DNS Traffic Analysis
Opportunities, Risks, and (Self-)Defenses

Utility for forensic investigations
Potential threats to privacy
New ideas for protection

Dr. Dominik Herrmann
Slides available at https://dhgo.to/dns-traffic-analysis
Motivation of monitoring DNS
– block known malicious domains (e.g. phishing)
– retain log of all DNS queries for later analysis

Why is DNS monitoring interesting for forensics?
analyzing hard disk not sufficient any more
(cloud, private browsing, disk encryption)

What can we infer from DNS query logs?
low storage needs

DNS log contains essential metadata:
2016-03-05  11:14:05.124  2.240.3.12  www.yahoo.com  A
date and time  user’s address  domain  type
Example 1: confirm source of traffic
Did incriminating traffic originate from Bob’s laptop?

Source of discrepancy? Rogue hardware?

DNS queries from Bob’s IP

<table>
<thead>
<tr>
<th>Date</th>
<th>Time</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>2016-03-05</td>
<td>06:46:01.383</td>
<td>aus5.mozilla.org</td>
</tr>
<tr>
<td>2016-03-05</td>
<td>09:46:01.455</td>
<td><a href="http://www.exploit-db.com">www.exploit-db.com</a></td>
</tr>
<tr>
<td>2016-03-05</td>
<td>10:22:01.814</td>
<td>time.apple.com</td>
</tr>
<tr>
<td>2016-03-05</td>
<td>10:22:01.950</td>
<td>b.config.skype.com</td>
</tr>
<tr>
<td>2016-03-05</td>
<td>14:17:09.663</td>
<td>notify5.dropbox.com</td>
</tr>
<tr>
<td>2016-03-05</td>
<td>14:17:10.411</td>
<td>ols.officeapps.live.com</td>
</tr>
<tr>
<td>2016-03-05</td>
<td>15:29:22.510</td>
<td>api.textmate.org</td>
</tr>
</tbody>
</table>
Example 2: reconstruct visited websites
– What websites did Eve visit before we fired her?
– Which users surfed to www.yahoo.com last week?

Searching for www.yahoo.com ...

09:41:20.242  ad4.adition.com
09:41:21.770  ads.nuggad.com
09:41:40.152  skypedata.akadns.net
09:42:41.985  dl-debug.dropbox.com
09:45:11.201  google.com  visited
09:46:00.033  www.heise.de  visited
09:46:00.133  dealbook.nytimes.com
09:46:00.134  pressroom.yahoo.net  DNS prefetching
09:46:00.169  www.yahoo.com
09:46:00.783  imagesrv.adition.com
09:46:00.989  ad.atdmt.com  advertisements & user tracking
09:46:00.989  ad.doubleclick.net
09:46:00.991  imagerv2.adition.com
09:46:01.017  jobs.heise.de  embedded image
Simple heuristics look promising ...
... but are not always accurate.

Heuristic search:
\[ \Delta t > 5 \text{ sec} \]

09:41:20.242 ad4.adition.com
09:41:21.770 ads.nuggad.com
09:41:40.152 skypedata.akadns.net
09:42:41.985 dl-debug.dropbox.com
09:45:11.201 google.com \[ true positive \]
09:46:00.033 www.heise.de \[ true positive \]
09:46:00.133 dealbook.nytimes.com
09:46:00.134 pressroom.yahoo.net
09:46:00.169 www.yahoo.com \[ true negative \]
09:46:30.812 [visit Yahoo website] \[ false negative \]

www.yahoo.com cached for 1–5 min
Can we use the **set of domains** to verify whether a website was visited?
Experimental approach: 1. Download websites with a browser  
2. Record resolved hostnames  
3. Determine $k$-identifiability of websites

Measurements indicate: 
many websites have a unique DNS pattern

- **visited home page**
  - **ALEXA** top 100,000 websites
  - $k = 1$: 99 %
  - $k \leq 5$: 99 %

- **inference of whole (!) URL**
  - **HEISE** 6283 news pages
  - 63 %
  - 76 %

Interesting problems:  
- robustness  
- threshold for match  
- influence of cache
Is DNS-based visited website verification still possible?

Logging of flow records (common practice)

DNS log might not be available (due to data protection obligations)

Only packet sizes are logged (no domain names)

However: DNS packet size correlated with domain name length

Yahoo’s DNS flow record fingerprint (multiset of 51 domain name lengths)
Measurements indicate:
domain lengths multiset is characteristic

- For domain names available in the ALEXA top 100,000 websites:
  - $k = 1$: 99%
  - $k \leq 5$: 99%

- For only domain name lengths:
  - 69% (ALEXA top 100,000 websites)
  - 77% (ALEXA top 1000: 75%)
useful for forensics

real-world accuracy?
utility for law enforcement?
probative value of evidence?

drawing inferences from DNS logs and flow records

privacy concerns

your own resolver
resolver of ISP
resolver “in the cloud”

e.g. by Google and OpenDNS
But third-party DNS resolvers cannot track their users — or can they?

Challenge:
IP address changes frequently (daily)
3 May 2015
spiegel.de 4 x
google.de 15 x
apple.com 1 x
airbus.com 3 x
mpg.de 2 x

re-identification via
resolved domains

Do users have
distinct habits?

4 May 2015
1 x spiegel.de
9 x google.de
0 x apple.com
6 x airbus.com
3 x mpg.de
Sessions are modelled as vectors that are compared with cosine similarity (nearest-neighbor classifier).

Yesterday

<table>
<thead>
<tr>
<th>Session 1</th>
<th>Session 2</th>
<th>Session 3</th>
<th>Session 4</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 0 0 0 0 2 0 0 1 0</td>
<td>0 0 1 0 6 0 0 4 0 0</td>
<td>0 9 7 3 0 0 0 0 0 0</td>
<td>0 0 0 2 0 1 0 0 9 0</td>
<td></td>
</tr>
</tbody>
</table>

Today

<table>
<thead>
<tr>
<th>Session 5</th>
<th>Session 6</th>
<th>Session 7</th>
<th>Session 8</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 0 0 0 0 3 0 0 1 0</td>
<td>0 0 0 0 3 0 0 1 0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Cosine similarity:

- \( \cos = 0.86 \)
How accurate is behavior-based tracking in practice?

Experimental approach:
1. Obtain DNS log with realistic traffic
2. Track users day to day (24h sessions)
3. Determine overall accuracy

re-identification accuracy [%]

<table>
<thead>
<tr>
<th>Raw</th>
<th>+8</th>
<th>+10</th>
<th>+2</th>
</tr>
</thead>
<tbody>
<tr>
<td>75</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

DNS Log
- 61 days
- >3800 users
- 5 mn. domains

with »ground truth« (pseudonymized)

Commonly applied transformations
How accurate is behavior-based tracking in practice?

re-identification accuracy [%]

<table>
<thead>
<tr>
<th>Errors</th>
<th>Ambiguous Predictions</th>
</tr>
</thead>
<tbody>
<tr>
<td>75</td>
<td>14</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Raw</th>
<th>Bigrams</th>
<th>Log</th>
<th>Idf</th>
</tr>
</thead>
<tbody>
<tr>
<td>54</td>
<td>+8</td>
<td>+2</td>
<td>+10</td>
</tr>
</tbody>
</table>

Ambiguous predictions ... can be resolved

Accuracy

Number of Users

Date: 05/08 to 05/29
How accurate is behavior-based tracking in practice?

Re-identification accuracy [%]

86
75
86
75
54
Bigrams
Log
Idf
Prune

Accuracy
Number of users
Date
05/08
05/15
05/22
05/29
0.9
0.7
3000
2000
1000
Application to network forensics:

*How accurate is user re-identification with flow records only?*

<table>
<thead>
<tr>
<th>Domain names</th>
<th>Re-identification accuracy [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>86</td>
<td></td>
</tr>
<tr>
<td>Raw</td>
<td>54</td>
</tr>
<tr>
<td>Bigrams</td>
<td>+8</td>
</tr>
<tr>
<td>Log</td>
<td>+10</td>
</tr>
<tr>
<td>Idf</td>
<td>+2</td>
</tr>
<tr>
<td>Prune</td>
<td>+11</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Domain name lengths</th>
<th>Re-identification accuracy [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>76</td>
<td></td>
</tr>
<tr>
<td>Raw</td>
<td>21</td>
</tr>
<tr>
<td>Log</td>
<td>+8</td>
</tr>
<tr>
<td>Idf</td>
<td>+8</td>
</tr>
<tr>
<td>Prune</td>
<td>+25</td>
</tr>
<tr>
<td>Ngrams</td>
<td>+4</td>
</tr>
<tr>
<td>With markers</td>
<td>+28</td>
</tr>
</tbody>
</table>

Idea of ngram markers:

- Observed: 15 30 [pause ≥ 5 s] 18
- Bigrams: 15–30 30–P P–18
Behavior-based re-identification is quite robust.

- only $N$ most popular domains (not all 5 million)
- number of days between sessions (instead of 1)

100

62%

76%

28

86%

for 50% of users >90% of sessions linked correctly

12,015  76%

number of users (instead of 3800)
behavior-based linkage of browsing sessions

significant because undetectable
threatens informational self-determination

accuracy improvements?

yes
work in progress

exploitable by ad-networks?

other applications?
forensics
authentication
anomaly detection

affordable protection?

yes
stay tuned
**What should a privacy-preserving DNS resolver look like?**

**Tailored solution: EncDNS**
repurpose resolver of ISP as a proxy for encrypted queries

**Challenge:**
limited space (255 bytes)
cryptobox of Bernstein’s NaCl library (Curve25519)

**Measurements indicate:**
fast and scalable (>6000 queries/sec)
We can exploit **peculiarities of DNS** to improve performance and privacy.

**Observation 1:**
few domains are very popular (power law)
top 10,000 domains: 80% of all queries

**Tailored solution: PushDNS Service**
send DNS records of most popular domains to connected clients

**Traffic requirements (10,000 domains):**
- resolving domains: 350 MB per day
- pushing updates: 0.8 KB/s per user

**Observation 2:**
most IPs constant over long time
for 50% of domains: TTL > 5 min

**Consequence:** majority of queries **unobservable** and resolved **instantaneously**
Protection against behavior-based tracking
... can be delegated to Internet Service Provider

<table>
<thead>
<tr>
<th>Session Duration</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 min</td>
<td>31 %</td>
</tr>
<tr>
<td>10 min</td>
<td>34 %</td>
</tr>
<tr>
<td>1 h</td>
<td>55 %</td>
</tr>
<tr>
<td>6 h</td>
<td>70 %</td>
</tr>
<tr>
<td>24 h</td>
<td>86 %</td>
</tr>
<tr>
<td>7 days</td>
<td>97 %</td>
</tr>
</tbody>
</table>

Change IP address frequently!

Chance for ISPs
Effortless protection with IPv6 Prefix Bouquets

ANON-Next
(BMBF, 2016 – 2019)
A Double-Edged Sword: Traffic Analysis in the Domain Name System

DNS patterns of software and websites
behavior-based tracking of users

INFERENCE IN NETWORKED SYSTEMS

PRIVACY ENHANCING TECHNOLOGIES
tailored protection tools promising
effortless tracking protection by delegation

https://dhgo.to/dh  Dr. Dominik Herrmann  https://dhgo.to/dns-traffic-analysis