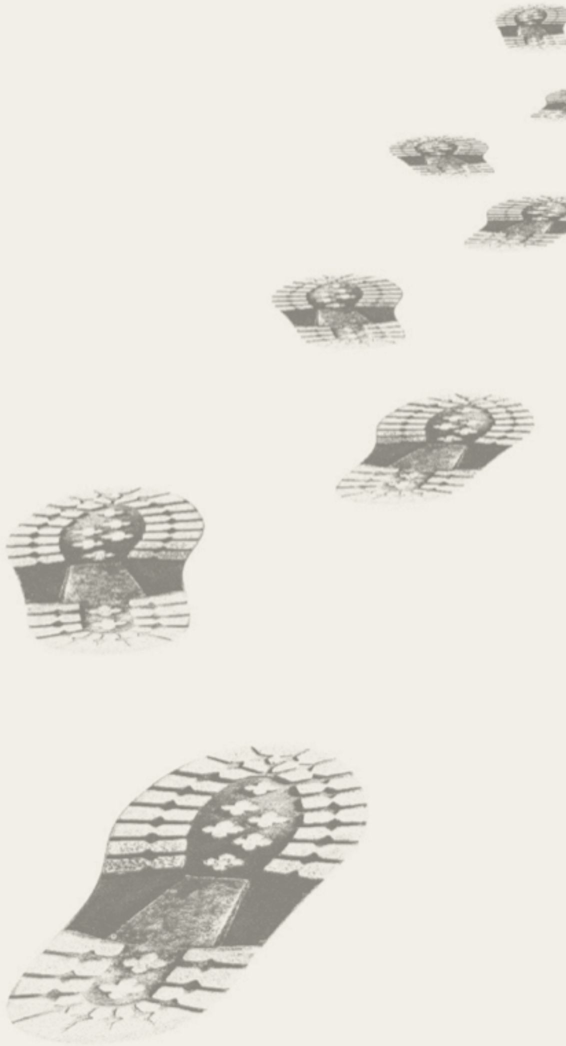
A Double-Edged Sword: **Metadata Collection in the Domain Name System (DNS)**

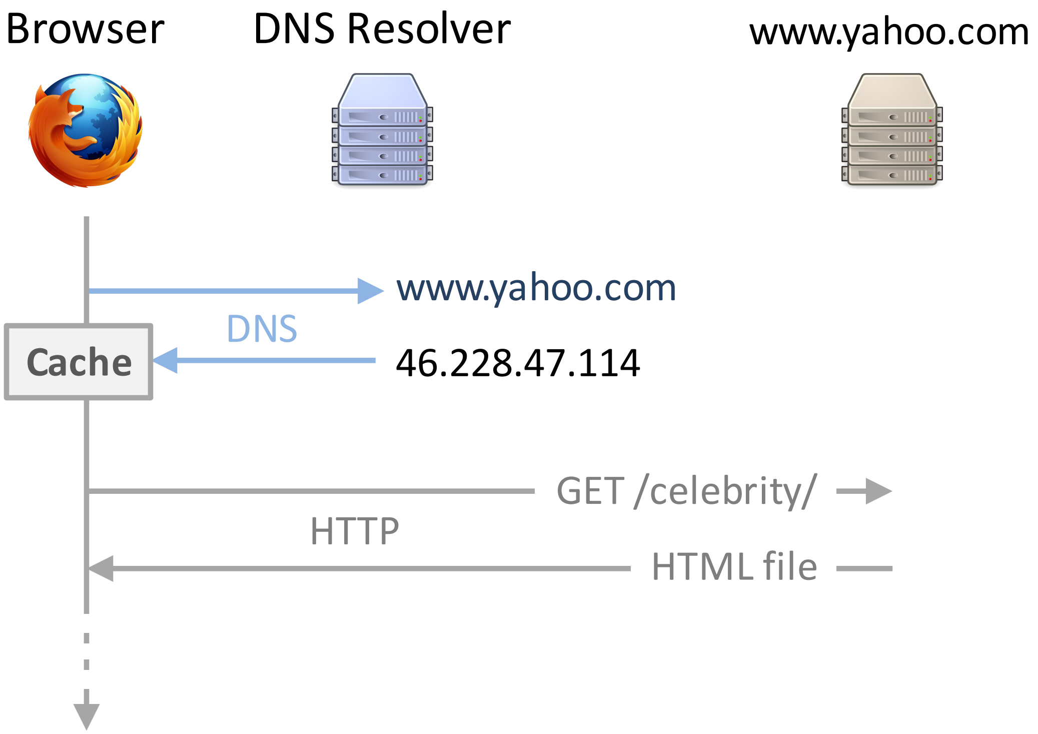
Utility for forensic investigations

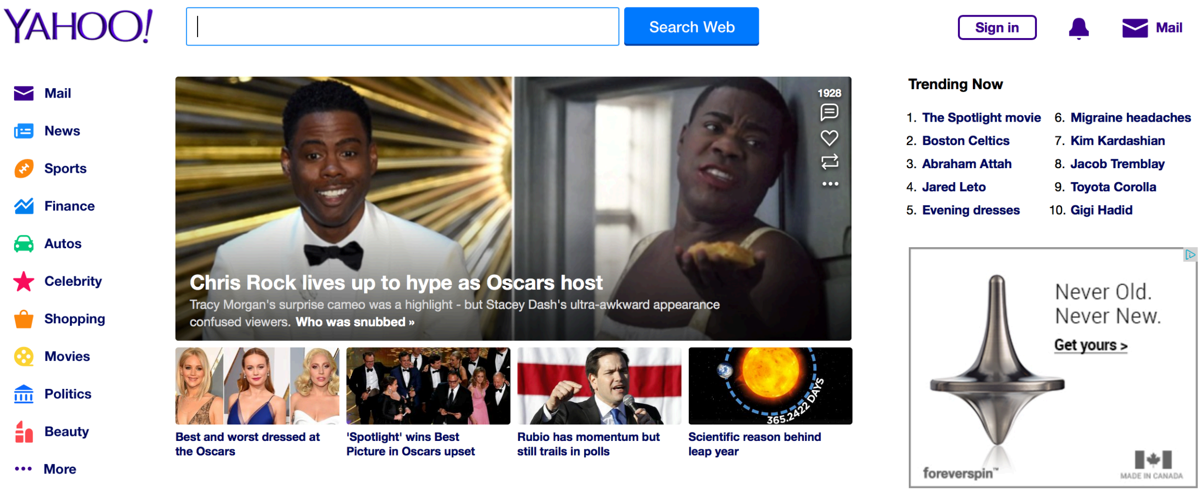
Potential threats to privacy

New ideas for protection



Dr. Dominik Herrmann





- Mail
- News
- Sports
- Finance
- Autos
- Celebrity
- Shopping
- Movies
- Politics
- Beauty
- ••• More

Chris Rock lives up to hype as Oscars host

Tracy Morgan's surprise cameo was a highlight - but Stacey Dash's ultra-awkward appearance confused viewers. **Who was snubbed »**

Best and worst dressed at the Oscars

'Spotlight' wins Best Picture in Oscars upset

Rubio has momentum but still trails in polls

Scientific reason behind leap year

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

Motivation of monitoring DNS

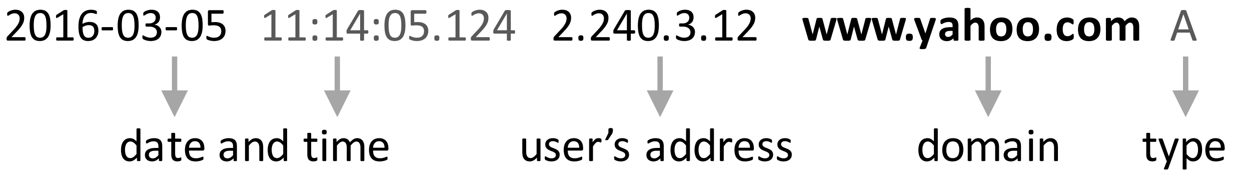
- block known malicious domains (e.g. phishing)
- retain log of all DNS queries for later analysis

OpenDNS

retained data	log size [%]	level of detail
HTTP(S) traffic	100.00	<html><head><title>Yahoo</title…
HTTP(S) URLs	0.81	http://www.yahoo.com/celebrity/
DNS names	**0.04**	**www.yahoo.com**

low storage needs

DNS log contains essential metadata:



2016-03-05 11:14:05.124 2.240.3.12 **www.yahoo.com** A

Motivation of monitoring DNS

- block known malicious domains (e.g. phishing)
- retain log of all DNS queries for later analysis

OpenDNS

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?

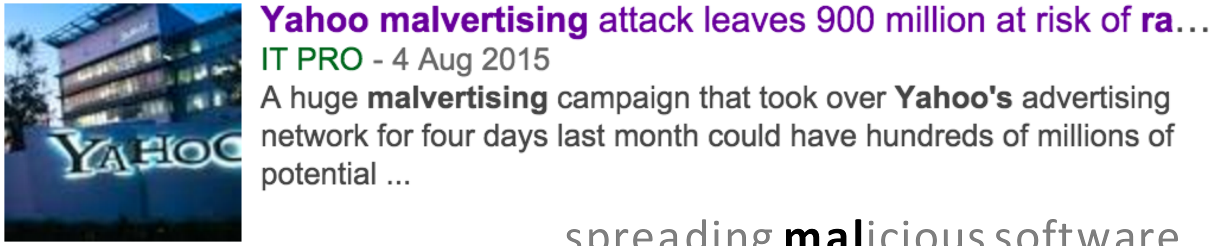
Source of discrepancy?
Rogue hardware?

DNS queries from Bob's IP

```
    2016-03-05   06:46:01.383   aus5.mozilla.org
    2016-03-05   09:41:37.263   c1.adform.net
    2016-03-05   09:46:01.455   www.exploit-db.com
    2016-03-05   09:41:37.708   l.betrad.com
?   2016-03-05   10:22:01.814   time.apple.com
    2016-03-05   09:41:38.262   lotame.nexac.com
    2016-03-05   10:22:01.950   b.config.skype.com
    2016-03-05   09:41:38.686   ping.chartbeat.net
    2016-03-05   14:17:09.663   notify5.dropbox.com
    2016-03-05   09:41:41.627   lb._dns-sd._udp.fritz.box
    2016-03-05   14:17:10.411   ols.officeapps.live.com
    2016-03-05   09:41:48.917   college.usatoday.com
?   2016-03-05   15:29:22.510   api.textmate.org
```

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 ...

spreading **mal**icious software
via online ad**vertising**

Example 2: reconstruct visited websites

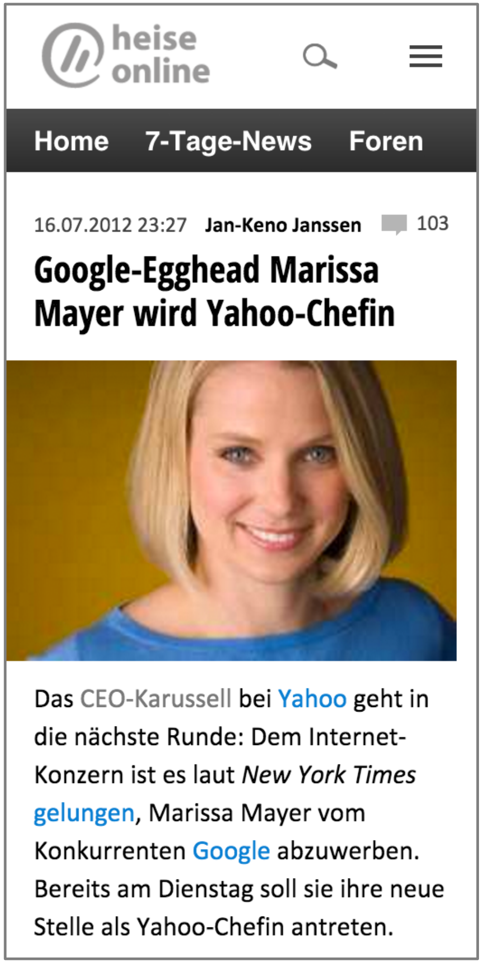
- **What websites** did *Eve* visit before we fired her?
- **Which users** surfed to *www.yahoo.com* last week?

Searching for www.yahoo.com …

```
2016-03-05   09:41:20.242   ad4.adition.com
2016-03-05   09:41:21.770   ads.nuggad.com
2016-03-05   09:41:40.152   skypedata.akadns.net
2016-03-05   09:42:41.985   dl-debug.dropbox.com
2016-03-05   09:45:11.201   google.com
2016-03-05   09:46:00.033   www.heise.de
2016-03-05   09:46:00.133   dealbook.nytimes.com
2016-03-05   09:46:00.134   pressroom.yahoo.net
2016-03-05   09:46:00.169   www.yahoo.com           false positive
2016-03-05   09:46:00.783   imagesrv.adition.com
2016-03-05   09:46:00.989   ad.atdmt.com
2016-03-05   09:46:00.989   ad.doubleclick.net
2016-03-05   09:46:00.991   imagerv2.adition.com
2016-03-05   09:46:01.017   jobs.heise.de
```

Example 2: reconstruct visited websites

- **What websites** did *Eve* visit before we fired her?
- **Which users** surfed to *www.yahoo.com* last week?



16.07.2012 23:27 Jan-Keno Janssen 103

Google-Egghead Marissa Mayer wird Yahoo-Chefin

Das CEO-Karussell bei Yahoo geht in die nächste Runde: Dem Internet-Konzern ist es laut *New York Times* gelungen, Marissa Mayer vom Konkurrenten Google abzuwerben. Bereits am Dienstag soll sie ihre neue Stelle als Yahoo-Chefin antreten.

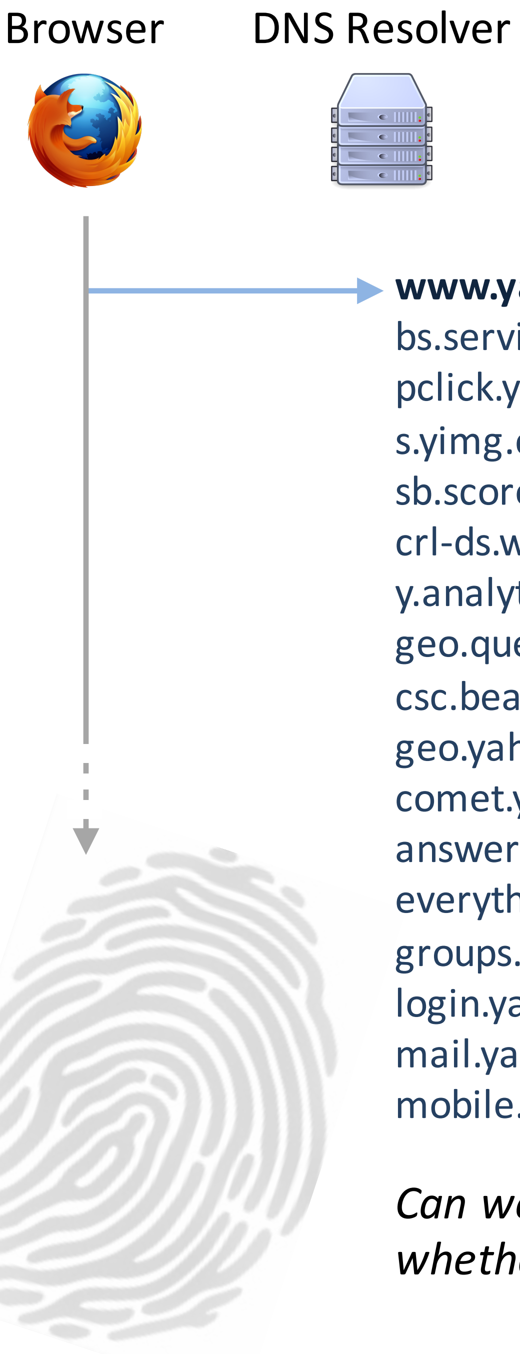
Searching for www.yahoo.com …

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 &**
09:46:00.989	**ad.doubleclick.net**	**user tracking**
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: **Δt > 5 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



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.advertisin…

search.yahoo.com
sports.yahoo.com
thinkprogress.org
sync.adap.tv
sync.adaptv.advertisin…
www.cbsnews.com
ads.yahoo.com
www.chicagotribune.…
www.foxnews.com
www.latimes.com
fonts.googleapis.com
tpc.googlesyndication…
cm.g.doubleclick.net
www.npr.org
www.politico.com
www.sbnation.com
www.upi.com

*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

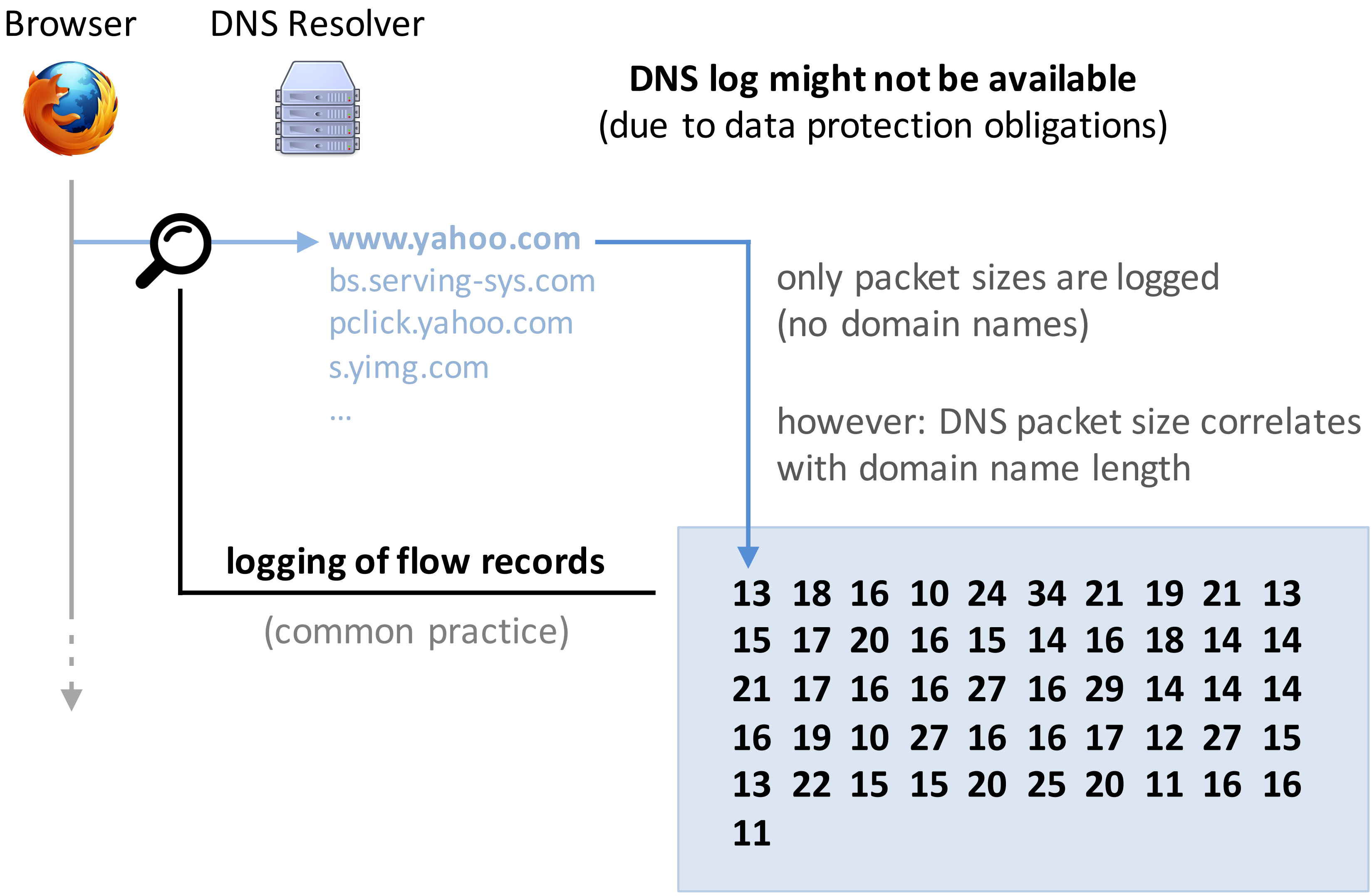
Measurements indicate:

many websites have a unique DNS pattern

	visited home page **ALEXA** top 100 000 websites	inference of **whole (!) URL** **HEISE** 6283 news pages
$k = 1$	99 %	63 %
$k \leq 5$	99 %	76 %

Interesting problems:

- robustness
- threshold for match
- influence of cache



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

only packet sizes are logged
(no domain names)

however: DNS packet size correlates
with domain name length

13	**18**	**16**	**10**	**24**	**34**	**21**	**19**	**21**	**13**
15	**17**	**20**	**16**	**15**	**14**	**16**	**18**	**14**	**14**
21	**17**	**16**	**16**	**27**	**16**	**29**	**14**	**14**	**14**
16	**19**	**10**	**27**	**16**	**16**	**17**	**12**	**27**	**15**
13	**22**	**15**	**15**	**20**	**25**	**20**	**11**	**16**	**16**
11									

Yahoo's DNS flow record fingerprint
(multiset of 51 domain name lengths)

Is DNS-based visited website verification still possible?

Measurements indicate:

domain lengths multiset is characteristic

	domain names available	only domain name lengths	
	ALEXA top 100 000 websites	**ALEXA** top 100 000 websites	
$k = 1$	99 %	69 %	(top 1000: 75%)
$k \leq 5$	99 %	77 %	

drawing inferences from DNS logs and flow records

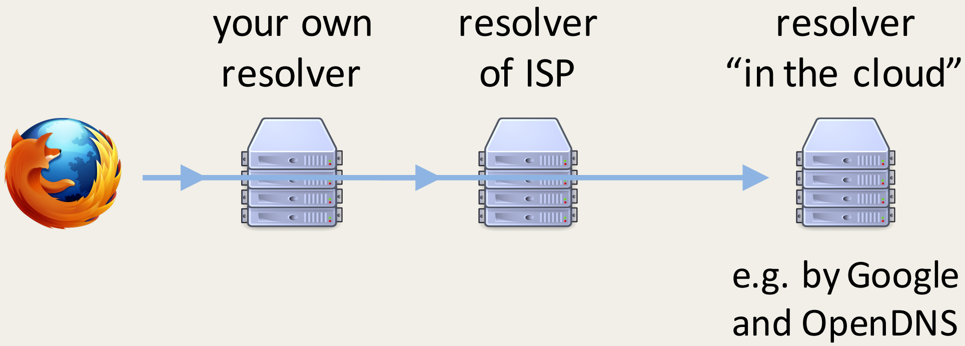
useful for forensics

real-world accuracy?

utility for law enforcement?

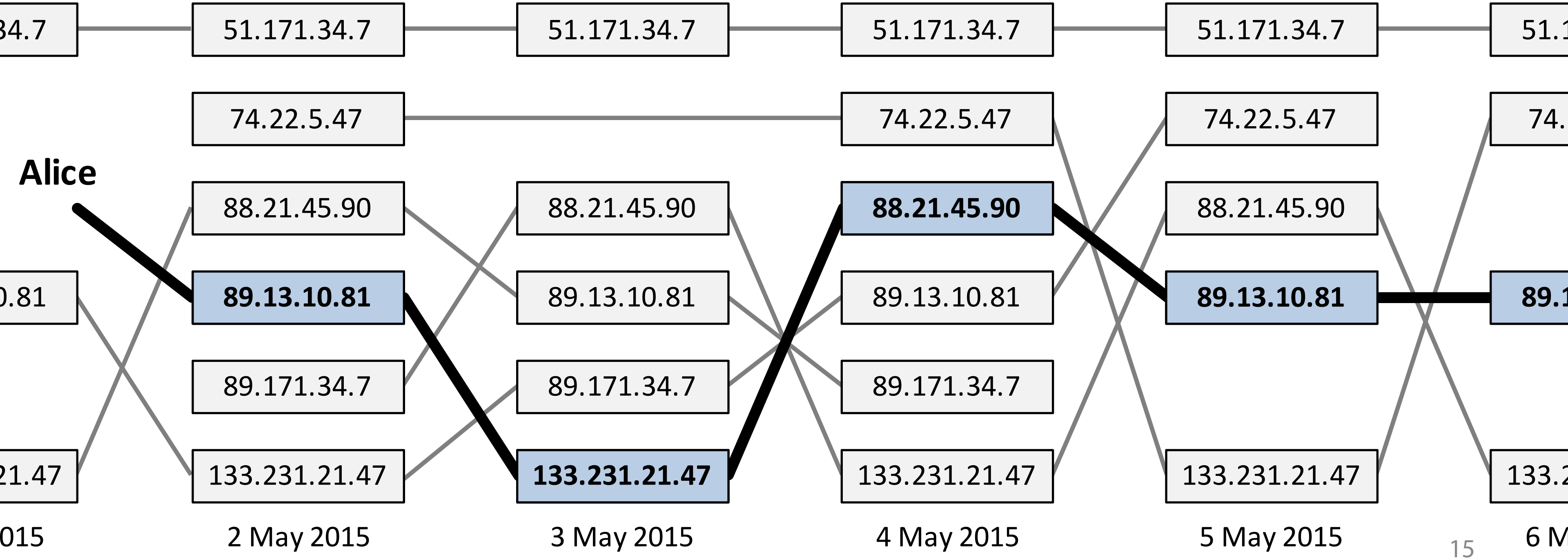
probative value of evidence?

privacy concerns



But third-party DNS resolvers
cannot track their users ***– or can they?***

Challenge:
IP address changes frequently (daily)

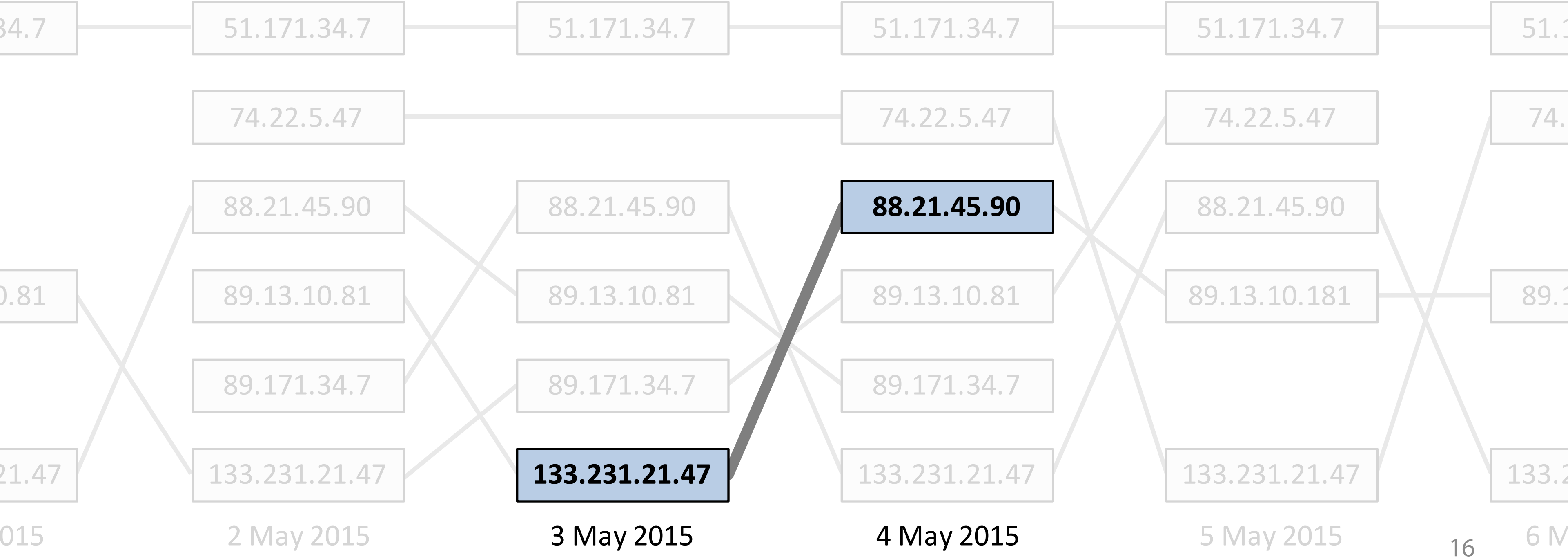


3 May 2015	
spiegel.de	4 x
google.de	15 x
apple.com	1 x
airbus.com	**3 x**
mpg.de	**2 x**

re-identification via resolved domains

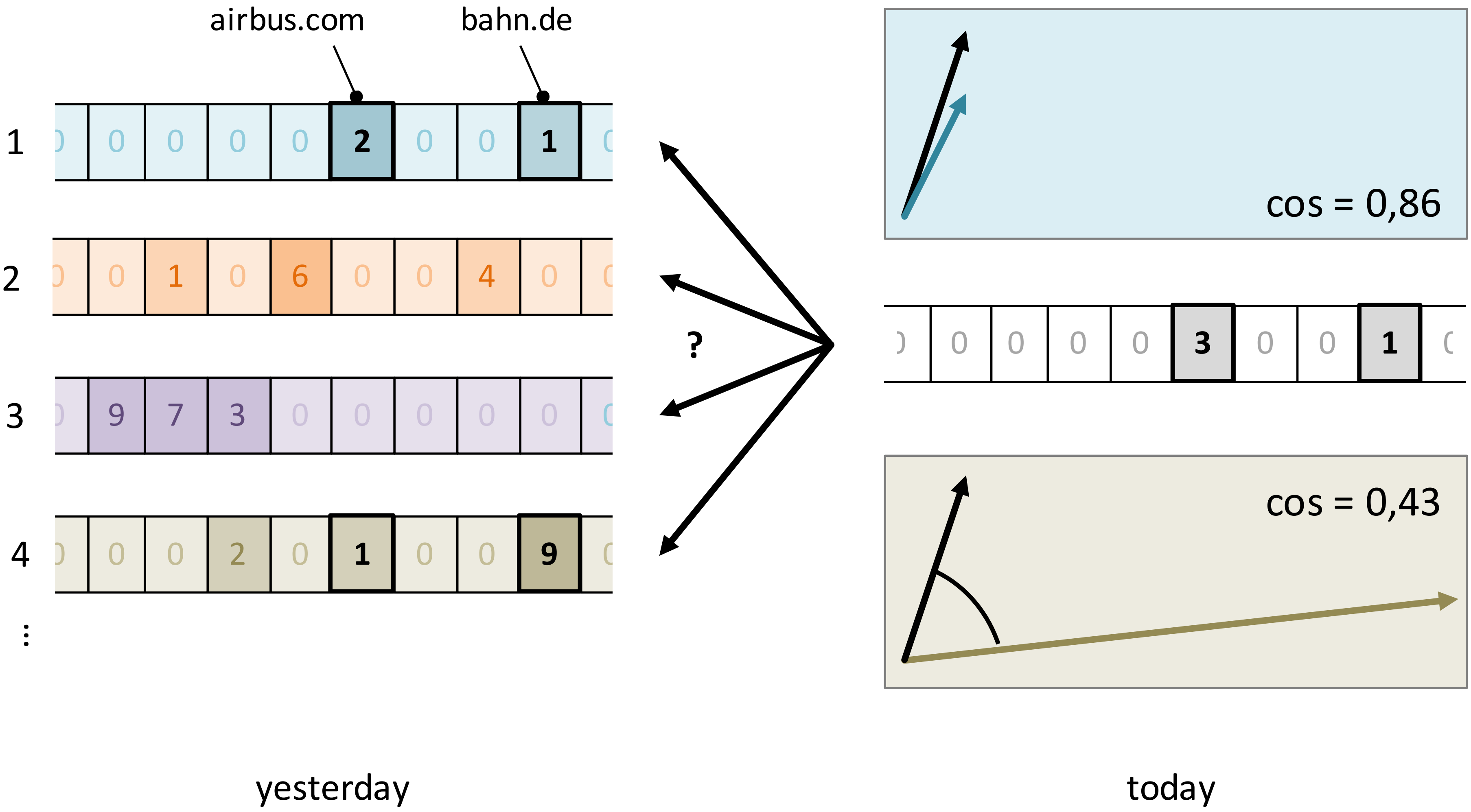
Do users have distinct habits?

	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)

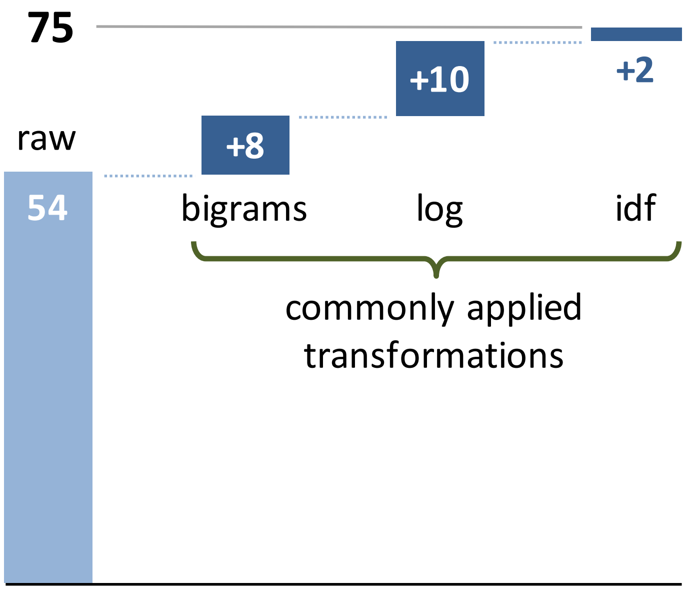


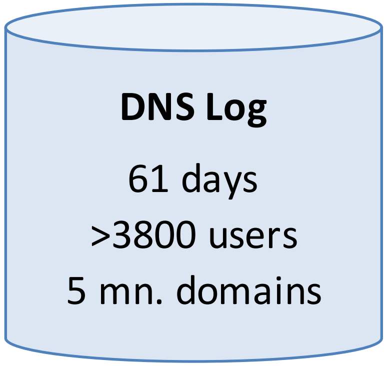
How accurate is behavior-based tracking in practice?

Experimental approach:

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

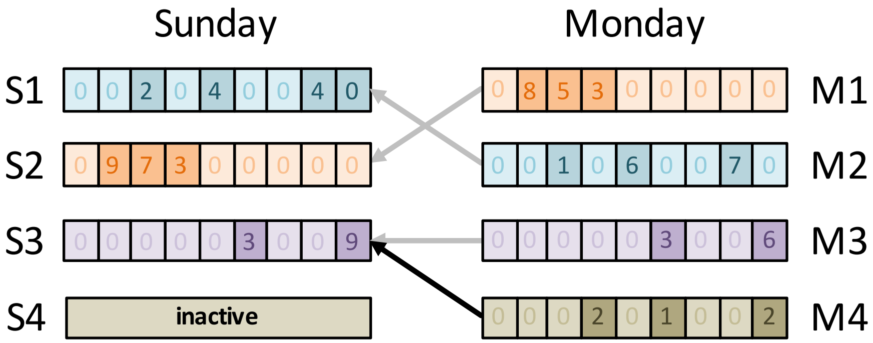
re-identification accuracy [%]





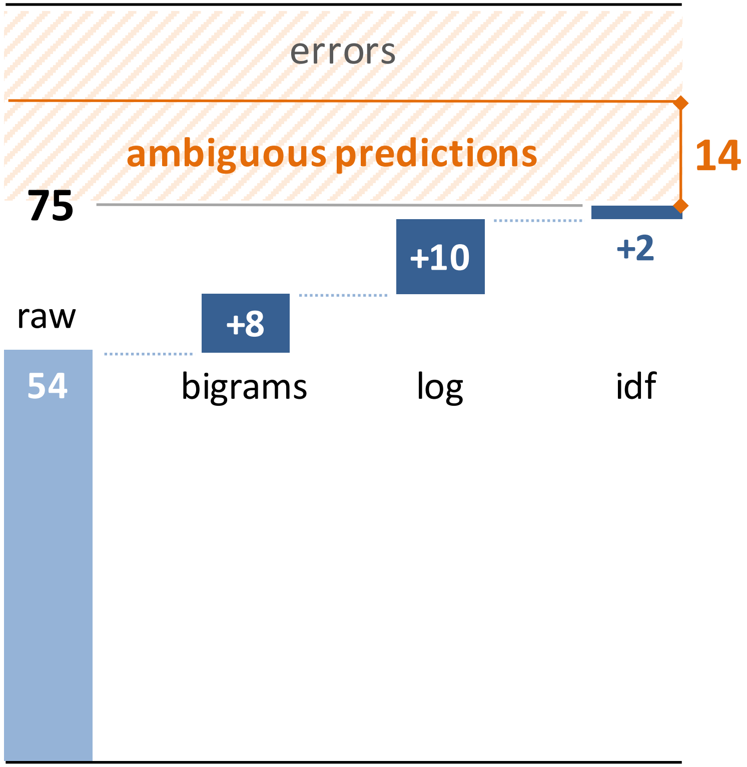
with »ground truth« (pseudonymized)

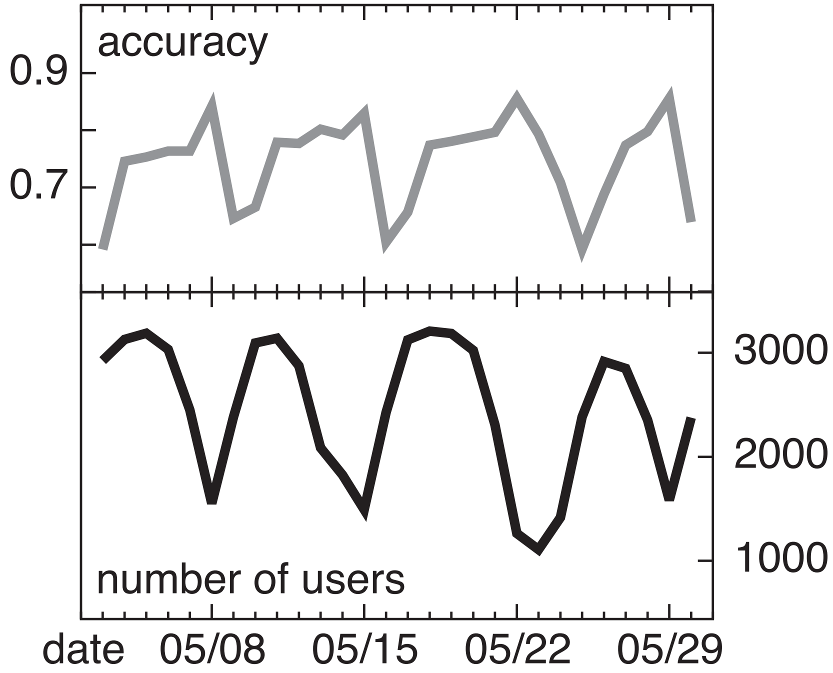
How accurate is behavior-based tracking in practice?



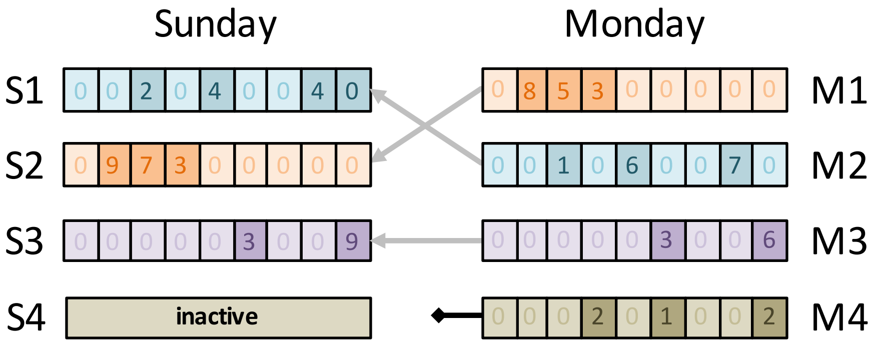
ambiguous prediction
… can be resolved

re-identification accuracy [%]

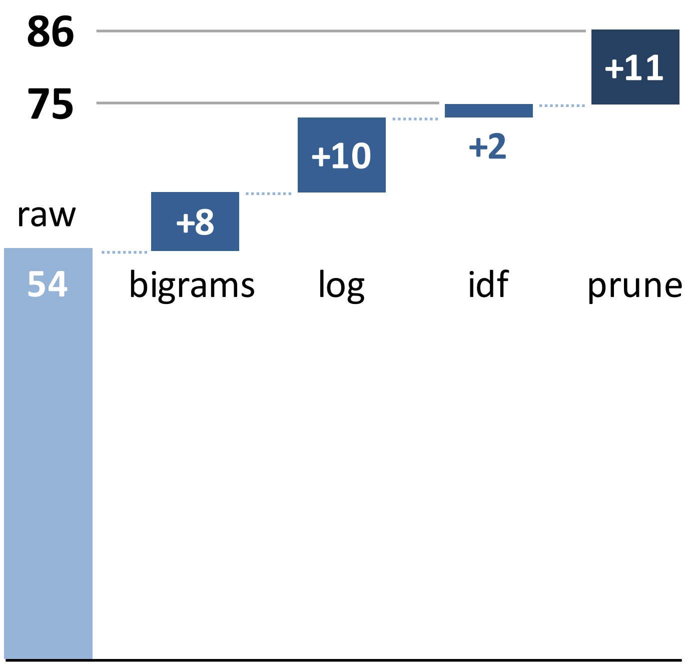


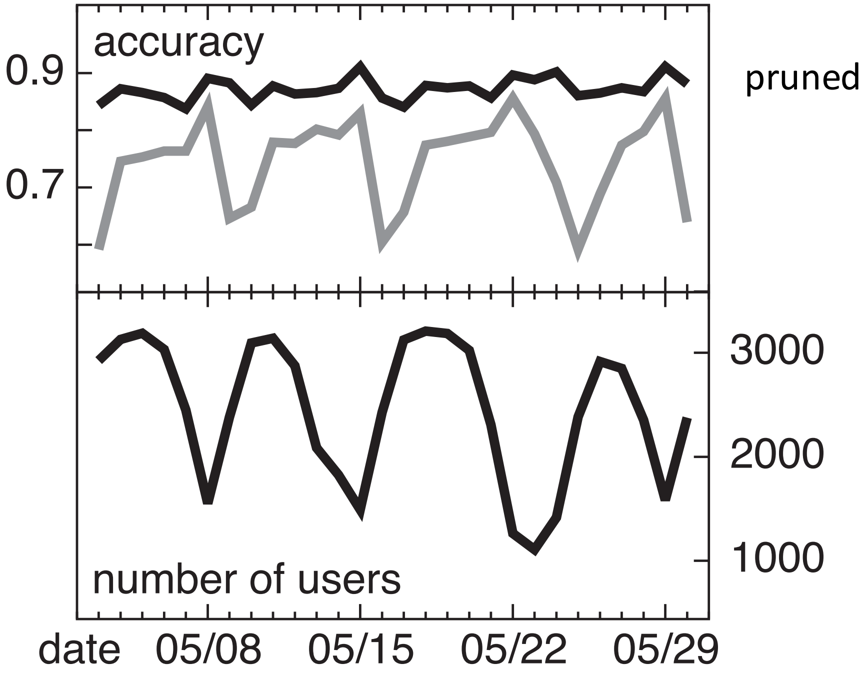


How accurate is behavior-based tracking in practice?



re-identification accuracy [%]





Application to network forensics:

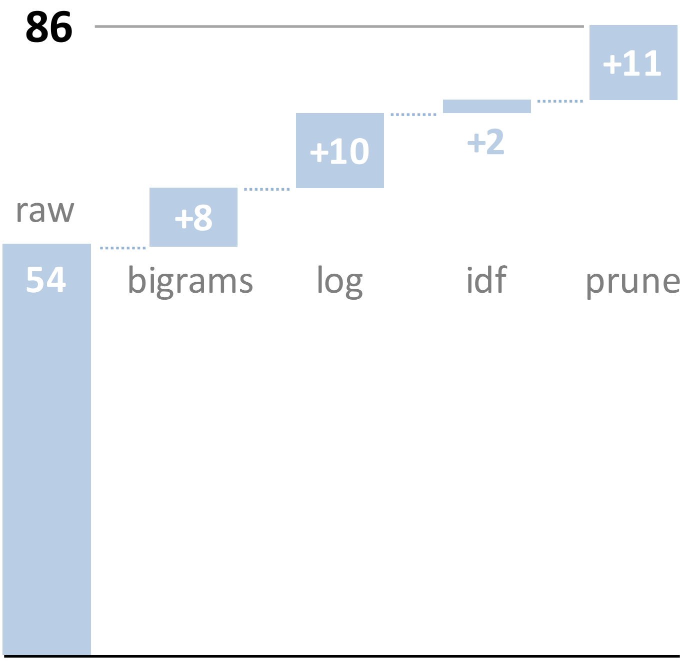
*How accurate is user re-identification with **flow records only**?*

Idea of ngram markers:

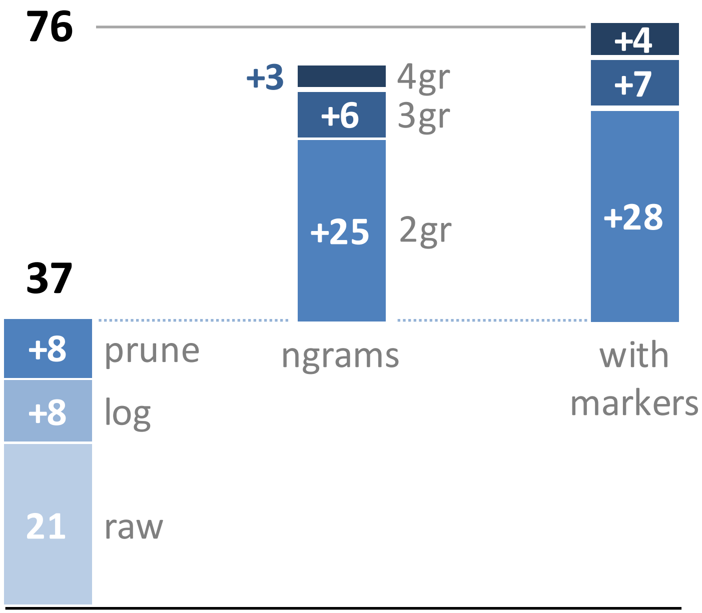
observed: **15 30 [pause ≥ 5 s] 18**

bigrams: **15–30 30–P P–18**

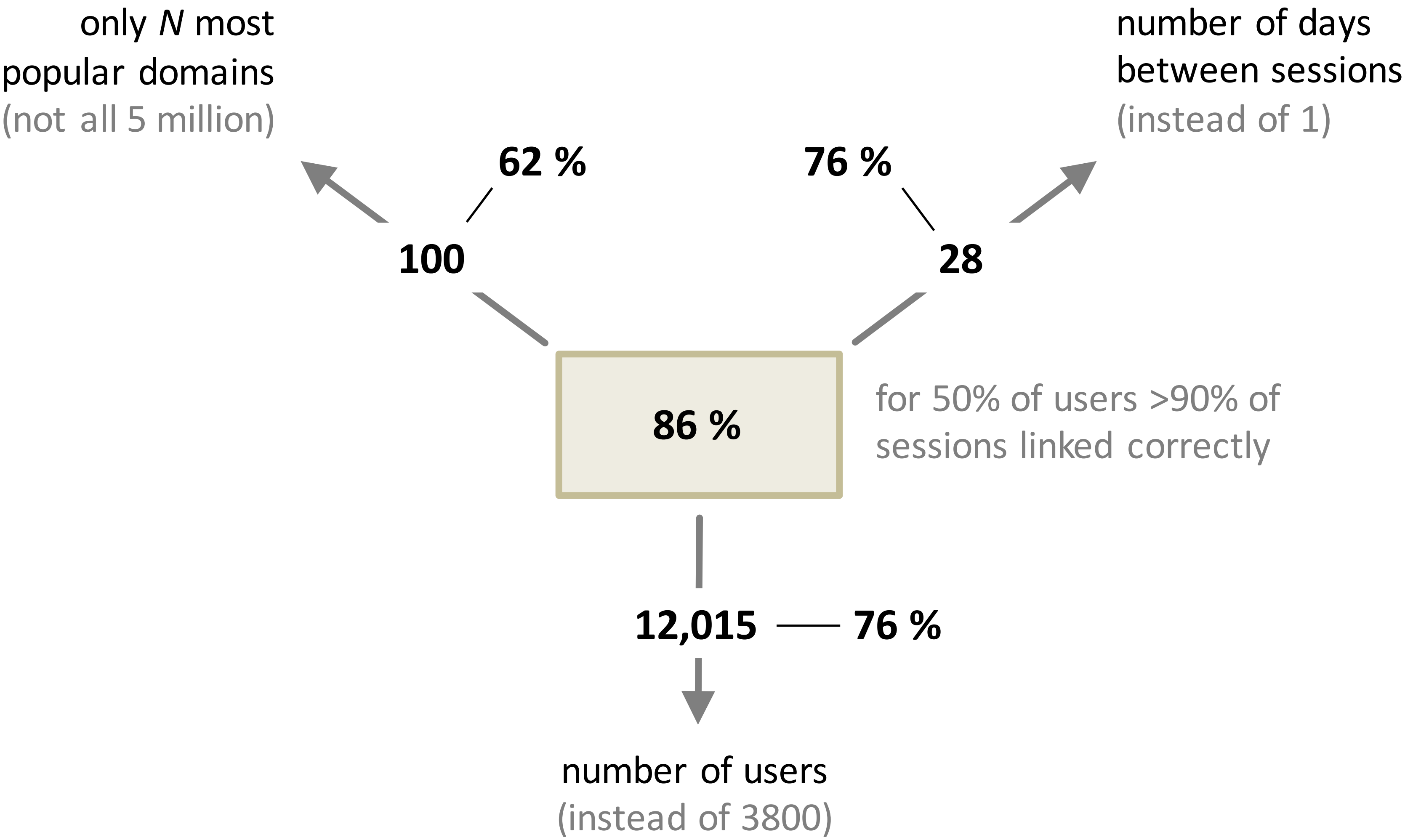
domain names
re-identification accuracy [%]



domain name lengths
re-identification accuracy [%]



Behavior-based re-identification is quite robust.



only *N* most popular domains
(not all 5 million)

62 %

100

number of days between sessions
(instead of 1)

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?

affordable protection?

yes
stay tuned

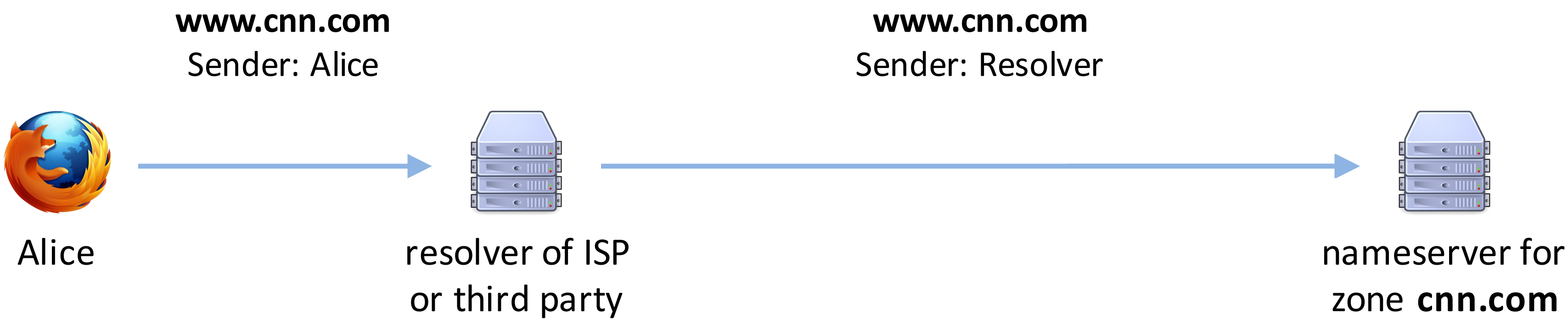
other applications?

forensics
authentication
anomaly detection

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



www.cnn.com
Sender: Alice

www.cnn.com
Sender: Resolver

Alice

resolver of ISP
or third party

nameserver for
zone **cnn.com**

What should a privacy-preserving DNS resolver look like?

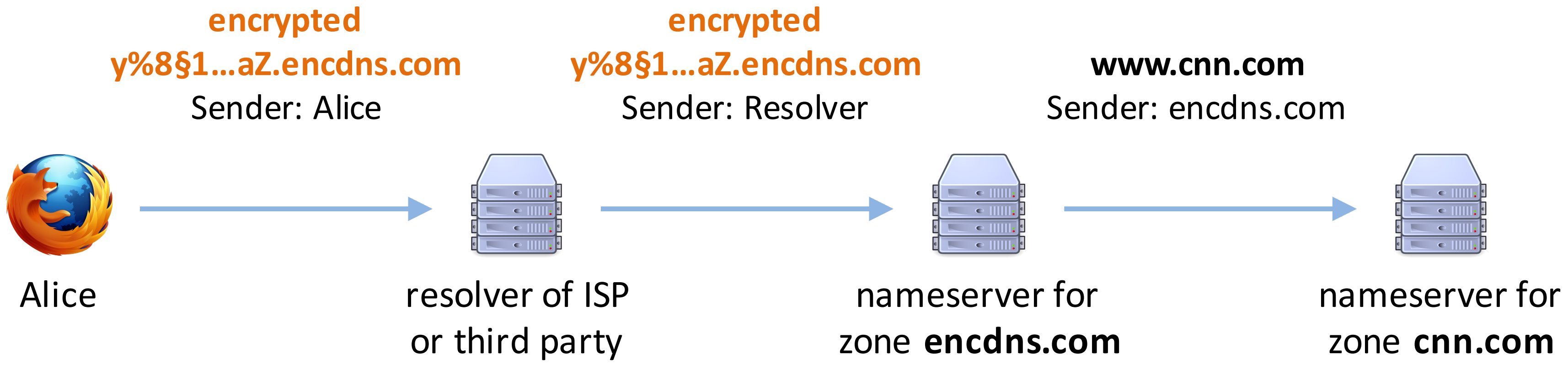
generic anonymization services (Tor) too slow

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)


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

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

www.cnn.com
Sender: encdns.com

Alice

resolver of ISP
or third party

nameserver for
zone **encdns.com**

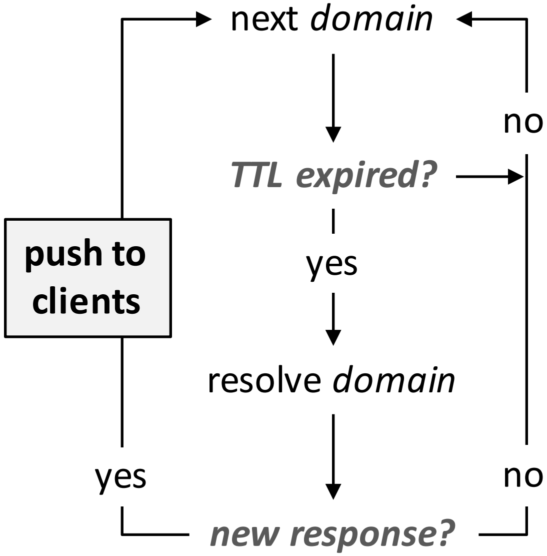
nameserver for
zone **cnn.com**

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

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

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

Consequence: majority of queries **unobservable** and resolved **instantaneously**

Protection against behavior-based tracking

… can be delegated to Internet Service Provider

session duration	accuracy
5 min	**31 %**
10 min	**34 %**
1 h	55 %
6 h	70 %
24 h	**86 %**
7 days	97 %

Change IP address frequently!

Chance for ISPs

Effortless protection with **IPv6 Prefix Bouquets**

ANON-Next

(BMBF, 2016 – 2019)

manitu

opportunity for forensics

A Double-Edged Sword: **Metadata Collection in the Domain Name System**

threat to privacy

time.apple.com

DNS patterns of software and websites

13 18 16 10 24 34
15 17 20 16 15 14

88.21.45.90
89.13.10.81
89.171.34.7
88.21.45.90
89.13.10.81
133.231.21.47
88.21.45.90
89.13.10.81
89.171.34.7

behavior-based tracking of users

0 2 0 1 0 0 2

INFERENCE IN NETWORKED SYSTEMS

PRIVACY ENHANCING TECHNOLOGIES

EncDNS

tailored protection tools promising

PushDNS

effortless tracking protection by delegation

IPv6 Prefix Bouquets

https://dhgo.to/dh

Dr. Dominik Herrmann

https://dhgo.to/dns-metadata