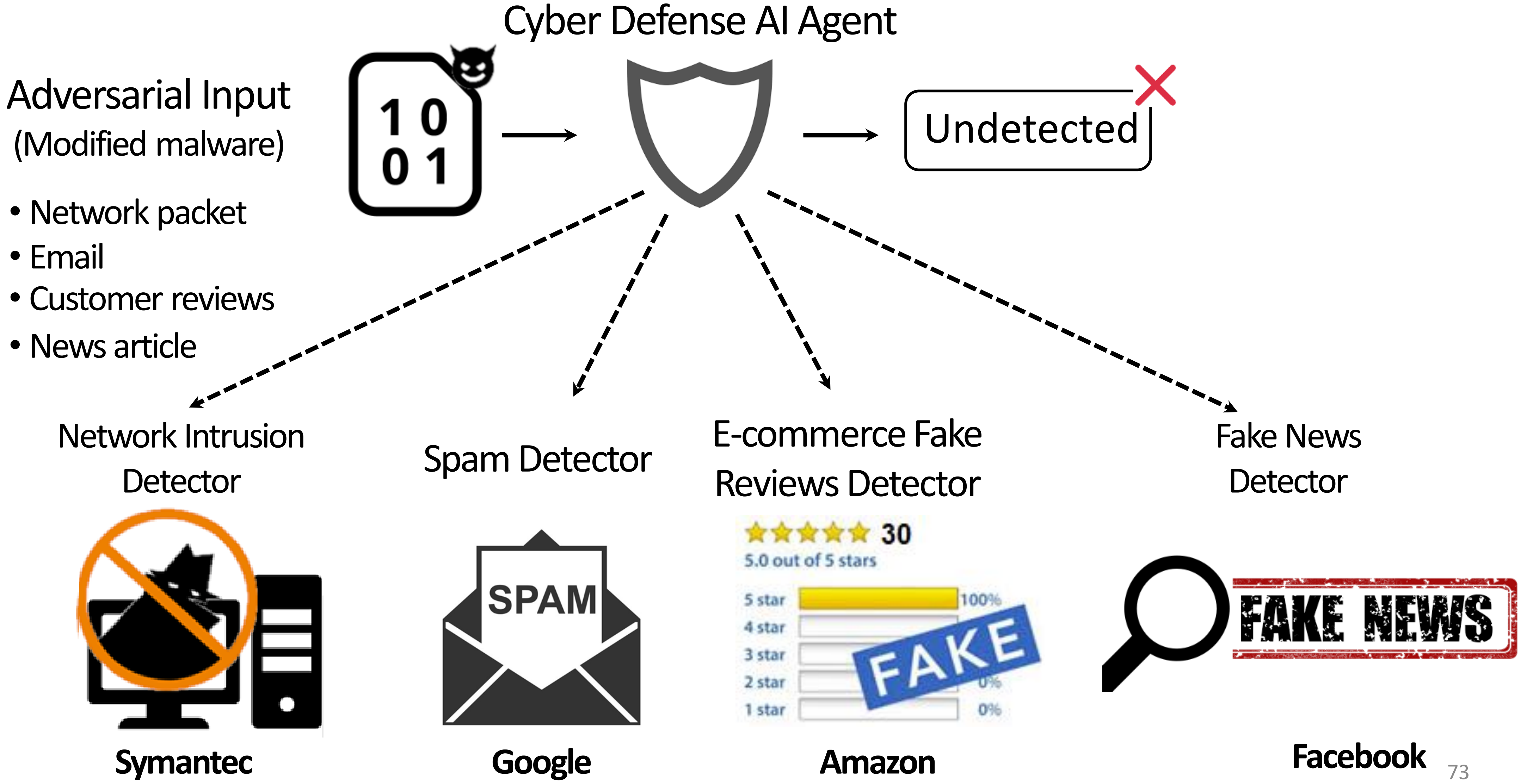
Detected as Stop Sign ✓

Detected as Speed Limit 45 ✗

(Eykholt et al. 2018)

- **How can we protect cyber defense AI agents?**

Defending Cybersecurity AI Agents



Cyber Defense AI Agent

Adversarial Input
(Modified malware)

- Network packet
- Email
- Customer reviews
- News article

Undetected

Network Intrusion Detector

Spam Detector

E-commerce Fake Reviews Detector

Fake News Detector

Symantec

Google

Amazon

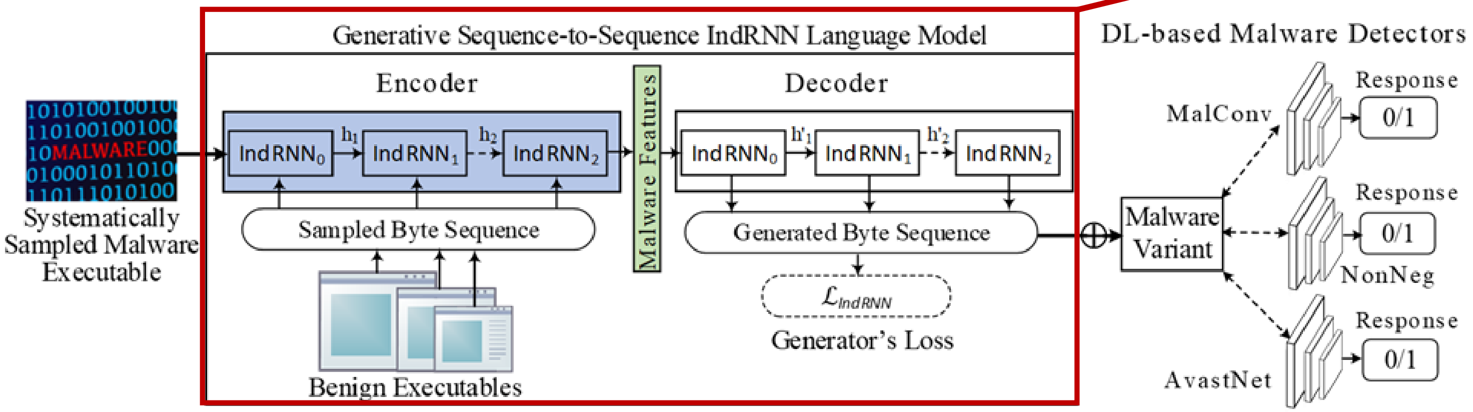
Facebook

Essay I: Learning to Protect Malware Detectors

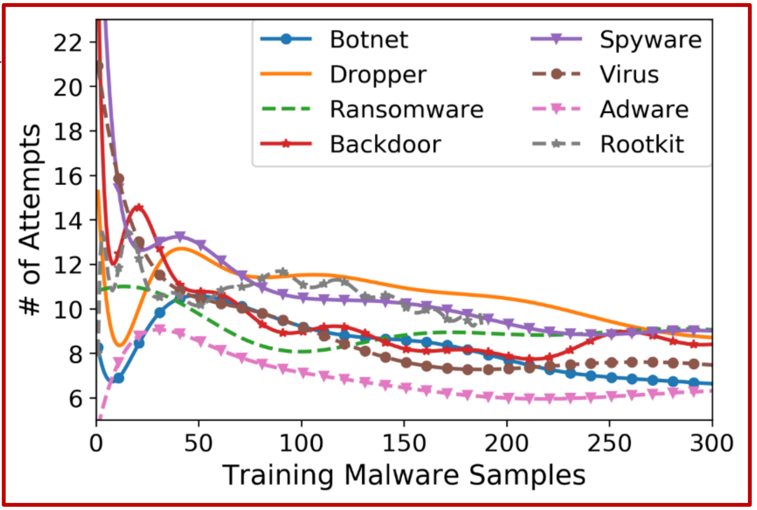
(JMIS, In sub.)

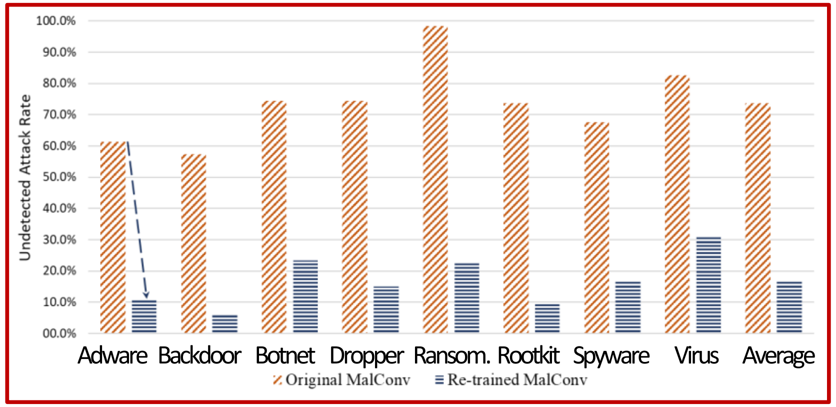
- Malware attack is #1 cause of damage to IT infrastructure (Bissell et al. 2019).
- Malware detector is the first line of defense. → Can be misled by adversarial inputs.
  - Language modeling helps emulate these inputs.

Adversarial Language Modeling



$$\underset{\delta\in\Delta}{\text{maximize}}\,\mathcal{L}(\mathcal{H}_\theta(x+\delta),y)$$



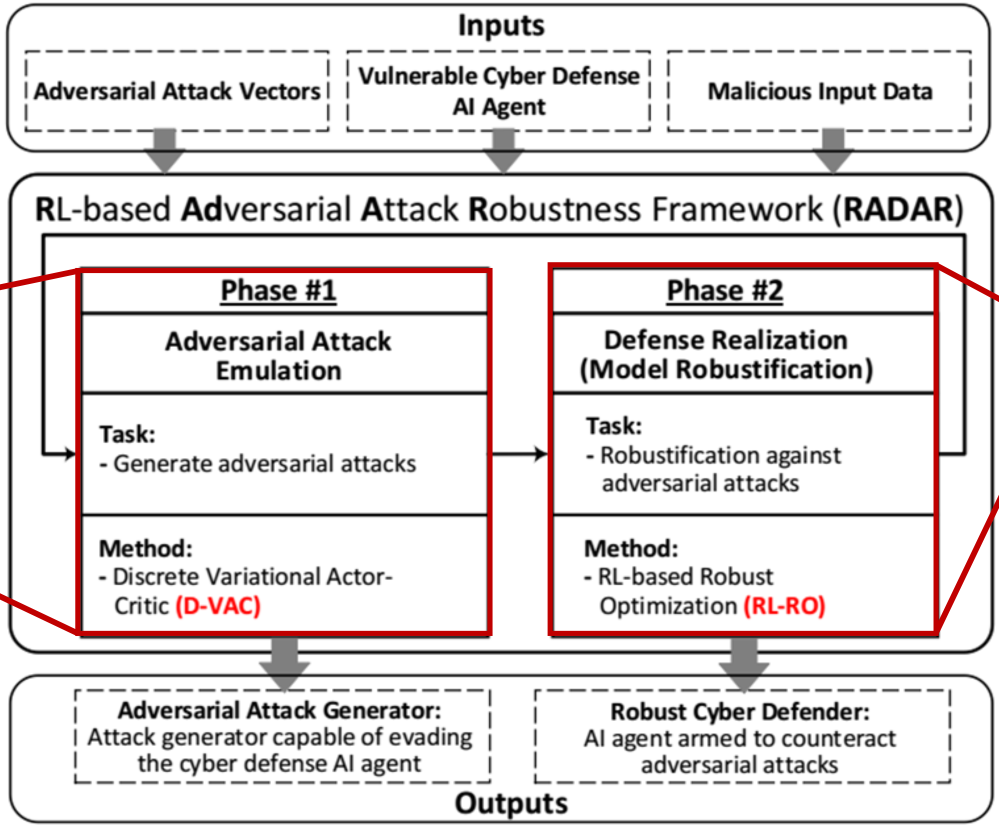


- Significantly improves the robustness of malware detectors against adversarial attacks.

Essay II: Learning to Protect any Defense AI Agent

(MISQ, 1st Round)

- Modern AI agents can be misled by adversarial attacks. → Emulating these attacks is vital for defense.
- A game between adversary and defender helps emulation.



Inputs

Adversarial Attack Vectors

Vulnerable Cyber Defense AI Agent

Malicious Input Data

RL-based **Ad**versarial **A**ttack **R**obustness Framework (**RADAR**)

Phase #1

Adversarial Attack Emulation

Task:
- Generate adversarial attacks

Method:
- Discrete Variational Actor-Critic **(D-VAC)**

Phase #2

Defense Realization (Model Robustification)

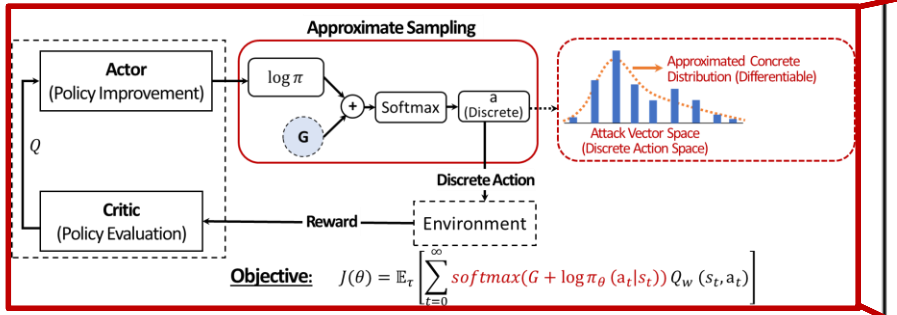
Task:
- Robustification against adversarial attacks

Method:
- RL-based Robust Optimization **(RL-RO)**

Adversarial Attack Generator: Attack generator capable of evading the cyber defense AI agent

Robust Cyber Defender: AI agent armed to counteract adversarial attacks

Outputs



Objective: $J(\theta) = \mathbb{E}_\tau\left[\sum_{t=0}^{\infty} softmax(G + \log \pi_\theta\,(a_t|s_t))\, Q_w\,(s_t, a_t)\right]$

Emulate Adversary with
Discrete Variational Actor-Critic
(D-VAC)

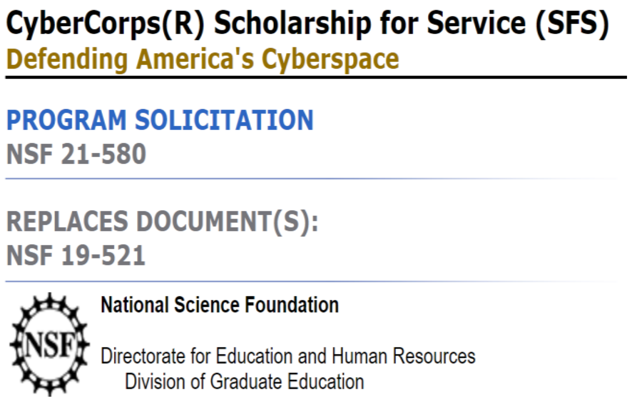
$$\underbrace{\min_{\phi}}_{\text{Estimate by Gradient Descent}}\left(\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell\left(h_\phi\left(\underbrace{\Pi_{\theta,h_\phi}(x)}_{\text{Adversarial Attack Sample Generated by D-VAC}}\right),y\right)\right]\right)$$

Strengthen Robustness with
RL-based Robust Optimization
(RL-RO)

- Strengthened the robustness of AI agents against adversarial attacks.

Smart Vulnerability Assessment: scientific workflows and OSS vulnerability analytics and mitigation

(CICI 2019-; SFS-2, 2019-)



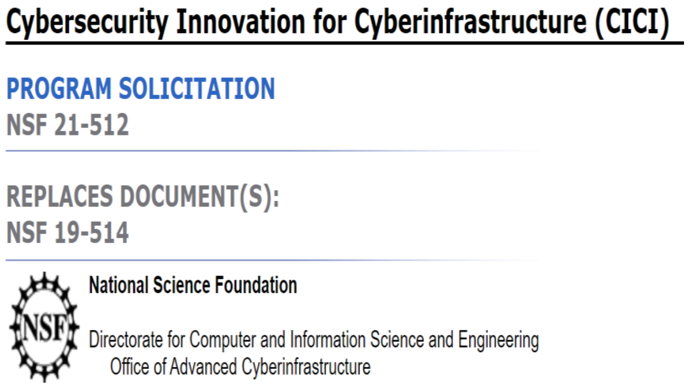
CyberCorps(R) Scholarship for Service (SFS)
Defending America's Cyberspace

PROGRAM SOLICITATION
NSF 21-580

REPLACES DOCUMENT(S):
NSF 19-521

National Science Foundation

Directorate for Education and Human Resources
Division of Graduate Education



Cybersecurity Innovation for Cyberinfrastructure (CICI)

PROGRAM SOLICITATION
NSF 21-512

REPLACES DOCUMENT(S):
NSF 19-514

National Science Foundation

Directorate for Computer and Information Science and Engineering
Office of Advanced Cyberinfrastructure

Linking Hacker Community Exploits to Known Vulnerabilities for Proactive Cyber Threat Intelligence: An Attention-based Deep Structured Semantic Model Approach

Sagar Samtani (MISQ, forthcoming)

Protecting Scientific Instruments and Cyberinfrastructure:

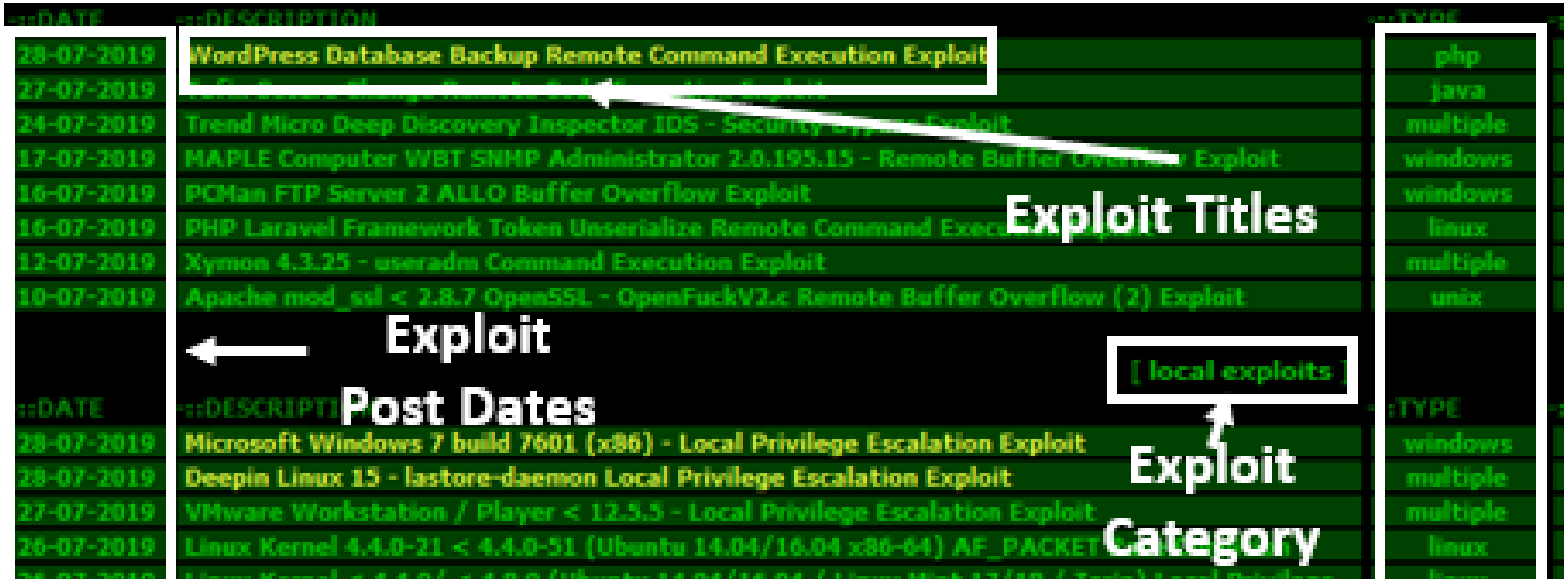
From iPlant/CyVerse (life sciences)
to BioSphere 2/LEO (earth sciences)…

a new UA/USF/AZSecure NSF CICI project, 2019-2022

iPlant Collaborative
Empowering a new plant biology

CYVERSE®

Hacker Forum Exploits



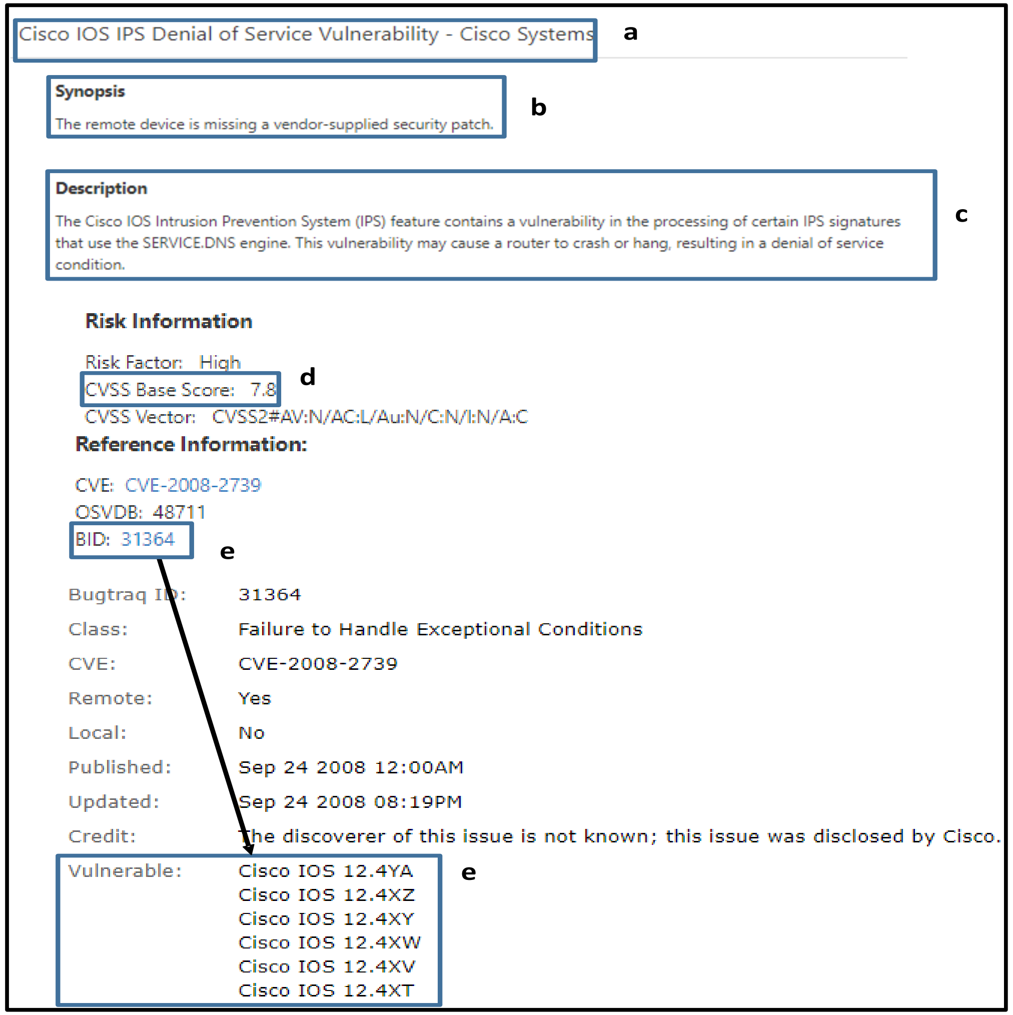
::DATE	::DESCRIPTION	::TYPE
28-07-2019	WordPress Database Backup Remote Command Execution Exploit	php
27-07-2019	[illegible]	java
24-07-2019	Trend Micro Deep Discovery Inspector IDS - Security Bypass Exploit	multiple
17-07-2019	MAPLE Computer WBT SNMP Administrator 2.0.195.15 - Remote Buffer Overflow Exploit	windows
16-07-2019	PCMan FTP Server 2 ALLO Buffer Overflow Exploit	windows
16-07-2019	PHP Laravel Framework Token Unserialize Remote Command Execution Exploit	linux
12-07-2019	Xymon 4.3.25 - useradm Command Execution Exploit	multiple
10-07-2019	Apache mod_ssl < 2.8.7 OpenSSL - OpenFuckV2.c Remote Buffer Overflow (2) Exploit	unix

[local exploits]

::DATE	::DESCRIPTION	::TYPE
28-07-2019	Microsoft Windows 7 build 7601 (x86) - Local Privilege Escalation Exploit	windows
28-07-2019	Deepin Linux 15 - lastore-daemon Local Privilege Escalation Exploit	multiple
27-07-2019	VMware Workstation / Player < 12.5.5 - Local Privilege Escalation Exploit	multiple
26-07-2019	Linux Kernel 4.4.0-21 < 4.4.0-51 (Ubuntu 14.04/16.04 x86-64) AF_PACKET [illegible]	linux

- **Key Characteristics:**
  1. Descriptive tool names (target, operations, etc.)
  2. Clear categories of exploits (e.g., target system)
  3. Post date of when exploit was posted

Vulnerability Assessment



Cisco IOS IPS Denial of Service Vulnerability - Cisco Systems

Synopsis

The remote device is missing a vendor-supplied security patch.

Description

The Cisco IOS Intrusion Prevention System (IPS) feature contains a vulnerability in the processing of certain IPS signatures that use the SERVICE.DNS engine. This vulnerability may cause a router to crash or hang, resulting in a denial of service condition.

Risk Information

Risk Factor: High
CVSS Base Score: 7.8
CVSS Vector: CVSS2#AV:N/AC:L/Au:N/C:N/I:N/A:C

Reference Information:

CVE: CVE-2008-2739
OSVDB: 48711
BID: 31364

```
Bugtraq ID:   31364
Class:        Failure to Handle Exceptional Conditions
CVE:          CVE-2008-2739
Remote:       Yes
Local:        No
Published:    Sep 24 2008 12:00AM
Updated:      Sep 24 2008 08:19PM
Credit:       The discoverer of this issue is not known; this issue was disclosed by Cisco.
Vulnerable:   Cisco IOS 12.4YA
              Cisco IOS 12.4XZ
              Cisco IOS 12.4XY
              Cisco IOS 12.4XW
              Cisco IOS 12.4XV
              Cisco IOS 12.4XT
```

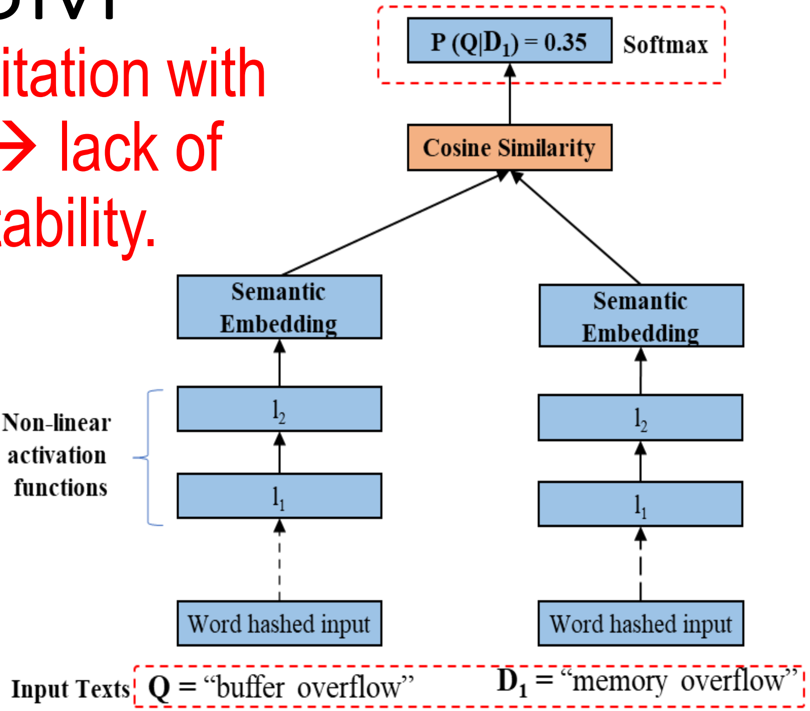
Category	Metadata	Description	Data Type
Description	Name	Short, descriptive name of vulnerability	Short text
1.	Family Name	Family vulnerability belongs to (e.g., Windows, etc.)	Categorical
	Description	Lengthy text description about vulnerability	Long text
	Synopsis	Short description of vulnerability	Short text
	Solution	Description or solution links	Short text
2.	Vulnerable Systems	List of systems susceptible to vulnerability	Short text (list)
Risk	CVSS	Value between 0.0-10.0 indicating vulnerability severity	Continuous
3.	Risk Factor	Categorical rating of risk (High, Low)	Categorical
	CVE	Vulnerability reference number	Categorical
	Publication Date	Date vulnerability was publicly published	Date

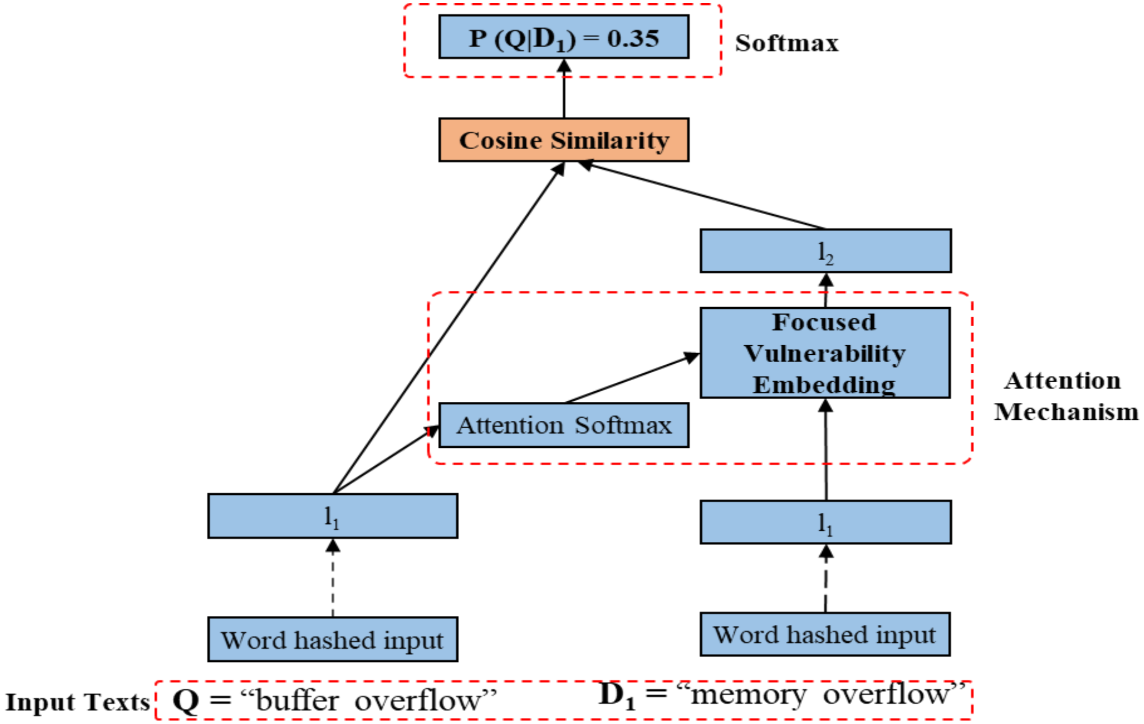
Key Attributes Returned by Modern Vulnerability Scanners

- **Key Characteristics:**
  1. Short, descriptive title of vulnerability
  2. List of systems susceptible to vulnerability
  3. Common Vulnerability Severity Score (0.0 – 10.0)

Proposed Exploit Vulnerability Attention-DSSM

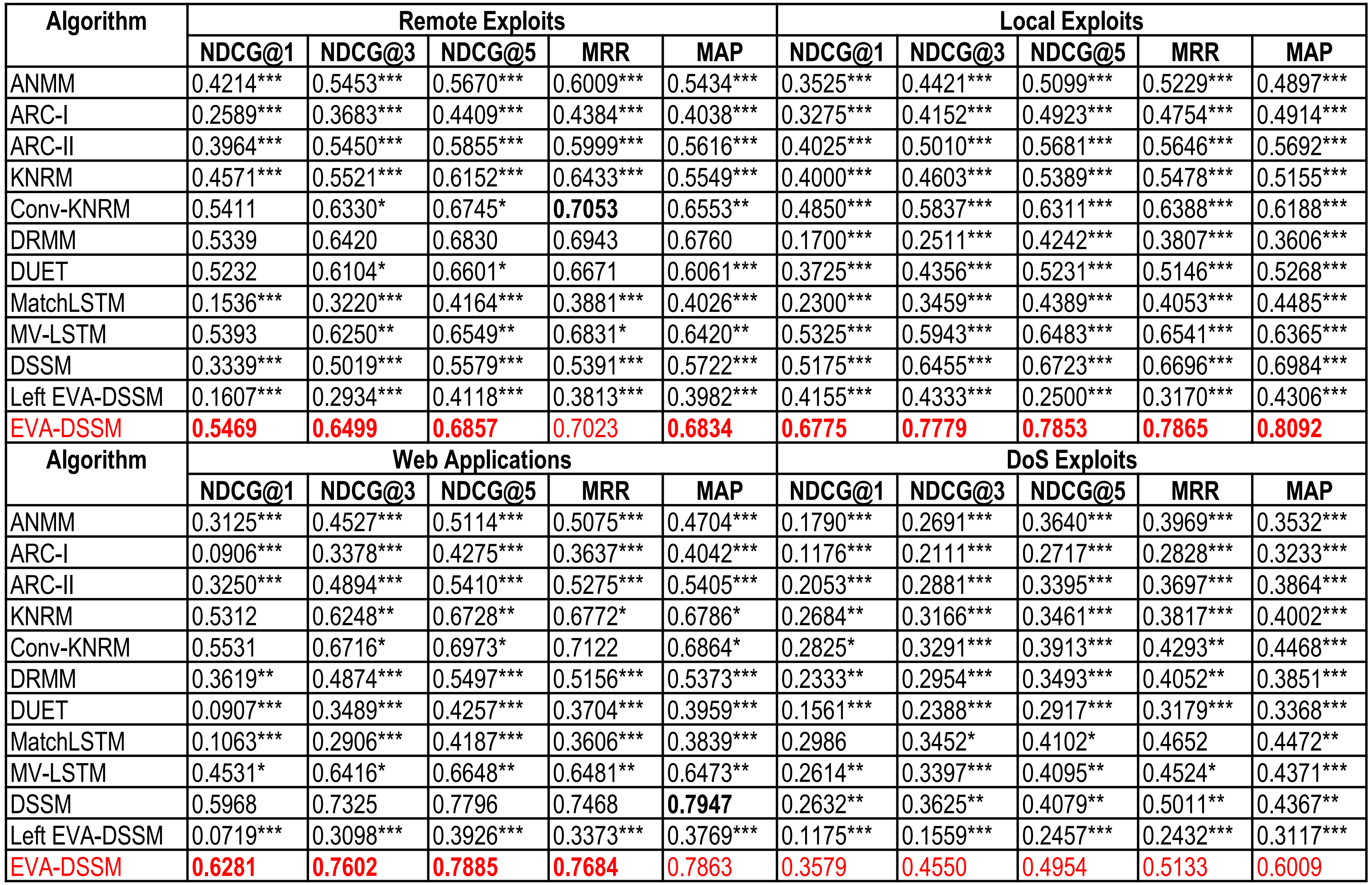
- Key Limitation with DSSM → lack of interpretability.





- **Contribution:** EVA-DSSM integrates an attention mechanism into the DSSM. Identifies and outputs key exploit features essential for creating links

Experiment Results: EVA-DSSM vs Deep Learning Matching Algorithms



Algorithm	Remote Exploits					Local Exploits				
	NDCG@1	NDCG@3	NDCG@5	MRR	MAP	NDCG@1	NDCG@3	NDCG@5	MRR	MAP
ANMM	0.4214***	0.5453***	0.5670***	0.6009***	0.5434***	0.3525***	0.4421***	0.5099***	0.5229***	0.4897***
ARC-I	0.2589***	0.3683***	0.4409***	0.4384***	0.4038***	0.3275***	0.4152***	0.4923***	0.4754***	0.4914***
ARC-II	0.3964***	0.5450***	0.5855***	0.5999***	0.5616***	0.4025***	0.5010***	0.5681***	0.5646***	0.5692***
KNRM	0.4571***	0.5521***	0.6152***	0.6433***	0.5549***	0.4000***	0.4603***	0.5389***	0.5478***	0.5155***
Conv-KNRM	0.5411	0.6330*	0.6745*	**0.7053**	0.6553**	0.4850***	0.5837***	0.6311***	0.6388***	0.6188***
DRMM	0.5339	0.6420	0.6830	0.6943	0.6760	0.1700***	0.2511***	0.4242***	0.3807***	0.3606***
DUET	0.5232	0.6104*	0.6601*	0.6671	0.6061***	0.3725***	0.4356***	0.5231***	0.5146***	0.5268***
MatchLSTM	0.1536***	0.3220***	0.4164***	0.3881***	0.4026***	0.2300***	0.3459***	0.4389***	0.4053***	0.4485***
MV-LSTM	0.5393	0.6250**	0.6549**	0.6831*	0.6420**	0.5325***	0.5943***	0.6483***	0.6541***	0.6365***
DSSM	0.3339***	0.5019***	0.5579***	0.5391***	0.5722***	0.5175***	0.6455***	0.6723***	0.6696***	0.6984***
Left EVA-DSSM	0.1607***	0.2934***	0.4118***	0.3813***	0.3982***	0.4155***	0.4333***	0.2500***	0.3170***	0.4306***
EVA-DSSM	**0.5469**	**0.6499**	**0.6857**	0.7023	**0.6834**	**0.6775**	**0.7779**	**0.7853**	**0.7865**	**0.8092**
Algorithm	Web Applications					DoS Exploits				
	NDCG@1	NDCG@3	NDCG@5	MRR	MAP	NDCG@1	NDCG@3	NDCG@5	MRR	MAP
ANMM	0.3125***	0.4527***	0.5114***	0.5075***	0.4704***	0.1790***	0.2691***	0.3640***	0.3969***	0.3532***
ARC-I	0.0906***	0.3378***	0.4275***	0.3637***	0.4042***	0.1176***	0.2111***	0.2717***	0.2828***	0.3233***
ARC-II	0.3250***	0.4894***	0.5410***	0.5275***	0.5405***	0.2053***	0.2881***	0.3395***	0.3697***	0.3864***
KNRM	0.5312	0.6248**	0.6728**	0.6772*	0.6786*	0.2684**	0.3166***	0.3461***	0.3817***	0.4002***
Conv-KNRM	0.5531	0.6716*	0.6973*	0.7122	0.6864*	0.2825*	0.3291***	0.3913***	0.4293**	0.4468***
DRMM	0.3619**	0.4874***	0.5497***	0.5156***	0.5373***	0.2333**	0.2954***	0.3493***	0.4052**	0.3851***
DUET	0.0907***	0.3489***	0.4257***	0.3704***	0.3959***	0.1561***	0.2388***	0.2917***	0.3179***	0.3368***
MatchLSTM	0.1063***	0.2906***	0.4187***	0.3606***	0.3839***	0.2986	0.3452*	0.4102*	0.4652	0.4472**
MV-LSTM	0.4531*	0.6416*	0.6648**	0.6481**	0.6473**	0.2614**	0.3397***	0.4095**	0.4524*	0.4371***
DSSM	0.5968	0.7325	0.7796	0.7468	**0.7947**	0.2632**	0.3625**	0.4079**	0.5011**	0.4367**
Left EVA-DSSM	0.0719***	0.3098***	0.3926***	0.3373***	0.3769***	0.1175***	0.1559***	0.2457***	0.2432***	0.3117***
EVA-DSSM	**0.6281**	**0.7602**	**0.7885**	**0.7684**	0.7863	0.3579	0.4550	0.4954	0.5133	0.6009

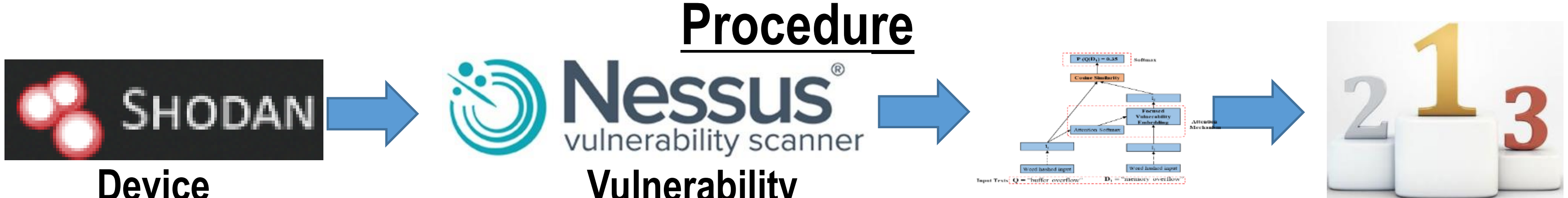
- EVA-DSSM outperforms all deep learning benchmarks
- Conv. or LSTM operations achieved lower performances
- Indicates that integrating an attention mechanism into the DSSM architecture does not deteriorate performance

Case Studies: SCADA and Hospitals

- 20,461 SCADA Devices from major vendors (e.g., Rockwell)
- **Motivation:** SCADA → control critical infrastructure

- 1,879 devices from top 8 US hospitals
- **Motivation:** Hospitals → popular target for hackers

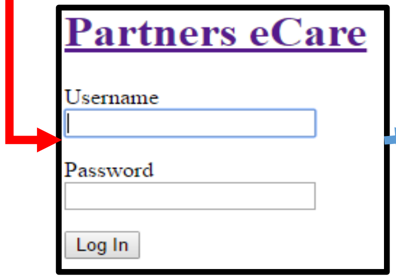
Procedure



Device Identification → Vulnerability Scanning → EVA-DSSM → DVSM

Hospital Case Study

Hospital Device Information		Device Severity Score Information for Selected Devices			
Hospital Name	**# of Vulnerable Devices/# of devices**	**Device Type**	**# of Vulnerabilities**	**Vulnerabilities**	**DVSM**
12x.x.x.x	133/808	FTP/SSH Server	3	FTP issues	4.591
19x.x.x.x	27/301	SSH Server	3	SSH issues	4.376
17x.x.x.x	31/274	eCare web portal	47	XSS, OpenSSL, buffer overflow, DoS	61.761
16x.x.x.x	59/160	Medical computing portal	5	PHP and SSH issues	4.863
14x.x.x.x	64/130	Web Server	3	SQL Injections	7.528
		Apple TV	2	Buffer overflow	5.381
14x.x.x.x	14/107	SSH/Web server	4	PHP and SSH issues	3.871
6x.x.x.x	9/52	Informational diabetes portal	3	SVN and Unix vulnerabilities	7.159
16x.x.x.x	7/47	Web Server	6	XSS, HTMLi	9.367
Total:	**344/1,879 (18.31%)**	**-**	**-**	**-**	**-**



Vulnerability Name (CVSS Score)	Exploit Name (Post Date)	Severity Score
"OpenSSL Unsupported" (10.0)	*"OpenSSL TLS Heartbeat Extension – Memory Disclosure"* (4/8/2014)	3.366
"Multiple XSS Vulnerabilities" (4.3)	*"Portal XSS Vulnerability"* (5/28/2010)	1.261
…	…	…
-	**-**	**Total: 61.761**

- Portals are a common avenue for hackers to access sensitive records (Ayala 2016).
- Analysis shows an eCare portal with a large attack surface: 47 vulnerabilities for a DVSM of 61.761.
- Network admins can prioritize this device when analyzing their weaknesses.

Grouping Vulnerable Virtual Machines in Scientific Cyberinfrastructure: A Multi-View Representation Learning Approach

Steven Ullman (MISQ, under review)

VM OSS Vulnerabilities

- Configurable VM images allow users to install Open-Source Software (OSS, i.e., applications) from third parties (e.g., GitHub) and manipulate file systems (e.g., permissions) to support their desired analytics.
- OSS can contain significant software-level vulnerabilities often missed by general-purpose scanners (e.g., Nessus) (Ullman et al., 2020).
- Misconfigured file permissions and file block organization can lead to kernel crash, and permission bypass (Cai et al., 2019).
- Scientific CIs often lack dedicated support staff to prioritize and remediate vulnerabilities (JASON 2019). Therefore, vulnerabilities can remain undetected for years (Osborne 2020).

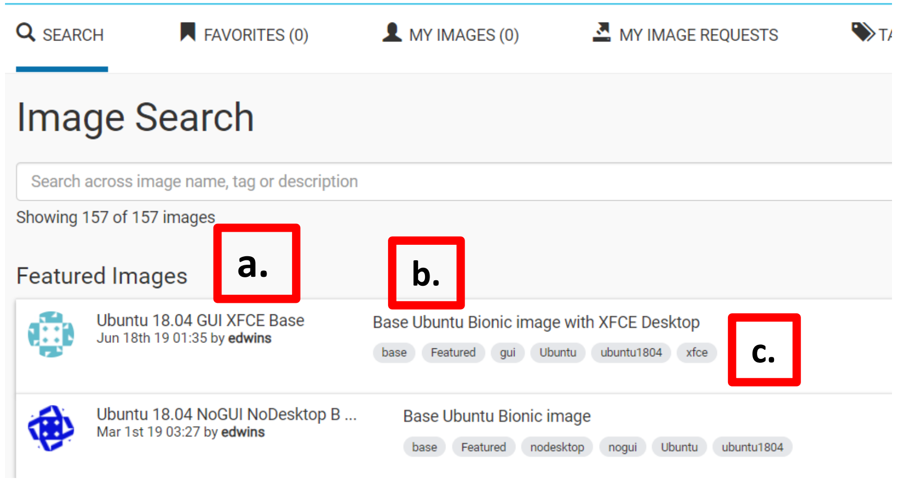


Image Search

Search across image name, tag or description

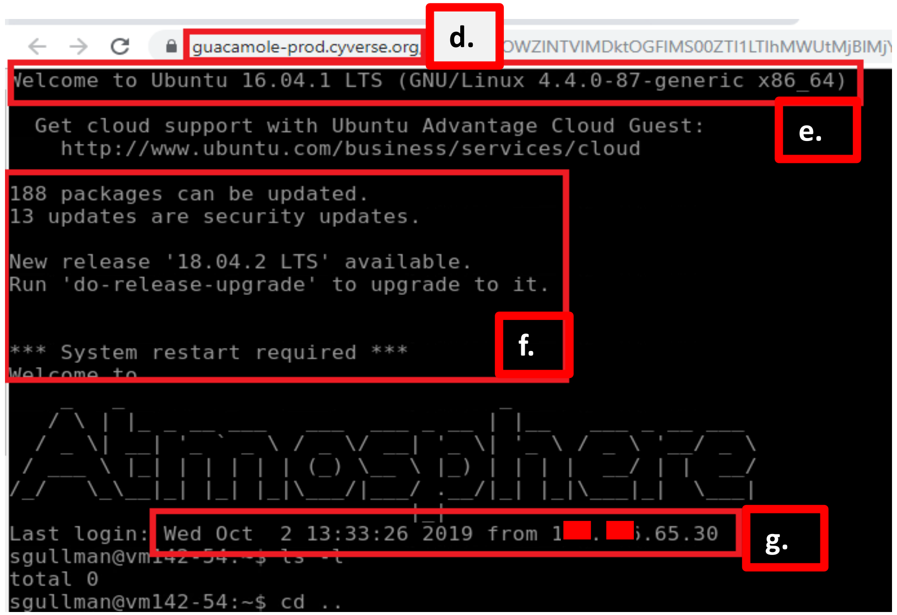
Showing 157 of 157 images

Featured Images

Ubuntu 18.04 GUI XFCE Base
Jun 18th 19 01:35 by **edwins**
Base Ubuntu Bionic image with XFCE Desktop
base · Featured · gui · Ubuntu · ubuntu1804 · xfce

Ubuntu 18.04 NoGUI NoDesktop B ...
Mar 1st 19 03:27 by **edwins**
Base Ubuntu Bionic image
base · Featured · nodesktop · nogui · Ubuntu · ubuntu1804

Figure 1. VM image details include (a) name and date of creation, (b) description, and (c) tags of included technologies

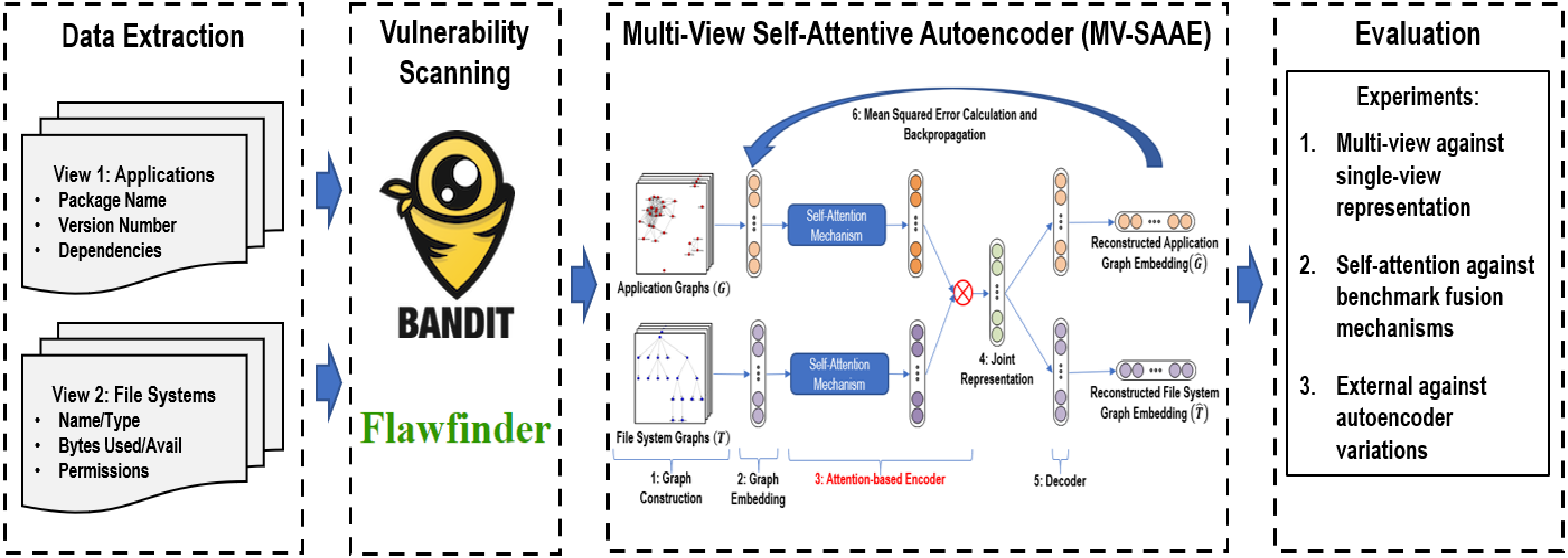


guacamole-prod.cyverse.org OWZINTVIMDktOGFIMS00ZTI1LTIhMWUtMjBIMj

```
Welcome to Ubuntu 16.04.1 LTS (GNU/Linux 4.4.0-87-generic x86_64)

  Get cloud support with Ubuntu Advantage Cloud Guest:
    http://www.ubuntu.com/business/services/cloud

188 packages can be updated.
13 updates are security updates.

New release '18.04.2 LTS' available.
Run 'do-release-upgrade' to upgrade to it.


*** System restart required ***
Welcome to

Last login: Wed Oct  2 13:33:26 2019 from 1[redacted].[redacted]5.65.30
sgullman@vm142-54:~$ ls -l
total 0
sgullman@vm142-54:~$ cd ..
```

Figure 2. VM image with (d) server name, (e) operating system and kernel version, (f) available updates, and (g) time of login with IP address

Research Design



Data Extraction

View 1: Applications
- Package Name
- Version Number
- Dependencies

View 2: File Systems
- Name/Type
- Bytes Used/Avail
- Permissions

Vulnerability Scanning

BANDIT

Flawfinder

Multi-View Self-Attentive Autoencoder (MV-SAAE)

6: Mean Squared Error Calculation and Backpropagation
Application Graphs (G)
Self-Attention Mechanism
Self-Attention Mechanism
File System Graphs (T)
4: Joint Representation
Reconstructed Application Graph Embedding($\hat{G}$)
Reconstructed File System Graph Embedding ($\hat{T}$)
1: Graph Construction
2: Graph Embedding
3: Attention-based Encoder
5: Decoder

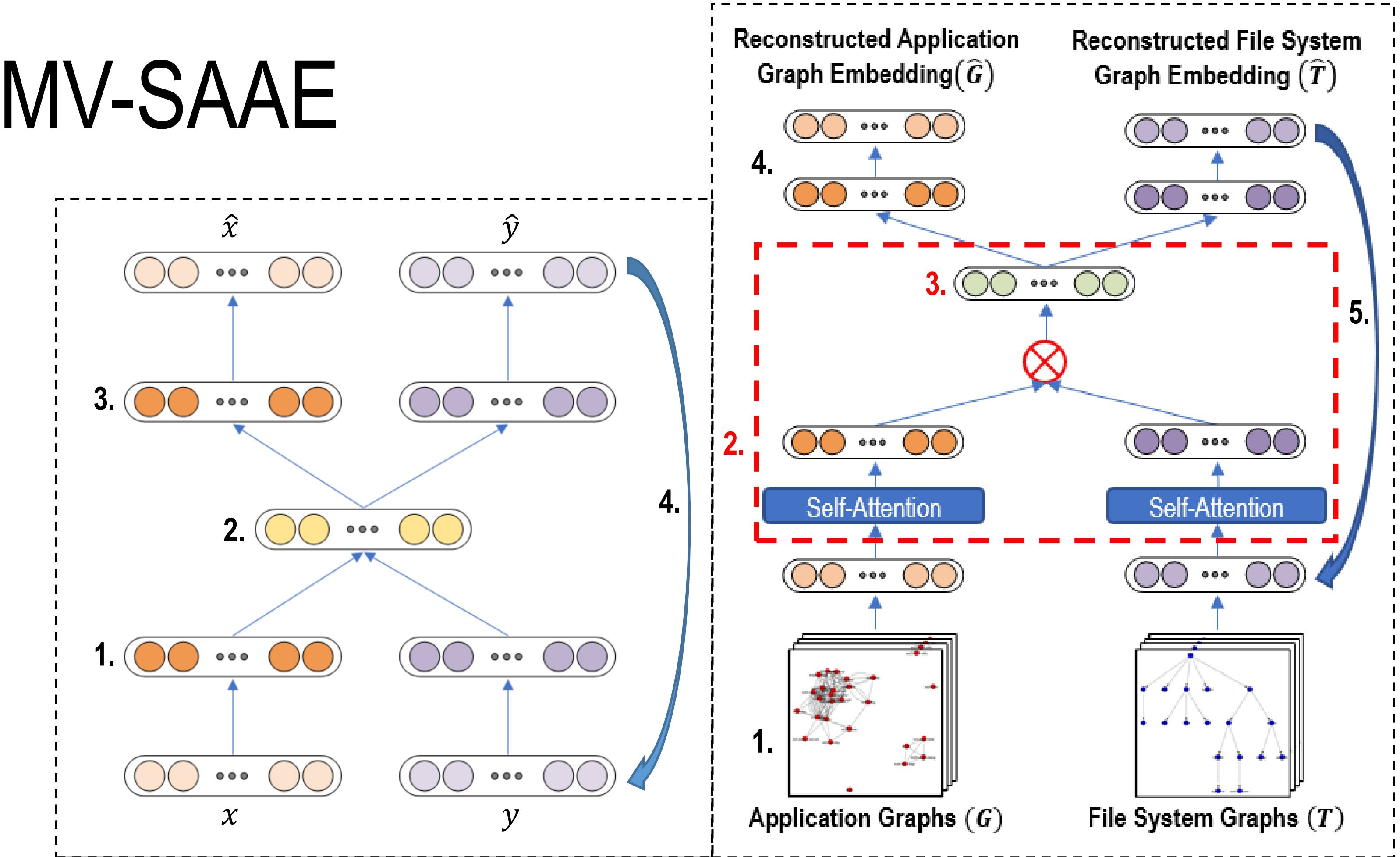
Evaluation

Experiments:

1. Multi-view against single-view representation
2. Self-attention against benchmark fusion mechanisms
3. External against autoencoder variations

Research Design: MV-SAAE

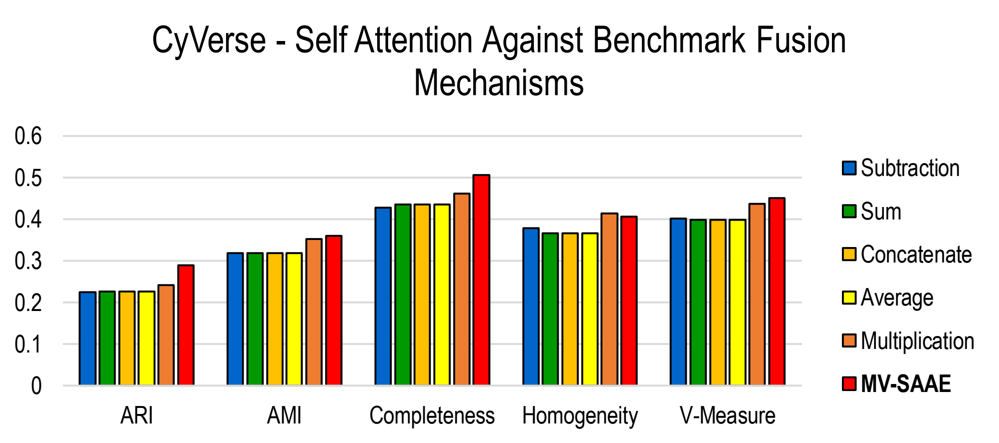
- Multi-View Self-Attentive Autoencoder (MV-SAAE) extends the MVA with an attention-based encoder and attention-based fusion operation.

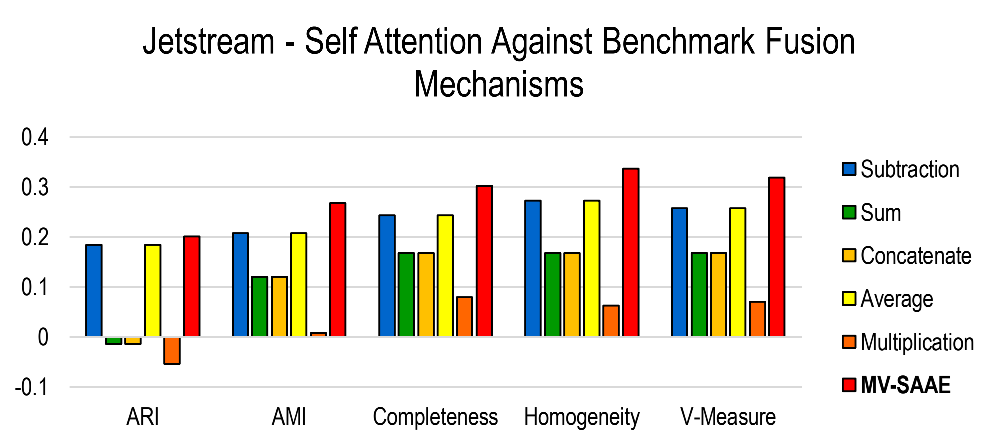


Conventional MVA Procedure:	MV-SAAE Procedure (Novelty in red):
1. Encoder	1. Graph Construction and Embedding
2. Fusion Operation	**2. Attention-Based Encoder**
3. Decoder	**3. Attention-Based Fusion Operation**
4. MSE Calculation and Backpropagation	4. Decoder
	5. MSE Calculation and Backpropagation

Figure 5. Comparison of MVA (left) vs. Proposed Multi-View Self-Attentive Autoencoder (MV-SAAE) (right)

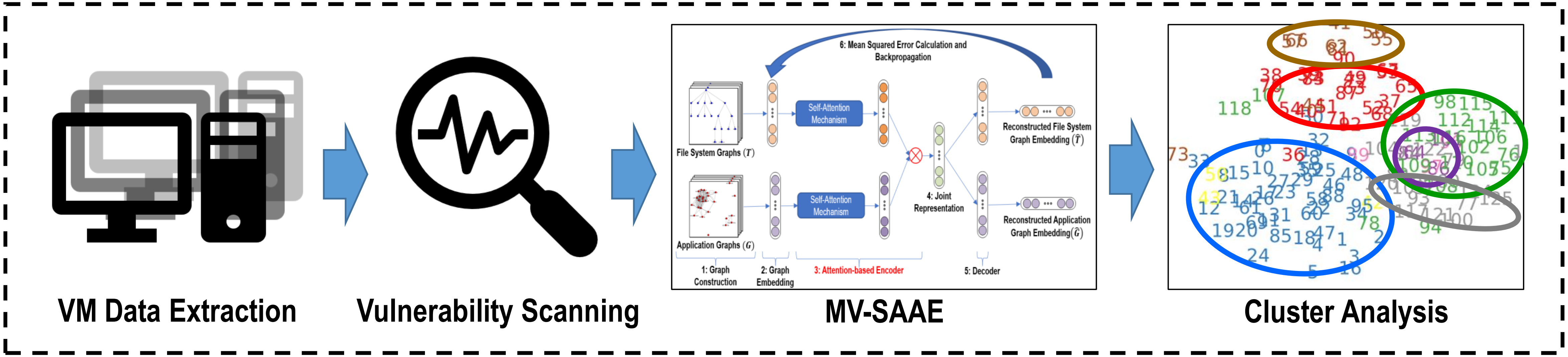
Results and Discussion


CyVerse - Self Attention Against Benchmark Fusion Mechanisms


Jetstream - Self Attention Against Benchmark Fusion Mechanisms

Dataset	**Method**	**Evaluation Metric**				
		ARI	**AMI**	**Completeness**	**Homogeneity**	**V-Measure**
CyVerse	Subtraction	0.225**	0.319**	0.428**	0.378**	0.402**
	Sum	0.226**	0.318**	0.435**	0.367**	0.398**
	Concatenate	0.226**	0.318**	0.435**	0.367**	0.398**
	Average	0.226**	0.318**	0.435**	0.367**	0.398**
	Multiplication	0.241**	0.353**	0.462**	**0.414**	0.436**
	MV-SAAE	**0.289**	**0.361**	**0.506**	0.406	**0.45**
Jetstream	Subtraction	0.185*	0.207**	0.244**	0.273**	0.258**
	Sum	-0.014**	0.12**	0.168**	0.168**	0.168**
	Concatenate	-0.014**	0.12**	0.168**	0.168**	0.168**
	Average	0.185*	0.207**	0.244**	0.273**	0.258**
	Multiplication	-0.054**	0.008**	0.08**	0.063**	0.071**
	MV-SAAE	**0.201**	**0.267**	**0.302**	**0.337**	**0.318**

Case Study: Clustering Similar Images & Vulnerabilities



6: Mean Squared Error Calculation and Backpropagation
File System Graphs (T)
Application Graphs (G)
Self-Attention Mechanism
Self-Attention Mechanism
4: Joint Representation
Reconstructed File System Graph Embedding ($\hat{T}$)
Reconstructed Application Graph Embedding ($\hat{G}$)
1: Graph Construction
2: Graph Embedding
3: Attention-based Encoder
5: Decoder

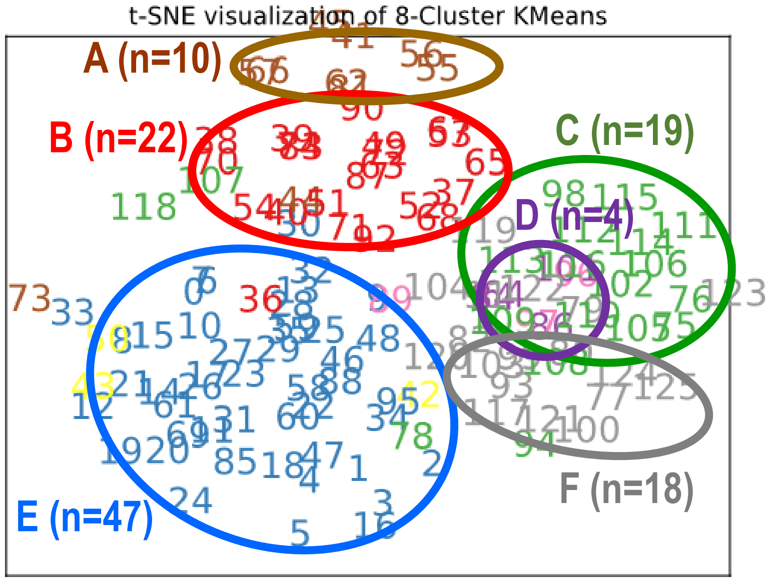
VM Data Extraction

Vulnerability Scanning

MV-SAAE

Cluster Analysis

Case Study – CyVerse Image Clusters & Vulnerabilities



Cluster	Vulnerability	Severity	Application	Count
Cluster C	Insecure Function	High	Libblockdev2	41
			Yadm	29
			Youker-assistant	28
			Pymca	27
	Insecure Input	High	Zabbix-cli	103
			Cupp	29
			Elastalert	13
	XSS Vulnerability	High	Libqt5webkit5	4
			Python3-spyder	3
			Python3-azure	2
Cluster E	Insecure Input	Low	node-gyp	34
			bup	31
			python-google-compute-engine	31
	Insecure Module	Low	bup	20
			python-google-compute-engine	17
			python-sympy	15
	Insecure Function	Medium	python-sympy	60
			python-html5lib	4
			cura-engine	4

Table 14. Selected Vulnerable Applications Within Clusters

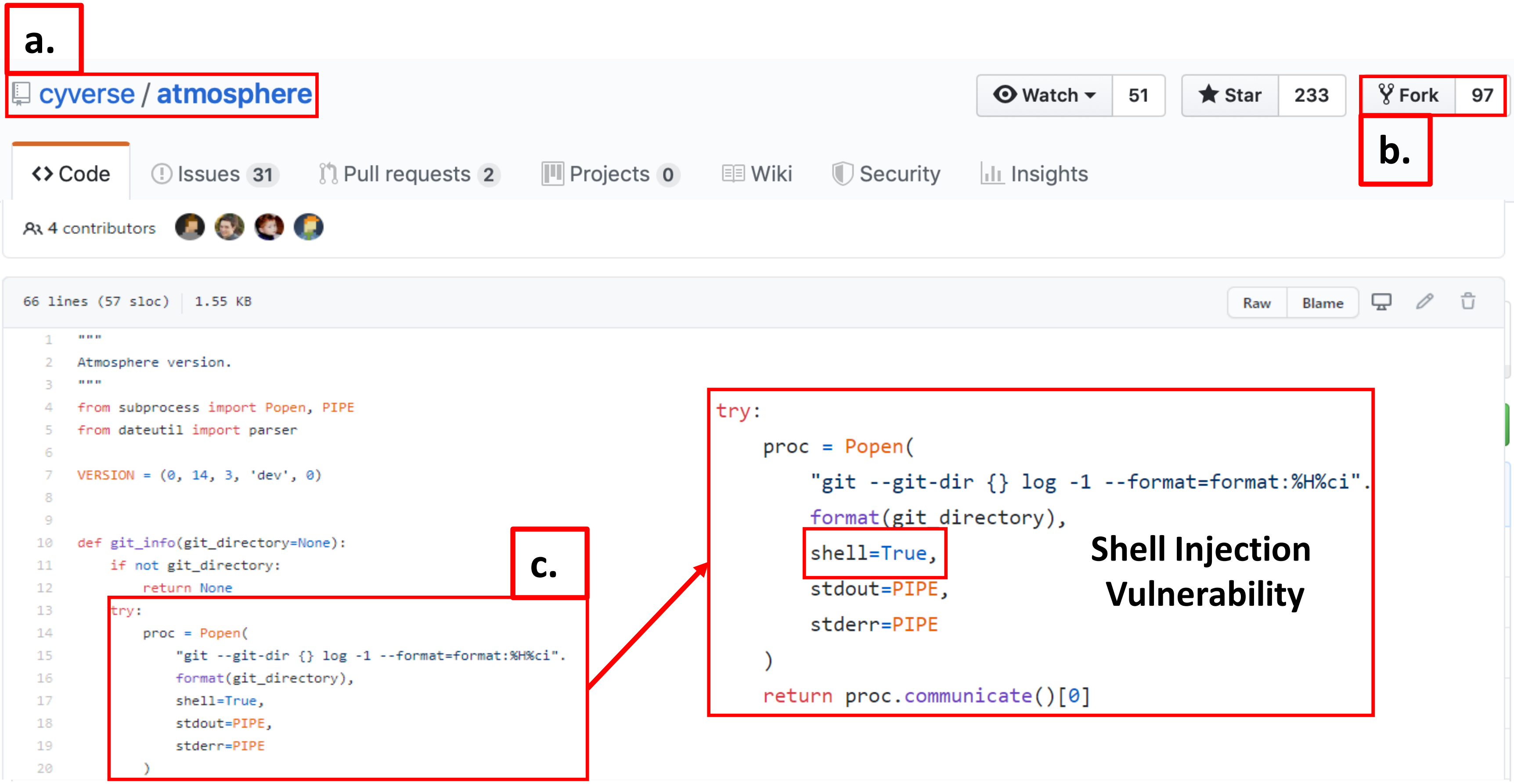
Detecting and Grouping Vulnerable GitHub Repositories in Scientific Cyberinfrastructure: An Unsupervised Graph Embedding Approach

Ben Lazarine (in preparation)

Scientific CI GitHub Vulnerabilities

- Illustrated in Figure 1 is an insecure source code snippet in a major scientific CI's GitHub repository that is susceptible to shell injection attacks.

- Insecure coding practices can lead to the spread of vulnerabilities (i.e., shell injection) that can disrupt scientific CI.

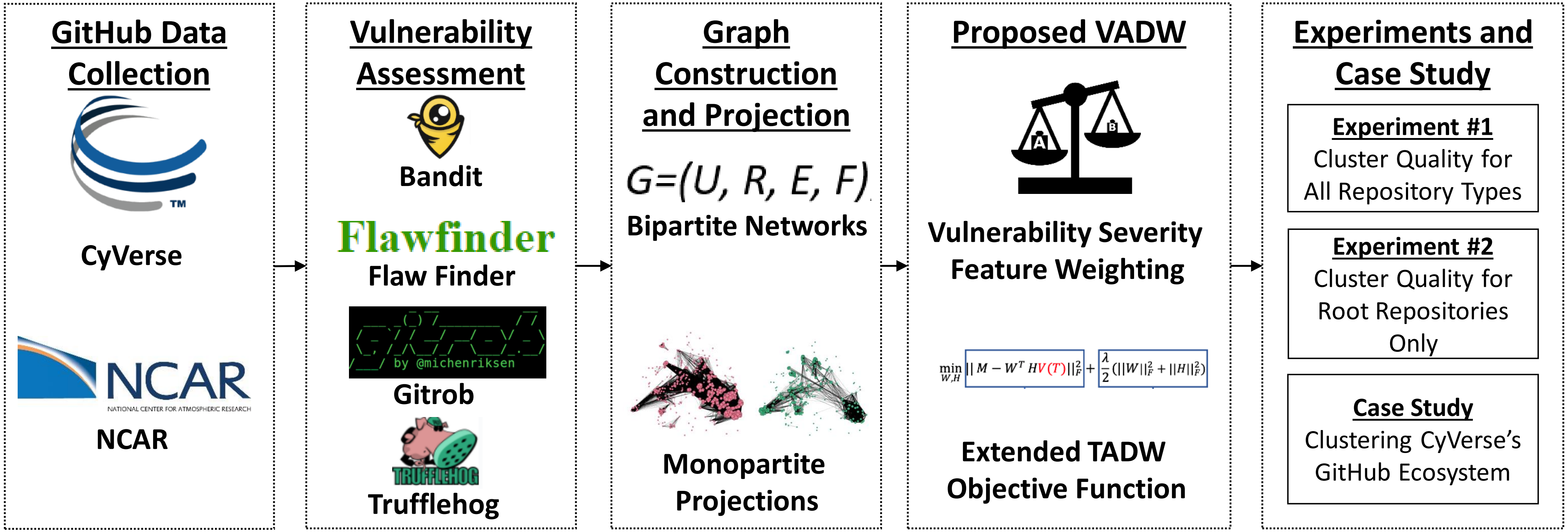


```
"""
Atmosphere version.
"""
from subprocess import Popen, PIPE
from dateutil import parser

VERSION = (0, 14, 3, 'dev', 0)


def git_info(git_directory=None):
    if not git_directory:
        return None
    try:
        proc = Popen(
            "git --git-dir {} log -1 --format=format:%H%ci".
            format(git_directory),
            shell=True,
            stdout=PIPE,
            stderr=PIPE
        )
```

```
try:
    proc = Popen(
        "git --git-dir {} log -1 --format=format:%H%ci".
        format(git directory),
        shell=True,
        stdout=PIPE,
        stderr=PIPE
    )
    return proc.communicate()[0]
```

Figure 1. GitHub Repository pages include (a) the name of the owner and repository, (b) the number of times the repository has been forked (copied by other users), and (c) the source code within the repository

Research Design



GitHub Data Collection

CyVerse

NCAR

Vulnerability Assessment

Bandit

Flawfinder

Flaw Finder

Gitrob

Trufflehog

Graph Construction and Projection

$G=(U, R, E, F)$

Bipartite Networks

Monopartite Projections

Proposed VADW

Vulnerability Severity Feature Weighting

$$\min_{W,H} \boxed{||M - W^T HV(T)||_F^2} + \boxed{\frac{\lambda}{2}(||W||_F^2 + ||H||_F^2)}$$

Extended TADW Objective Function

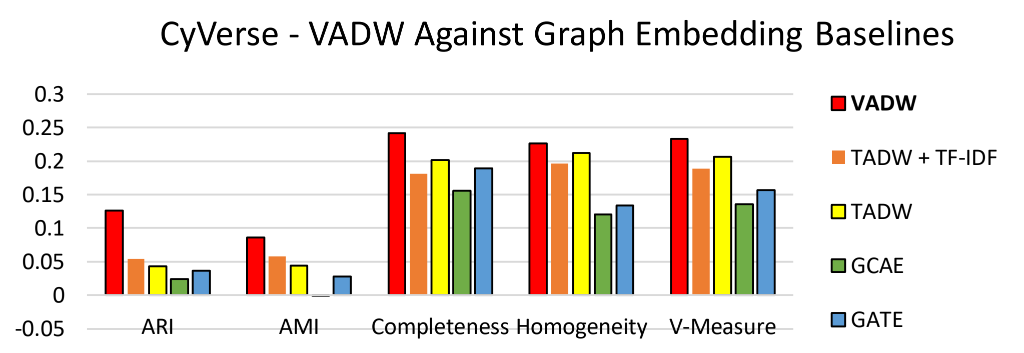
Experiments and Case Study

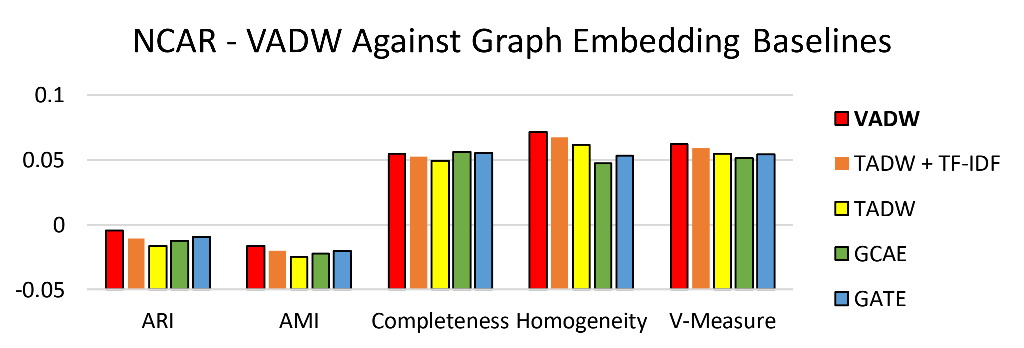
Experiment #1
Cluster Quality for All Repository Types

Experiment #2
Cluster Quality for Root Repositories Only

Case Study
Clustering CyVerse's GitHub Ecosystem

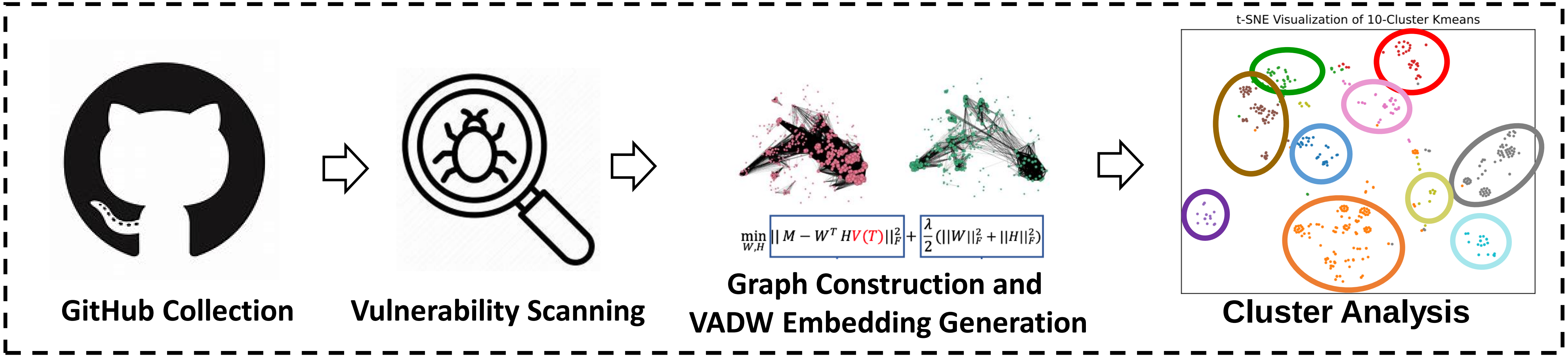
Results and Discussion


CyVerse - VADW Against Graph Embedding Baselines


NCAR - VADW Against Graph Embedding Baselines

		Evaluation Metric				
Dataset	**Method**	**ARI**	**AMI**	**Completeness**	**Homogeneity**	**V-Measure**
CyVerse	**VADW**	**0.126**	**0.086**	**0.242**	**0.227**	**0.233**
	TADW + TF-IDF	0.054***	0.058	0.181***	0.196***	0.188***
	TADW	0.044***	0.044***	0.201***	0.212**	0.206***
	GCAE	0.024***	-0.001***	0.155***	0.121***	0.136***
	GATE	0.036***	0.028***	0.190***	0.134***	0.157***
NCAR	**VADW**	**-0.004**	**-0.016**	0.055	**0.072**	**0.062**
	TADW + TF-IDF	-0.011	-0.019	0.053	0.067	0.059
	TADW	-0.016**	-0.025**	0.049	0.062**	0.055*
	GCAE	-0.013**	-0.022*	**0.056**	0.047***	0.051***
	GATE	-0.010	-0.020	0.055	0.053***	0.054**

Case Study: Clustering Similar CI/GitHub Repositories



$$\min_{W,H} \boxed{||M - W^T HV(T)||_F^2} + \boxed{\frac{\lambda}{2}(||W||_F^2 + ||H||_F^2)}$$

GitHub Collection

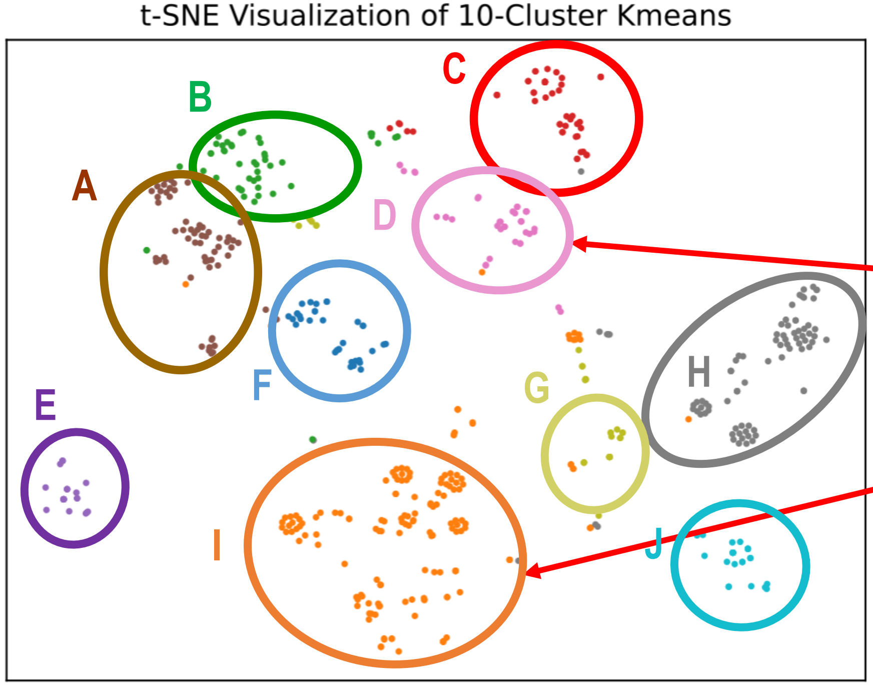
Vulnerability Scanning

Graph Construction and VADW Embedding Generation

Cluster Analysis

Figure 7. VADW Vulnerability Grouping Procedure

Case Study: Clustering Similar Repositories


t-SNE Visualization of 10-Cluster Kmeans

Cluster	Vulnerability	Severity	Repository	Count
Cluster D (n = 40)	Password	Low **1.**	**Cyverse-archive/DE**	**4,096**
			Angrygoat/DE	255
			Johnworth/DE	43
	Secret	High	**Cyverse-archive/DE**	**1,717**
			Angrygoat/DE	89
	Insecure Function	High	**CyVerse-learning-materials/container_camp_workshop_2019**	**685**
			julianpistorius/container_camp_workshop_2019	685
			mcutshall/atmosphere	139
Cluster I (n = 163)	Insecure Input	High **2.**	**Cyverse/irods-legacy**	**7,744**
			niravmerchant/Visual_Interactive_Computing_Environment	39
			CyVerse-learning-materials/Visual_Interactive_Computing_Environment	39
	Insecure Function	High	**niravmerchant/Visual_Interactive_Computing_Environment**	**685**
			CyVerse-learning-materials/Visual_Interactive_Computing_Environment	685
			Cyverse/irods-legacy	556
	File Permission	High	**Cyverse/irods-legacy**	**271**
			cyverse/ansible	16
			steve-gregory/ansible	16

1. DE is a repository that contains code for CyVerse's Discovery Environment life science research web portal that provides access to the data store and compute resources of the CI. Contains secret and password vulnerabilities and has been forked 10 times, indicating the vulnerabilities have propagated.
2. The irods-legacy repository contains data management software. Contains 167 high severity insecure input, insecure functions, and file permissions C/C++ vulnerabilities.

Some Advice for Junior Faculty and Ph.D. Students: Journals and Grants

MIS Quarterly

NSF

Major Journals: i-School, c-School, b-School

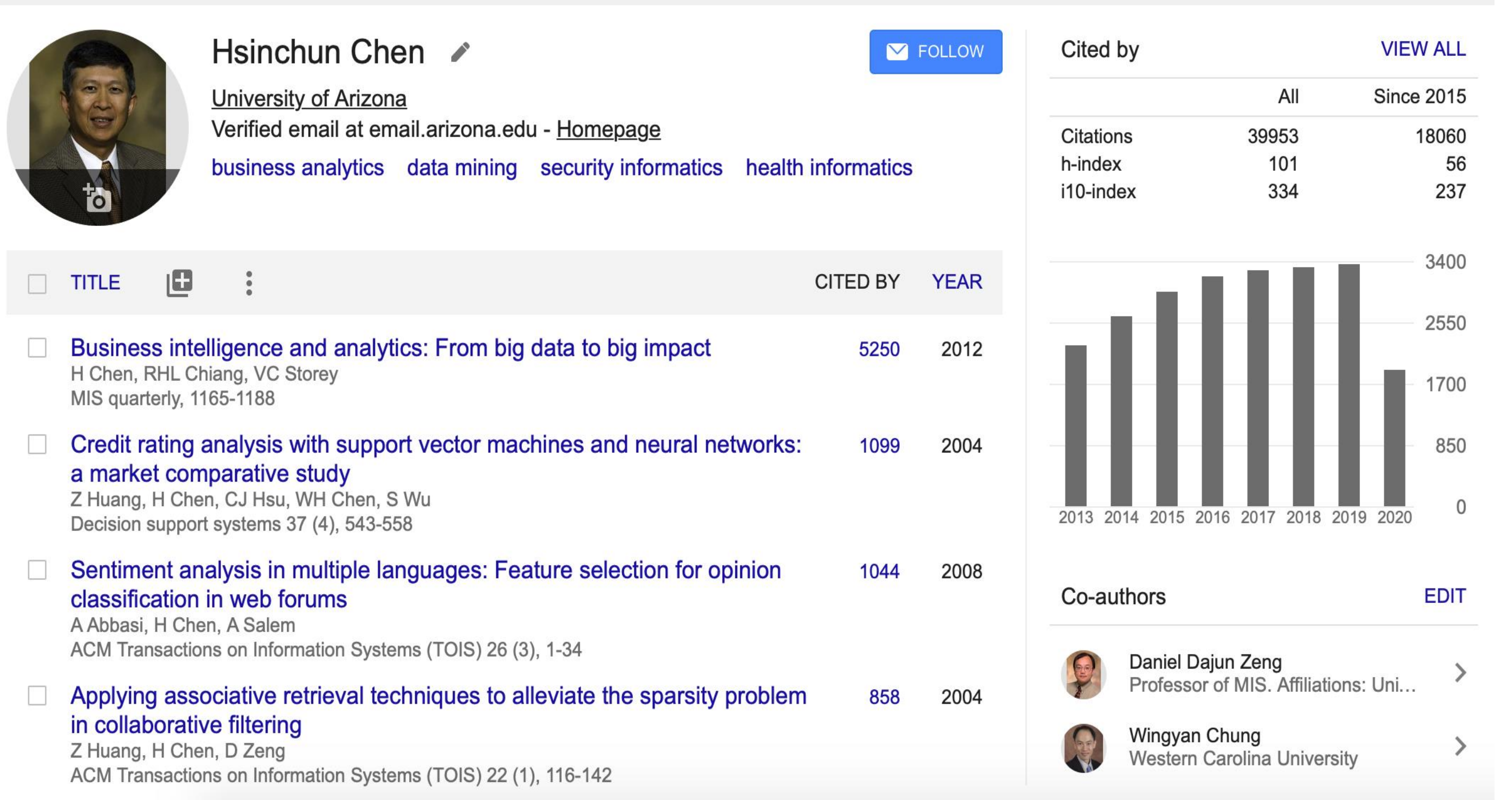
- i-School ($80K) & health informatics Journals: JASIST, ACM TOIS; JAMIA, JBI ➔ “informatics” (text) focused, system driven; helpful for NSF & NIH/NLM funding

- c-School ($100K) Journals: ACM TOIS, IEEE TKDE, CACM, IEEE IS, IEEE Computer, IEEE SMC ➔ algorithm/computing focused, data driven; helped significantly with NSF funding (same for major CS conferences)

- b-School ($180K) Journals: MISQ, ISR, JMIS, MS, ACM TMIS, DSS ➔ “design science” focused, managerial framework/principle/knowledge base; helped get jobs in major b-schools (little federal funding)

Major Journals: Chen, i-, c-, b-school, CISE

- Work hard; be persistent; colleagues & students help a lot; a little bit of luck helps



Hsinchun Chen

FOLLOW

University of Arizona

Verified email at email.arizona.edu - Homepage

business analytics data mining security informatics health informatics

□ TITLE	CITED BY	YEAR
□ Business intelligence and analytics: From big data to big impact H Chen, RHL Chiang, VC Storey MIS quarterly, 1165-1188	5250	2012
□ Credit rating analysis with support vector machines and neural networks: a market comparative study Z Huang, H Chen, CJ Hsu, WH Chen, S Wu Decision support systems 37 (4), 543-558	1099	2004
□ Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums A Abbasi, H Chen, A Salem ACM Transactions on Information Systems (TOIS) 26 (3), 1-34	1044	2008
□ Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering Z Huang, H Chen, D Zeng ACM Transactions on Information Systems (TOIS) 22 (1), 116-142	858	2004

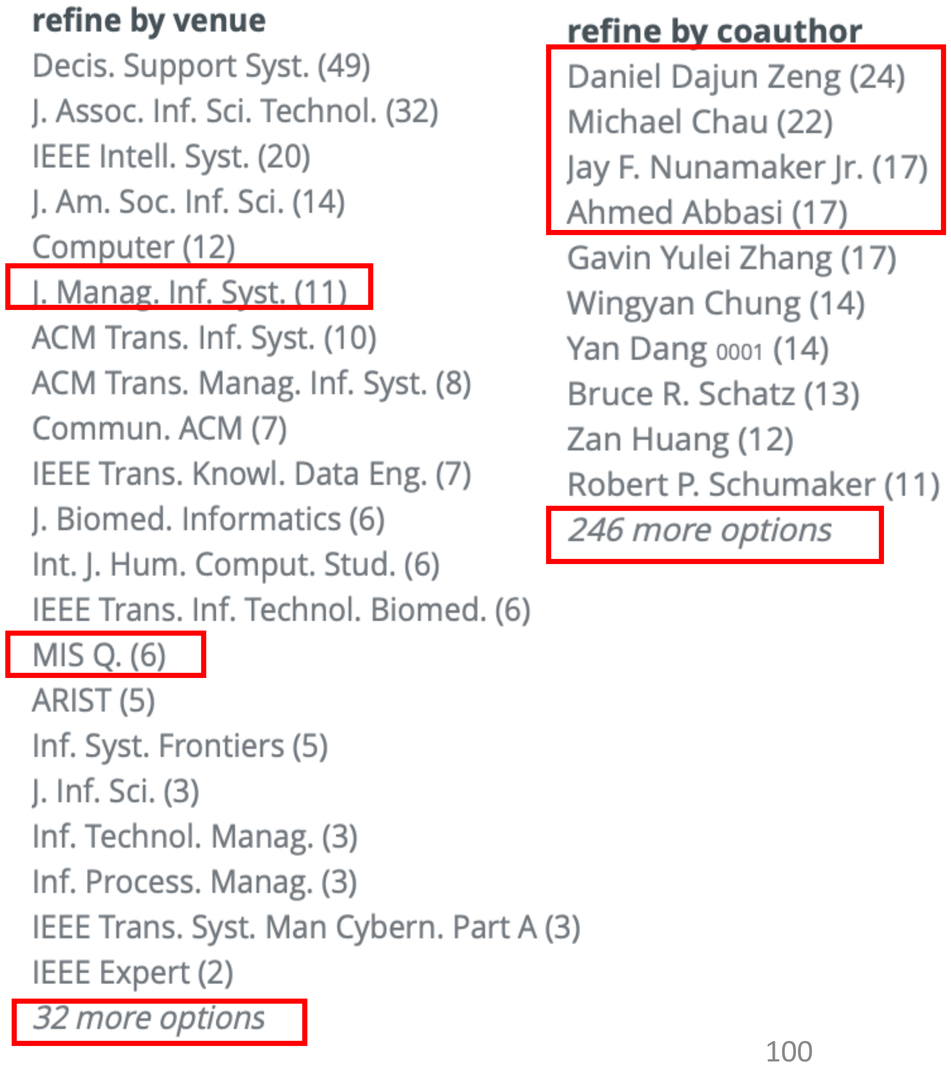
Cited by VIEW ALL

	All	Since 2015
Citations	39953	18060
h-index	101	56
i10-index	334	237

Co-authors EDIT

Daniel Dajun Zeng
Professor of MIS. Affiliations: Uni...

Wingyan Chung
Western Carolina University



refine by venue

Decis. Support Syst. (49)
J. Assoc. Inf. Sci. Technol. (32)
IEEE Intell. Syst. (20)
J. Am. Soc. Inf. Sci. (14)
Computer (12)
J. Manag. Inf. Syst. (11)
ACM Trans. Inf. Syst. (10)
ACM Trans. Manag. Inf. Syst. (8)
Commun. ACM (7)
IEEE Trans. Knowl. Data Eng. (7)
J. Biomed. Informatics (6)
Int. J. Hum. Comput. Stud. (6)
IEEE Trans. Inf. Technol. Biomed. (6)
MIS Q. (6)
ARIST (5)
Inf. Syst. Frontiers (5)
J. Inf. Sci. (3)
Inf. Technol. Manag. (3)
Inf. Process. Manag. (3)
IEEE Trans. Syst. Man Cybern. Part A (3)
IEEE Expert (2)
32 more options

refine by coauthor

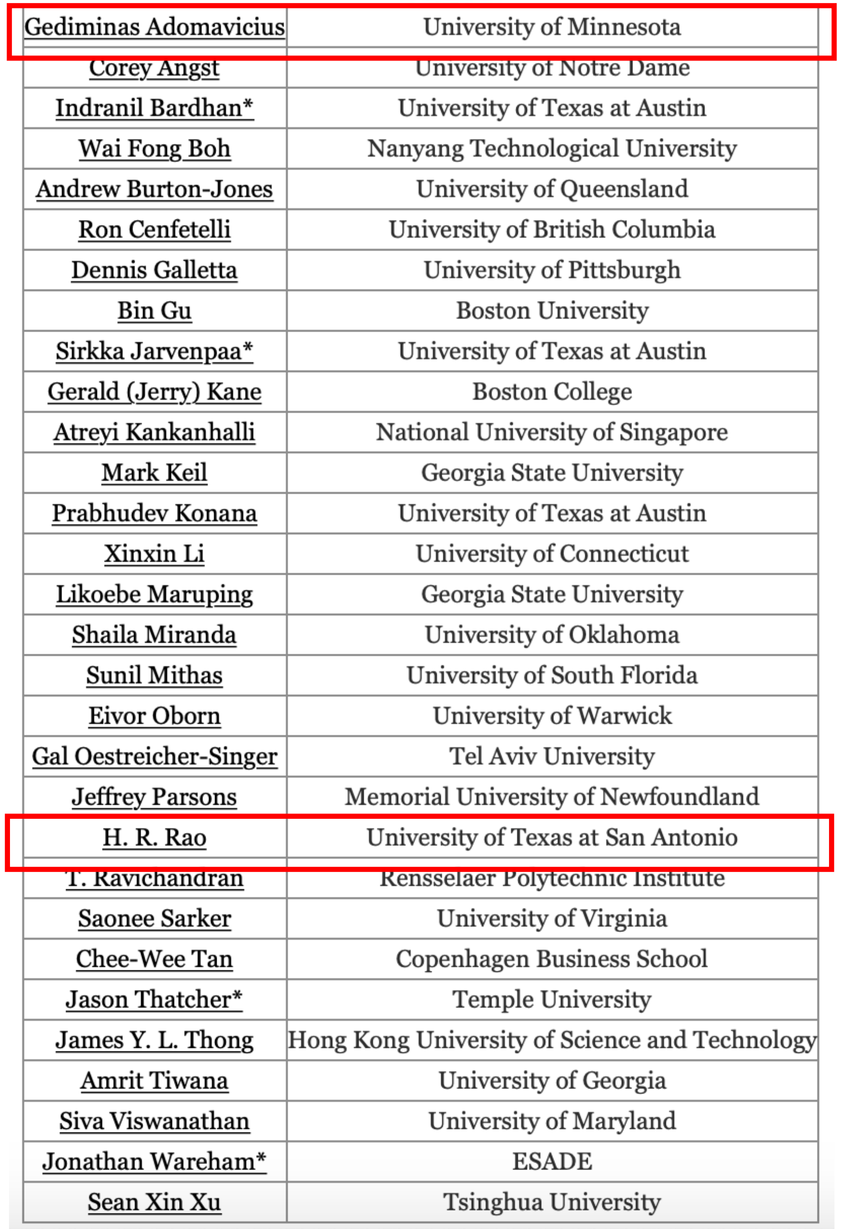
Daniel Dajun Zeng (24)
Michael Chau (22)
Jay F. Nunamaker Jr. (17)
Ahmed Abbasi (17)
Gavin Yulei Zhang (17)
Wingyan Chung (14)
Yan Dang 0001 (14)
Bruce R. Schatz (13)
Zan Huang (12)
Robert P. Schumaker (11)
246 more options

Major Journals: MISQ & JMIS

- MISQ: A+ journal, #1 in MIS
  - behavior/management focused traditionally (most SEs)
  - recent focus in business analytics & data sciences (SEs: HRR, GA, IB, PK, JP) ➔ selecting the right SEs/AEs
  - Computational design science: application-inspired novelty (algorithm, representation, framework, HCI) + societal impact ➔ significant content & mature writing (40+ pages)
  - MIS-specific lit review + methodology/framework/design "theory" + contribution to KB + principles (research abstraction) ➔ right packaging
- JMIS: A journal, #3 in MIS
  - Same as above; more system driven
  - Zwass + Nunamaker; HICSS special issue

MISQ SEs

MIS Quarterly



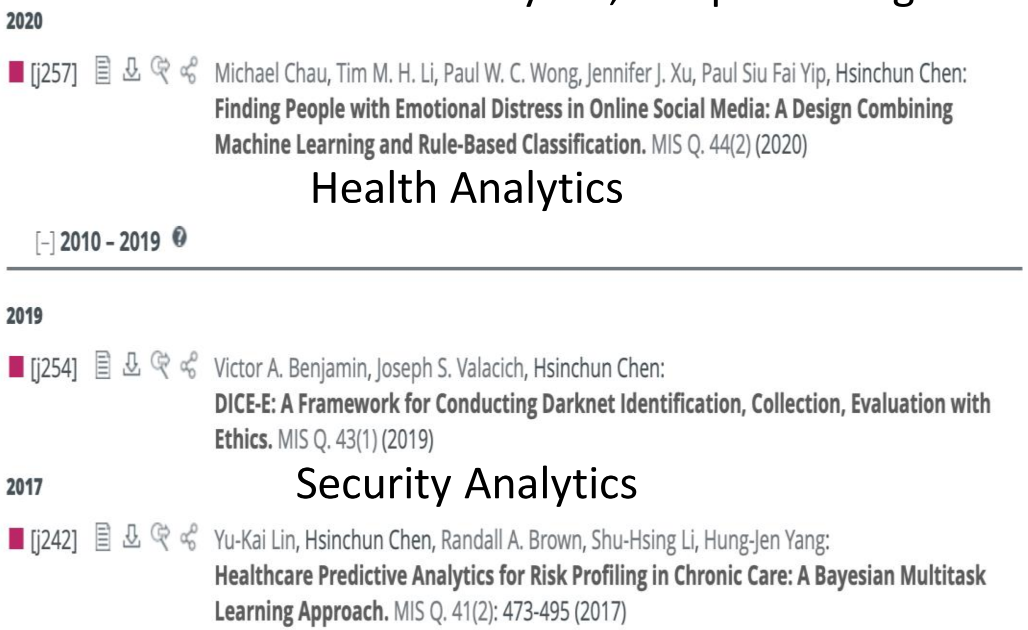
Gediminas Adomavicius	University of Minnesota
Corey Angst	University of Notre Dame
Indranil Bardhan*	University of Texas at Austin
Wai Fong Boh	Nanyang Technological University
Andrew Burton-Jones	University of Queensland
Ron Cenfetelli	University of British Columbia
Dennis Galletta	University of Pittsburgh
Bin Gu	Boston University
Sirkka Jarvenpaa*	University of Texas at Austin
Gerald (Jerry) Kane	Boston College
Atreyi Kankanhalli	National University of Singapore
Mark Keil	Georgia State University
Prabhudev Konana	University of Texas at Austin
Xinxin Li	University of Connecticut
Likoebe Maruping	Georgia State University
Shaila Miranda	University of Oklahoma
Sunil Mithas	University of South Florida
Eivor Oborn	University of Warwick
Gal Oestreicher-Singer	Tel Aviv University
Jeffrey Parsons	Memorial University of Newfoundland
H. R. Rao	University of Texas at San Antonio
T. Ravichandran	Rensselaer Polytechnic Institute
Saonee Sarker	University of Virginia
Chee-Wee Tan	Copenhagen Business School
Jason Thatcher*	Temple University
James Y. L. Thong	Hong Kong University of Science and Technology
Amrit Tiwana	University of Georgia
Siva Viswanathan	University of Maryland
Jonathan Wareham*	ESADE
Sean Xin Xu	Tsinghua University

Major Journals: Chen, AI Lab Computational Design Science (CDS) Papers in MISQ, 2008+

A Deep Learning Approach for Recognizing Activity of Daily Living (ADL) for Senior Care: Exploiting Interaction Dependency and Temporal Patternsn
Hongyi Zhu, Sagar Samtani, Randall A. Brown, and Hsinchun Chen

Forthcoming, 2020

Health Analytics; Deep Learning



2020

■ [j257] Michael Chau, Tim M. H. Li, Paul W. C. Wong, Jennifer J. Xu, Paul Siu Fai Yip, Hsinchun Chen:
Finding People with Emotional Distress in Online Social Media: A Design Combining Machine Learning and Rule-Based Classification. MIS Q. 44(2) (2020)

Health Analytics

[–] **2010 - 2019**

2019

■ [j254] Victor A. Benjamin, Joseph S. Valacich, Hsinchun Chen:
DICE-E: A Framework for Conducting Darknet Identification, Collection, Evaluation with Ethics. MIS Q. 43(1) (2019)

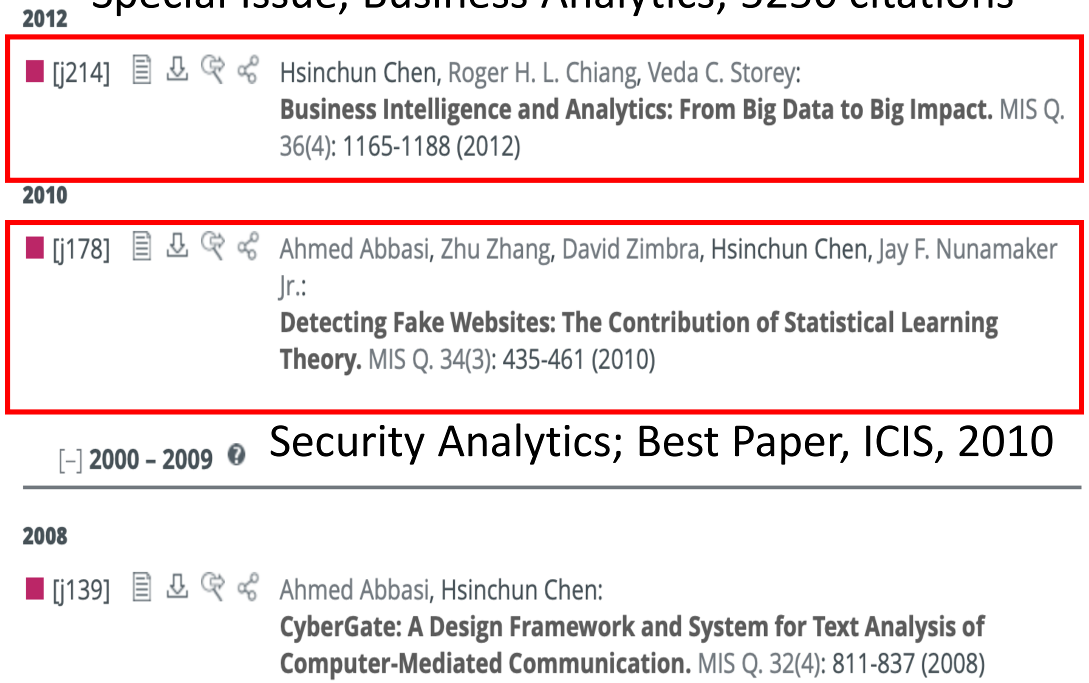
Security Analytics

2017

■ [j242] Yu-Kai Lin, Hsinchun Chen, Randall A. Brown, Shu-Hsing Li, Hung-Jen Yang:
Healthcare Predictive Analytics for Risk Profiling in Chronic Care: A Bayesian Multitask Learning Approach. MIS Q. 41(2): 473-495 (2017)

Health Analytics

Special Issue, Business Analytics; 5250 citations



2012

■ [j214] Hsinchun Chen, Roger H. L. Chiang, Veda C. Storey:
Business Intelligence and Analytics: From Big Data to Big Impact. MIS Q. 36(4): 1165-1188 (2012)

2010

■ [j178] Ahmed Abbasi, Zhu Zhang, David Zimbra, Hsinchun Chen, Jay F. Nunamaker Jr.:
Detecting Fake Websites: The Contribution of Statistical Learning Theory. MIS Q. 34(3): 435-461 (2010)

Security Analytics; Best Paper, ICIS, 2010

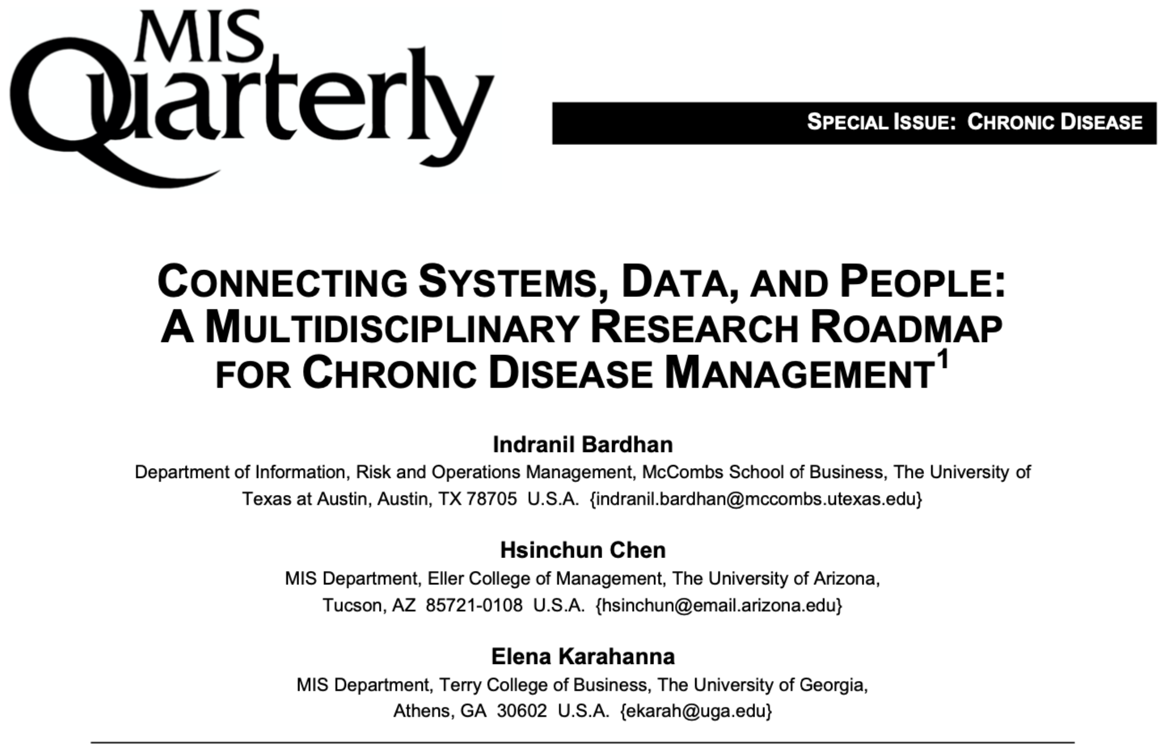
[–] **2000 - 2009**

2008

■ [j139] Ahmed Abbasi, Hsinchun Chen:
CyberGate: A Design Framework and System for Text Analysis of Computer-Mediated Communication. MIS Q. 32(4): 811-837 (2008)

Social Media Analytics

Major Journals: Health IT & Analytics Special Issue, March 2020



MIS Quarterly

SPECIAL ISSUE: CHRONIC DISEASE

CONNECTING SYSTEMS, DATA, AND PEOPLE: A MULTIDISCIPLINARY RESEARCH ROADMAP FOR CHRONIC DISEASE MANAGEMENT[1]

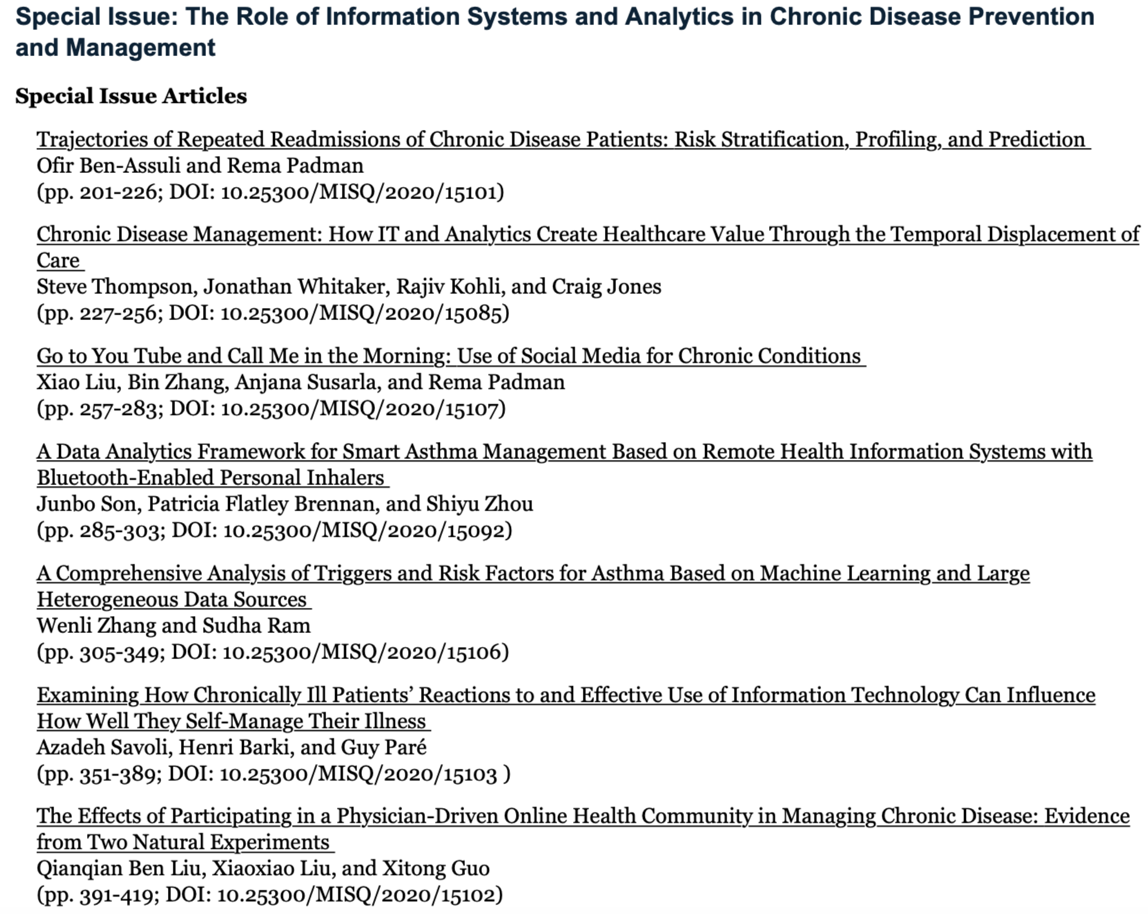
Indranil Bardhan
Department of Information, Risk and Operations Management, McCombs School of Business, The University of Texas at Austin, Austin, TX 78705 U.S.A. {indranil.bardhan@mccombs.utexas.edu}

Hsinchun Chen
MIS Department, Eller College of Management, The University of Arizona,
Tucson, AZ 85721-0108 U.S.A. {hsinchun@email.arizona.edu}

Elena Karahanna
MIS Department, Terry College of Business, The University of Georgia,
Athens, GA 30602 U.S.A. {ekarah@uga.edu}

DOI: 10.25300/MISQ/2020/14644

MIS Quarterly Vol. 44 No. 1, pp. 185-200March 2020



Special Issue: The Role of Information Systems and Analytics in Chronic Disease Prevention and Management

Special Issue Articles

Trajectories of Repeated Readmissions of Chronic Disease Patients: Risk Stratification, Profiling, and Prediction
Ofir Ben-Assuli and Rema Padman
(pp. 201-226; DOI: 10.25300/MISQ/2020/15101)

Chronic Disease Management: How IT and Analytics Create Healthcare Value Through the Temporal Displacement of Care
Steve Thompson, Jonathan Whitaker, Rajiv Kohli, and Craig Jones
(pp. 227-256; DOI: 10.25300/MISQ/2020/15085)

Go to You Tube and Call Me in the Morning: Use of Social Media for Chronic Conditions
Xiao Liu, Bin Zhang, Anjana Susarla, and Rema Padman
(pp. 257-283; DOI: 10.25300/MISQ/2020/15107)

A Data Analytics Framework for Smart Asthma Management Based on Remote Health Information Systems with Bluetooth-Enabled Personal Inhalers
Junbo Son, Patricia Flatley Brennan, and Shiyu Zhou
(pp. 285-303; DOI: 10.25300/MISQ/2020/15092)

A Comprehensive Analysis of Triggers and Risk Factors for Asthma Based on Machine Learning and Large Heterogeneous Data Sources
Wenli Zhang and Sudha Ram
(pp. 305-349; DOI: 10.25300/MISQ/2020/15106)

Examining How Chronically Ill Patients' Reactions to and Effective Use of Information Technology Can Influence How Well They Self-Manage Their Illness
Azadeh Savoli, Henri Barki, and Guy Paré
(pp. 351-389; DOI: 10.25300/MISQ/2020/15103)

The Effects of Participating in a Physician-Driven Online Health Community in Managing Chronic Disease: Evidence from Two Natural Experiments
Qianqian Ben Liu, Xiaoxiao Liu, and Xitong Guo
(pp. 391-419; DOI: 10.25300/MISQ/2020/15102)

Major Journals: MISQ CDS Common Issues

- MISQ, My Experience: no paper/involvement before 2008 (no SE in design science); Abbasi 2008 (CyberGate), 2010 (AZProtect, ICIS best paper); Guest Editor, BI&A special issue, 2010-2012 (Straub); SE 2016-2019 (Rai); Guest Editor, Health IT/Analytics special issue, 2016-2020 (Rai)
- Design Science paper common issues:
  - Where is the theory? Is this MIS? (early reviewers' critiques)
  - Few qualified/sympathetic design science SEs, AEs, reviewers. (overly critical)
  - Long review cycle (2-4 rounds/years) and uncertainty (rejection at late round).

  ➔ but
  - BI&A and data sciences are hot, in society and in b-school curriculum!
  - Young MIS CDS scholars need 1-2 MISQ/JMIS papers accepted or in deep round.
  - Mid-career MIS CDS scholars need 3-5 MISQ/JMIS papers for tenure.

Major Journals: MISQ CDS Paper Template

- Computational design science (Chen in Rai, 2017): application-inspired novelty (algorithm, representation, framework, HCI) + emerging high-impact problems
- Significant content & mature writing (40+ pages)
- MIS-specific lit review (3-4 pages) ➔ Who/what had (been) published in MISQ/ISR/JMIS (10-20 MIS references, taxonomy, analytics relevance)
- Methodology/framework/design "theory" (2-3 pages) ➔ underlying methodological foundation (not behavioral theory of +/- hypotheses), e.g., Systematic Functional Linguistic Theory, Kernel Learning Theory, etc.
- Contribution to KB + principles (research abstraction; 2-3 pages) ➔ What have been learned about the design, use and general knowledge gained?

➔ Carefully study sample MISQ DS papers, e.g., (Abbasi, 2008; 2010).

Major Journals: MISQ CDS Review Process

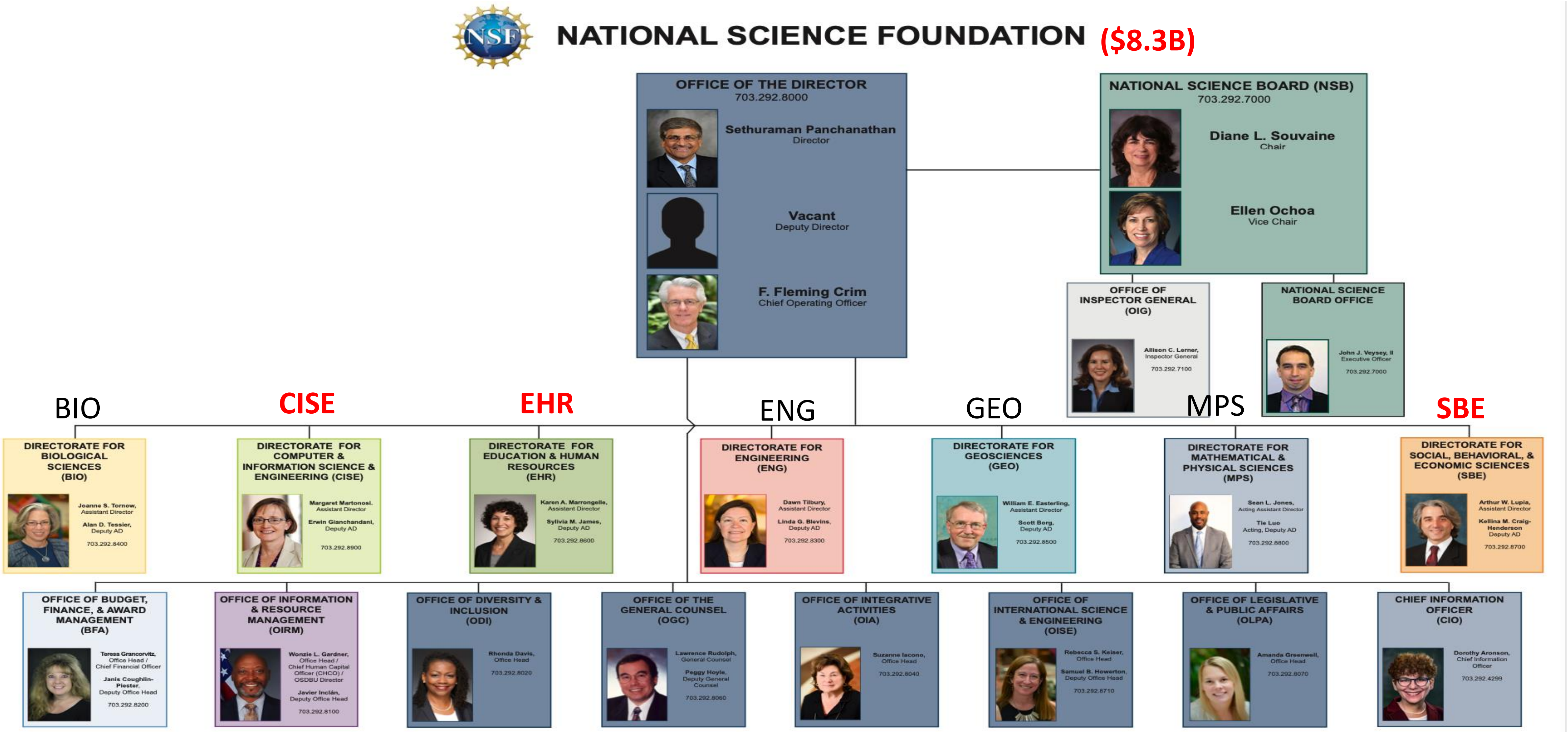
- Make sure the research fits. ➔ Emerging high-impact problems + some (not a lot) application-inspired novelty
- Make sure writing is mature. ➔ Error-free! (40+ pages)
- Select the “right” SEs and AEs. ➔ Recruit and/or consult a senior experienced MISQ DS scholar.
- 1st-round review; hope for the best after 6 months. ➔ Getting Major Revision is good (10+ pages of feedback is common)! Now their demands are clear!
- 1st-round revision is important; in 6 months. ➔ Showing appreciation, respect and tangible revision actions. Don’t fight/argue! (50+ pages of response letter!)
- 2nd/3rd/4th round review/revision ➔ Removing one critical reviewer at a time; more Minor Revision and/or Accept over time
- Final decision; 2-4 years later ➔ Eventually the SE needs to make a decision. Everyone is tired after so many years!

Major Grants: NIH, DARPA, DHS, IARPA

- NIH: NLM is informatics-focused; “translational” research with some application-inspired health-related novelty; need pubs and networking in AMIA/JAMIA; strong health informatics (NLM) tradition and turf (strong personality) ➔ Chen as NLM Scientific Counselor, 2002-2006
- DOD/DARPA: was innovative, basic/foundational, long-term (ARPA Net); now mission-critical, system-driven, short-term; commercial company (defense contractor) as prim, academic as sub; bi-monthly milestones/metrics/reporting ➔ Chen early success with DARPA/IARPA/DHS for COPLINK/Dark Web research
- DHS, IARPA: similar to DARPA, but aspiring; lesser scientific quality (strong personality)

➔ Not my focus any more! (Need to smell like them.)

Major Grants: NSF Org Chart



NATIONAL SCIENCE FOUNDATION ($8.3B)

OFFICE OF THE DIRECTOR
703.292.8000

Sethuraman Panchanathan
Director

Vacant
Deputy Director

F. Fleming Crim
Chief Operating Officer

NATIONAL SCIENCE BOARD (NSB)
703.292.7000

Diane L. Souvaine
Chair

Ellen Ochoa
Vice Chair

OFFICE OF INSPECTOR GENERAL (OIG)

Allison C. Lerner,
Inspector General
703.292.7100

NATIONAL SCIENCE BOARD OFFICE

John J. Veysey, II
Executive Officer
703.292.7000

BIO

DIRECTORATE FOR BIOLOGICAL SCIENCES (BIO)

Joanne S. Tornow,
Assistant Director
Alan D. Tessier,
Deputy AD
703.292.8400

CISE

DIRECTORATE FOR COMPUTER & INFORMATION SCIENCE & ENGINEERING (CISE)

Margaret Martonosi.
Assistant Director
Erwin Gianchandani,
Deputy AD
703.292.8900

EHR

DIRECTORATE FOR EDUCATION & HUMAN RESOURCES (EHR)

Karen A. Marrongelle,
Assistant Director
Sylivia M. James,
Deputy AD
703.292.8600

ENG

DIRECTORATE FOR ENGINEERING (ENG)

Dawn Tilbury,
Assistant Director
Linda G. Blevins,
Deputy AD
703.292.8300

GEO

DIRECTORATE FOR GEOSCIENCES (GEO)

William E. Easterling,
Assistant Director
Scott Borg,
Deputy AD
703.292.8500

MPS

DIRECTORATE FOR MATHEMATICAL & PHYSICAL SCIENCES (MPS)

Sean L. Jones,
Acting Assistant Director
Tie Luo
Acting, Deputy AD
703.292.8800

SBE

DIRECTORATE FOR SOCIAL, BEHAVIORAL, & ECONOMIC SCIENCES (SBE)

Arthur W. Lupia,
Assistant Director
Kellina M. Craig-Henderson
Deputy AD
703.292.8700

OFFICE OF BUDGET, FINANCE, & AWARD MANAGEMENT (BFA)

Teresa Grancorvitz,
Office Head /
Chief Financial Officer
Janis Coughlin-Piester,
Deputy Office Head
703.292.8200

OFFICE OF INFORMATION & RESOURCE MANAGEMENT (OIRM)

Wonzie L. Gardner,
Office Head /
Chief Human Capital Officer (CHCO) /
OSDBU Director
Javier Inclán,
Deputy Office Head
703.292.8100

OFFICE OF DIVERSITY & INCLUSION (ODI)

Rhonda Davis,
Office Head
703.292.8020

OFFICE OF THE GENERAL COUNSEL (OGC)

Lawrence Rudolph,
General Counsel
Peggy Hoyle,
Deputy General Counsel
703.292.8060

OFFICE OF INTEGRATIVE ACTIVITIES (OIA)

Suzanne Iacono,
Office Head
703.292.8040

OFFICE OF INTERNATIONAL SCIENCE & ENGINEERING (OISE)

Rebecca S. Keiser,
Office Head
Samuel B. Howerton,
Deputy Office Head
703.292.8710

OFFICE OF LEGISLATIVE & PUBLIC AFFAIRS (OLPA)

Amanda Greenwell,
Office Head
703.292.8070

CHIEF INFORMATION OFFICER (CIO)

Dorothy Aronson,
Chief Information Officer
703.292.4299

Major Grants: NSF CISE/IIS/III

CISE

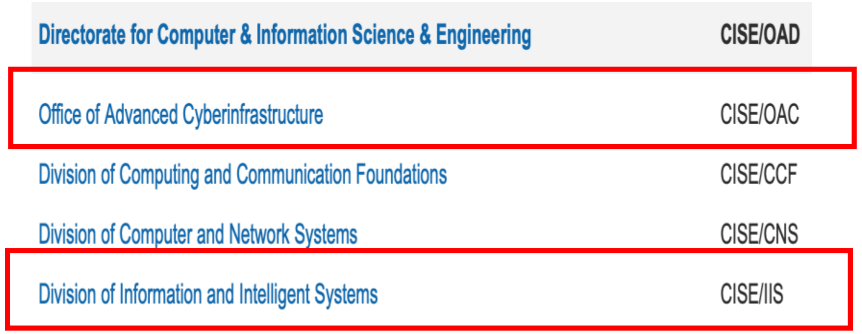
DIRECTORATE FOR COMPUTER & INFORMATION SCIENCE & ENGINEERING (CISE)

Margaret Martonosi,
Assistant Director

Erwin Gianchandani,
Deputy AD

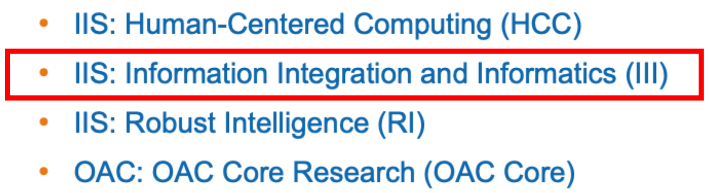
703.292.8900

IIS/OAC



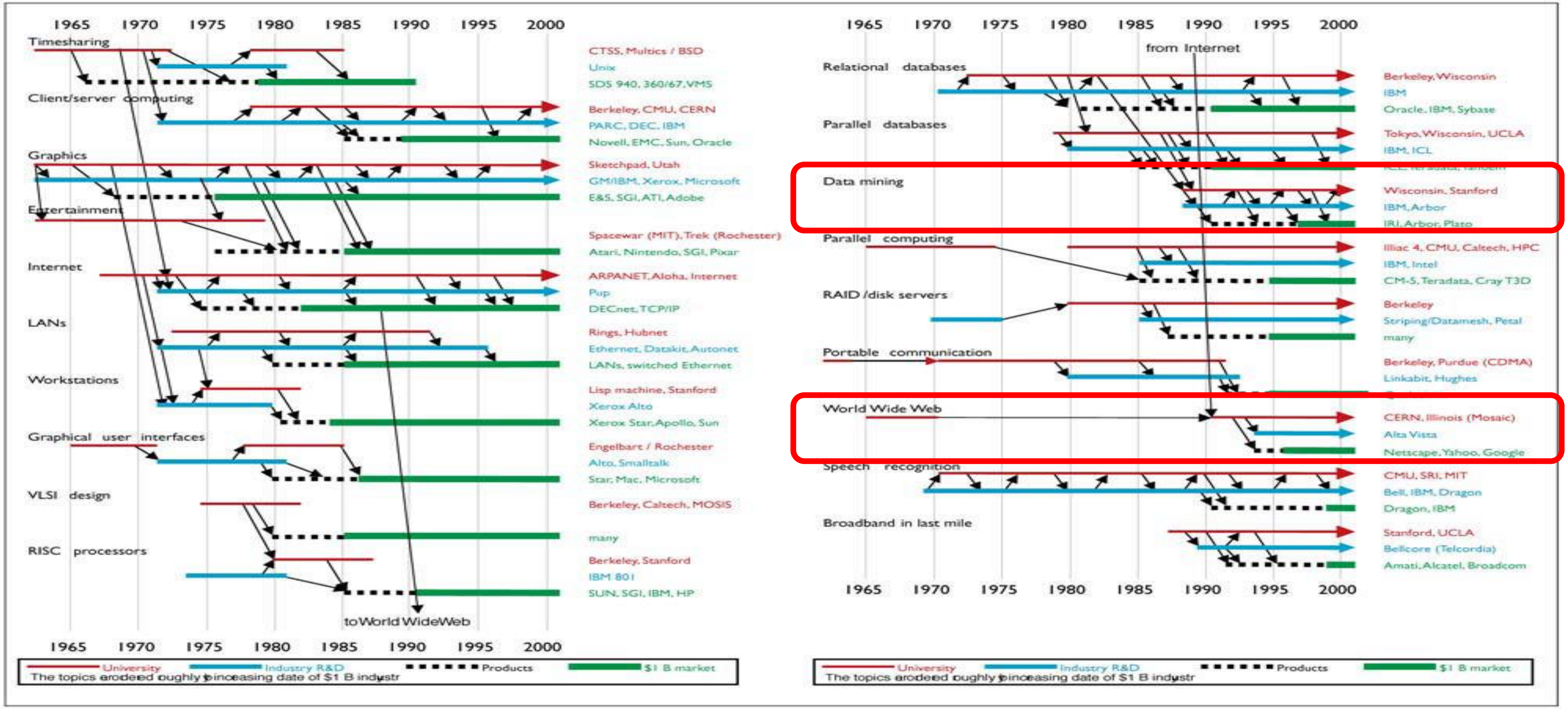
Directorate for Computer & Information Science & Engineering	CISE/OAD
Office of Advanced Cyberinfrastructure	CISE/OAC
Division of Computing and Communication Foundations	CISE/CCF
Division of Computer and Network Systems	CISE/CNS
Division of Information and Intelligent Systems	CISE/IIS

III



- IIS: Human-Centered Computing (HCC)
- IIS: Information Integration and Informatics (III)
- IIS: Robust Intelligence (RI)
- OAC: OAC Core Research (OAC Core)

Major Grants: NSF CISE/IT Societal Impacts (NAS)



Source: From [6], reprinted with permission from the National Academy of Sciences, courtesy of the National Academies Press, Washington D.C. © 2003.

University research ➔ Industry R&D ➔ **Products** ➔ $1B Market (job and wealth creation)

Major Grants: NSF Programs

- CORE: NSF CISE/IIS/III CORE most relevant to fundamental research in AI, machine learning, WWW, data sciences, NLP; acceptance rate 6-8%, highly competitive, critical young CS reviewers ➔ IIS Core ($100M/yr)
- OAC: NSF CISE/OAC relevant to applied cyberinfrastructure for sciences; acceptance rate 20-30%, less competitive, reviewers including CS, SBE, and domain sciences ➔ DIBBs, CICI ($25M-30M/yr; my focus)
- Applied Programs: Many emerging cross-directorate (e.g., EHR, SBE, CISE) and cross-agency (e.g., NSF, NIH, DOD) high-impact applied research programs (e.g., security, health); acceptance rate 15-20%, less competitive, reviewers including CS, SBE, and SME ➔ SaTC, SFS, CCRI, SCH, BIGDATA, I-DSN, National AI Institutes ($50M-100M/yr; my focus)
- Young Scholars: Many opportunities for early-career scholars; acceptance rate 10-20%, competitive, for early career; valuable for obtaining tenure! ➔ CRII, CAREER + EAGER ($200K-$1M for each award)

Major Grants: NSF Proposal Observations

- Computational Design Science (CDS) has excellent chance for successful proposals (CISE). ➔ in general, not so much for behavioral or economics MIS researchers (SBE; too basic, too incremental, not novel).
- “Business” (finance, accounting, marketing) school research is not considered STEM. ➔ need to position for larger societal/STEM problems.
- CDS research needs to compete with CS researchers (“locusts” in emerging technical fields); deep & novel domain application for emerging societal problems could be viable. ➔ my approach at least, for the past 30 years: digital library, intelligence, health, cybersecurity, etc.
- Need application or domain-inspired novelty for applied cross-directorate programs. ➔ senior Ph.D. students; last 1-2 dissertation chapters
- A lab or center can help with sustainable advantage and funding. ➔ developing collection, prototype system, etc.; structure & organizational memory

Major Grants: NSF Proposal Template

- Proposal title: short and succinct; need a multi-disciplinary team
- Project summary: Summarize problems and approach; include IM + BI
- Main text (15 pages)
  - Need mature writing; good diagrams
  - Need methodological/algorithmic novelty (IM, 60%); need strong impacts (BI, 40%)
  - Need good lit review (state-of-the-art) & promising preliminary results
- CV: need relevant ACM/IEEE references; MISQ/ISR pubs help very little
- Others: Good to have office support, e.g., budget, facilities, DMP, routing, etc.

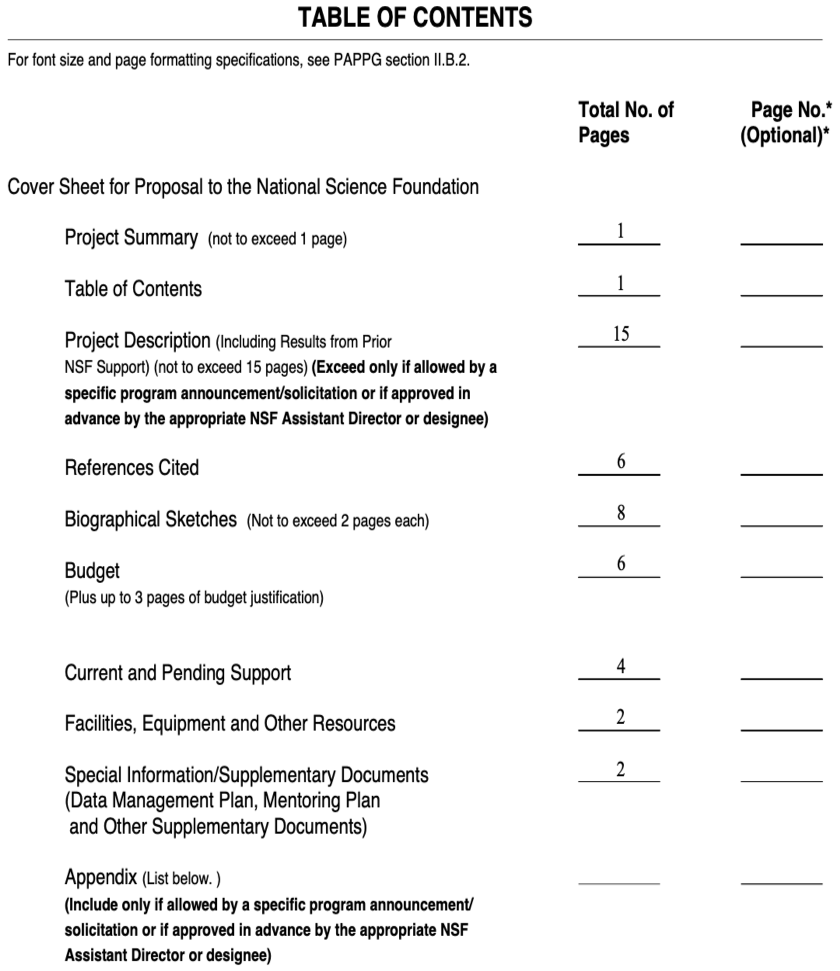


TABLE OF CONTENTS

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)**	15	
References Cited	6	
Biographical Sketches (Not to exceed 2 pages each)	8	
Budget (Plus up to 3 pages of budget justification)	6	
Current and Pending Support	4	
Facilities, Equipment and Other Resources	2	
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)**		

Major Grants: NSF Proposal Reviews

- As a reviewer/panelist:
  - Asked to review 7-8 proposals (out of 20-22) in 2-3 weeks
  - One-page review, overall rating: E, VG, G, F, P (few with E or VG)
  - Only 2-3 proposals received Competitive or Highly Competitive (fundable; 90% proposal) in each panel (2/20)
  - On-site panel discussion critical for outcome (vocal panelist)

- As a proposer/PI:
  - Will receive 4-5 reviews, varying from VG, G, F (aiming 80%; rarely receiving E).
  - Common IM critiques: lack of novelty, poor lit review, missing preliminary results
  - Common BI critiques: value unclear, lack of diversity/education plan
  - Other significant critiques: lack of track record, poor team, lack of collaboration plan, etc.
  - Need to improve from 10% success rate (60% proposal) to 30% (80% proposal) over time in 2-3 tries.
  - Learn the process and grantsmanship for future proposals.

Major Grants: NSF General Advice for CDS Scholars

- Develop methodological novelty and application-specific strengths over your career. ➔ world-class excellence vs. other CS scholars
- Train your Ph.D. students well. ➔ their last 2 dissertation chapters could be fundable; they can be trained to write proposals (scale & efficiency)
- Build a center/lab/group. ➔ more sustainable and impressive (common in CS, ECE, MED)
- Improve your grantsmanship. ➔ get to know your PDs and become frequent NSF panelists (getting into their heads)
- Improve your success rate to 30% (one in 3). ➔ target repeating programs for re-submissions
- Monitor and anticipate current and emerging programs. ➔ prepare the next proposals; repeat the cycle!

Parting Thoughts: Hard Work + A Bit of Luck

- Societal Impact > Academic Impact
  - Looking for high-impact societal problems (NYT, WSJ, The Economists)
- IT > MIS
  - MIS is a smaller subfield within broader IT/computing.
- CISE > SBE
  - Computational Design Science can make a difference.
- New > Old
  - Looking for new, interesting, unknown problems
- EQ > IQ
  - Hard work, discipline, aspiration, etc. always beat raw talent. Plus a bit of luck!

For questions and comments

hchen@eller.Arizona.edu

http://ai.Arizona.edu