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?
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?
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
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

- ~ 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 3.5 min 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 → 6 GB Parquet (some filtering too)
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?

(I) Pull new domains  
(II) Classify into 1st-timer and re-regs.  
(III) Feature extraction  
(IV) k-means  
(V) Registrar Notification

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
- $\sum \text{Req}$: how popular it is
- $\sum \text{IPs}$: resolver’s diversity
- $\sum \text{CC}$: countries’ diversity
- $\sum \text{ASes}$: 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

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

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

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Size</th>
<th>$\sum$ Req</th>
<th>$\sum$ IPs</th>
<th>$\sum$ CC</th>
<th>$\sum$ ASes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>132,425</td>
<td>4.31</td>
<td>3.06</td>
<td>1.64</td>
<td>1.43</td>
</tr>
<tr>
<td>Suspicious</td>
<td>2,956</td>
<td>55.03</td>
<td>27.87</td>
<td>4.99</td>
<td>7.43</td>
</tr>
</tbody>
</table>

**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
Discussion

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

- Most counterfeit product = \(~ 40\%\) 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)
Summary

1. A DSW delivers the performance 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
Questions?

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

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

