Big data security on .nl: infrastructure and one application

Giovane C. M. Moura, Maarten Wullink,
Moritz Mueller, and Cristian Hesselman
SIDN Labs
first.lastname@sidn.nl

IEPG Meeting – IETF 94
Yokohama, Nov. 1st, 2015

SIDN labs
Internet Research & Innovation
SIDN

Open .nl DNS datasets

http://stats.sidnlabs.nl/

- SIDN is the .nl registry; nonprofit
  - 5.6M domains registered; 5th ccTLD in zone size
  - 2.5M DNSSec signed domains; 1st worldwide
- aggregated .nl auth servers data (DNS/IP/DNSSEC...)
  - 18 months + ; daily updated
- open for research collab: talk to me
  - giovane.moura@sidn.nl

SIDN labs
Internet Research & Innovation
SIDN

Background

- SIDN is the .nl registry; nonprofit
- SIDN Labs $\rightarrow$ research arm
- This presentation: big data security
- Contains parts of material submitted to NOMS 2016 and PAM 2016 conferences
- Mini-bio: joined SIDN last May; in academia before

Introduction

- Newly registered malicious domains have an abnormal initial DNS lookup [1]

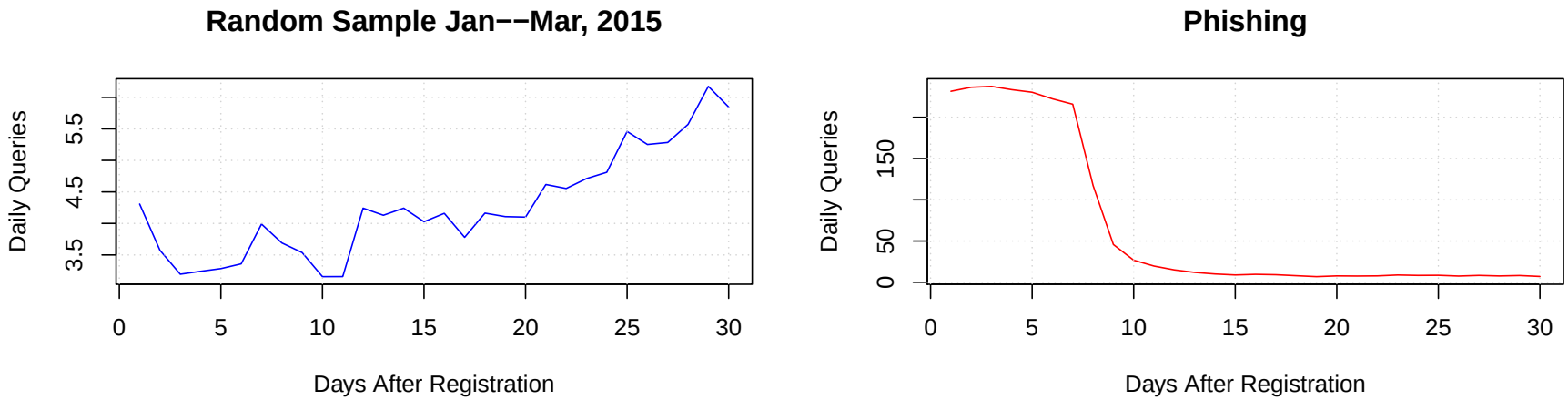


Figure: .nl DNS lookups - 20K Random vs Netcraft Phishing

SIDN labs
Internet Research & Innovation
SIDN

Introduction

- Why is that?
  - Assumption: spam-based business model
  - Automated
  - Maximize profit before being taken down
- Question: can we use this to improve security in the.nl zone?
  - Or build an early warning system for newly registered domains?
- We have both registration and DNS traffic data
  - Registry role
  - Privacy framework and board that oversees it

Introduction

- What we need:
  1. High-performance data analytics platform
  2. Efficient algorithm that can be used in production
- This presentation covers both things

ENTRADA: our big data platform

- Hadoop cluster data streaming warehouse $\rightarrow$ interactive response times
- 5K USD/EUR per node; low cost
- Store traffic data from .nl auth servers
- Enable production applications
- Based on open-source, can be deployed even in a cloud environment
  - only one part (one converter) we develop in house
  - studying open-source it

ENTRADA: our big data platform

- By definition, a data-streaming warehouse must deliver interactive response times
  - pcap storage& analysis wouldn't fly
  - not at low cost
- So, what are the alternatives?
- Our requirements:
  - Usability = SQL
  - Extensibility: no vendor lock-in
  - Security:
  - Dependability
  - Low cost
  - High performance

ENTRADA: our big data platform

Engine	Usab.	Exten.	Perf.	Scal.	Dep.
HBase(HDFS)	0	0	1	1	0
Elasticsearch	0	0	1	1	1
MongoDB	0	0	1	1	1
Hadoop+MapReduce(HDFS)	1	0	0	1	1
PostgreSQL	1	0	0	1	0
Impala+Parquet(HDFS)	1	1	1	1	1

Table: Comparison of Data Query Engines (1 = matches our requirements, 0 does not match)

- Two Core parts:
  1. Optimized Apache Parquet file format (based on Google's Dremel [2])
     - Column-based storage; reads only necessary columns
     - convert pcap to parquet
  2. MPP query engine (Impala [3])
     - multi parallel queries

SIDN labs
Internet Research & Innovation
SIDN

ENTRADA: data flow

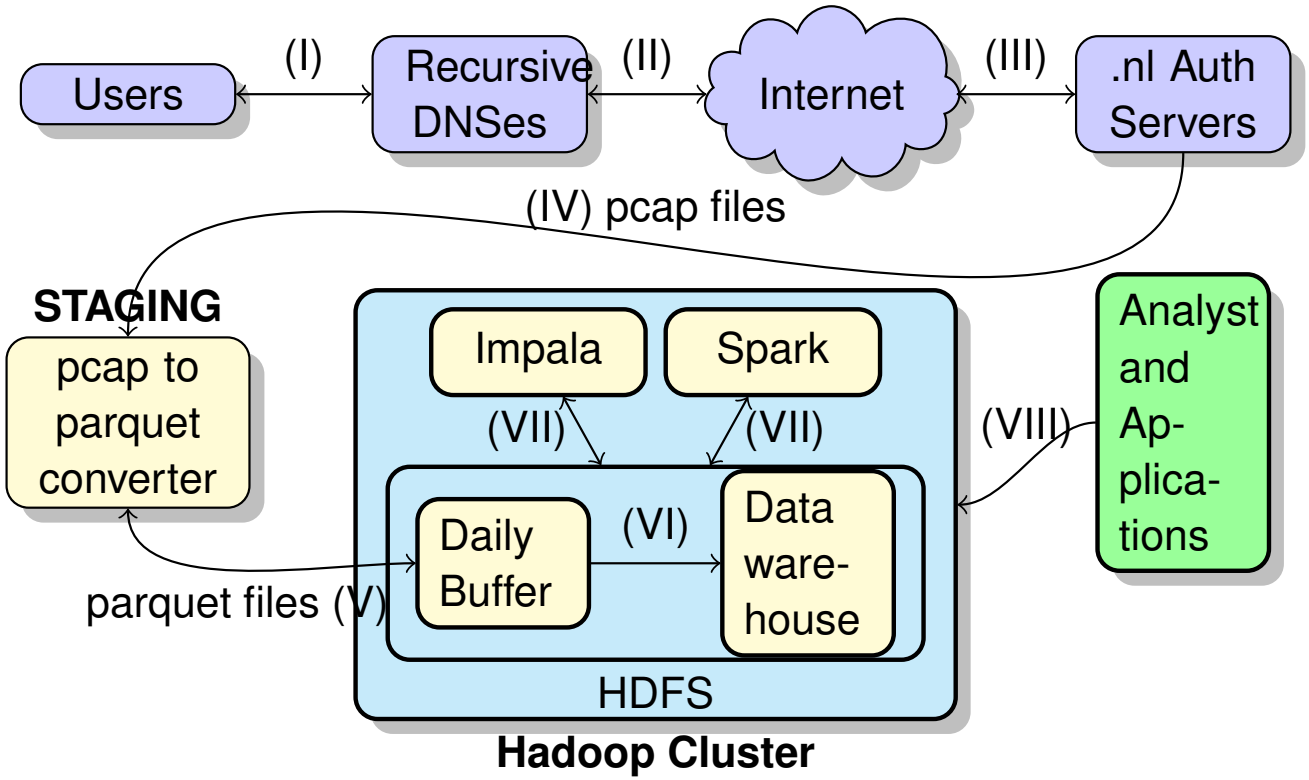


Figure: ENTRADA data sequence flow

ENTRADA: evaluation

- Query: `select concat_ws('-',day,month,year), count(1) from dns.queries where ipv=4 and year = 'X' and month = 'Y' and day='Z' group by concat_ws('-',day,month,year).`
- 10 parallel queries (1 per day)
- $\sim$ 52 TB of pcap data = 2.2 TB of parquet;
- Time: 3.5 minutes on 4 data nodes
- Conclusion: fast, and cheap; and open-source

Part 2: Early Warning System

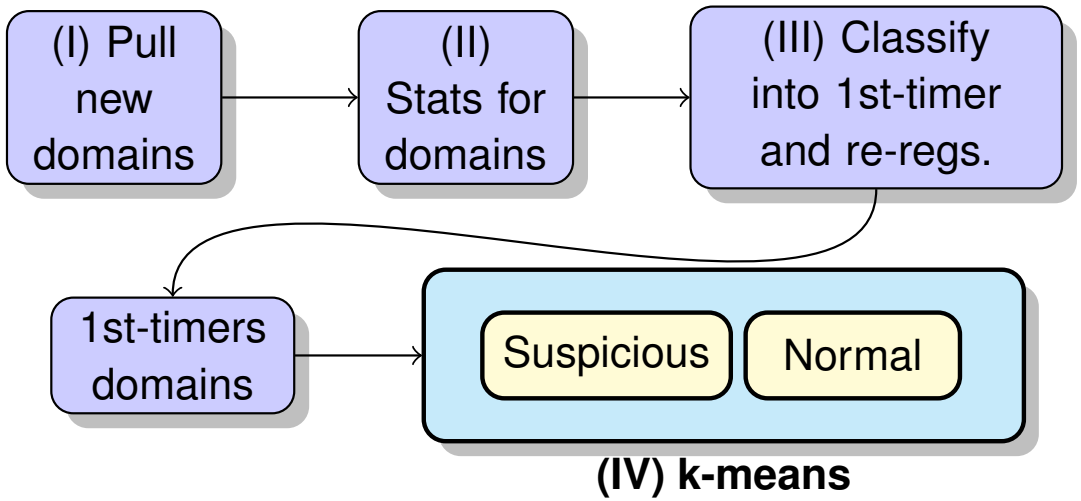


Figure: nDEWS Architecture

- "Bad" domains are likely to be more popular
- k-means clustering algorithm: unsupervised, classifies according to features
- $\sum Req$, $\sum IPs$, $\sum CC$, $\sum ASes$
- Run it every day

SIDN labs
Internet Research & Innovation
SIDN

Evaluation

- 1,5+ years of DNS data on ENTRADA
- 78B DNS request/responses
- All registration database

Key	Value
Interval	Jan 1st, 2015 to Aug 30th 2015
Average .nl zone size	$\sim$ 5,500,000
$\sum$ new domains	586,201
New domains - first timers	476,040(81.2%)
New domains - re-registered	110,161 (18.8%)
Total DNS Requests	32,864,402,270
DNS request new domains (24h)	826,740
DNS request new domains - first-timers (24h)	420,362

Table: Evaluated datasets (from one .nl auth server)

Evaluation

Cluster	**Size**	$\sum Req$	$\sum IPs$	$\sum CC$	$\sum ASes$
Normal	132,425	4.31	3.06	1.64	1.43
Suspicious	2,956	55.03	27.87	4.99	7.43

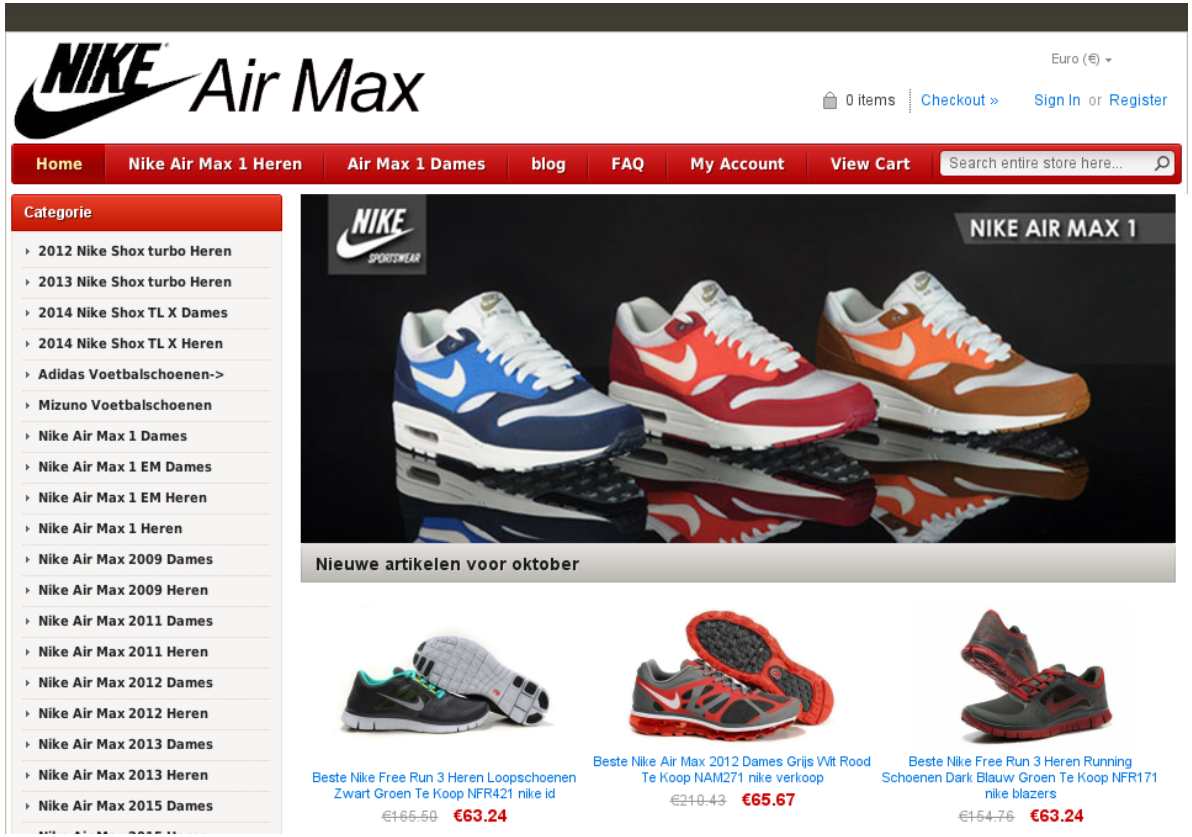
Table: Mean values for features and clusters - excluding domains with 1 request - 1st Timers

Validation

- Were those “suspicious” domains really malicious?
- Very hard to verify on historical data: if they had pages; they might be gone or diff by now
- Results on historical data:
  - Content analysis: 148 “shoes stores” , 17 adult/malware
  - 19 phishing domains (out of 49 reported by Netcraft on the same period)
  - VirusTotal: 25 domains matched
- Results on current data:
  - By far the “shoes” sites dominate it
  - Adult and malware is also detected; we now download screenshots and content as we classify
  - False positives: rapidly popular political websites and others

Discussion

- Why so many (5–10) new shoes stores per day?
- Probably concocted websites [4]
- Automatically created; spam based



Why shoes?

- Most counterfeit product = $\sim$ 40% of US Border seizures [5]
- Large demand
- Re-current registration suggest profitability; one site down does not affect operations
- Online fraud is the NL: 5.3 billion EUR in 2 years; many from site websites [6]
- Evade industry's tools/techniques:
  - Solutions for phishing and malware exist
  - Users left unprotected
- Shoes are a smart play: high demand, and low penalties
- Currently: studying how to share/handle this

Summary

- We showed ENTRADA, our data streaming warehouse
  - Fast & cheap
  - Anyone can deploy it; even on a cloud
- We showed one application of it: new malicious domains early warning system
  - Under the radar abuse form (shoes)
  - Can be detected by their lookup patterns
- Run it on a daily basis; have to reduce false positives
- Studying pilot studies to handle that information
- More big-data based security applications to come

Questions?

- Contact:
  - `http://sidnlabs.nl`
  - `giovane.moura@sidn.nl`
- Looking for collaboration to :
  - build and validate systems to improve security;
  - write measurement papers
- Thank you for your attention

SIDN labs
Internet Research & Innovation
SIDN

Bibliography I

- Hao, Shuang and Feamster, Nick and Pandrangi, Ramakant, "Monitoring the initial dns behavior of malicious domains," in *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*, ser. IMC '11. New York, NY, USA: ACM, 2011, pp. 269–278.
- S. Melnik, A. Gubarev, J. J. Long, G. Romer, S. Shivakumar, M. Tolton, and T. Vassilakis, "Dremel: Interactive Analysis of Web-scale Datasets," *Proc. VLDB Endow.*, vol. 3, no. 1-2, pp. 330–339, Sep. 2010.
- M. Kornacker, A. Behm, V. Bittorf, T. Bobrovytsky, C. Ching, A. Choi, J. Erickson, M. Grund, D. Hecht, M. Jacobs *et al.*, "Impala: A modern, open-source SQL engine for Hadoop," in *Proceedings of the Conference on Innovative Data Systems Research (CIDR'15)*, 2015.

Bibliography II

- A. Abbasi and H. Chen, “A comparison of tools for detecting fake websites,” *Computer*, no. 10, pp. 78–86, 2009.
- N. Schmidle, “Inside the Knockoff-Tennis-Shoe Factory - The New York Times,” http://www.nytimes.com/2010/08/22/magazine/22fake-t.html, 2010.
- FraudHelpdesk.nl, “Ruim miljoen Nederlanders opgelicht (in Dutch),” https://www.fraudehelpdesk.nl/nieuws/ruim-miljoen-nederlanders-opgelicht-2/, Dec 2014.