A Double-Edged Sword: Metadata Collection in the Domain Name System (DNS)

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

Dr. Dominik Herrmann
Browser → DNS Resolver → www.yahoo.com

Cache

www.yahoo.com

46.228.47.114

HTTP

GET /celebrity/

HTML file

Yahoo!

Mailing List

News

Sports

Finance

Autos

Celebrity

Shopping

Movies

Politics

Beauty

More

Trending Now

1. The Spotlight movie
2. Boston Celtics
3. Abraham Attah
4. Jared Leto
5. Evening dresses
6. Migraine headaches
7. Kim Kardashian
8. Jacob Tremblay
9. Toyota Corolla
10. Gigi Hadid

Never Old. Never New. Get yours >

foreverspin™

MADE IN CANADA
Motivation of monitoring DNS
– block known malicious domains (e.g. phishing)
– retain log of all DNS queries for later analysis

<table>
<thead>
<tr>
<th>retained data</th>
<th>log size [%]</th>
<th>level of detail</th>
</tr>
</thead>
<tbody>
<tr>
<td>HTTP(S) traffic</td>
<td>100.00</td>
<td>&lt;html&gt;&lt;head&gt;&lt;title&gt;Yahoo&lt;/title&gt;...</td>
</tr>
<tr>
<td>HTTP(S) URLs</td>
<td>0.81</td>
<td><a href="http://www.yahoo.com/celebrity/">http://www.yahoo.com/celebrity/</a></td>
</tr>
<tr>
<td>DNS names</td>
<td>0.04</td>
<td><a href="http://www.yahoo.com">www.yahoo.com</a></td>
</tr>
</tbody>
</table>

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
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?
Example 1: confirm source of traffic
Did incriminating traffic originate from Bob’s laptop?

DNS queries from Bob’s IP

<table>
<thead>
<tr>
<th>Date</th>
<th>Time</th>
<th>Domain</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?

**Yahoo malvertising attack leaves 900 million at risk of ra...**

IT PRO - 4 Aug 2015

A huge *malvertising* campaign that took over *Yahoo's* advertising network for four days last month could have hundreds of millions of potential ...
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* ...

<table>
<thead>
<tr>
<th>Date</th>
<th>Time</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>2016-03-05</td>
<td>09:41:20.242</td>
<td>ad4.adition.com</td>
</tr>
<tr>
<td>2016-03-05</td>
<td>09:41:21.770</td>
<td>ads.nuggad.com</td>
</tr>
<tr>
<td>2016-03-05</td>
<td>09:41:40.152</td>
<td>skypedata.akadns.net</td>
</tr>
<tr>
<td>2016-03-05</td>
<td>09:42:41.985</td>
<td>dl-debug.dropbox.com</td>
</tr>
<tr>
<td>2016-03-05</td>
<td>09:45:11.201</td>
<td>google.com</td>
</tr>
<tr>
<td>2016-03-05</td>
<td>09:46:00.033</td>
<td><a href="http://www.heise.de">www.heise.de</a></td>
</tr>
<tr>
<td>2016-03-05</td>
<td>09:46:00.133</td>
<td>dealbook.nytimes.com</td>
</tr>
<tr>
<td>2016-03-05</td>
<td>09:46:00.134</td>
<td>pressroom.yahoo.net</td>
</tr>
<tr>
<td><strong>2016-03-05 09:46:00.169</strong></td>
<td><strong><a href="http://www.yahoo.com">www.yahoo.com</a></strong></td>
<td><strong>false positive</strong></td>
</tr>
<tr>
<td>2016-03-05</td>
<td>09:46:00.783</td>
<td>imagesrv.adition.com</td>
</tr>
<tr>
<td>2016-03-05</td>
<td>09:46:00.989</td>
<td>ad.atdmt.com</td>
</tr>
<tr>
<td>2016-03-05</td>
<td>09:46:00.989</td>
<td>ad.doubleclick.net</td>
</tr>
<tr>
<td>2016-03-05</td>
<td>09:46:00.991</td>
<td>imagerv2.adition.com</td>
</tr>
<tr>
<td>2016-03-05</td>
<td>09:46:01.017</td>
<td>jobs.heise.de</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?

51 domains resolved when Yahoo’s home page is visited

- www.yahoo.com
- bs.serving-sys.com
- pclick.yahoo.com
- s.yimg.com
- sb.scorecardresearch.
- crl-ds.ws.symantec.co.
- y.analytics.yahoo.com
- geo.query.yahoo.com
- csc.beap.bc.yahoo.com
- geo.yahoo.com
- comet.yahoo.com
- answers.yahoo.com
- everything.yahoo.com
- groups.yahoo.com
- login.yahoo.com
- mail.yahoo.com
- mobile.yahoo.com
- shopping.yahoo.com
- www.flickr.com
- www.tumblr.com
- beap.gemini.yahoo.com
- finance.yahoo.com
- ftw.usatoday.com
- geo.um.btrll.com
- googleads.g.doublecli.
- match.adsrvr.org
- pagead2.googlesyndic.
- help.yahoo.com
- info.yahoo.com
- news.yahoo.com
- na.ads.yahoo.com
- pr-bh.ybp.yahoo.com
- r.turn.com
- rmx.pxl.ace.advertisi...
- search.yahoo.com
- sports.yahoo.com
- thinkprogress.org
- sync.adap.tv
- sync.adaptv.advertisi...
- www.cbsnews.com
- ads.yahoo.com
- www.chicagotribune....
- www.foxnews.com
- www.latimes.com
- www.fonts.googleapis.com
- tpc.googlesyndication...
- cm.g.doubleclick.net
- www.npr.org
- www.politico.com
- www.sbnation.com
- www.upi.com
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

![Diagram showing visited home page and inference of whole URL]

<table>
<thead>
<tr>
<th></th>
<th>ALEXA</th>
<th>HEISE</th>
<th>Interesting problems:</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k = 1$</td>
<td>99 %</td>
<td>63 %</td>
<td>– robustness</td>
</tr>
<tr>
<td>$k \leq 5$</td>
<td>99 %</td>
<td>76 %</td>
<td>– threshold for match</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>– influence of cache</td>
</tr>
</tbody>
</table>

ALEXA top 100,000 websites
HEISE 6283 news pages
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)

- www.yahoo.com
- bs.serving-sys.com
- pclick.yahoo.com
- s.yimg.com
- ...

Only packet sizes are logged (no domain names)

However: DNS packet size correlates with domain name length

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

\[
\begin{align*}
k = 1 & \quad 99 \% \\
k \leq 5 & \quad 99 \% \\
\end{align*}
\]
Drawing inferences from DNS logs and flow records

Useful for forensics

Real-world accuracy?

Utility for law enforcement?

Probative value of evidence?

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)
Do users have distinct habits?

re-identification via resolved domains
**Sessions** are modelled as **vectors** that are compared with **cosine similarity** (nearest-neighbor classifier)

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

- **airbus.com**
- **bahn.de**

\[ \text{cos} = 0.43 \]

\[ \text{cos} = 0.86 \]

\[ \text{cos} = 0.43 \]
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 [%]

75

- Borrowed from raw
- +8 bigrams
- +10 log
- +2 idf

commonly applied transformations

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

with »ground truth« (pseudonymized)
How accurate is behavior-based tracking in practice?

re-identification accuracy [%]

errors

ambiguous predictions

14

75

+10

+2

raw

+8

bigrams

log

idf

ambiguous prediction
... can be resolved

number of users

accuracy

date

05/08 05/15 05/22 05/29

0.7

0.9

3000

2000

1000

How accurate is behavior-based tracking in practice?
How accurate is behavior-based tracking in practice?

re-identification accuracy [%]

<table>
<thead>
<tr>
<th></th>
<th>Sunday</th>
<th>Monday</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>0 0 2 0 4 0 0 4 0</td>
<td>0 8 5 3 0 0 0 0</td>
</tr>
<tr>
<td>S2</td>
<td>0 9 7 8 0 0 0 0 0</td>
<td>0 0 1 0 6 0 0 7 0</td>
</tr>
<tr>
<td>S3</td>
<td>0 0 0 3 0 0 0 9 0</td>
<td>0 0 0 3 0 0 6 0 0</td>
</tr>
<tr>
<td>S4</td>
<td>inactive</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>date</th>
<th>05/08</th>
<th>05/15</th>
<th>05/22</th>
<th>05/29</th>
</tr>
</thead>
<tbody>
<tr>
<td>number of users</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pruned</td>
<td>3000</td>
<td>2000</td>
<td>1000</td>
<td>3000</td>
</tr>
</tbody>
</table>

86
75

86
75

raw

+8
+10
+2

86
75

86
75

raw

+8
+10
+2

How accurate is behavior-based tracking in practice?
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>Domain name lengths</td>
<td>Re-identification accuracy [%]</td>
</tr>
</tbody>
</table>

### Domain Names

<table>
<thead>
<tr>
<th>Domain name lengths</th>
<th>Re-identification accuracy [%]</th>
</tr>
</thead>
</table>

---

Idea of ngram markers:

- **observed**: 15 30 [pause ≥ 5 s] 18
- **bigrams**: 15–30 30–p P–18

---

<table>
<thead>
<tr>
<th>Raw</th>
<th>Bigrams</th>
<th>Log</th>
<th>Idf</th>
<th>Prune</th>
</tr>
</thead>
<tbody>
<tr>
<td>54</td>
<td>+8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>86</td>
<td>+10</td>
<td>+2</td>
<td></td>
<td>+11</td>
</tr>
</tbody>
</table>

---

<table>
<thead>
<tr>
<th>Raw</th>
<th>Prune</th>
<th>Ngrams</th>
<th>With markers</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>+8</td>
<td>+25</td>
<td>+28</td>
</tr>
<tr>
<td>37</td>
<td>+8</td>
<td>+6</td>
<td>+7</td>
</tr>
<tr>
<td>76</td>
<td>+3</td>
<td>4gr</td>
<td>+4</td>
</tr>
<tr>
<td>54</td>
<td></td>
<td>3gr</td>
<td></td>
</tr>
</tbody>
</table>

---
Behavior-based re-identification is quite robust.

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

- 62% for 100
- 76% for 28
- 86% for 50% of users >90% of sessions linked correctly

- 76% for 12,015
- 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

generic anonymization services (Tor) too slow
What should a privacy-preserving DNS resolver look like?

generic anonymization services (Tor) too slow

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)

Alice

encrypted
y%8§1...aZ.encdns.com
Sender: Alice

resolver of ISP or third party

encrypted
y%8§1...aZ.encdns.com
Sender: Resolver

nameserver for zone encdns.com

www.cnn.com
Sender: encdns.com

nameserver for zone cnn.com
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

<table>
<thead>
<tr>
<th>next domain</th>
<th>TTL expired?</th>
</tr>
</thead>
<tbody>
<tr>
<td>no</td>
<td>push to clients</td>
</tr>
<tr>
<td>yes</td>
<td>resolve domain</td>
</tr>
<tr>
<td></td>
<td>new response?</td>
</tr>
</tbody>
</table>

**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: Metadata Collection 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

IPv6 Prefix Bouquets

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