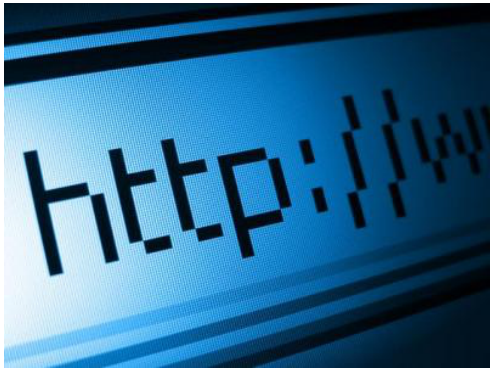
HTTP(S)-Based Clustering for Assisted Cybercrime Investigations

Marco `embyte` Balduzzi, *Vincenzo Ciangaglini and Robert McArdle*

TREND MICRO™

Ingredients



Who am I?

- Italian, (?)
- M.Sc. in Comp. Engineering, Ph.D. in System Security
- 10+ years experience
- Now with Trend Micro Research
- Bridge academia and industry

@embyte

RoadMap

- Intro
- Target Attacks
- Detection
- System Overview
- *SPuNge*
- Experiments
- Conclusions

Security is hot.. is burning!

- Symantec's Internet Security Threat Report:
- Spam volume is decreased, but...
- Web-based attacks increased **30%**
- 5,291 new vulnerabilities discovered in 2012
- The number of phishing sites spoofing social networks increased 125%
- **42%** increase in targeted attacks in 2012

Targeted

- Shift from a world dominated by widespread malware that infects indiscriminately, to a more **selectively targeted approach**

- Just-for-fun era is over?
- Espionage, nation-driven, criminal organizations
- Specific target / industry, e.g. the energy sector

Edward Snowden: U.S., Israel 'Co-Wrote' Cyber Super Weapon Stuxnet

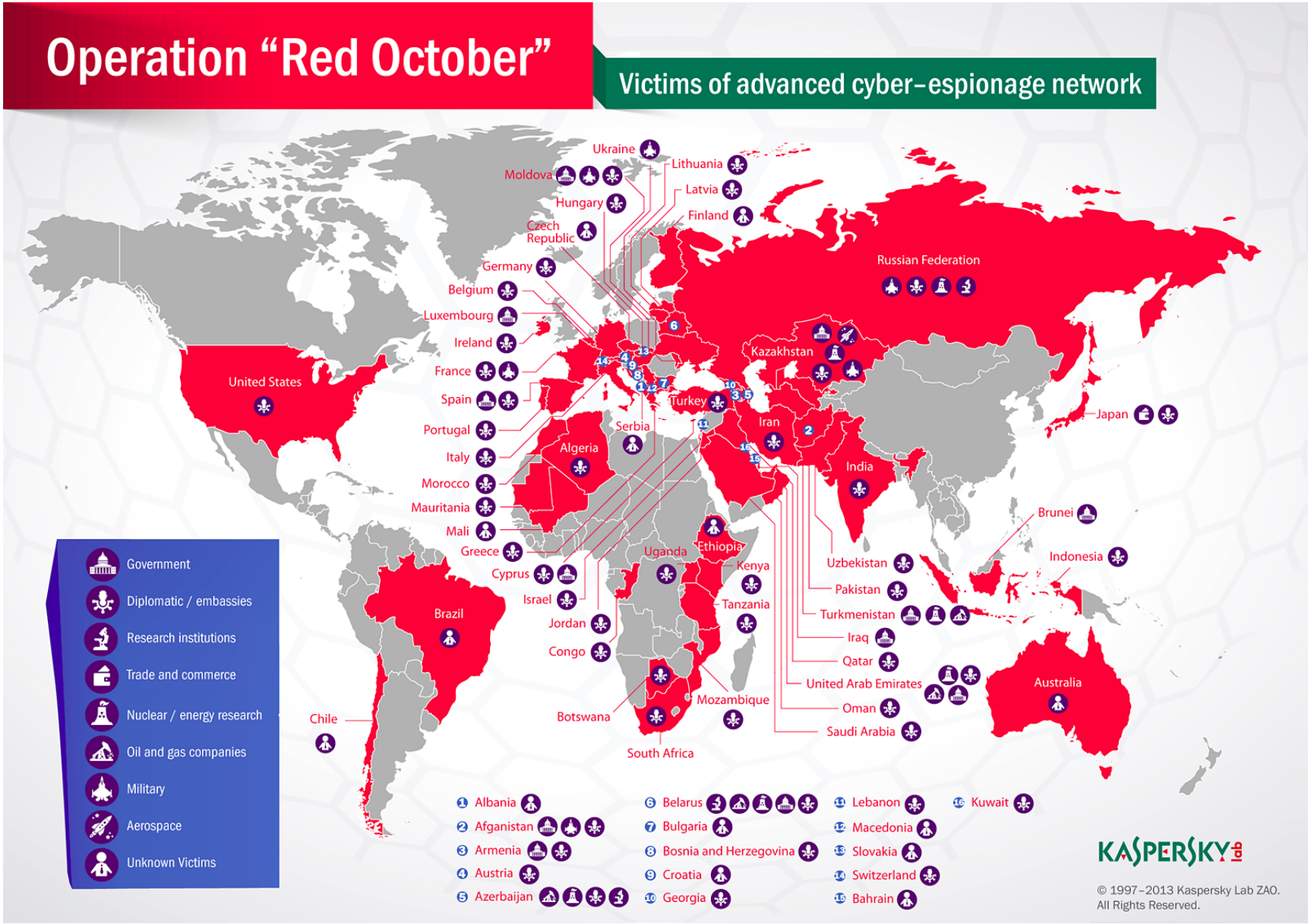
By Lee Ferran @leeferran
Jul 9, 2013 2:22pm

Kirit Radia @KiritRadia_ABC

The former National Security Agency contractor on the run from U.S. authorities halfway around the world said that Stuxnet, an unprecedented cyber weapon that targeted Iran's nuclear program, was the product of a joint American-Israeli secret operation.

Before Edward Snowden became a household name, he conducted an interview via encrypted emails with cyber security expert Jacob Appelbaum and was asked about the game-changing computer code, according to the interview published in the German newspaper Der Spiegel Monday.

"NSA [U.S. National Security Agency] and Israel co-wrote it," Snowden said.


Operation “Red October”
Victims of advanced cyber–espionage network
Ukraine
Moldova
Lithuania
Latvia
Hungary
Finland
Czech Republic
Germany
Belgium
Luxembourg
Ireland
France
Spain
Portugal
Italy
Morocco
Mauritania
Mali
Greece
Cyprus
Israel
Jordan
Congo
Botswana
South Africa
Mozambique
Chile
Brazil
United States
Algeria
Serbia
Turkey
Russian Federation
Kazakhstan
Iran
Japan
India
Brunei
Indonesia
Uzbekistan
Pakistan
Turkmenistan
Iraq
Qatar
United Arab Emirates
Oman
Saudi Arabia
Australia
Ethiopia
Uganda
Kenya
Tanzania
Government
Diplomatic / embassies
Research institutions
Trade and commerce
Nuclear / energy research
Oil and gas companies
Military
Aerospace
Unknown Victims
1 Albania
2 Afganistan
3 Armenia
4 Austria
5 Azerbaijan
6 Belarus
7 Bulgaria
8 Bosnia and Herzegovina
9 Croatia
10 Georgia
11 Lebanon
12 Macedonia
13 Slovakia
14 Switzerland
15 Bahrain
16 Kuwait
KASPERSKY lab
© 1997–2013 Kaspersky Lab ZAO.
All Rights Reserved.

NEWS

Targeted cyber attacks cost up to £1.6m

Warwick Ashford
Thursday 25 July 2013
09:37

Targeted cyber attacks could cost up to £1.6m, the 2013 Global Corporate IT Security Risks survey by B2B International and security firm Kaspersky Lab has revealed.

According to the report, £1.4m stems directly from the incident itself in losses from critical data leakages, business interruptions and expenses for remediation specialist services.

Companies face an additional bill of about £146,000 for actions taken to prevent such incidents from taking place again in the future, including updating software and hardware, and hiring and training staff.

Company losses resulting from targeted attacks on small and medium enterprises (SMEs) are lower, at around £60,000 per incident.

How to detect them?

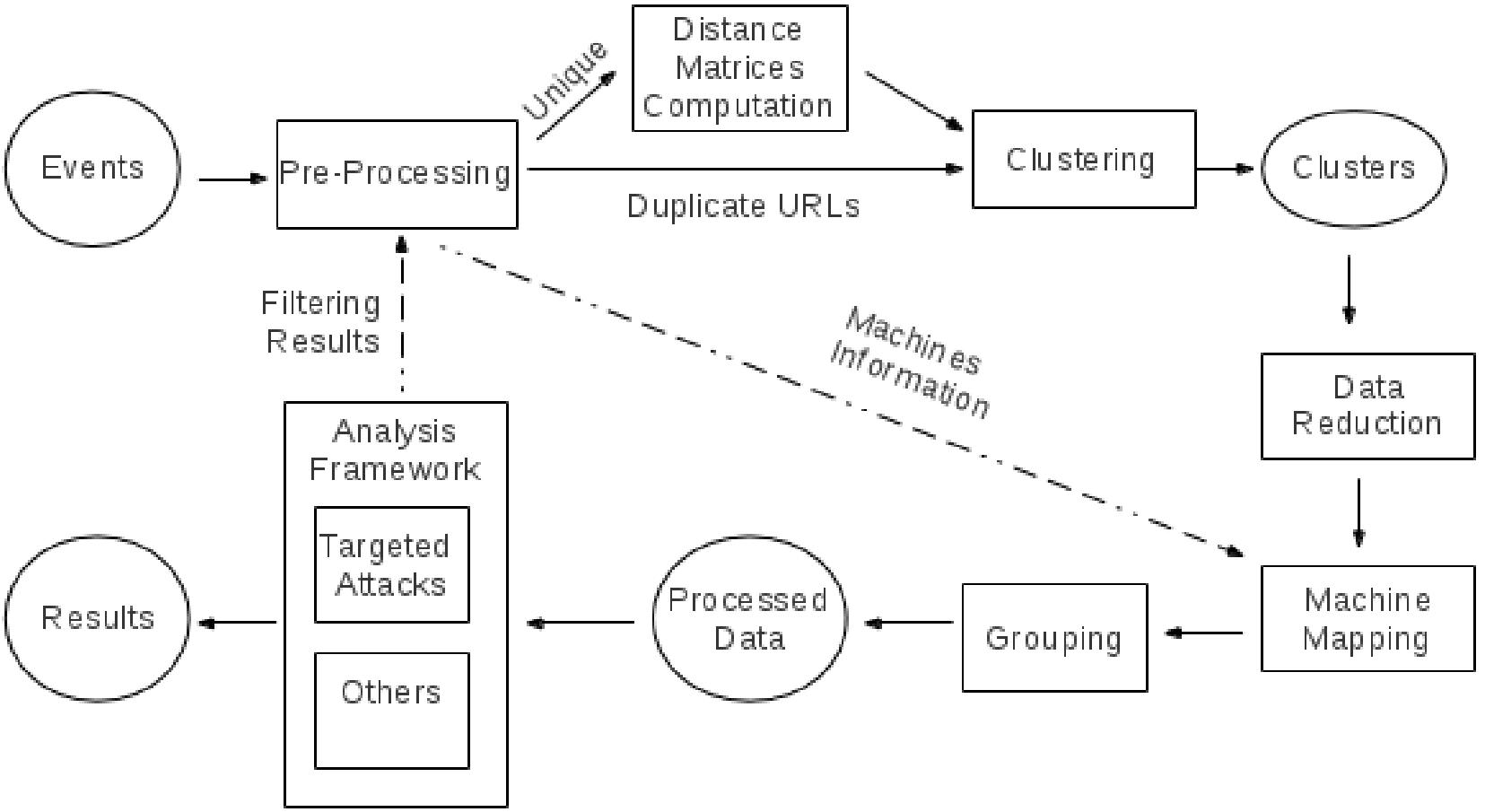
- Difficult to detect
- Generic detection → difficult to distinguish from “traditional” widespread attacks
- **Same techniques**, different methodologies

- Assist cybercrime investigations
- To **reduce** the number of normal incidents down to a more manageable amount for **further in-deep analysis**.

Idea!

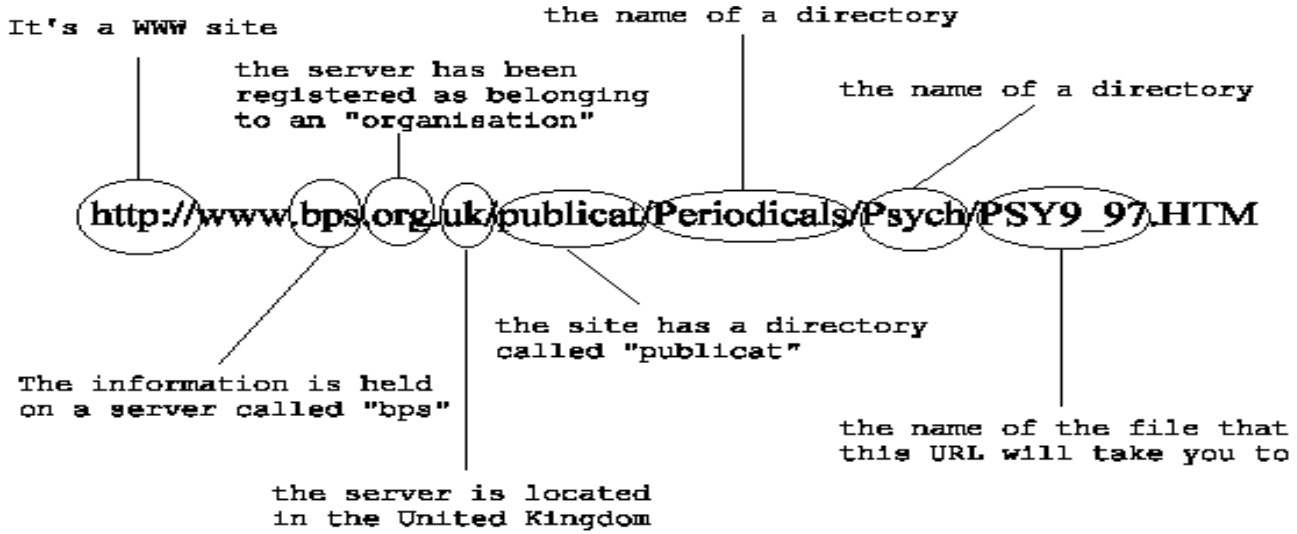
- Identify groups of similar machines
- Share a **common network behavior**
  - With respect to the **malicious resources** they access/request
  - e.g. exploit kits, drive-by-downloads, C&C servers
- Correlate **location** and **industry** information
- Build “context”

General Overview



Working Data

- HTTP & HTTPS network traces
  - Population of ~20,000,000 installations
- Collected at **proxy-level**, client-side
- Already-known malicious URLs
  - Drive-by / web-based malware, fakeAVs, C&C servers, etc...

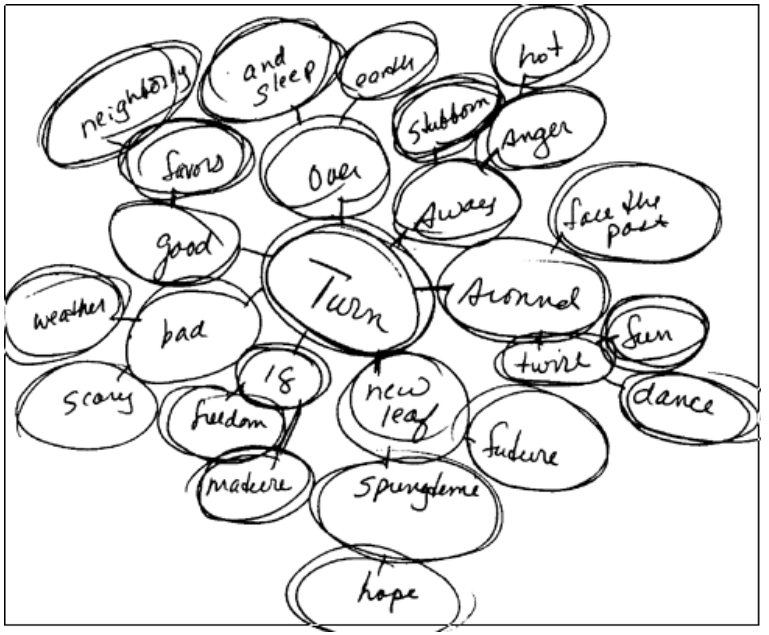


Pre-Processing

- Classification: Ignore parental controlled URLs
- Network sampling: Keep a single “candidate” event per network (Class B)
- Event sampling: Remove multiple identical requests from single machine
  - E.g. Botnet -controlled machines
- Duplicates identification: Remove URLs widely requested (e.g. >50 networks) → Widespread
- Whitelisting: Remove entries known to be useless (by previous iterations)

Step 1: Clustering

- Given a set of arbitrary elements, *without prior information*, identifies and assigns the elements to **groups** (called clusters)
- **Patterns** in the collected data (URLs)
- Group malicious URLs according to similar Hostname or Request (Path + Query String) – or both



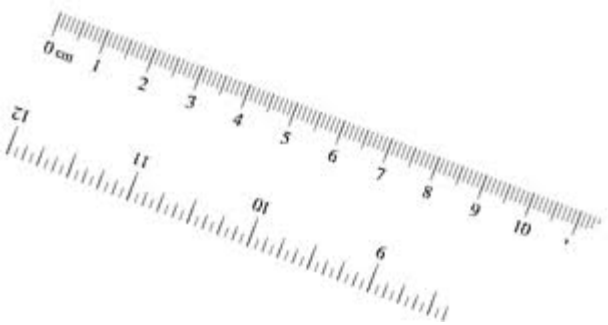
Host vs Request Clustering

	cr5aigslist.com	craigsli8st.com	crauglist.com	craeglist.com	google.com
cr5aigslist.com	0	0.0666	0.1428	0.1428	0.520
craigsli8st.com	0.0666	0	0.1428	0.1428	0.520
crauglist.com	0.1428	0.1428	0	0.0769	0.478
craeglist.com	0.1428	0.1428	0.0769	0	0.478
google.com	0.520	0.520	0.478	0.478	0

TABLE III. EXAMPLE OF DISTANCE MATRIX FOR HOSTNAMES (NORMALIZED LEVENSHTEIN).

Exploit Kit	URL's Host	URL's Request
Blackhole	http://77.79.13.88	/content/w.php?f=52&e=4
Blackhole	http://188.127.249.241	/image/l.php?f=553&e=2
Blackhole	http://brown.mydomxd.org	/root/w.php?f=2293&e=6
Nuclear	http://zeak.rghil.info	/a456gh/9493af39692e[...].jar
Nuclear	http://163.1.32.2	/1rg54e/55c2b44e0c8a[...].jar
Nuclear	http://31.184.244.9	/6ju9a2/bb136b125774[...].jar

TABLE II. EXAMPLE OF URLS USED BY THE BLACKHOLE AND NUCLEAR EXPLOIT KITS AND DETECTED WITH SPUNGE.

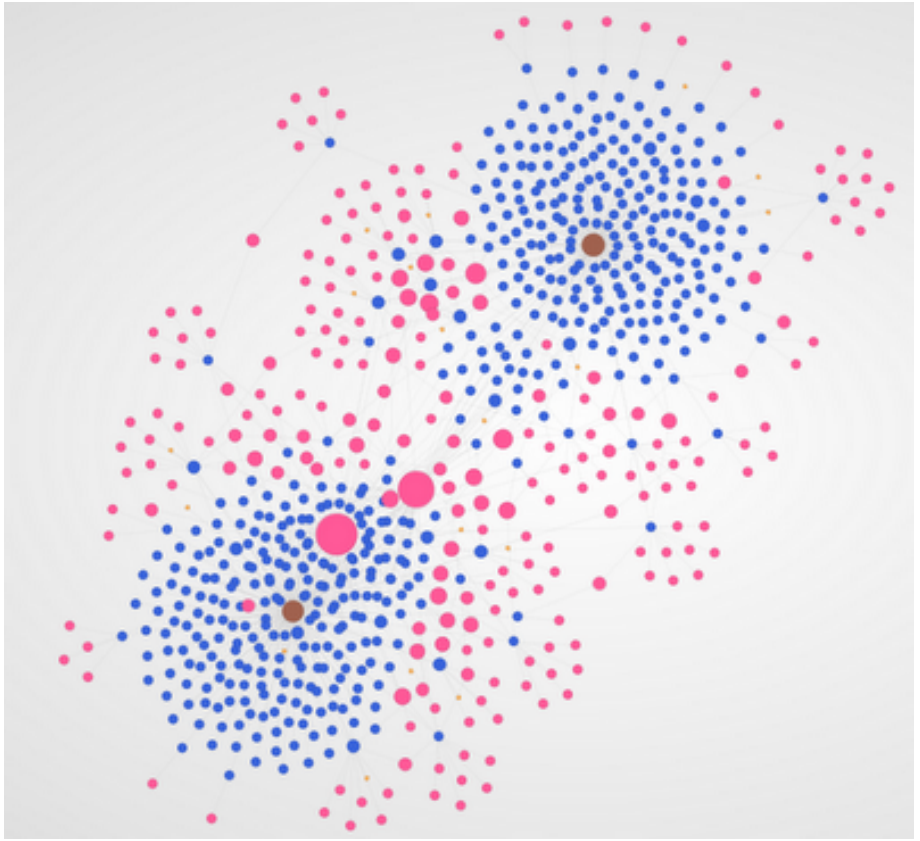


Distance Function

- Hostname
  - Levenshtein = distance between strings
  - Robert → Roger : Ro**b**ert → Ro**g**ert, Roger**t** → Roger
- Request
  - Path: Levenshtein
  - Query String: Jaccard
    - # parameters in common (ignore values)
  - *http://[hostname]/path1.php?a=10&b=20&c=30*
  - *http://[hostname]/path2.php?a=100&b=200*

$$d_{req}(e_1, e_2) = \sqrt{d_{path}(e_1, e_2)^2 + (WeightFactor \times d_{qsl}(e_1, e_2))^2}$$

Bubble View?



- Red = Hostname
- Blue = Request

- Violet = Both (?)

Step 2: Labeling and Merging

- Merge “similar” clusters, subsets
- Assign label to clusters (H/R)

Clusters	Cluster Label	Event	URL
C_1	H zfmudav4aaq33r5.com >: R /get2.php?c=BLMEUGUBd=266 >: R /CZ4ODY9MzImdHA9MCZmbD0w0	e_1 e_2 e_3 e_4	zfmudav4aaq33r5.com/get2.php?c=BLMEUGUBd=266 zfmudav4aaq35r5.com/get.php?c=ZLXULJNRd=266 zfmudav3aap36r5.com/CZ4ODY9MzImdHA9MCZmbD0w0 zfmudav2acq35r4.com/CZ4ODY9MzImdHB9MCZmbD0w1
C_2	H facebookc.com	e_5 e_6 e_7 e_8	facebookc.com facaebook.com faceboook.com facebopok.com
C_3	H h-aelameftzgj4vxient.com =: R /qKA0rO4d8I7qBhS7Y2xrPTQu	e_9 e_{10} e_{11} e_{12}	h-aelameftzgj4vxient.com/qKA0rO4d8I7qBhS7Y2xrPTQu h-aelameftxcd5vxient.com/IkG1yP3L8q5YPtU7Y2xrPTQu h-aelameftssd6vxient.com/BAq3T78d8I5Q7bs0Y2xrPTQu h-aelanfftzgj1vxient.com/pA71gKND6P5MTIs9Y2xrPTQu

Step 3: Machines Mapping

- Map URLs into machines → IP addresses

Cluster	Cluster Label	Event	Source Machine
C_1	H zfmudav4aaq33r5.com >: R /get2.php?c=BLMEUGUBd=266 >: R /CZ4ODY9MzImdHA9MCZmbD0w0	e_1 e_2 e_3 e_4	M_1 M_2 M_3 M_4
C_2	H facebookc.com	e_5 e_6 e_7 e_8	M_1 M_2 M_5 M_6
C_3	H h-aelameftzgj4vxient.com =: R /qKA0rO4d8I7qBhS7Y2xrPTQu	e_9 e_{10} e_{11} e_{12}	M_3 M_4 M_5 M_7

- Exercise:
  - M1 to which cluster belongs to? M2?

Step 4: Grouping

- Identify machines that belong to the same cluster (**>=1**).
- Machines that share a **similar malicious behavior**
- Scenario: Drive-by-download infection
  - 1. The victim is redirected to the malicious page
  - 2. Served with the right exploit.
- **2** Clusters

Step 4: Grouping

- Looking for similar victims
- Groups of machines (IPs) and clusters (URLs)

Source Machine	Clusters
M_1	C_1, C_2
M_2	C_1, C_2
M_3	C_1, C_3
M_4	C_1, C_3
M_5	C_2, C_3
M_6	C_2
M_7	C_3

Groups	Machines Set	Clusters Set
G_1	M_1, M_2	C_1, C_2
G_2	M_3, M_4	C_1, C_3
G_3	M_5	C_2, C_3
G_4	M_6	C_2
G_5	M_1, M_7	C_3

Last step: Analysis & Reporting

- Correlation: **industry, country, etc...**
- Two type of analysis:
  - **Clusters**: N+ machines, operating in the same industry or country, reaching our a cluster of similar URLs (1 cluster)
  - **Groups**: N+ machines sharing C+ clusters
- 2 <= N,C <= 5
- Automated reporting for threat analysts

Findings

Experiments settings

- Python 2.7 prototype, multi-core
- Process data in daily batch (nighttime)
- Two machines: Processing and Final Analysis
- Evaluation run over 1 week of data

# of	Sun. 11	Mon. 12	Tue. 13	Wed. 14	Thu. 15	Fri. 16	Sat. 17
Raw Events (Million)	2.792	5.170	5.584	5.685	5.225	4.911	2.628
Events	387,339	536,524	256,270	221,954	230,758	269,103	329,458
Clusters	4,106	8,825	8,195	7,825	7,196	7,281	3,869
Machines	10,866	15,581	15,413	15,391	14,165	14,364	8,406
Groups	2,144	3,941	3,579	3,528	2,679	2,896	1,069

```
Cluster 7543 - H 146.185.246.116 >:R /p98a.exe >:R /dd.exe

http://146.185.246.111/p98a.exe    NET 1   notepad.exe 2012-11-13 09:50:35
http://146.185.246.116/p18a.exe    NET 1   notepad.exe 2012-11-13 09:50:37
[...]
http://146.185.246.121/mailsa.exe NET 1   notepad.exe 2012-11-13 09:50:24
http://146.185.246.101/lmqa.exe    NET 1   notepad.exe 2012-11-13 09:50:26
http://146.185.246.63/dd.exe       NET 2   svchost.exe 2012-11-13 11:45:27
http://146.185.246.63/dd.exe       NET 3   svchost.exe 2012-11-13 20:58:55
http://146.185.246.104/dqs.exe     NET 1   notepad.exe 2012-11-13 09:47:36

NETWORK 1        Technology        Mexico    Windows 5.1
NETWORK 2        Technology        Turkey    Windows 5.1
NETWORK 3        Technology        Morocco   Windows 5.1
```

Listing 1.1. RBN Example - Technology Industry

- Victims:
  - 3 Organizations, 3 distinct countries
  - Operating in the technology sector (manufacture)
- Malware injection into memory space to avoid easy detection (persistent)
- Netblock → Russian Business Network, known to provide support for targeted attacks

```
Group 1245, 2 Clusters, 2 Networks

Cluster 1725,Label:R /list.php?c=140C3[...] =:H w.nucleardiscover.com:888
E1: http://w.nucleardiscover.com:888/list.php?c=140C34E31DAB3B9746[...]&t=0.689831&v=2
E2: http://w.nucleardiscover.com:888/list.php?c=D8C08B5CD1670FA396[...]&v=1&t=0.9288141

Cluster 1932, Label:R /gggg_r.jpg?t=0.1424164
E1: http://61.147.99.179:81/gggg_r.jpg?t=0.1424164
E2: http://ru.letmedo.net:2011/myck.jpg?t=0.3245672

NETWORK1:    Oil and Gas    Malaysia    Windows 5.1 r18nwn.exe 2012-11-14
NETWORK2:    Oil and Gas    Malaysia    Windows 5.1 r18nwn.exe 2012-11-14
```

Listing 1.2. Example of Cluster Group - Oil&Gas Industry.

- Victims:
  - 2 Organizations, Malaysia
  - Oil&Gas Industry
- C&C servers reached out by *r18nwn.exe*
- Malware for Industrial Environments
- Domains → Registered by a person in China, associated with Targeted Attacks Operations

Conclusions

- Increasing number of Targeted Attacks
- Difficult to spoil, similarities with traditional attacks

- SPuNge: Clustering-based techniques to identify *potential* targeted attack from threat data
- Future work
  - On-line processing
  - GPU-assisted processing
  - Enhance clustering, more features (e.g. process name, hash)

Thanks!