Malicious domains: Automatic Detection with DNS traffic analysis

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

Network Machine Learning Research Group (NMLRG)
IETF 95
April 7, 2016
Buenos Aires, Argentina

Introduction

- DNS provides a simple label for hosts, services, applications on the Internet
- Often, it is misused in malicious activities
  - phishing campaigns
  - malware
  - spam
- For phishing:
  1. Compromised domains (majority) - easier
  2. Malicious domains (new domains) - more effective?

SIDN LABS

Introduction

- Newly registered malicious domains have an abnormal initial DNS lookup [1]
- We see the same on the .nl TLD

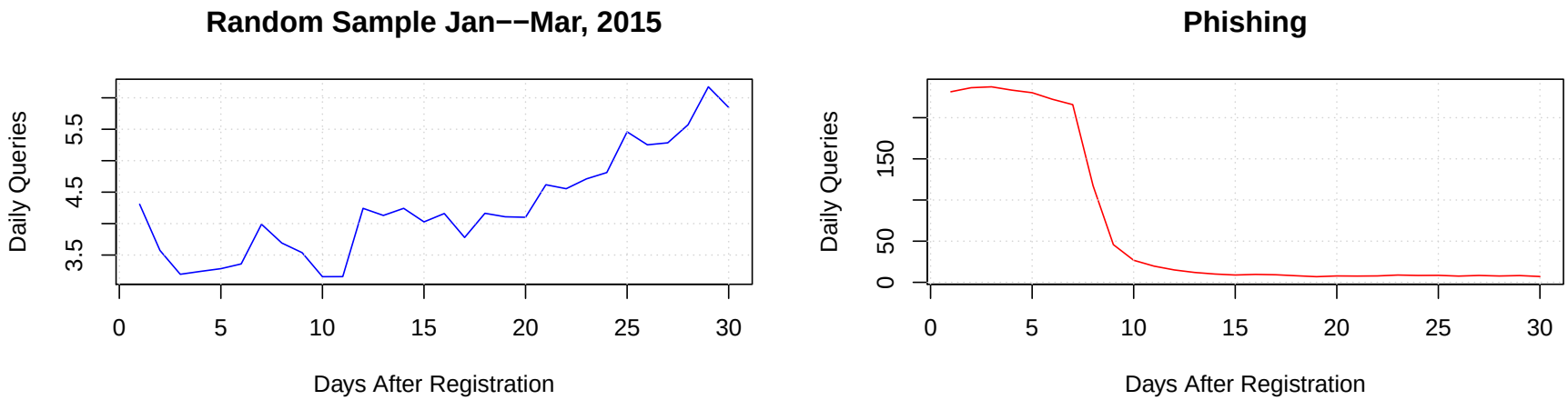


Figure: .nl DNS lookups - 20K Random vs Netcraft Phishing

“Popular” new domains

- Why phishing is more popular?
  - Assumption: spam-based business model
  - Automated
  - Maximize profit before being taken down
- Question: **can we detect these domains based on DNS traffic as soon as possible?**

SIDN LABS

Early Detection of Malicious Domains

What we need:

1. “Centralized” data (TLD point-of-view)
   - As A TLD registry, we observe a fraction of all .nl TLD traffic (due to caching)
   - Plus, we have registration information
2. High-performance data analytics platform (ENTRADA [2])
   - Our open-source solution – `http://entrada.sidnlabs.nl`
   - Allows quick hypothesis test : 53 TB of equivalent pcap analysis under 3.5 min (4 data node cluster)
   - In short: pcap analysis is either too slow or too expensive
3. Efficient algorithm that can be used in production

SIDN LABS

DNS and TLD traffic: “centralized” data

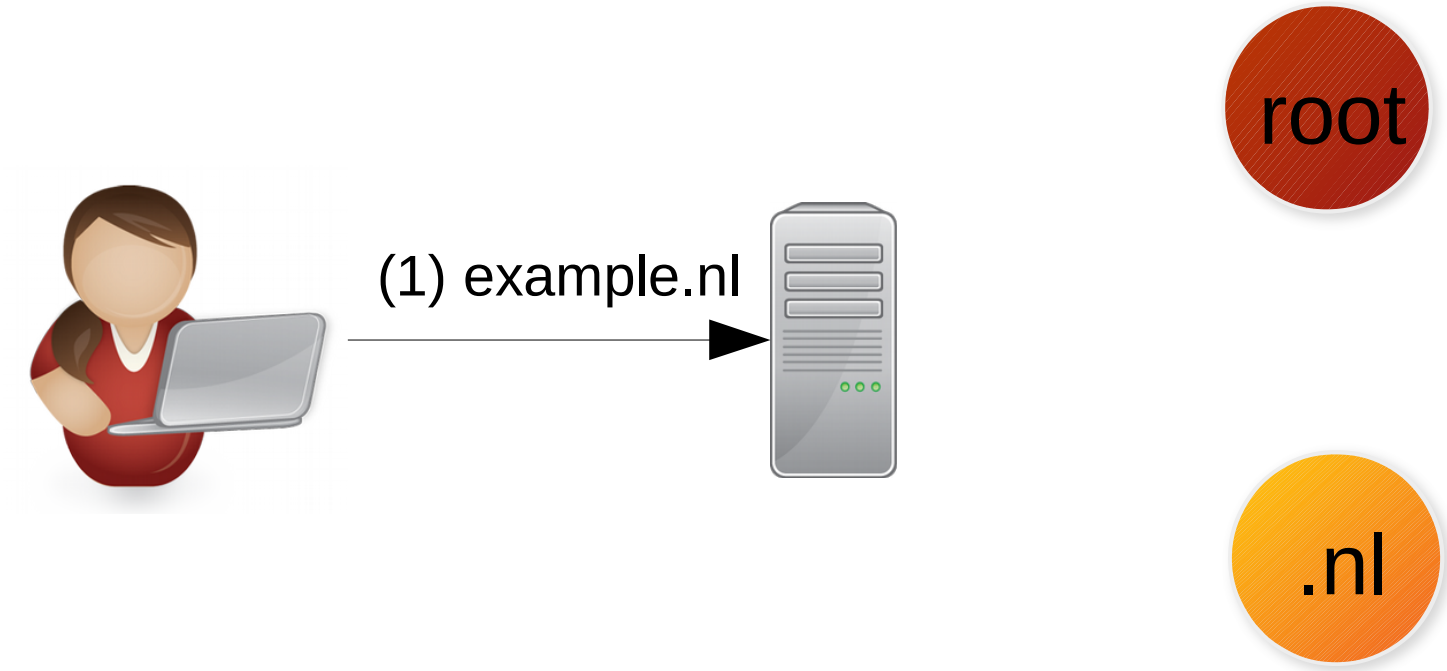


Figure: Resolving a Name

DNS and TLD traffic: “centralized” data

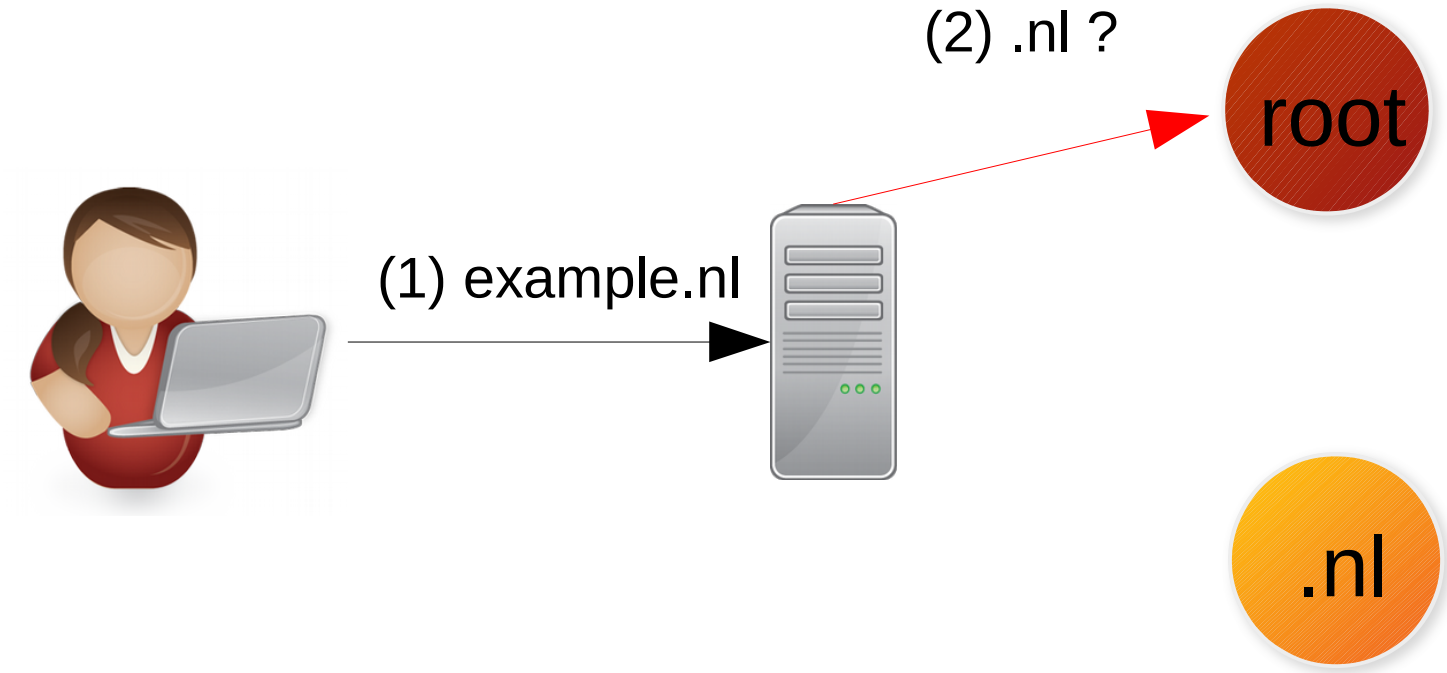


Figure: Resolving a Name

DNS and TLD traffic: “centralized” data

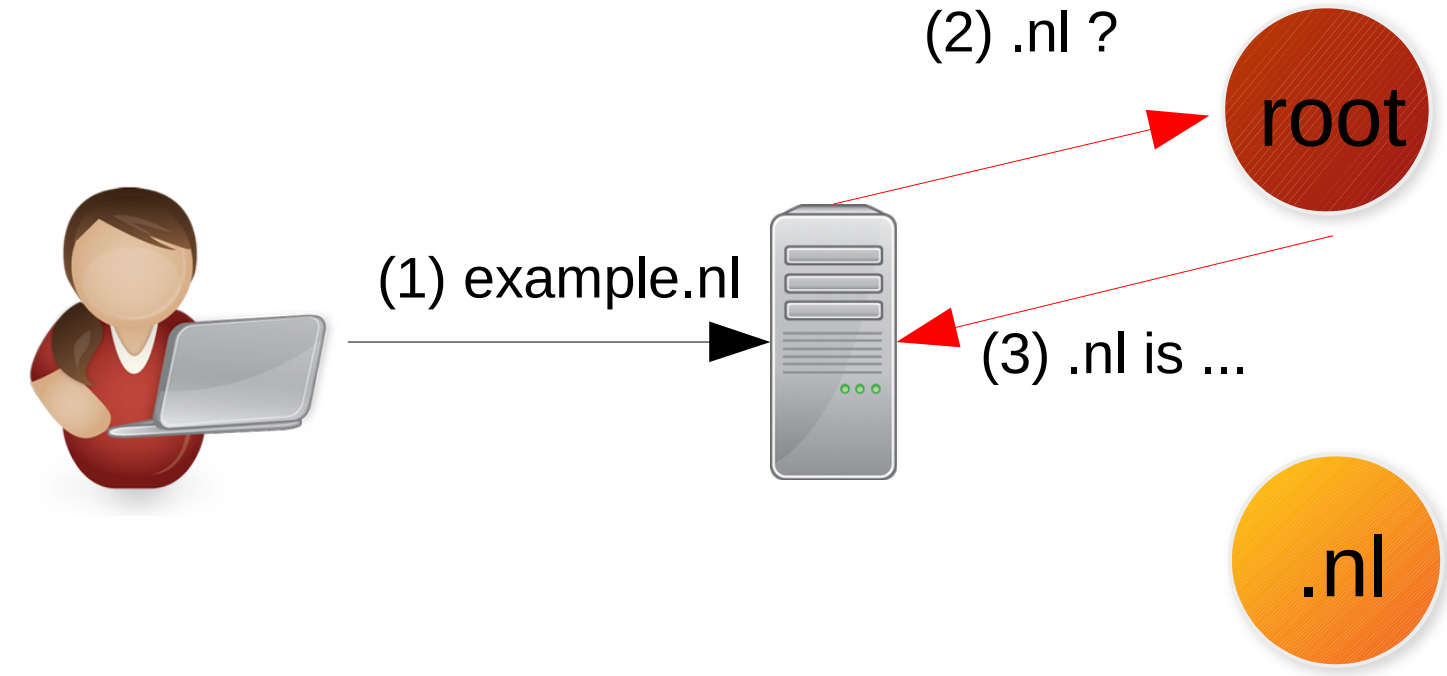


Figure: Resolving a Name

DNS and TLD traffic: “centralized” data

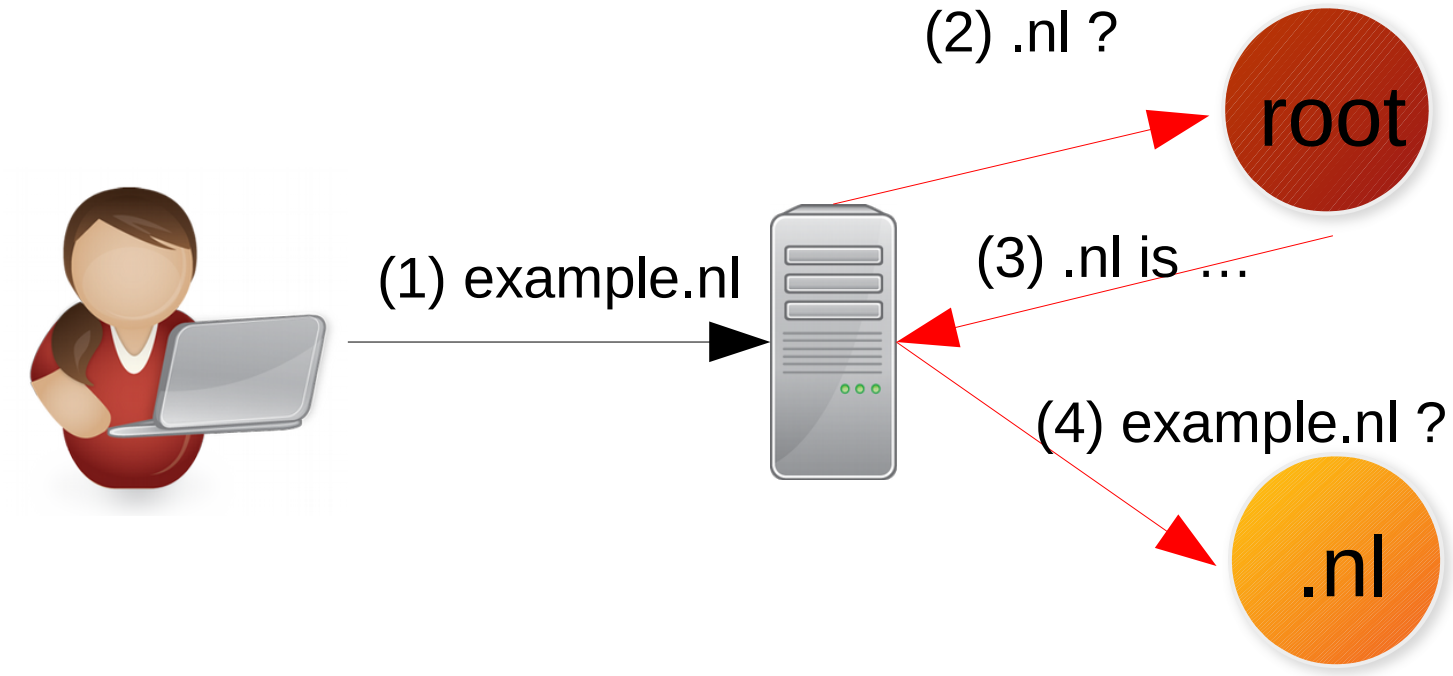


Figure: Resolving a Name

DNS and TLD traffic: “centralized” data

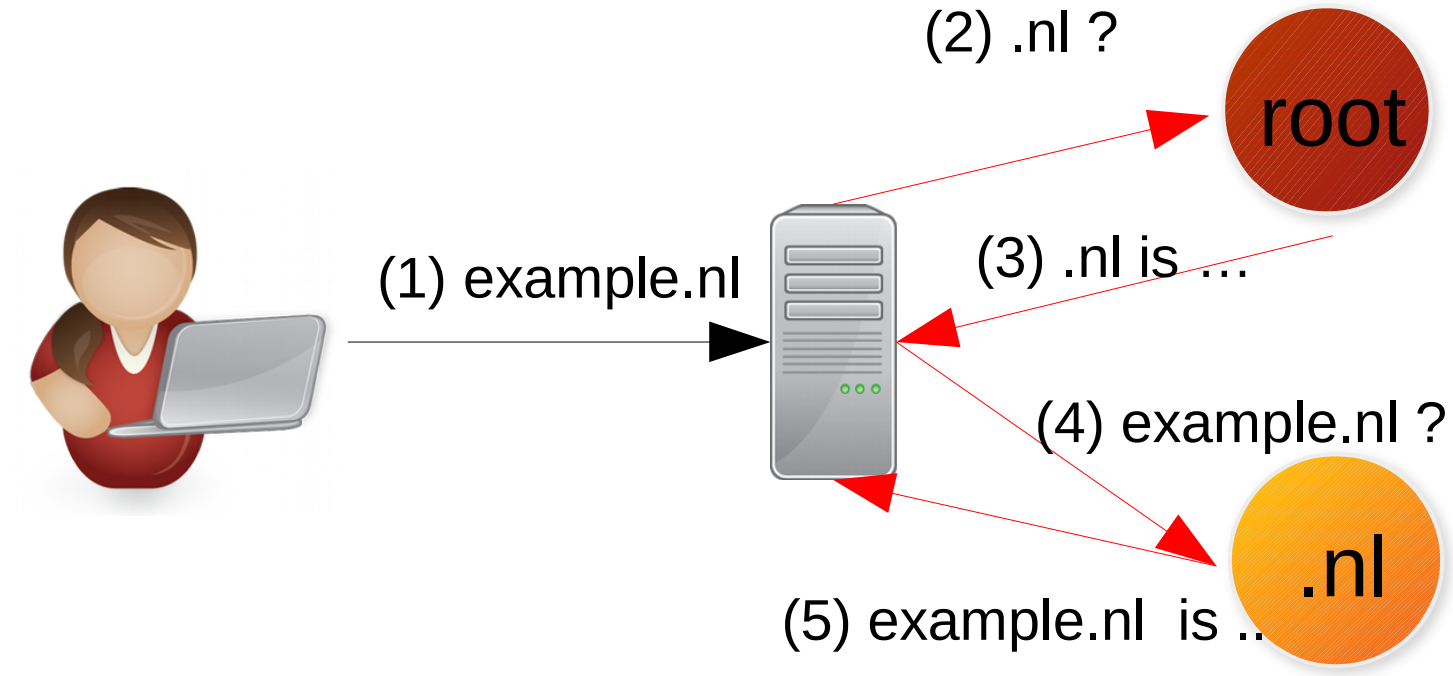


Figure: Resolving a Name

DNS and TLD traffic: “centralized” data

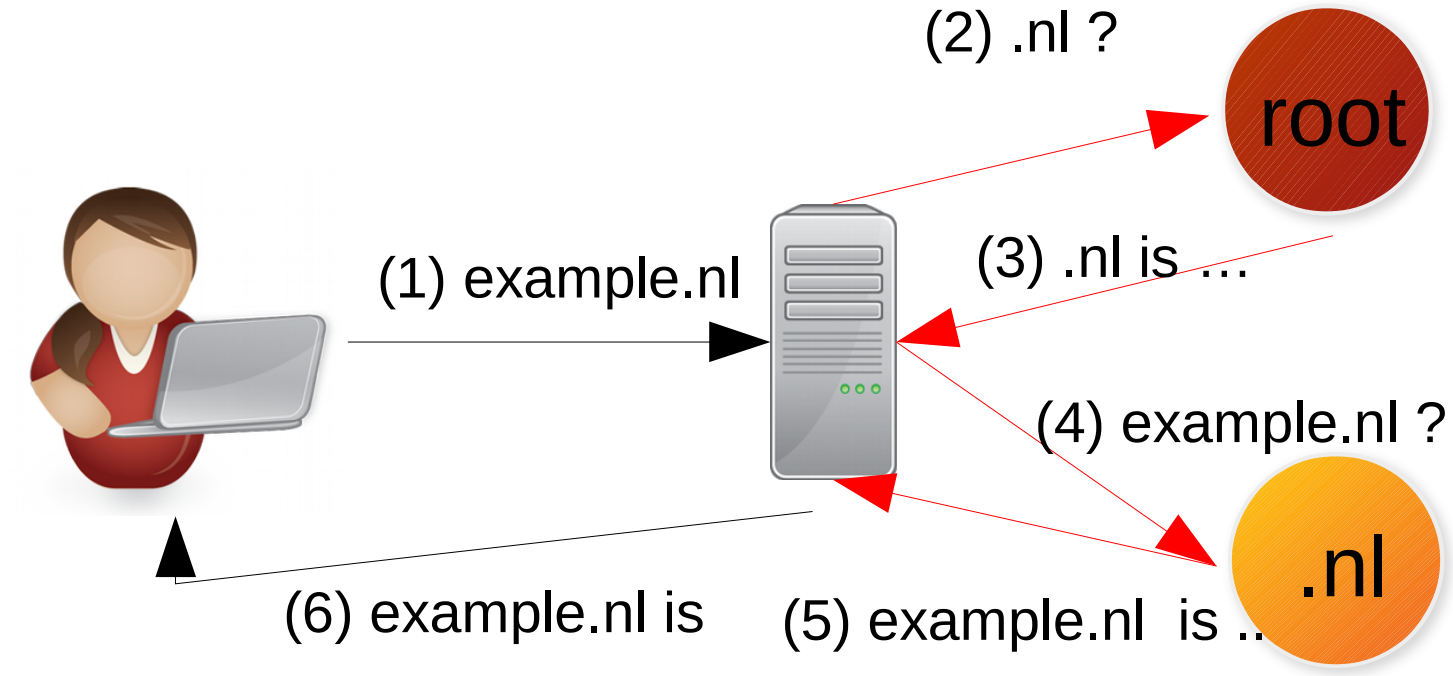


Figure: Resolving a Name

OK, we've got the data... now analyze it

- $\sim$ 85 GB of pcap per day, per auth name server
- You can map/reduce it, but it's gonna be costly or slow
- CSV, DBRMS have their own limitations
  Still it would be very hard to deliver interactive response times (< few minutes)

OK, so what can we do?

- Build your data streaming warehouse (DSW)
- ENTRADA, ours, is a DSW
- Open-source: `http://entrada.sidnlabs.nl`
- Analyze 53 TB of pcap data in less than 3.5min in a small 4-data node cluster!
- Used in operation for 2+ years; 100 Billion+ DNS records
- Our case: DNS analysis

How? Why?

Three main reasons:

1. Efficient file format (Apache Parquet)
2. Efficient query engine (Cloudera Impala - SQL)
3. Hadoop cluster beneath the hood

ENTRADA Data Flow

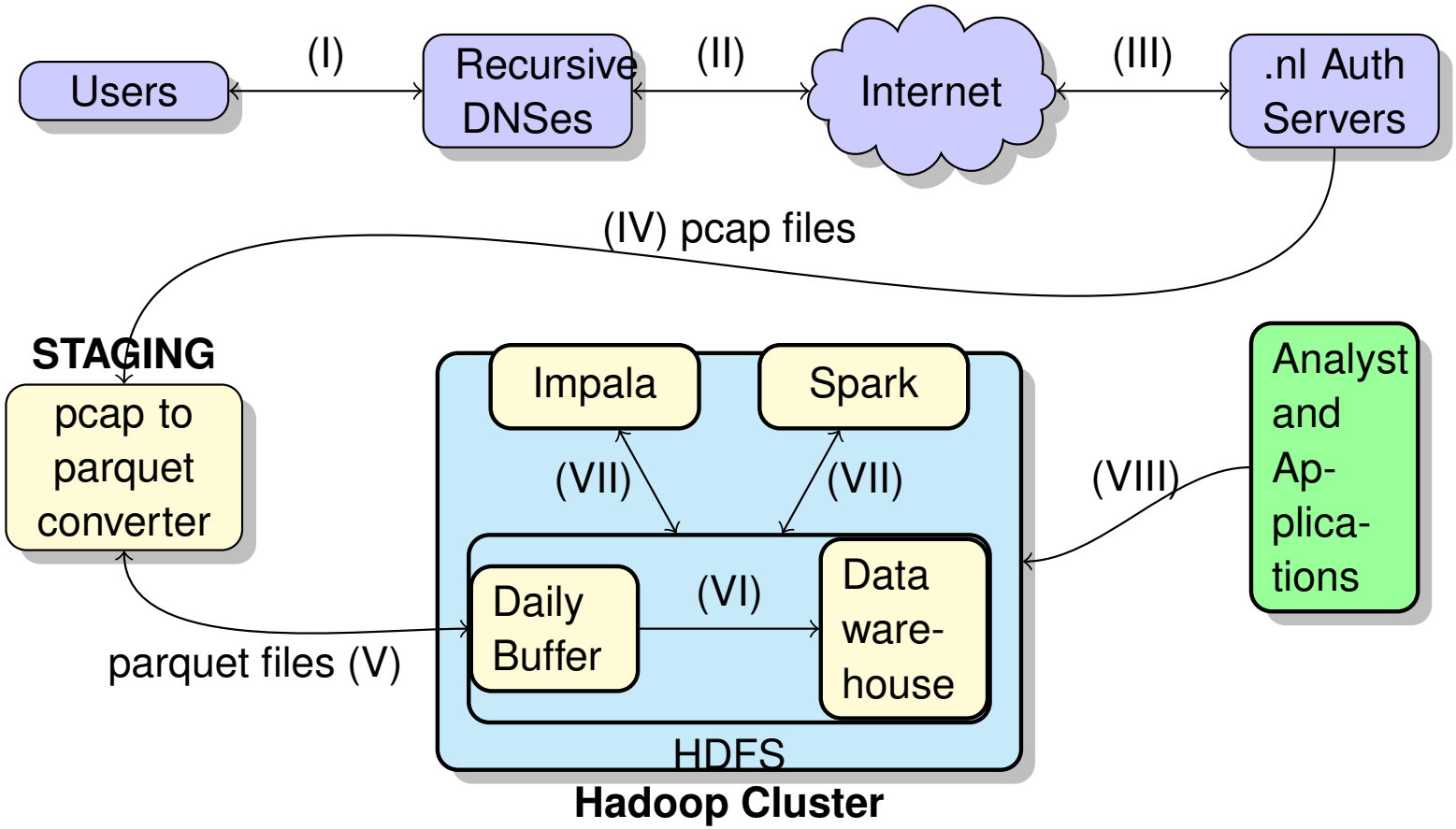


Figure: ENTRADA DNS data flow [2]

1st: File format - Apache Parquet

- Google Dremel: optimized format for aggregation type queries
- Parquet: based on Dremel (Apache)
- It combines columnar storage
  - Fields stored separately
- Partition pruning !
- Compression
- 85 GB DNS pcap $\rightarrow$ 6 GB Parquet (some filtering too)

SIDN LABS

2nd: Query Engine: Cloudera Impala

- SQL support
  - no more awk
- Run daemons on each node; parallel queries
- Parquet-file compatible
- Note: there were other options; please refer to paper [2]

3rd: Hadoop Cluster

- Scalability
- HDFS
- Redundancy

Ok, we've got the data and the platform. What's next?

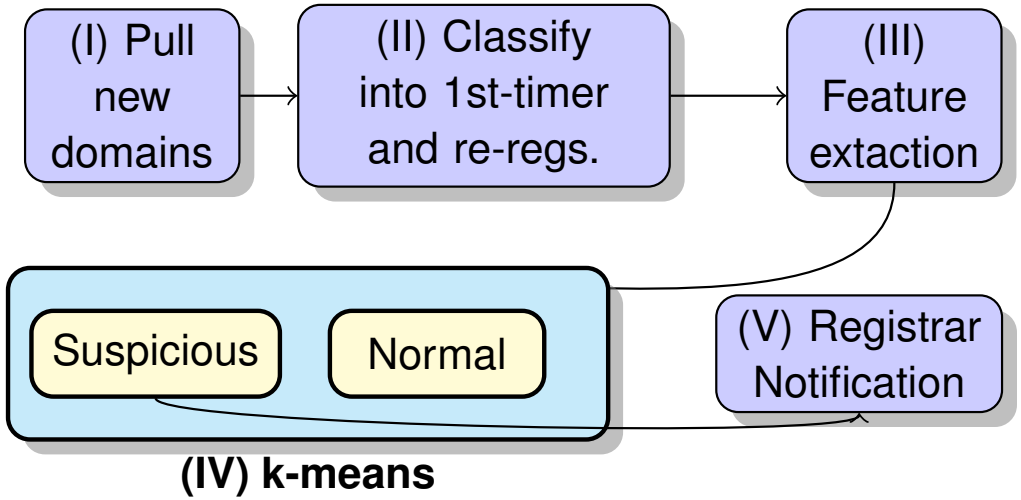


Figure: nDEWS Architecture [3]

- Work to be presented at AnNET 2016/IEEE NOMS 2016 [3]
- "Bad" domains are likely to be more popular
- k-means clustering algorithm: unsupervised, classifies according to features
- Run it daily, for all newly added domains on the .nl zone

Feature selection

- Empirically chosen
- $\sum$ *Req* : how popular it is
- $\sum$ *IPs*: resolver's diversity
- $\sum$ *CC*: countries' diversity
- $\sum$ *ASes*: ASes diversity

- Domains involved in phishing tend to score high on all of them
- Why? spam knows no borders
- We choose two cluster: "normal" and "suspicious"

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)

SIDN LABS

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

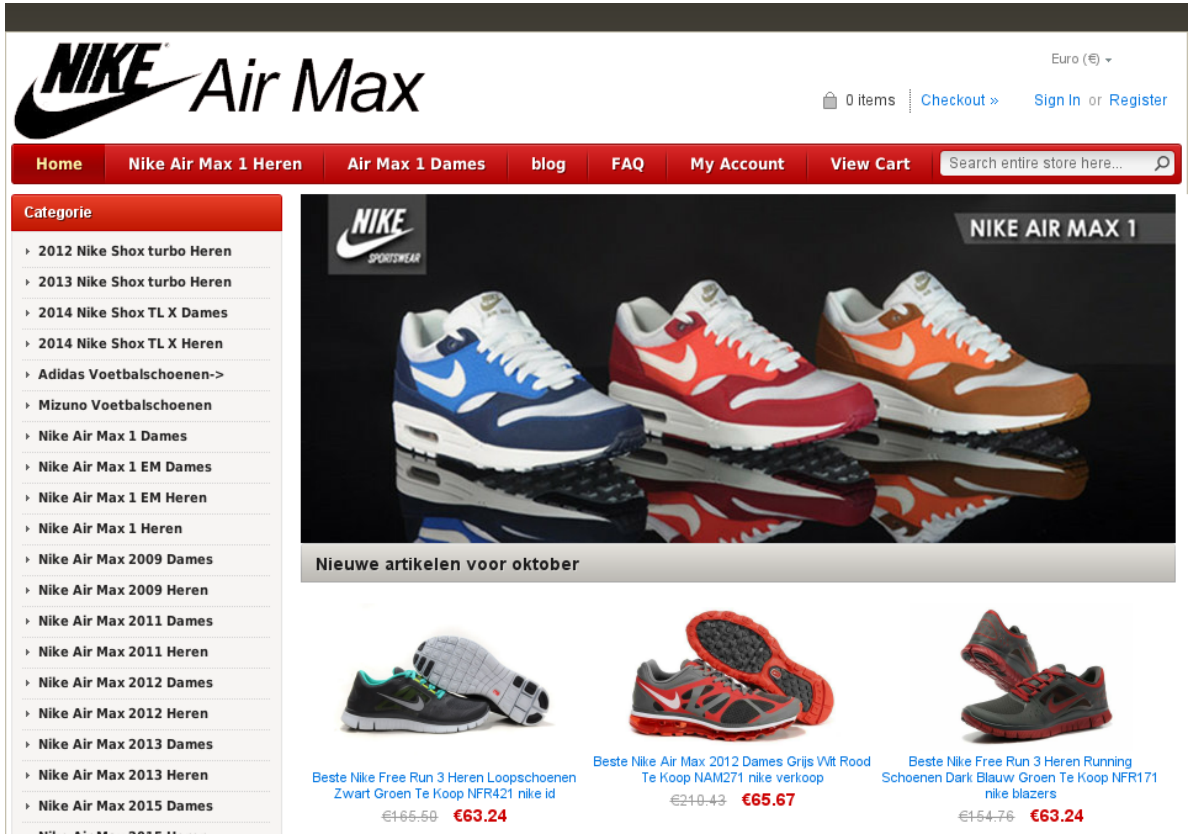
Validation: historical data

- 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

SIDN LABS

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]
- 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

Validation on current data

- “Shoes” sites dominate it, depending on the day
- Adult and malware is also detected; we now download screenshots and content as we classify
- False positives: rapidly popular political websites and others
  - work on reducing this now
- Working on making it in near real-time (currently 24h delay)

SIDN LABS

Summary

1. A DSW delivers the performanced needed for ML on network traffic
   - Ours is open-source: `https://entrada.sidnlabs.nl`
   - Test hypothesis on large datasets within seconds
2. We presented nDEWS
   - Early Warning system for new domains
   - Uses k-means to classify each domain based on network traffic features
   - It monitors all new domains on the .nl zone, daily
   - We notify registrars about it
3. Future work:
   - making it near real-time
   - incorporate time-series analysis
   - evaluate all the domains, and not only the new ones

SIDN LABS

Questions?

- Contact:
  - `http://sidnlabs.nl`
  - `giovane.moura@sidn.nl`
- Thank you for your attention

Download our software at: `http://entrada.sidnlabs.nl`

Bibliography I

[1] 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.

[2] Maarten Wullink, Giovane C. M. Moura, Moritz Muller, and Cristian Hesselman, "ENTRADA: a High Performance Network Traffic Data Streaming Warehouse," in *Network Operations and Management Symposium (NOMS), 2016 IEEE (to appear)*, April 2016. [Online]. Available: https://www.sidnlabs.nl/downloads/sidn-noms2016_EN.pdf

SIDN LABS

Bibliography II

[3] Giovane C. M. Moura, Moritz Muller, Maarten Wullink, and Cristian Hesselman, “nDEWS: a New Domains Early Warning System for TLDs,” in *IEEE/IFIP International Workshop on Analytics for Network and Service Management (AnNet 2016), co-located with IEEE/IFIP Network Operations and Management Symposium (NOMS 2016)*, April 2016. [Online]. Available: https://www.sidnlabs.nl/downloads/presentations/sidn-annet2016.pdf

[4] A. Abbasi and H. Chen, “A comparison of tools for detecting fake websites,” *Computer*, no. 10, pp. 78–86, 2009.

[5] N. Schmidle, “Inside the Knockoff-Tennis-Shoe Factory - The New York Times,” http://www.nytimes.com/2010/08/22/magazine/22fake-t.html, 2010.

Bibliography III

[6] FraudHelpdesk.nl, “Ruim miljoen Nederlanders opgelicht (in Dutch),” https://www.fraudehelpdesk.nl/nieuws/ruim-miljoen-nederlanders-opgelicht-2/, Dec 2014.