Security monitoring in Internet: the use case of Phishing

Jérôme François
jerome.francois@inria.fr

MADYNES
Nancy – Grand Est
Research Centre

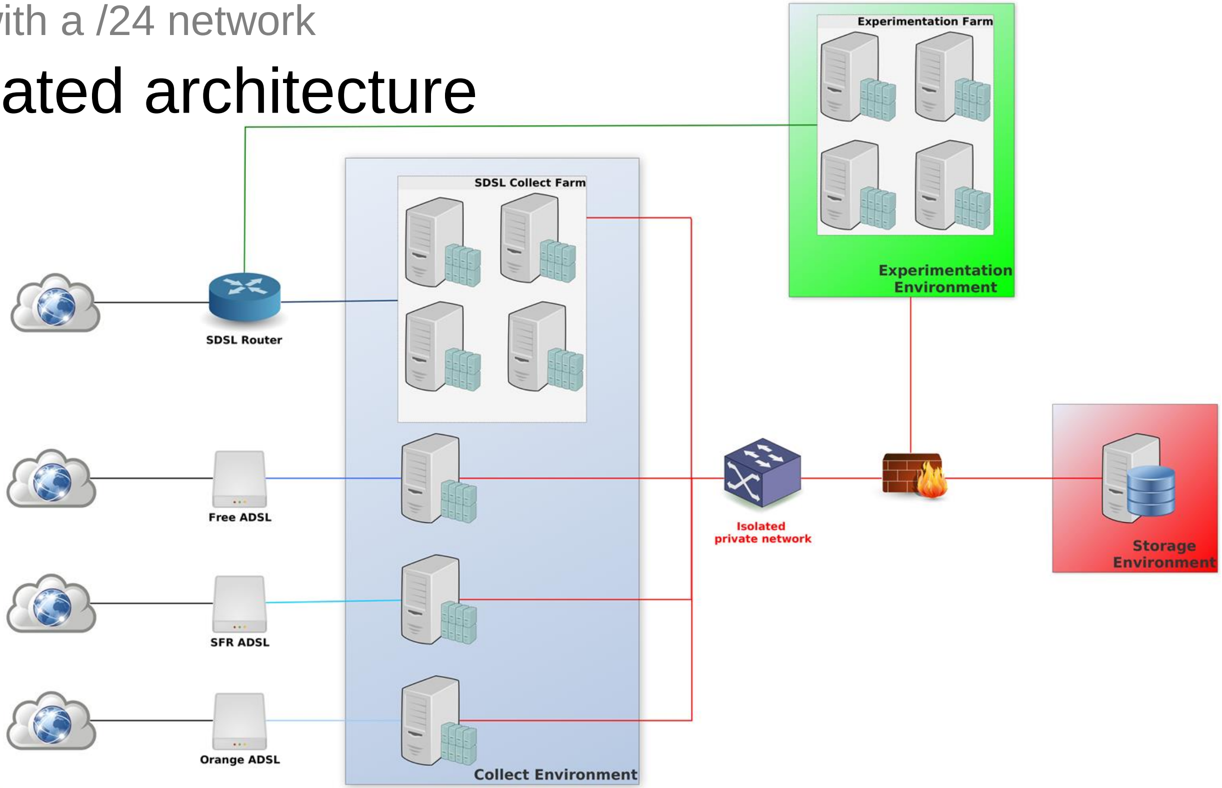
1

Overview

How we monitor security in Internet?

Network telescope

- Objective: collect attack traces from Internet without being seen as a research institute
  - Multi-provider architecture
    - 3 public ADSL with different providers
    - 1 SDSL 2Mbits with a /24 network
  - Virtual and isolated architecture



Honeypots and sensors

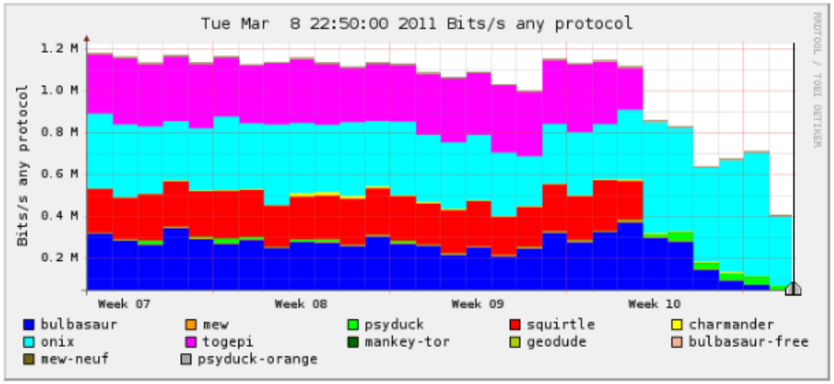
- Being attacked and monitor them
  - Expose vulnerabilities (honeypots)
    - **1 instance of each deployed** in the current deployment
    - **Dionaea:** RPC/Netbios, HTTP, FTP/TFTP, SIP/VoIP, MSSQL
    - **Amun:** Vulnerabilities emulated via python plugins
    - **Kippo:** Brute-force SSH always works and access to minimalistic shell sessions and brute-force attempts are logged
    - **Conpot:** ICS/SCADA Honeypot
    - **Glastopf:** WEB applications honeypot
  - + monitoring sensors
    - **Snort + snort_hpfeeds:** Intrusion detection on the whole SDSL /24 IP range, Collector for shipping snort alerts using hpfeeds
    - **Network Traffic**: PCAP, Netflow
    - Syslog

Some numbers

- **Operational since the 09th of September 2008**

- **Total (29/10/2014)**
  - 901 832 393 attacks
  - **368 984 073 malicious attacks**
  - 38 878 269 malwares captured
  - **301 013 unique binaries**

- **Daily (on a 800 Kbit/s bandwidth)**
  - 500 000 attacks - 300 000 malicious
  - **25 000 binaries captured**

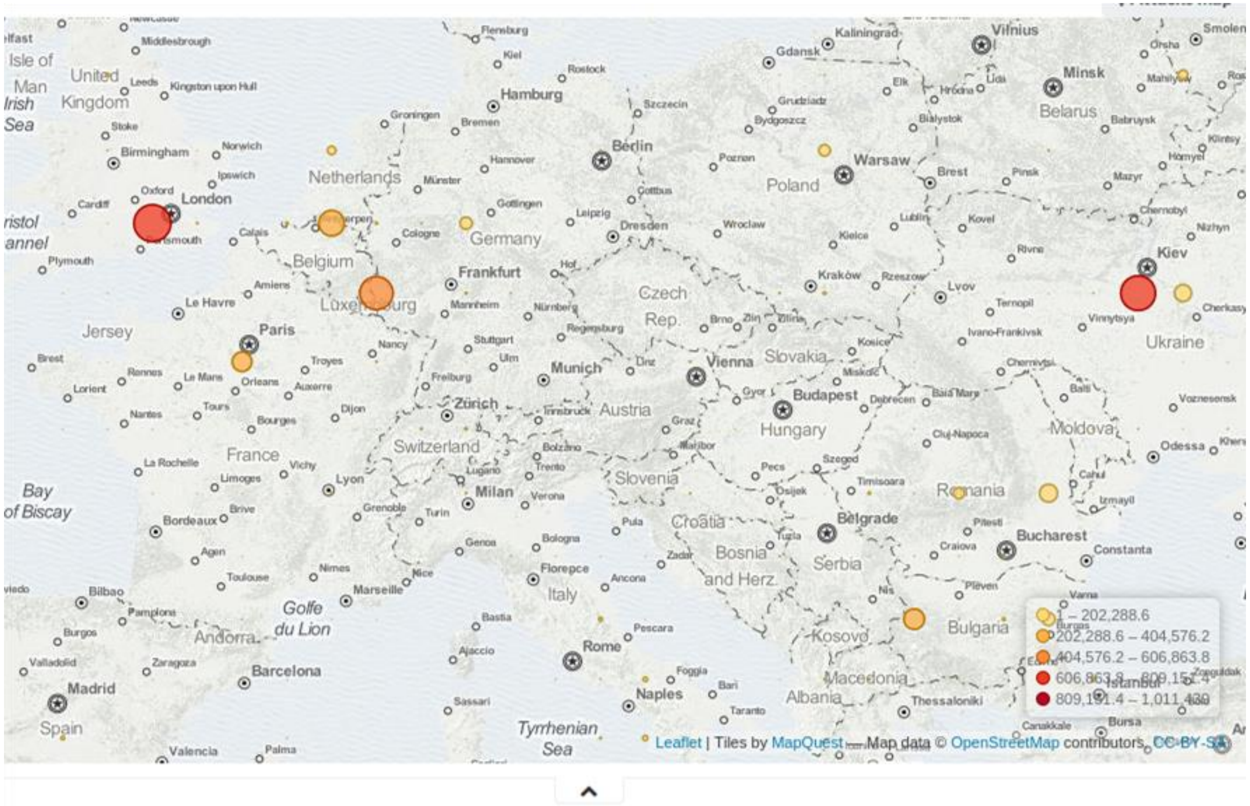


- **Network traces**
  - **15 To of PCAP traces**
  - **240 Go of NetFlow flows (v5 et v9)**
  - **6 Go of anonymized Tor flows**

Dashboard

password	Count
123456	7320
!@	5470
password	3641
1234	2481
ubnt	2071
12345	1707
123	1673
	1384
test	1375
1	1243
admin	1120
qwerty	1109
123qwe	1059

- **Geographic location of attacks**



- **Most used SSH passwords**

2

Towards proactive monitoring

The use case of phishing

A Joint work with the Univ. of Luxembourg – SnT (Samuel Marchal, Radu State, Thomas Engel)

What is Phishing ?

- Use of technical subterfuges and social engineering to steal any kind of valuable Internet users' data:
- Cause billions of dollars of loss every year
- Blacklists exist but updates might appear too late
  - Unknown URL → predict in advance them
  - URL verification in progress → speed up the process

Phishing URLs characteristics

www.paypal.creasconsultores.com/www.paypal.com/Resolutioncenter.php

shevkun.org/css/paypal.com/cgi-bin/cmd%3D_login-submit/css/websc.php

us-mg6.mail.yahoo.com.dwarkamaigroup.com/Yahoo.html

emailoans.hostingventure.com.au/bankofamerica.com

nitkowski.pl/components/wellsfargo/questions.php

The registered domain has no relationship with the rest of the URL

http://4ld.3ld.mld.ps /path1/path2?key1=value1&key2=value2

- Most parts of URLs can be freely defined
- Except the registered domain: main level domain + public suffix

Proposition for phishing URLs detection

Assumptions:

- Components of legitimate URLs are all related
- Registered domains (mld.ps) of phishing URLs are not related to the remaining of the URL
- URL vocabulary ~ Internet vocabulary: differs from natural text

→ **Analyse relatedness between *mld.ps* and the remaining part of a URL : Intra-URL relatedness**

URL splitting

URL label extraction:

http://4ld.3ld.mld.ps/path1/path2?key1=value1&key2=value2

"mld" & "mld.ps"

Basic splitting

login.paypal.com/securepayment

- RD_{url} = {paypal; paypal.com}
- REM_{url} = {login; secure; payment}

Intra-URL relatedness evaluation

sezopostos.com/paypalitlogin/us/websrc.html?cmd=_login-run

URL label extraction

$RD_{url} = \{sezopostos, sezopostos.com\}$

$REM_{url} = \{paypal, it, login, us, web, src, html, cmd\}$

Search engine query data
***Term* computation**

$Term_{paypal} = \{\{amazon, paypal\}, \{paypal, fees\}, \{ebay, uk\}, \{paypal, login\}\}$

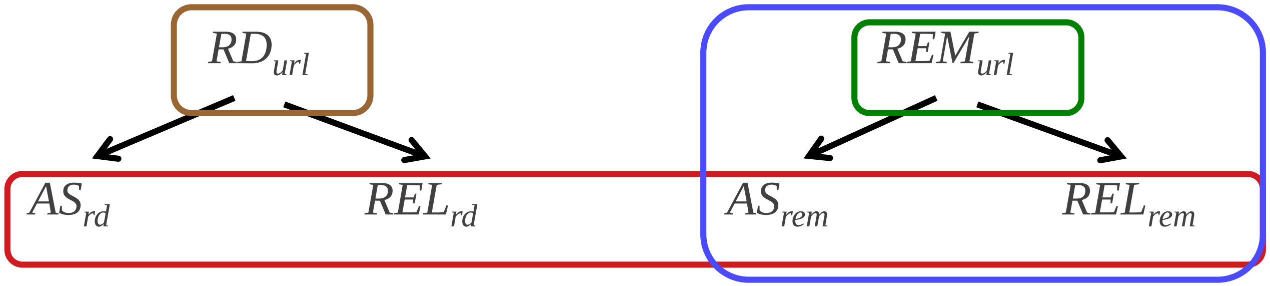
$AS_{rem} = \{amazon, fees, login\}$

$REL_{rem} = \{amazon, paypal, fees, ebay, uk, login\}$

AS_{rd} ***Associated* words**

REL_{rd} ***Related* words**

Features set



Word set relatedness
(Jaccard index)

J_{RR} J_{RA} J_{AA}

J_{AR} J_{ARrd} J_{ARrem}

Popularity of words in URL

$ratio_{Arem}$

$ratio_{Rrem}$

Words embedded in URL

$card_{rem}$

Popularity of the registered domain

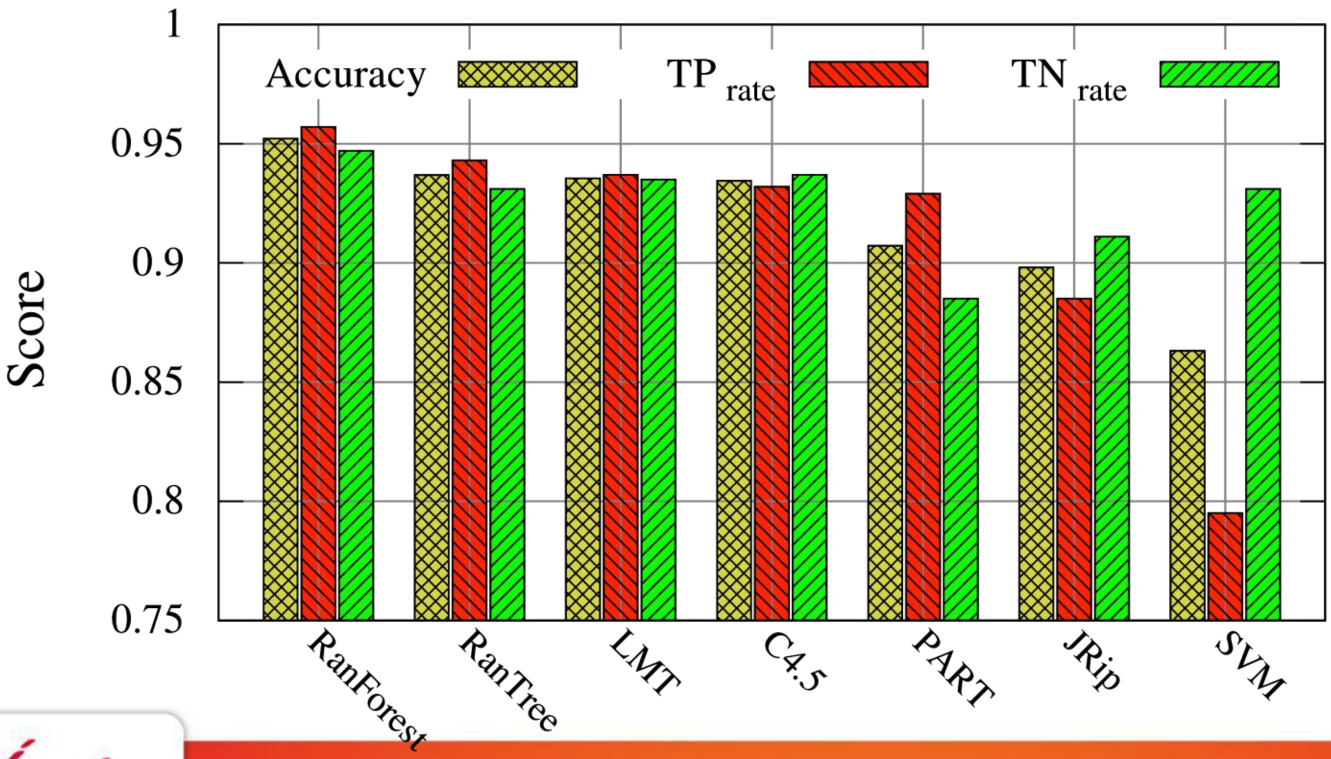
mld_{res}

$mld.ps_{res}$

$ranking$

URL classification

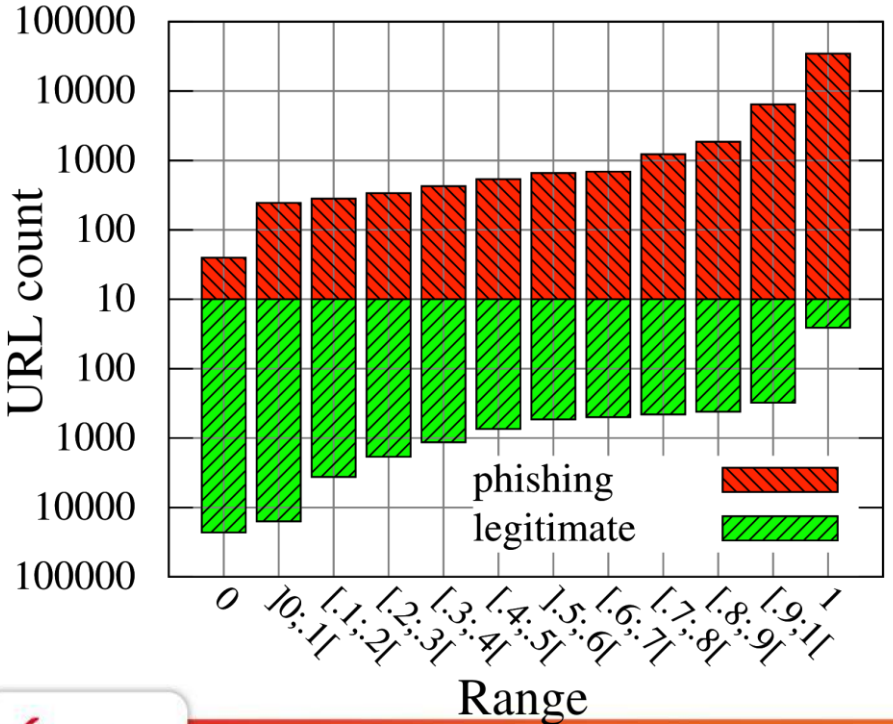
- Machine learning approach:
  - 48,009 phishing URLs (source: PhishTank)
  - 48,009 legitimate URLs (source DMOZ)
  - Determine the best classifier to identify phishing URLs
  - 7 classifiers tested: Random Forest, C4.5, JRip, SVM, etc.
  - 10-fold cross-validation on the presented feature set (96,016 URLs)



- Random Forest:
  94.91% accuracy
  1.44% FP_{rate}

URLs rating

- 7 classifiers tested: Random Forest, C4.5, JRip, SVM, etc.
- 10-fold cross-validation on 96,016 URLs (legitimate / phishing)
- Random Forest based rating system:
  - Strong decision: 95.66% accuracy
  - Processing time < 1 sec/URL



- **0:** 22,863 legitimate // 40 phishing
- **1:** 26 legitimate // 34,790 phishing

99.89% accuracy on 60.11% of the dataset

- [0;0.1] and [0.9;1]

99.22% accuracy on 83.97% of the dataset

3

Conclusion

Conclusion

- Semantic analysis is not always fully discriminative
    - URL rating system: >99% accuracy for > 80% URLs
    - Guide URL verification
        - Need to be coupled with more in-depth analysis of web page content (code inspection, binary download, visual perceptions, etc)
        - our approach ~ a filter to focus (and so speed up the analysis)

- References

- *PhishScore: Hacking phishers' minds*. CNSM 2014
- *PhishStorm: Detecting Phishing With Streaming Analytics. IEEE Transactions on Network and Service Management* (2014)

Security monitoring in Internet: the use case of Phishing

Jérôme François
jerome.francois@inria.fr

MADYNES
Nancy – Grand Est
Research Centre