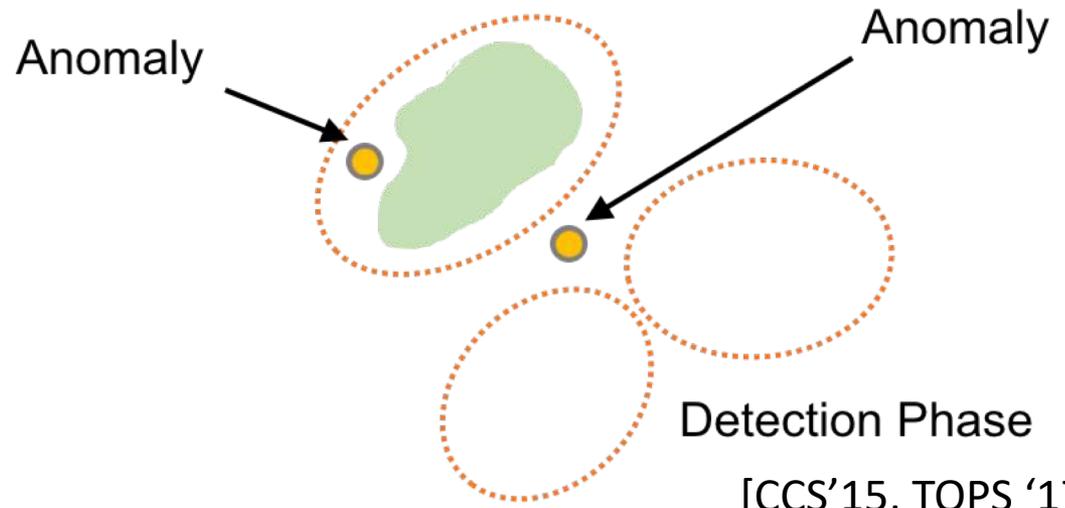
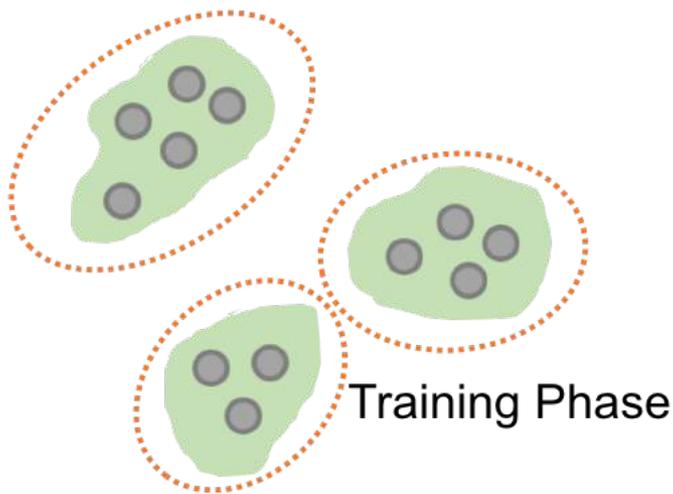


Data Breach and Multiple Points to Stop It

Danfeng (Daphne) Yao
Department of Computer Science
Virginia Tech



[CCS'15, TOPS '17]



MORGAN & CLAYPOOL PUBLISHERS

Anomaly Detection as a Service

*Challenges, Advances,
and Opportunities*

Danfeng (Daphne) Yao
Xiaokui Shu
Long Cheng
Salvatore J. Stolfo

SYNTHESIS LECTURES ON
INFORMATION SECURITY, PRIVACY, AND TRUST

Elisa Bertino & Ravi Sandhu, Series Editors

```

1 // Create a trust manager that does not validate certificate
2 TrustManager[] trustAllCerts = new TrustMana
3     new X509TrustManager() {
4         public java.security.cert.X509Certificate
5             getAcceptedIssuers() {return null;}
6         public void checkClientTrusted(...) {}
7         public void checkServerTrusted(...) {}
8     }
9 // Install the all-trusting trust manager
10 try {
11     SSLContext sc = SSLContext.getInstance("SSL")
12     sc.init(null, trustAllCerts, new java.secu
13         SecureRandom());
14     HttpURLConnection.setDefaultSSLSocketFacto
15         .getSocketFactory());
16 } catch (Exception e) {}

```

[ICSE '18, SecDev '17]

Acknowledgment



Xiaokui Shu
(IBM Research)



Fang Liu
(Palo Alto Networks)



Jing Zhang
(AMD)



Elisa Bertino
(Purdue)



Ali Butt
(VT)



Wu Feng
(VT)

Ford pickup truck F-150 has 150 million LOC

Ford GT has over 10 million lines of code

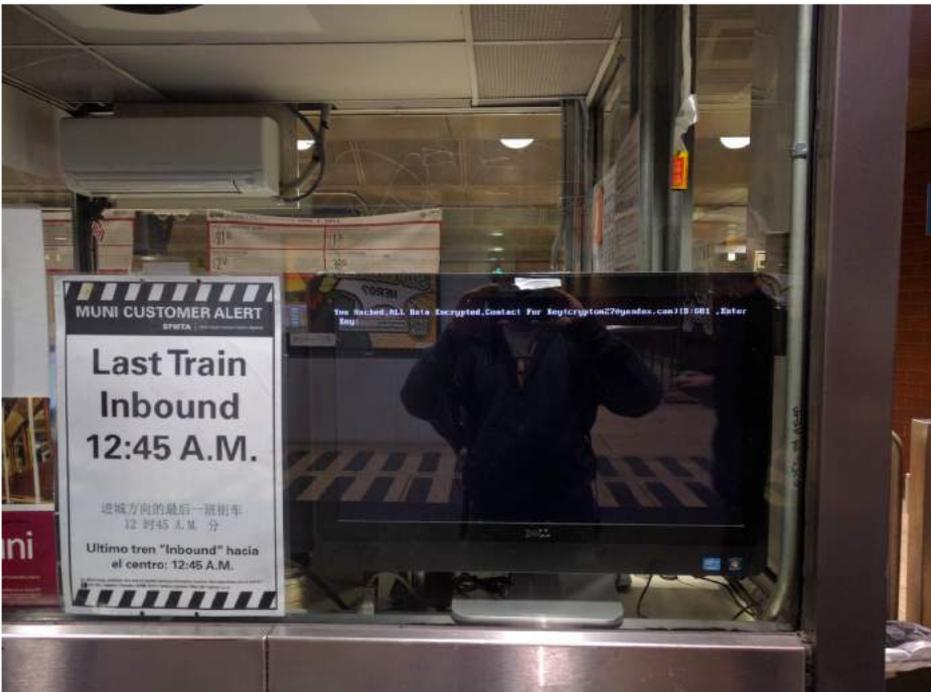
F-22 Raptor has 2 million lines of code

Boeing 787 Dreamliner has 7 million lines of code

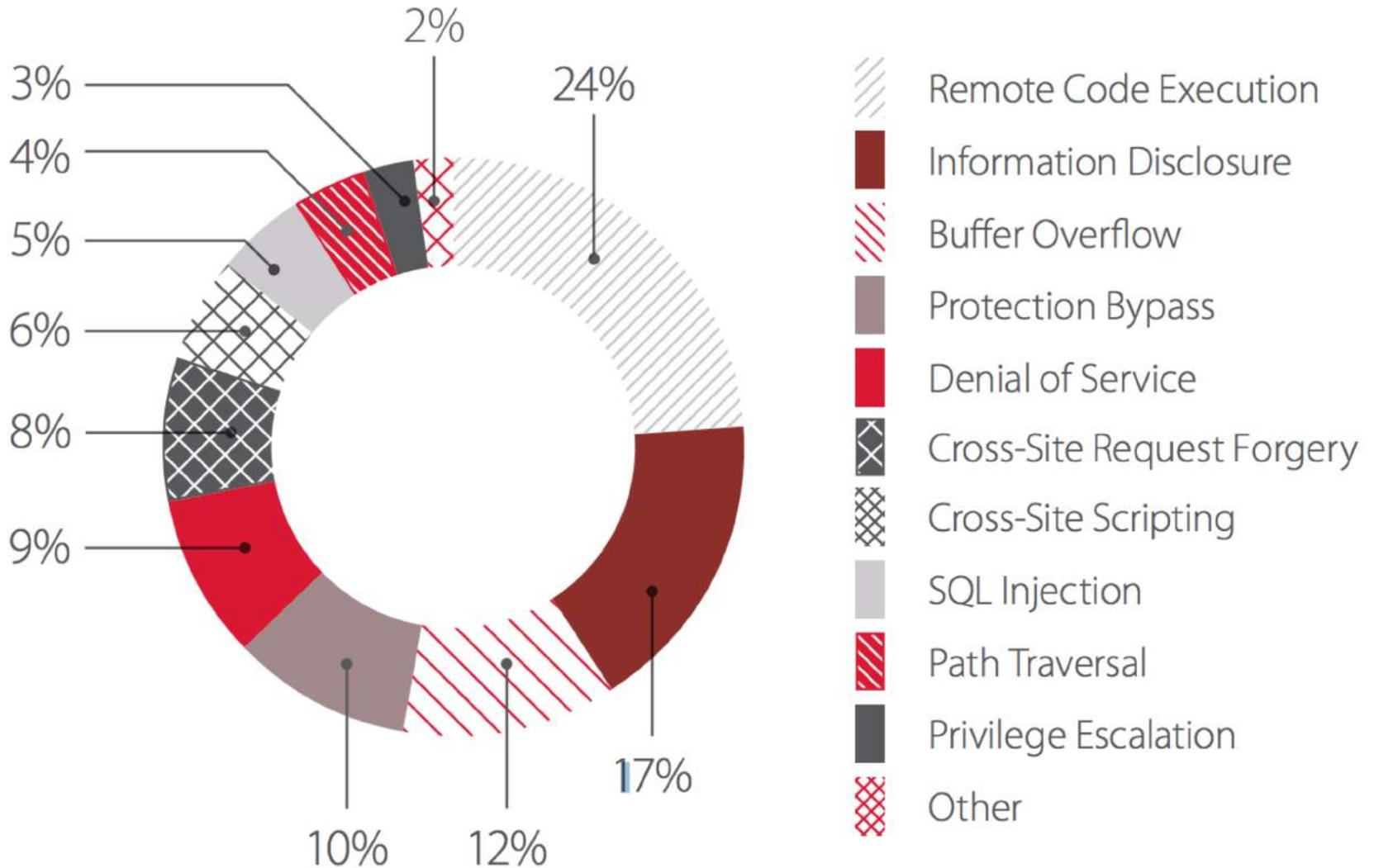


Ransomware attack on San Francisco public transit gives everyone a free ride

"You Hacked, ALL Data Encrypted. Contact For Key(cryptom27@yandex.com)ID:681,Enter."



Types of vulnerabilities in industrial control systems



<https://www.ptsecurity.com/upload/corporate/ww-en/analytics/ICS-Security-2017-eng.pdf>

<https://www.infosecurity-magazine.com/news/critical-infrastructure-more/>



HOME >> SECURITY

SECURITY

JAN
31
2017

How 3 Local Governments Mitigated Ransomware Attacks



Planning and education help local governments blunt the effects of ransomware attacks.

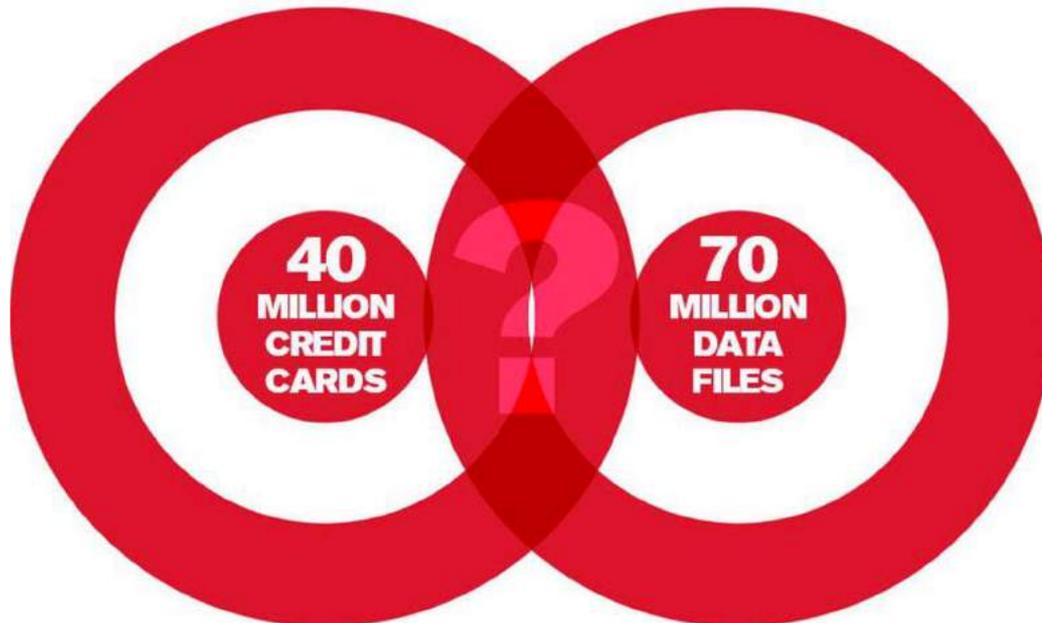
Target data breach



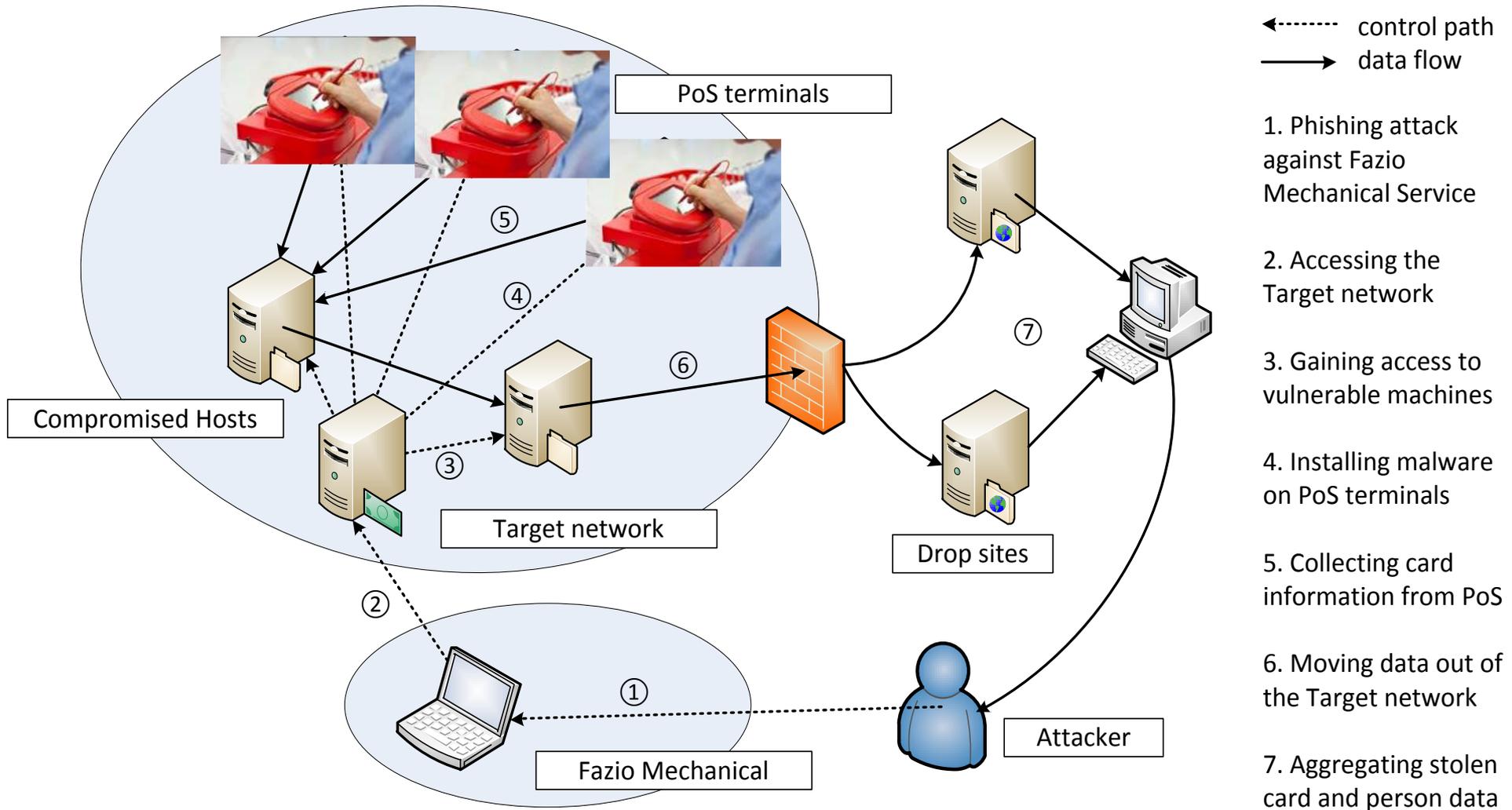
BUSINESS NEWS

Target Settles 2013 Hacked Customer Data Breach For \$18.5 Million

by Reuters / May.24.2017 / 10:49 AM ET / Source: Reuters



Target data breach (Nov. 27 to Dec. 15, 2013)



ree4@exploit.im: <http://plasmon.rghost.ru/44699041/image.png>

hidden: how does it keep the data (intercepted credit cards)?

reed4@exploit.im: from left side it is files, time.txt, then you click on it and you will find dumps in browser in plaintext

hidden: are there any differences in terms of infected Point-of-Sale systems?

ree4@exploit.im: no, but there are some nuances, for examples it doesn't work on Verifone

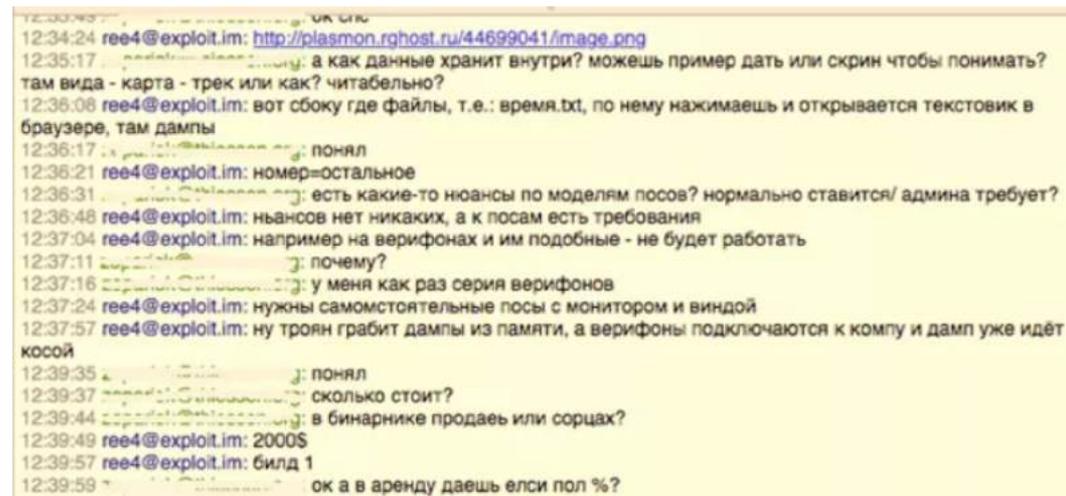
hidden: really? I have Verifones ...

reed4@exploit.im: it grabs dumps from memory, Verifone can be connected to PC, but it will be "secured", you need standalone Point-of-Sale terminals with monitor and Windows

hidden: how much?

ree4@exploit.im: 2000 USD

March 23, 2013



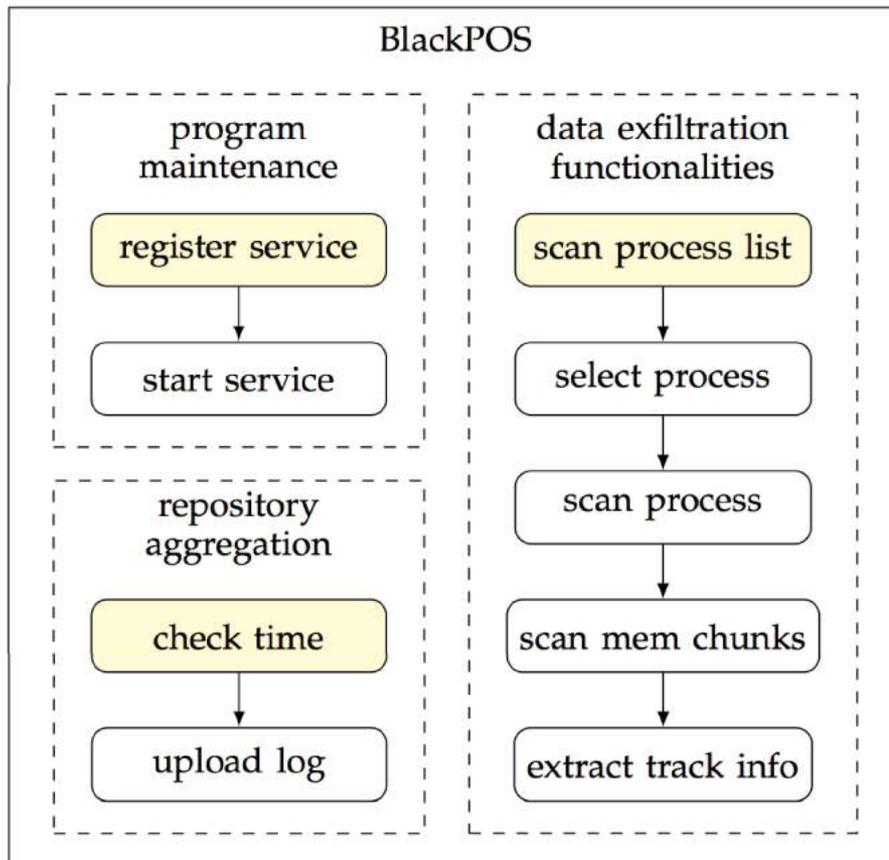
The screenshot shows a chat log with the following text:

```
12:34:24 ree4@exploit.im: http://plasmon.rghost.ru/44699041/image.png
12:35:17 ...: а как данные хранит внутри? можешь пример дать или скрин чтобы понимать?
там вида - карта - трек или как? читабельно?
12:36:08 ree4@exploit.im: вот сбоку где файлы, т.е.: время.txt, по нему нажимаешь и открывается текстовик в
браузере, там дампы
12:36:17 ...: понял
12:36:21 ree4@exploit.im: номер=остальное
12:36:31 ...: есть какие-то нюансы по моделям посов? нормально ставится/ админа требует?
12:36:48 ree4@exploit.im: нюансов нет никаких, а к посам есть требования
12:37:04 ree4@exploit.im: например на верифонах и им подобные - не будет работать
12:37:11 ...: почему?
12:37:16 ...: у меня как раз серия верифонов
12:37:24 ree4@exploit.im: нужны самостоятельные посы с монитором и виндой
12:37:57 ree4@exploit.im: ну троян грабит дампы из памяти, а верифоны подключаются к компу и дамп уже идёт
косой
12:39:35 ...: понял
12:39:37 ...: сколько стоит?
12:39:44 ...: в бинарнике продаёшь или сорцах?
12:39:49 ree4@exploit.im: 2000$
12:39:57 ree4@exploit.im: билд 1
12:39:58 ...: ок а в аренду даёшь елси пол %?
```

<https://securityaffairs.co/wordpress/21337/cyber-crime/blackpos-malware.html>

BlackPOS (memory scrapper malware)

- Runs as a Windows service “POSWDS”
- Scans a list of processes that interact with the card reader
- Uploads credit cards to a compromised server (internal network repository)



<https://blog.trendmicro.com/trendlabs-security-intelligence/new-blackpos-malware-emerges-in-the-wild-targets-retail-accounts/>

How can a HVAC vendor's credential access Target's internal networks?



[Home](#)

[About Us](#)

[Services](#)

[Our Work](#)

[Careers](#)

[Contact](#)



Ross E. Fazio

President



Ross A. Fazio

Executive Vice President



Jeff Rupert

Vice President of

“Fazio Mechanical does not perform remote monitoring of or control of heating, cooling and refrigeration systems for Target,” Fazio said (Feb. 2014).

Fazio's credential also had access to other portals in Target



Sign In

Email Address

Password

Remember Me

Sign In

[Forgot your password?](#)

[Interested in becoming a Business Partner? | Learn More](#)

A banner for the SAP Ariba Procure-to-Pay Software Solution. The banner features the SAP Ariba logo in the top left corner, a search icon and a menu icon in the top right corner, and a background image of a woman looking down. The text on the banner reads: "SAP Ariba Procure-to-Pay Software Solution" and "Provide your users with a fast, guided buying experience with the leading procure-to-pay software solution." There is a "Contact Us" button at the bottom left of the banner.

SAP Ariba

Solutions

SAP Ariba Procure-to-Pay Software Solution

Provide your users with a fast, guided buying experience with the leading procure-to-pay software solution.

Contact Us

SAP Ariba Billing System

But how can a billing system credential access Target's internal networks?

FireEye's IDS



FireEye alerts



Target's security team in Bangalore received FireEye alerts; sent alerts to Target headquarters

FireEye's auto-malware-delete function was turned off

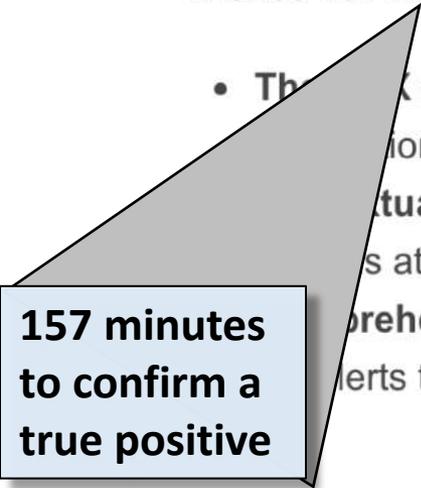
FireEye makes alerts worthwhile again



It takes 157 minutes for an expensive expert analyst to correctly identify a true positive alert. That's a lot of time and money.

- The **Next-Gen engine** identifies true positive alerts without volumes of alerts or false positives. Since security operations leaves them free for more important tasks. It even finds signs of threats for previously undetected threats. **Contextual intelligence** accompanies validated alerts to help your analysts quickly prioritize alerts based on attacker profile, threat severity and attack scale and scope.

Comprehensive visibility across the entire lifecycle to reduce alerts by up to 76 percent. By seeing threats that would be generated from subsequent stages of the attack (e.g. callbacks) and alerts that



**157 minutes
to confirm a
true positive**

"We haven't seen any false positives and the alerts are going on across our whole infrastructure. And by getting them to minimize wasting resources on having to clean up a bad posture is even more valuable for us."

- SCOTT ADAMS, MANAGER (

Research opportunities: better warning design so admins & analysts pay attention

“FireEye ... is cutting edge. But it takes love and care and feeding. You have to watch it and monitor it.”

-- John Strand, Black Hills InfoSec (regarding Target data breach)

1. Fear
2. Obedience
3. Greed
4. Helpfulness

Hacker psychology: Understanding the 4 emotions of social engineering

And some key considerations for better positioning your employees against falling prey to these types of attacks



By Austin Whipple, Senior Security Engineer, BetterCloud

Network World | MAY 13, 2016 1:07 PM PT

PCI Compliance is just a baseline



"Target was certified as meeting the standard for the payment card industry (PCI) in Sept. 2013."

-- Gregg Steinhafel (Target then CEO, stepped down in 2014)



PCI data security standard is a standard for securing electronic payments

DISCOVER NETWORK ATTESTATION OF COMPLIANCE STATUS WITH DISCOVER NETWORK'S SECURITY REQUIREMENTS

Discover Network requires all Merchants, Acquirers, Third Party Processors and Payment Service Providers ("Company") to comply with the Payment Card Industry Data Security Standard ("PCI DSS") located at www.discovernetwork.com and/or www.pcisecuritystandards.org as well as any additional security requirements and all related compliance requirements promulgated by Discover Network from time to time. This document will serve as your attestation of compliance with Discover Network's Security Requirements. The information below must be completed in its entirety, signed by an authorized officer of Company and submitted to Discover Network according to the instructions in Section 5.

Section 1 - Company Contact Information	
Date	
Company Legal Name	
Compliance Contact Name	
Compliance Contact Phone Number	(XXX)XXX-XXXX
Compliance Contact E-mail Address	

Section 2 - Company's PCI Compliance Status	
(Name/Title of Officer) certifies the following compliance status (select one):	
<input type="checkbox"/> COMPLIANT	(Company) has achieved full compliance with the PCI DSS as of (date of compliance). Name of Qualified Security Assessor (if applicable): Proceed to Section 4.
<input type="checkbox"/> NON-COMPLIANT	(Company) has not achieved full compliance with the PCI DSS as of (date). Company plans to achieve full compliance on: (date). Company is required to complete Section 3.

Section 3 - Summary of Company's Compliance with PCI DSS Requirements
Please select the appropriate "Compliance Status" for each requirement. If you answer "NO" to any of the requirements, you are required to provide the date Company will be compliant with the requirement and a brief description of the actions being taken to meet the requirement.

PCI Req.	Description of Requirement	Compliance Status (select one)		Remediation Date and Actions (if "Non-Compliant" was selected in the "Compliance Status" column)
		Compliant	Non-Compliant	
1	Install and maintain a firewall configuration to protect cardholder data	<input type="checkbox"/>	<input type="checkbox"/>	
2	Do not use vendor-supplied defaults for system passwords and other security parameters	<input type="checkbox"/>	<input type="checkbox"/>	
3	Protect stored cardholder data	<input type="checkbox"/>	<input type="checkbox"/>	
4	Encrypt transmission of cardholder data across open, public networks	<input type="checkbox"/>	<input type="checkbox"/>	
5	Use and regularly update anti-virus software	<input type="checkbox"/>	<input type="checkbox"/>	
6	Develop and maintain secure systems and applications	<input type="checkbox"/>	<input type="checkbox"/>	
7	Restrict access to cardholder data by cryptographic controls and to-know	<input type="checkbox"/>	<input type="checkbox"/>	
8	Assign a unique ID to each person with computer access	<input type="checkbox"/>	<input type="checkbox"/>	

Protect stored cardholder data

Regularly test security systems and processes



Multi-factor authentication -- A lesson learned by PCI from the Target breach

8.3 Secure all individual non-console administrative access and all remote access to the CDE using multi-factor authentication.

Note: Multi-factor authentication requires that a minimum of two of the three authentication methods (see Requirement 8.2 for descriptions of authentication methods) be used for authentication. Using one factor twice (for example, using two separate passwords) is not considered multi-factor authentication.

8.3.1 Incorporate multi-factor authentication for all non-console access into the CDE for personnel with administrative access.

Note: This requirement is a best practice until January 31, 2018, after which it becomes a requirement.

8.3.2 Incorporate multi-factor authentication for all remote network access (both user and administrator, and including third party access for support or maintenance) originating from outside the entity's network.



PCI merchant levels

LEVEL 1	LEVEL 2	LEVEL 3	LEVEL 4
<p>6M +</p> <p>Process more than 6 million Visa transactions per year, regardless of channel.</p> <p>Be identified as Level 1 by any card association.</p>	<p>1-6M</p> <p>Process 1 to 6 million credit card transactions annually across all channels.</p>	<p>20K-1M</p> <p>Process 20,000 to 1 million e-commerce credit card transactions annually.</p>	<p><20K</p> <p>Process fewer than 20,000 e-commerce transactions annually, or process fewer than 1 million credit card transactions annually across all channels.</p>
SECURITY REQUIREMENTS			
<p>Complete a ROC annually by a Qualified Security Assessor (QSA) *. This means an on-site audit needs to occur every year.</p> <p>Quarterly scans by an Approved Scanning Vendor (ASV) *.</p> <p>An AOC that verifies everything meets PCI standards.</p>	<p>Conduct an annual Self-Assessment Questionnaire (SAQ) *.</p> <p>Quarterly scans by an Approved Scanning Vendor (ASV).</p> <p>An AOC that verifies everything meets PCI standards.</p>	<p>Conduct an annual Self-Assessment Questionnaire (SAQ) *.</p> <p>Quarterly scans by an Approved Scanning Vendor (ASV).</p> <p>An AOC that verifies everything meets PCI standards.</p>	<p>Conduct an annual Self-Assessment Questionnaire (SAQ) *.</p> <p>Quarterly scans by an Approved Scanning Vendor (ASV).</p> <p>An AOC that verifies everything meets PCI standards.</p>

PCI approved scanning vendors



COMPANY	PLACE OF BUSINESS	PRODUCT NAME	EMAIL CONTACT	LOCATIONS SERVED	CERTIFICATION NUMBER
AccessIT Group, Inc **In Remediation**	United States	AccessIT Group ASV	Petem@		5086-01-01
Alert Logic, Inc.	United States	Alert Logic PCI	sales@alertlogic.com	North America, Europe, Japan	4222-01-12
Aperia	United States	Aperia Pro Scan	jnix@aperiasolutions.com	Global	5051-01-07
AppSec Consulting	United States	AppSec Certified	info@appsecconsulting.com	North America	3834-01-12
AT&T Consulting Solutions	United States	AT&T	pci@att.com	Global	5024-

But security guarantees are often vague

Or looks rather basic

Test Scope

The vulnerability scanning service covers all machines in the given internet address range from which responses were detected. For each machine detected, the services and characteristics of the machine are analysed.

TCP/IP characteristics	ICMP responses and other TCP/IP characteristics of the machine are examined. These are used to report the detected operating system (often including the version) and system uptime where available.
TCP services	A table of available TCP services and relevant further information is produced. Netcraft's tests identify the network service on each port — in particular, standard network services running on non-standard ports are identified and fully tested.
UDP services	A table of UDP ports which are believed to be open, and any information obtained from them. Note that due to the design of the UDP protocol, false positives are common in identifying active UDP ports, especially if firewalls are filtering content from these ports. If filtering is in place, our

Sophistication of the approval process for PCI scanners?



PCI council allows infinite retesting,
Training materials available

Specialized scan

Advanced threats

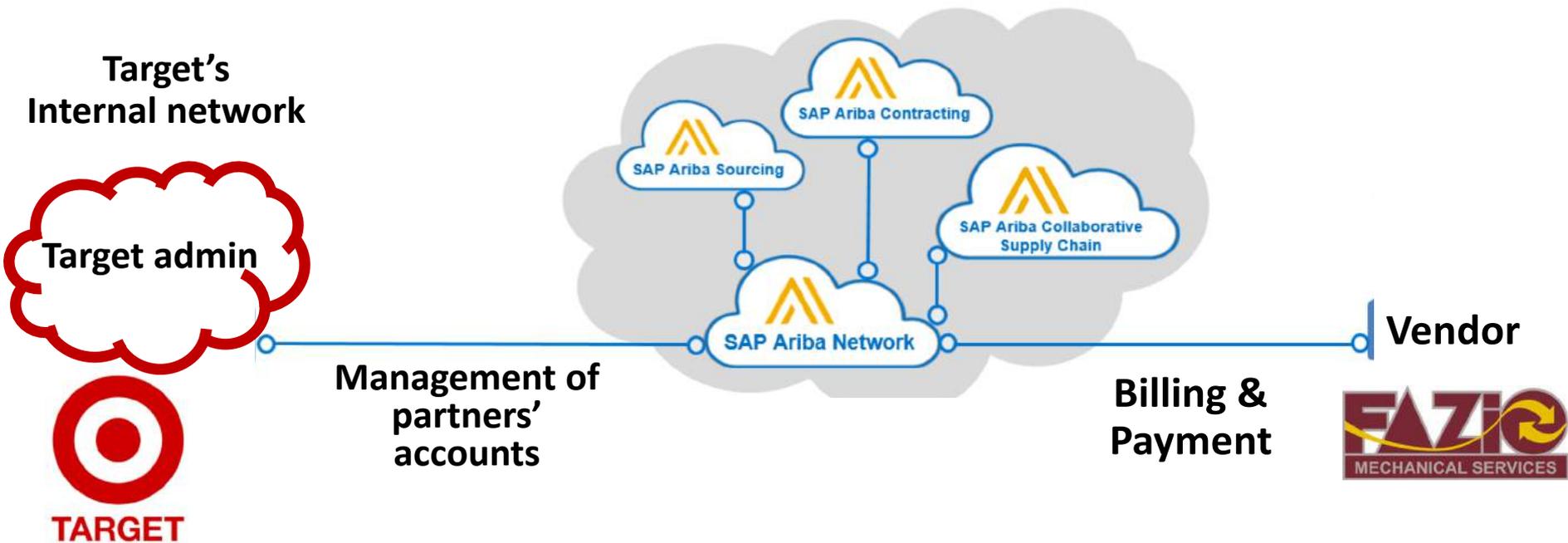
How can researchers help?

In-depth scan

Deployable tools

Transparency

Other technical vulnerabilities? Theory 1 – Issues in how Target admins managed Ariba



Theory 2

1. Php scripts uploaded as invoices to Target's billing portals

2. Web server attempted to open it; code got executed



[https://www.owasp.org/index.php/Unrestricted File Upload](https://www.owasp.org/index.php/Unrestricted_File_Upload)

<https://aroundcyber.files.wordpress.com/2014/09/aorato-target-report.pdf>

Missed opportunities



Lack of transparency makes it difficult to learn from past failures

Target's improvements (April 29, 2014)



Improved monitoring and logging of system activity

✓ Installed application whitelisting POS systems and

Implemented POS management tools

Improved firewall rules and policies

✓ Limited or disabled vendor access to their network

Disabled, reset, or reduced privileges on over 445,000 Target personnel and contractor accounts

✓ Expanded the use of two-factor authentication and password vaults

Trained individuals on password rotation

Target also joined 2 cybersecurity threat-sharing initiatives



Financial Services
Information Sharing and
Analysis Center

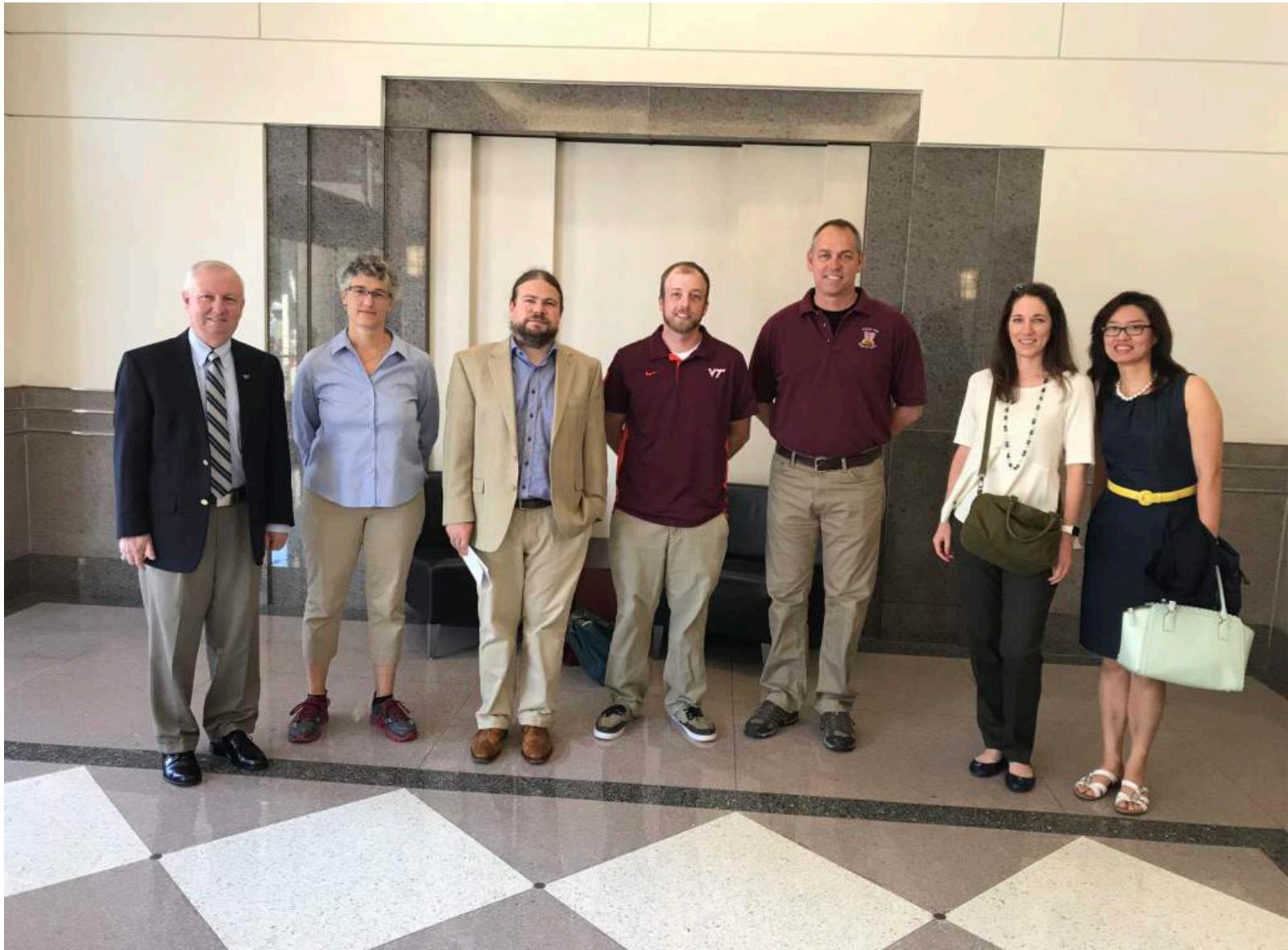


Retail Cyber Intelligence
Sharing Center

National Cybersecurity and Communications Integration Center



DHS NCCIC welcomes you to visit



Threat intelligence – How can researchers help?

Analyze security guarantees

Understand security limitations

What's useful beyond eye candies?

Know the types of actionable items

Equifax data breach --145.5 million consumers affected

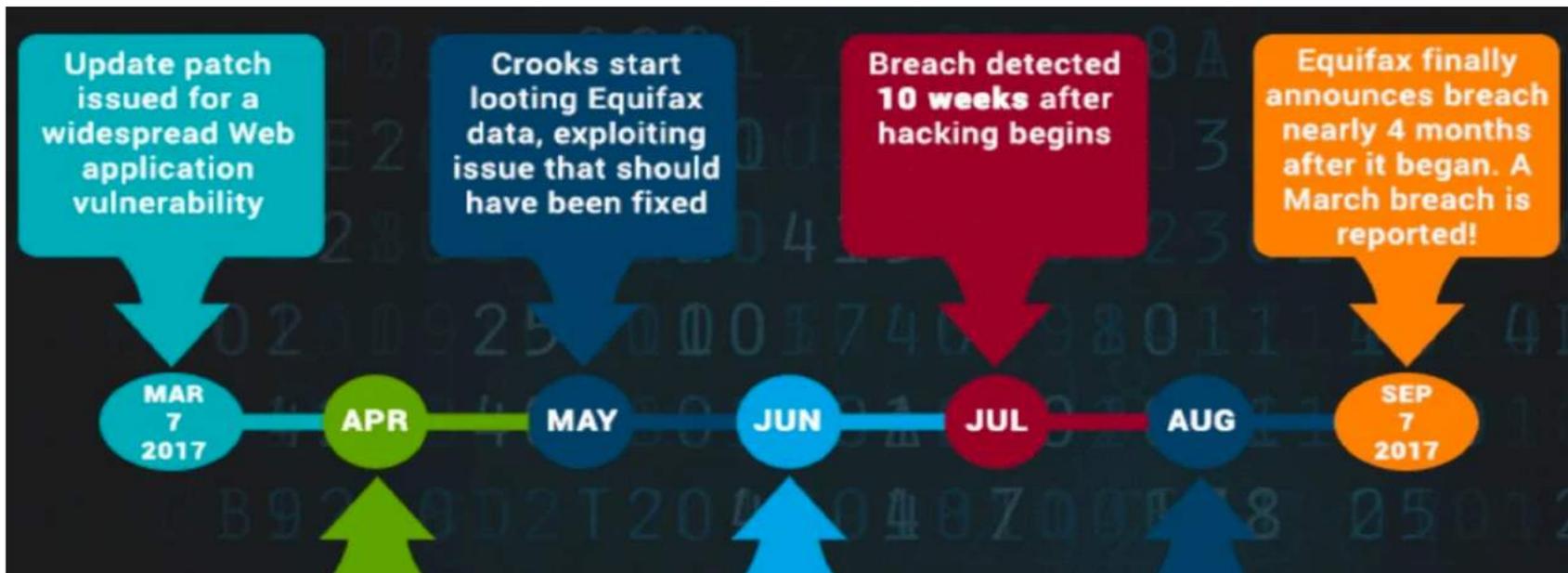
Apache Struts Vulnerability (CVE-2017-5638)

2017-03-06: vulnerability announced on along with a patch

2017-03-07: an exploit released

2017-07-30: Equifax patched

146 days: Time to patch at Equifax

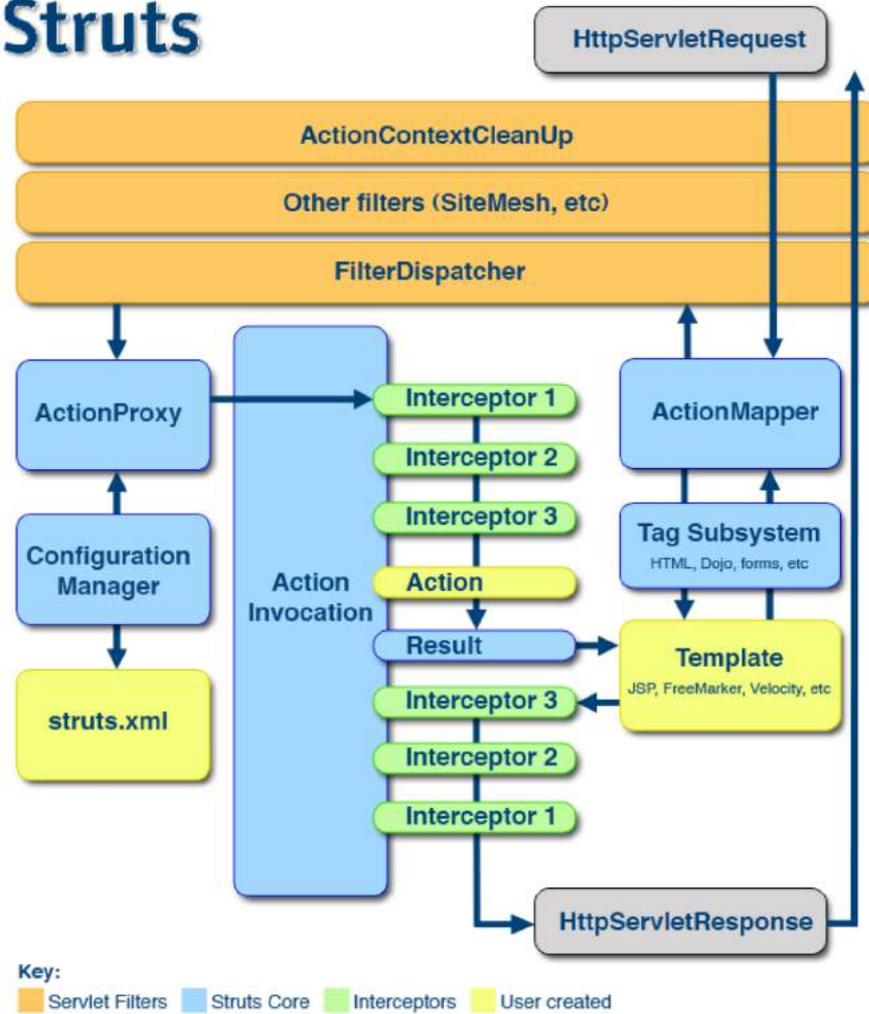


<https://www.gracefulsecurity.com/equifax-breach-timeline/>

<https://blog.blackducksoftware.com/equifax-apache-struts-cve-2017-5638-vulnerability>

Vulnerability allows remote attackers to execute commands

Struts



Apache Struts: an open-source web application framework for Java EE web applications

Apache Struts FileUploadInterceptor class

For error-handling during file upload
(e.g., parsing & size errors)

```
1. if (multiWrapper.hasErrors()) {  
2.   for (LocalizedMessage error : multiWrapper.getErrors()) {  
3.     if (validation != null) {  
4.       validation.addActionError(LocalizedTextUtil.findText(error.getClazz(),  
error.getTextKey(), ActionContext.getContext().getLocale(),  
error.getDefaultMessage(), error.getArgs())));  
       }  
     }  
  }  
}
```

Problem: Struts' error message
render engine shows untrusted
properties back to the user

An attack header

```
Content-Type: %{(#_='multipart/form-data')
...
. (#cmd='ls -l')
...
. (#ros.flush()) }
```

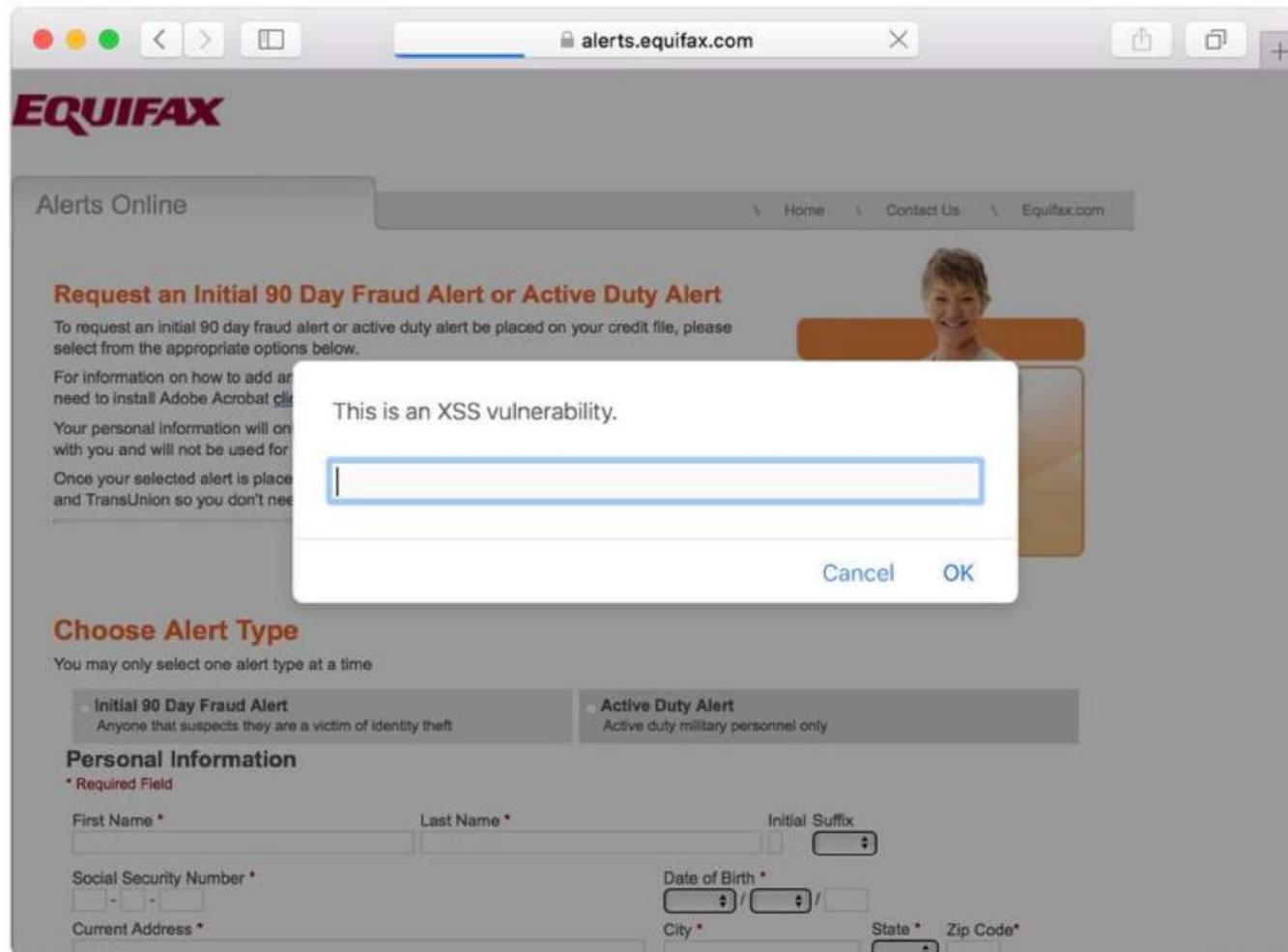
What happens:

1. FileUploadInterceptor cannot parse the header;
2. It attempts to put together an error message;
3. It evaluates/executes the OGNL expression from the attacker.

```
root@sh:~/struts2-S2-045# python exploit.py http://127.0.0.1:8080/2.3.15.1-showcase/showcase.action "ls -l"
[*] CVE: 2017-5638 - Apache Struts2 S2-045
[*] cmd: ls -l

total 12
lrwxrwxrwx 1 root    root      12 Nov 15 09:37 conf -> /etc/tomcat8
drwxr-xr-x 2 tomcat8 tomcat8 4096 Nov 15 09:37 lib
lrwxrwxrwx 1 root    root      17 Nov 15 09:37 logs -> ../../log/tomcat8
drwxr-xr-x 2 root    root      4096 Mar  7 00:55 policy
drwxrwxr-x 3 tomcat8 tomcat8 4096 Mar  7 01:34 webapps
lrwxrwxrwx 1 root    root      19 Nov 15 09:37 work -> ../../cache/tomcat8
```

XSS Negligence at Equifax



In addition, no Intrusion Detection Systems

Equifax's freeze PIN is the timestamp -- predictable



Tony Webster ✓

@webster

Follow



OMG, Equifax security freeze PINs are worse than I thought. If you froze your credit today 2:15pm ET for example, you'd get PIN 0908171415.

7:38 PM - 8 Sep 2017

3,797 Retweets 5,036 Likes



212 3.8K 5.0K





Tony Webster ✓ @webster · 8 Sep 2017

Verified PIN format w/ several people who froze today. And I got my PIN in 2007 —same exact format. Equifax has been doing this for A DECADE.

“admin/admin” login for Equifax Argentina employee portal

Id	Apellido	Nombre	Usuario	documento	Email	Estado	Perfil		
1859471	A	Marcela	m		ma	INACTIVO	USUARIO	Eliminar	Editar
1859475	A	Yeimy	ya		ye	INACTIVO	USUARIO	Eliminar	Editar
1271524	A	Maria Belen	ba		ma	INACTIVO	USUARIO	Eliminar	Editar
274804	A	Martin	m		ma	INACTIVO	USUARIO	Eliminar	Editar
527	A	Marita	m		me	INACTIVO	ADMINISTRADOR	Eliminar	Editar
1358701	A	Eugenia	ea		Eu	INACTIVO	USUARIO	Eliminar	Editar
1859467	A	Alejandra	aa		ale	INACTIVO	USUARIO	Eliminar	Editar
1572254	A	Mariela	m		ma	ACTIVO	USUARIO	Eliminar	Editar
2025633	A	Carlos	ca		ca	INACTIVO	USUARIO	Eliminar	Editar
2025667	A	Carlos	ca		ca	INACTIVO	USUARIO	Eliminar	Editar
2025660	A	Jose Pablo	jp		Jo	INACTIVO	USUARIO	Eliminar	Editar
709	E	Marcelo	m		ml	ACTIVO	USUARIO	Eliminar	Editar
1572338	E	Gaston	gt		ga	INACTIVO	USUARIO	Eliminar	Editar
1789253	E	Priscila	pt		pis	INACTIVO	USUARIO	Eliminar	Editar
1536812	E	Martin	m		ma	INACTIVO	USUARIO	Eliminar	Editar
711	E	Oscar	ob		ob	ACTIVO	USUARIO	Eliminar	Editar
334837	C	Alejandra	ac		ale	INACTIVO	USUARIO	Eliminar	Editar
123392	C	Guillermo	gc		gu	INACTIVO	USUARIO	Eliminar	Editar
1433356	D	Laura	ld		lau	INACTIVO	USUARIO	Eliminar	Editar
1702095	D	Eliana	ed		eli	INACTIVO	USUARIO	Eliminar	Editar

Would PCI compliance have saved Equifax?

PCI DSS Requirement 6

- 6.2** Protect all system components and software from known vulnerabilities by installing applicable vendor-supplied security patches. **Install critical security patches within one month of release.**

PCI DSS Requirement 11

- 11.4** Use network intrusion detection and/or intrusion prevention techniques to detect and/or prevent intrusions into the network. Monitor all traffic at the perimeter of the cardholder data environment

Research opportunities: Can these and other PCI DSS requirements be automatically checked?

Formatting Excel files and accidental data leak

[Feb. 2017] An employee emailed a company spreadsheet to his spouse, who didn't work at Boeing

36,000 Boeing employees' data is leaked

- names
- social security numbers
- dates of birth
- places of birth
- employee ID numbers
- accounting department codes



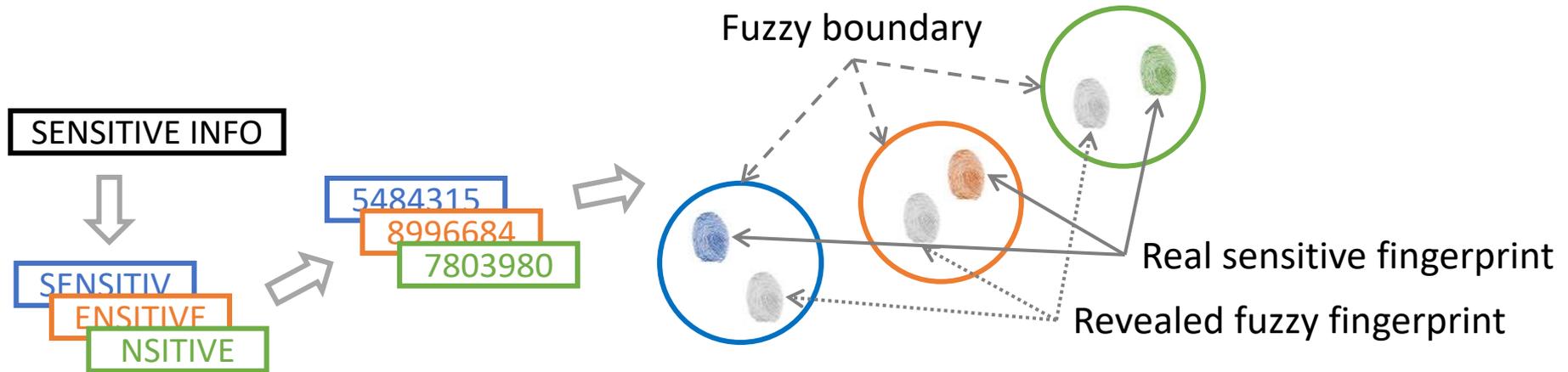
	Globally	Russia etc.	China	N. America	W. Europe	E. Markets	APAC	Mid-East	Japan
Base	4,438	518	208	400	1,576	611	822	105	198
Vulnerabilities / flaws in existing software	36%	50%	38%	33%	32%	37%	37%	23%	26%
Accidental leaks/sharing of data by staff	29%	34%	42%	26%	26%	25%	34%	25%	23%
Loss/theft of mobile devices by staff	26%	19%	27%	22%	29%	24%	29%	25%	28%
Intentional leaks/sharing of data by staff	21%	22%	32%	12%	18%	21%	30%	18%	14%
Information leaked/inappropriately shared on a mobile device	20%	18%	30%	16%	18%	22%	27%	13%	11%
Security failure by third party supplier	16%	10%	25%	14%	15%	17%	23%	11%	10%
Fraud by employees	16%	17%	18%	11%	14%	18%	21%	15%	11%
None	17%	14%	9%	26%	19%	14%	11%	30%	27%

Data Leak Detection as a Service?

Threat model:
accidental leak; a DLP provider is a semi-honest adversary



Fuzzy Fingerprints



DLD Provider

Data Owner

2. Release fingerprints

1. Preprocess and prepare fuzzy fingerprints

3. Monitor outbound network traffic

4. Detect

5. Report all data-leak alerts

6. Postprocess and identify true leak instances

PPDLD

- 1) Pre-process
- 2) Release
- 3) Monitor
- 4) Detect
- 5) Report
- 6) Post-process

Data owner sends sensitive collections and content collections to DLD provider

N-gram generation & Rabin fingerprints



Detection of transformed accidental data leak?

Auto-formatting (WordPress)

The application layer contains the higher-level protocols used by most applications for network communication. Examples of application layer protocols include the File Transfer Protocol (FTP) and the Simple Mail Transfer Protocol (SMTP). [19] Data coded according to application layer protocols are then encapsulated into one or (occasionally) more transport layer protocols (such as TCP or UDP), which in turn use lower layer protocols to effect actual data transfer.

The application layer contains the higher-level protocols used by most applications for network communication. Examples of application layer protocols include the File Transfer Protocol (FTP) and the Simple Mail Transfer Protocol (SMTP). [19] Data coded according to application layer protocols are then encapsulated into one or (occasionally) more transport layer protocols (such as TCP or UDP), which in turn use lower layer protocols to effect actual data transfer.

Partial source code leak

```
def encode(msg, pubkey, verbose=False):
    chunksize = int(log(pubkey.modulus, 256))
    outchunk = chunksize + 1
    outfmt = '%%0%dx' % (outchunk * 2,)
    bmsg = msg if isinstance(msg, binary_type) else msg
    result = []
    for start in range_func(0, len(bmsg), chunksize):
        chunk = bmsg[start:start + chunksize]
        chunk += b'\x00' * (chunksize - len(chunk))
        plain = int(hexlify(chunk), 16)
        coded = pow(plain, *pubkey)
        bcoded = unhexlify((outfmt % coded).encode())
        if verbose:
            print('Encode:', chunksize, chunk, plain, c
    result.append(bcoded)
```

```
return b''.join(result).rstrip(b'\x00').decode('utf-8')

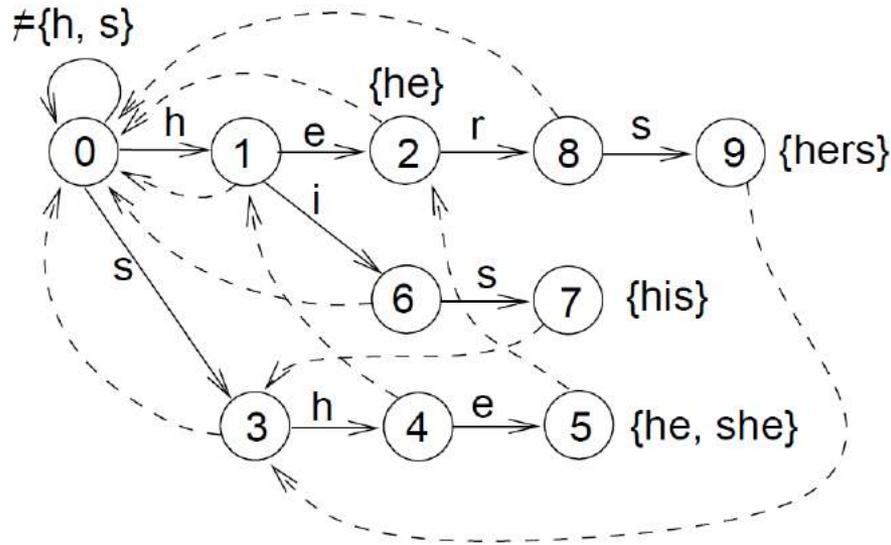
def __delitem__(self, item):
    self._remove_from_dict(item)
    self.heap = [(v,k) for v,k in self.heap if k != ite
    chunk += b'\x00' * (chunksize - len(chunk))
    heapq.heapify(self.heap)

def pop(self):
    _, smallest = heapq.heappop(self.heap)
    self._remove_from_dict(smallest)
    return smallest
```

How about string match? How about shorter n-gram?

Automata has some encoding flexibility, but ...

A keyword tree for $P = \{\text{he; she; his; hers}\}$



An automaton for *Hamlet*

- 4,042 lines
- 29,551 words
- Approximate 192,081 characters

Shorter n-grams increase **false positives** (i.e., accidental matches)

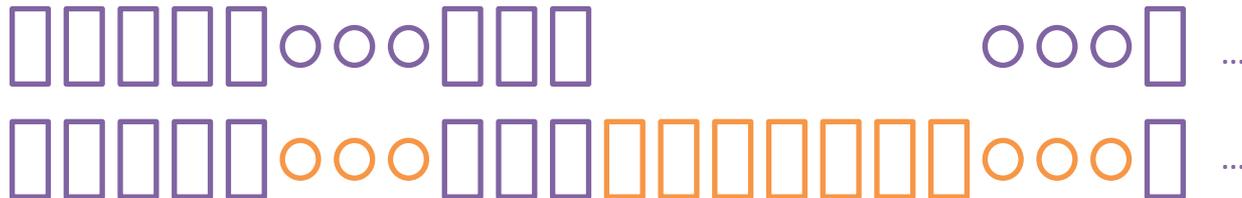
Transformed data leak – Our sequence-alignment based detection

between them.<ref>RFC 1812

["bet", "etw", "twe", "wee", "een", "en", "n t", "th", "the", "htm" ...]

[7872, 3408, 2410, 1961, 4238, 2383, 2158, 8131, 3271, 1833 ...]

Alignment
Result



[7872, 3408, 2410, 1961, 4238, 4340, 1666, 1008, 3271, 1833...]

["bet", "etw", "twe", "wee", "een", "en+", "n+t", "+th", "the", "htm" ...]

between+them.%26lt%3Bref%26gt%3BRFC+1812

Also invented a smart sampling algorithm

2 identical input streams:

```
1, 9, 4, 5, 3, 5, 9, 7, 6, 6, 3, 3, 7, 1  
1, 9, 4, 5, 3, 5, 9, 7, 6, 6, 3, 3, 7, 1
```

Output of random sampling:



```
1, -, 4, -, 3, 5, -, 7, -, 6, -, -, 7, 1  
-, 9, -, 5, -, 5, -, 7, -, 6, 3, -, -, 1
```

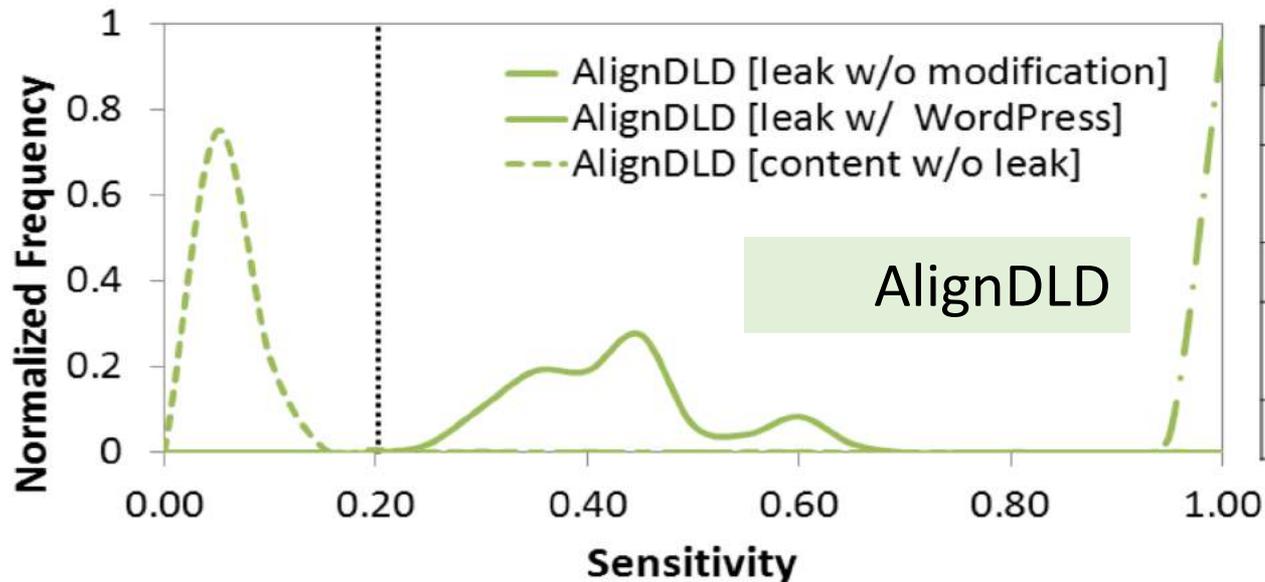
Output of our comparable sampling:



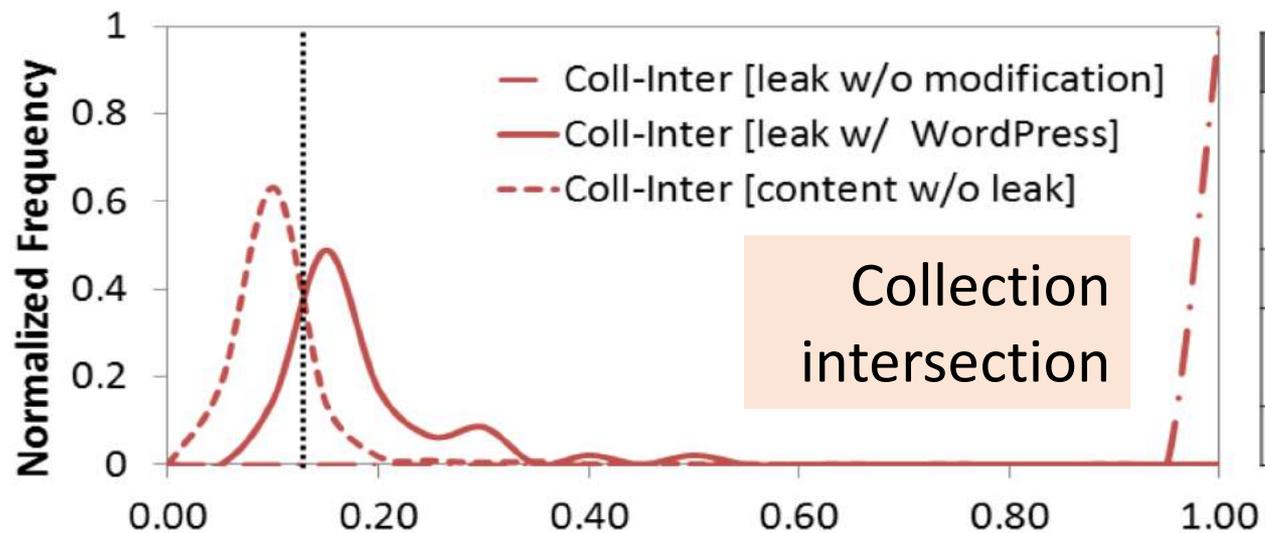
```
1, -, 4, -, 3, 5, -, -, -, -, 3, 3, -, 1  
1, -, 4, -, 3, 5, -, -, -, -, 3, 3, -, 1
```

If x is a substring of y , then x' (the sample of x) is a substring of y' (the sample of y).

Transformed leak stands out in AlignDLD



Best Threshold	
	0.2
Recall	
Leak w/ WordPress	100%
False Positive	
Content w/o leak	0.8%

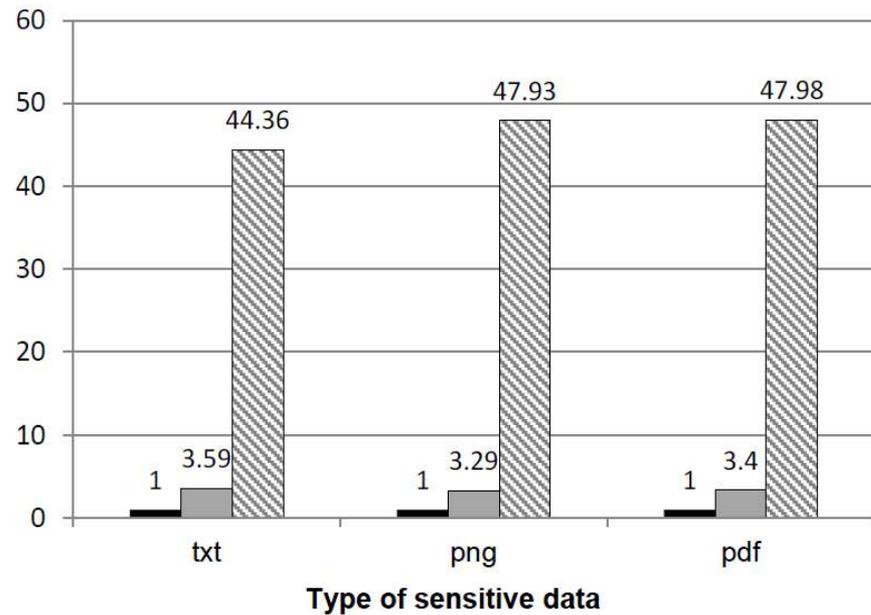


Best Threshold	
	0.14
Recall	
Leak w/ WordPress	63.8%
False Positive	
Content w/o leak	8.9%

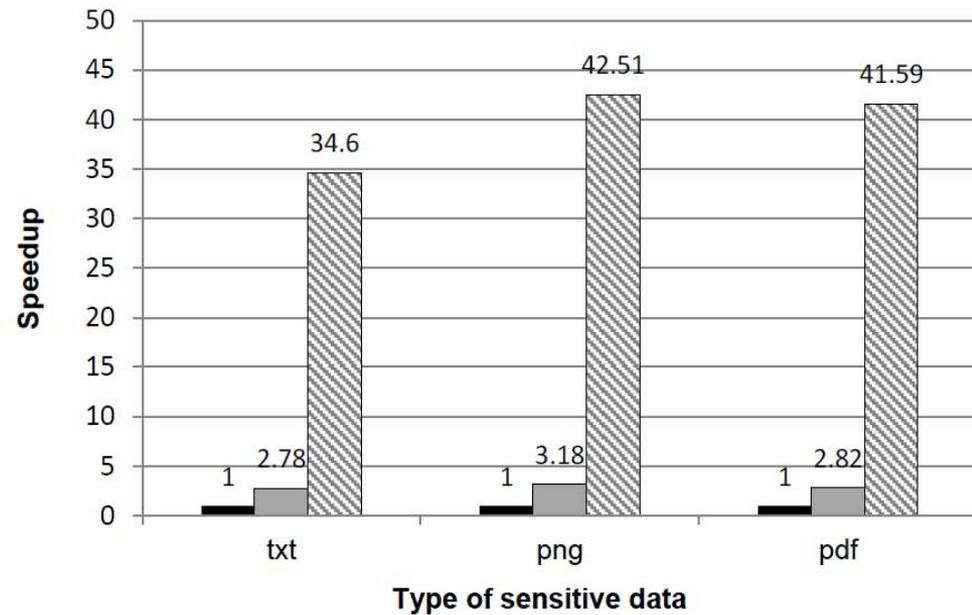
Enron dataset (2.6GB): 150 users, 517,424 emails. 3-grams.

GPU acceleration of AlignDLD

■ single-threading CPU ■ multithreading CPU ▨ GPU



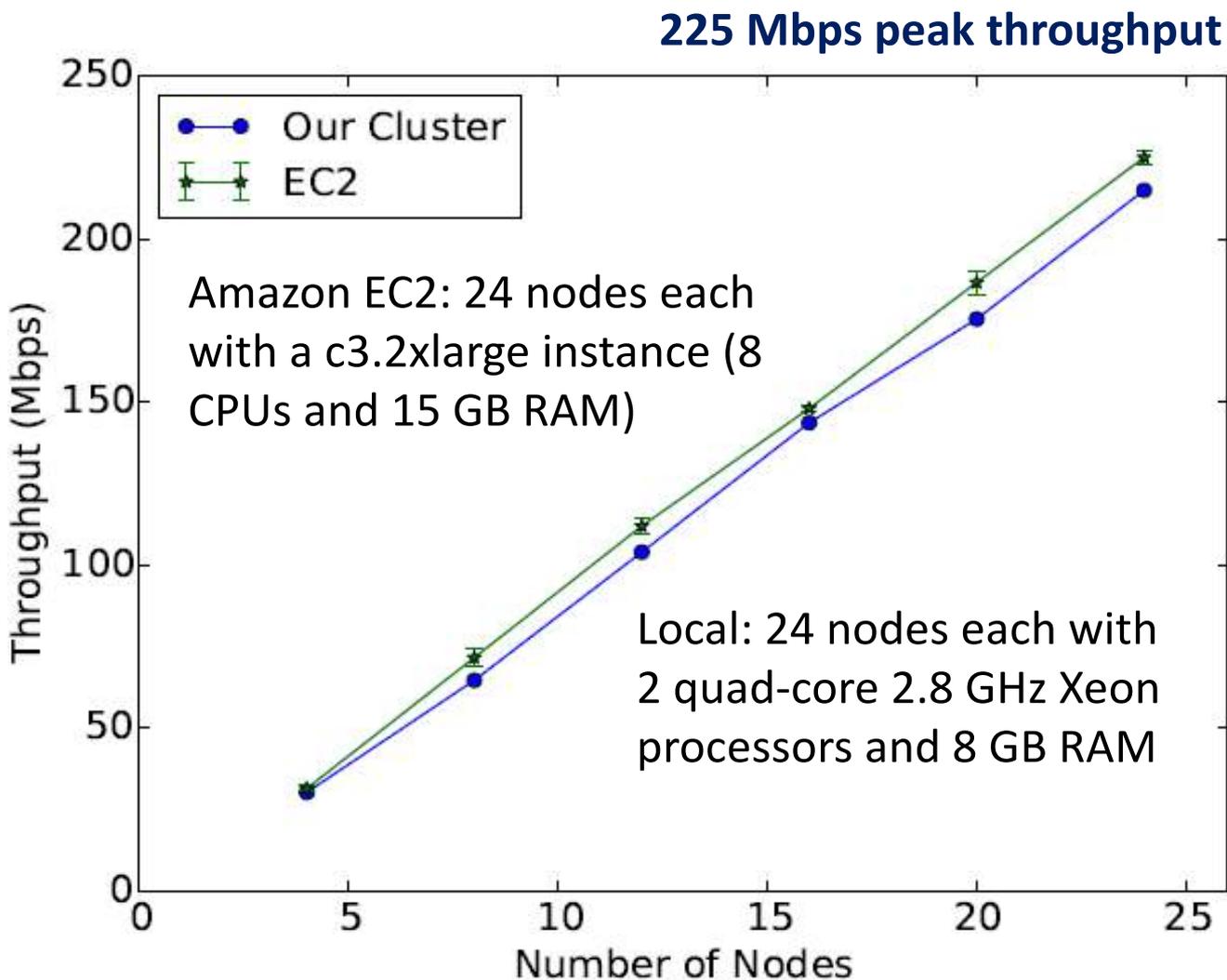
(a) Enron



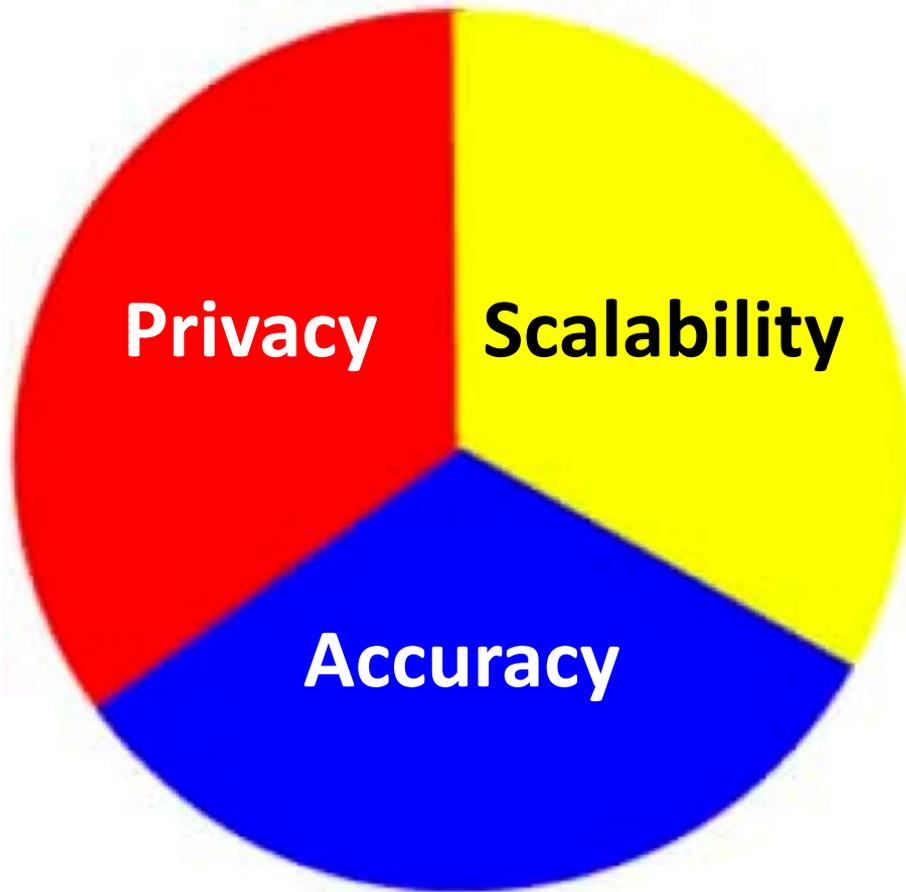
(b) MiscNet

Testing Platforms	# of Cores
CPU	
Intel Core i5 2400, Sandy-Bridge microarchitecture	4
GPU (single)	
Nvidia Tesla C2050, Fermi architecture	448

Hadoop (distributed hashtable) implementation



37 GB Enron Email Corpus as content, 10 papers as sensitive data



System-side solutions:
E.g., whole-system data
provenance tracking
-- Adam Bates, UIUC

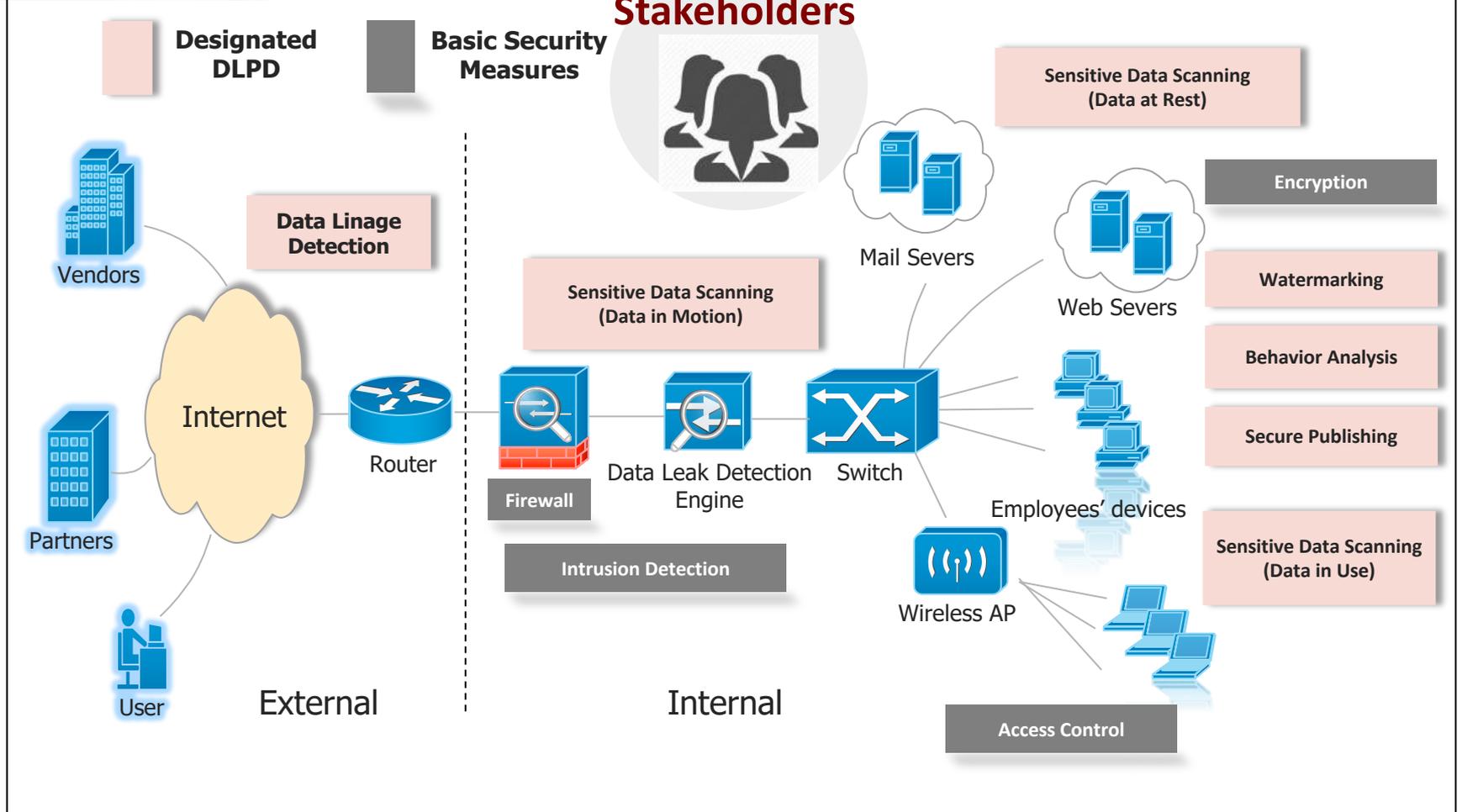
Data management papers in SACMAT '18!

Many opportunities to make impact



Security analysts
Developers Consumers Operators

Stakeholders



How much science in cybersecurity industry?



IBM X-Force Command Center

<https://www.youtube.com/watch?v=sHrgVqKW1RQ>



Questions?