Anomaly detection in DNS traffic

Clustering-based approach

Maciej Andziński • maciej.andzinski@nic.cz • 11.05.2019
Anomaly detection

- What is an anomaly?
  - Hard to define
    - Scanners?
    - Monitors?
    - Misconfigured resolvers?
  - Unusual behaviour
Data flow

.CZ DNS servers → PCAP parser → HDFS

Hadoop

Parquet

Spark

Data aggregator

RStudio
Data aggregation

1) Group DNS queries by source IP address
2) For each source IP address compute statistics (features)
   - Take only IP addresses which send min. 100 queries daily
   - Time window = 1 day
Features #1

- Entropy (normalised Shannon Index)
  - Source port
  - Transaction ID
- Coefficient of variation ($C_v = \frac{\sigma}{\mu}$)
  - Idletime
  - Packet length
Features #2

- Amplification factor
- Mean domain name length
- Domain name diversity
Features #3

- Observed DNS QTYPEs
  - A + AAAA
  - NS
  - DNSSEC RRs
  - Popular RRs
  - Weird RRs
Features #4

- **Observed DNS RCODEs**
  - NOERROR
  - NXDOMAIN

- **Observed DNS FLAGs**
  - RD
  - EDNS0 DO
Features #5

- Observed DNS QCLASSes
  - IN
- Observed DNS OPCODEs
  - QUERY
Features – an example

- **217.31.204.130 on 23 October 2018 (CZ.NIC open DNS resolver)**

```
<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>srcp_sh_ix_n</td>
<td>0.9876179</td>
</tr>
<tr>
<td>id_sh_ix_n</td>
<td>0.9879639</td>
</tr>
<tr>
<td>idletime_cv</td>
<td>0.709647</td>
</tr>
<tr>
<td>dn_len_mean</td>
<td>11.57786</td>
</tr>
<tr>
<td>dn_perc</td>
<td>0.292901</td>
</tr>
<tr>
<td>rcode_noerror_perc</td>
<td>0.9828929</td>
</tr>
<tr>
<td>rcode_nxdomain_perc</td>
<td>0.01710712</td>
</tr>
<tr>
<td>qtype_common_perc</td>
<td>0.975649</td>
</tr>
<tr>
<td>qtype_weird_perc</td>
<td>0.001121778</td>
</tr>
<tr>
<td>qtype_dnssec_perc</td>
<td>0.1577407</td>
</tr>
<tr>
<td>qtype_ns_perc</td>
<td>0.01025492</td>
</tr>
<tr>
<td>qtype_addr_perc</td>
<td>0.9247106</td>
</tr>
<tr>
<td>qclass_in_perc</td>
<td>1</td>
</tr>
<tr>
<td>edns_do_perc</td>
<td>1</td>
</tr>
<tr>
<td>flag_rd_perc</td>
<td>0</td>
</tr>
<tr>
<td>ampl_factor</td>
<td>4.743633</td>
</tr>
<tr>
<td>len_cv</td>
<td>0.1132715</td>
</tr>
<tr>
<td>opcode_query_perc</td>
<td>1</td>
</tr>
</tbody>
</table>
```
Anomaly detection concept
Anomaly detection concept

Heatmap: mean domain name length vs amplification factor

(number of DNS resolvers: 1000, 100, 10, 1)

(based on DNS traffic from: 11 Sep 2018 - 31 Oct 2018)
Anomaly detection concept

Heatmap: a simple classification example (k=2, 2 features)

Amplification factor

Mean domain name length

% of anomalies

100%
75%
50%
25%
0%

(based on DNS traffic from: 11 Sep 2018 - 31 Oct 2018)
Anomaly detection concept

Heatmap: final model results (k=13, 18 features)

% of anomalies
- 100%
- 75%
- 50%
- 25%
- 0%

(based on DNS traffic from: 11 Sep 2018 - 31 Oct 2018)
Model

- Spark MLlib
- K-means clustering
  - UDF to compute distance from cluster center
- MinMaxScaler
  - Entire dataset used for scaling (some features in training set were meaningful but had “near zero” variance)
Model

- Training set
  - **Real DNS resolvers** (each RIPE Atlas probe was employed to query its local DNS resolver for whoami.akamai.net)
    - Gathered 3,430 unique IP addresses
    - 51 days = 137,701 observations
    - Filtered out weird observations
Model

• Test/Validation set
  • Difficult to measure anomaly detection performance
  • Needed for grid search to select best model parameters (best F-score)
Model

- Test/Validation set #1
  - Known anomalies
    - DNSMON
    - Domain name scanners
    - Misconfigured DNS resolvers
Model

- Test/Validation set #2
  - Real DNS resolvers
    - DNS resolvers of RIPE Atlas probes
    - Google Public DNS
    - Cloudflare
    - Quad9
    - OpenDNS (Cisco)
    - Dyn
    - Level3
    - Yandex
    - CZ.NIC
Model

- Model parameters
  - $k = 13$
  - Threshold (maximal distance from cluster center) = $3 \times Q_3$
    (third quartile)
Model performance

F-score: 0.9894033

• Known anomalies

<table>
<thead>
<tr>
<th>dataset</th>
<th>total</th>
<th>anomaly</th>
<th>%anomaly</th>
</tr>
</thead>
<tbody>
<tr>
<td>dnsmon</td>
<td>38</td>
<td>38</td>
<td>100.0 %</td>
</tr>
<tr>
<td>scanners</td>
<td>25</td>
<td>25</td>
<td>100.0 %</td>
</tr>
<tr>
<td>scanners2</td>
<td>100</td>
<td>100</td>
<td>100.0 %</td>
</tr>
<tr>
<td>misconfigured</td>
<td>99</td>
<td>99</td>
<td>100.0 %</td>
</tr>
<tr>
<td>dnsviz</td>
<td>1</td>
<td>0</td>
<td>0.0 %</td>
</tr>
</tbody>
</table>
### Model performance

**F-score:** 0.9894033

- **Real DNS resolvers**

<table>
<thead>
<tr>
<th>dataset</th>
<th>total</th>
<th>anomaly</th>
<th>%anomaly</th>
</tr>
</thead>
<tbody>
<tr>
<td>atlas_resolvers</td>
<td>3193</td>
<td>48</td>
<td>1.5 %</td>
</tr>
<tr>
<td>google</td>
<td>1250</td>
<td>0</td>
<td>0.0 %</td>
</tr>
<tr>
<td>quad9</td>
<td>224</td>
<td>0</td>
<td>0.0 %</td>
</tr>
<tr>
<td>opendns</td>
<td>107</td>
<td>2</td>
<td>1.9 %</td>
</tr>
<tr>
<td>dyn</td>
<td>107</td>
<td>3</td>
<td>2.8 %</td>
</tr>
<tr>
<td>level3</td>
<td>160</td>
<td>4</td>
<td>2.5 %</td>
</tr>
<tr>
<td>cloudflare</td>
<td>180</td>
<td>2</td>
<td>1.1 %</td>
</tr>
<tr>
<td>yandex</td>
<td>82</td>
<td>2</td>
<td>2.4 %</td>
</tr>
<tr>
<td>cznic</td>
<td>2</td>
<td>0</td>
<td>0.0 %</td>
</tr>
</tbody>
</table>
Results

- DNS traffic from 11 Sept 2018 - 31 Oct 2018
  - 737 729 out of 9 918 267 observations (7.4%) were classified as anomaly
  - 8 649 294 465 out of 40 073 507 471 queries (17.7%) were originated from anomalous source
Findings in one of the big ISPs

- 93.6 % of queries originated from anomalous sources
Findings in one of the big ISPs

- 93.6% of queries originated from **anomalous sources**
  - 8 biggest sources classified as **anomaly**
    - Each using only 256 different source ports for UDP queries (!)
Findings in ..one of the big ISPs

- 93.6 % of queries originated from anomalous sources
  - 8 biggest sources classified as anomaly
    - Each using only 256 different source ports for UDP queries (!)

- Feedback
  - They based on software vendor / OS recommendations
Findings in AS15169 (Google LLC)

- 27.4 % of observations classified as anomaly
- 12 345 unique IP addresses
  - Only 1 387 IP addresses belonged to Google Public DNS
    - No observations classified as anomaly
Findings in AS25192 (CZ.NIC, z.s.p.o.)

- 5th biggest in terms of query number
- 2,496 observations (128 unique IP addresses)
  - 525 (21%) classified as anomaly (22 unique IP addresses)
  - 1,731,755,782 queries
- 87,432,646 (5%) from anomalous sources
Findings in AS25192 (CZ.NIC, z.s.p.o.)

- 32 out of 128 IP addresses were observed every day
  - 19 were *never anomalous* (0%)
  - 5 were *almost never anomalous* (<5%)
  - 7 were *always anomalous* (100%)
  - 1 was *almost always anomalous* (>90%)
Findings in AS25192 (CZ.NIC, z.s.p.o.)

- Always classified as anomaly (100%)
  - Incigna monitoring system (IPv4+IPv6)
  - Domain name crawler
  - RIPE Atlas anchor (IPv4 + IPv6)
  - A monitoring system without name
  - DNS resolver for Hadoop cluster (IPv6)
- Almost always classified as anomaly (>90%)
  - DNS resolver for Hadoop cluster (IPv4)
Findings in AS25192 (CZ.NIC, z.s.p.o.)

- **Never classified as anomaly (0%)**
  - Real DNS resolvers

- **Occasionally classified as anomaly (<5%)**
  - DNS resolver for mail server
    - A configuration issue was discovered
  - NAT gateways
Future work

- Add more classes
  - Scanner, monitor, misconfigured, under attack, etc.
- Extend / modify feature set
- Try different algorithms
- Collect better ground truth
- Visualise results
Thank You

Maciej Andziński • maciej.andzinski@nic.cz