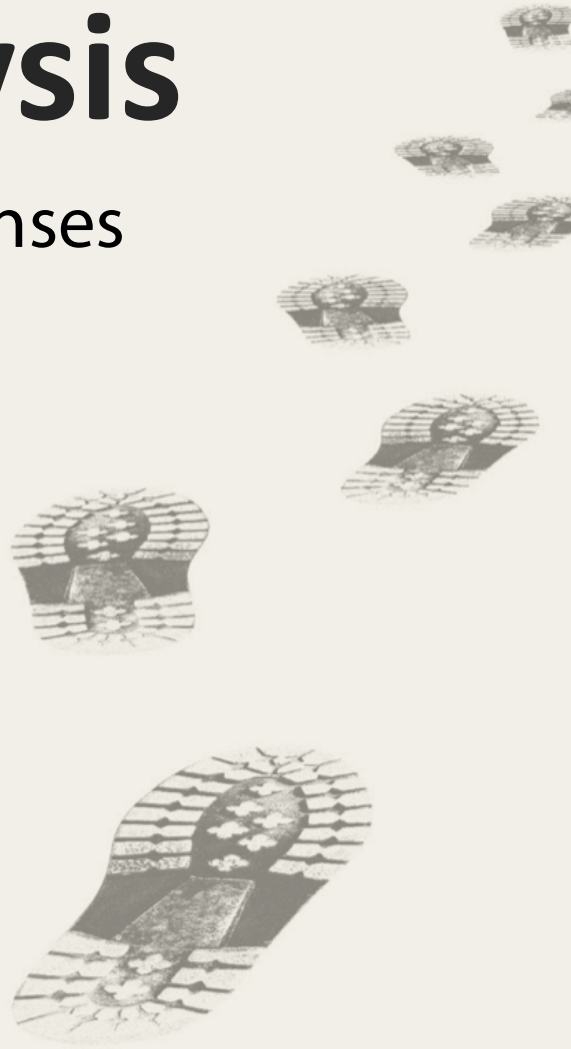
DNS Traffic Analysis

Opportunities, Risks, and (Self-)Defenses

Utility for forensic investigations

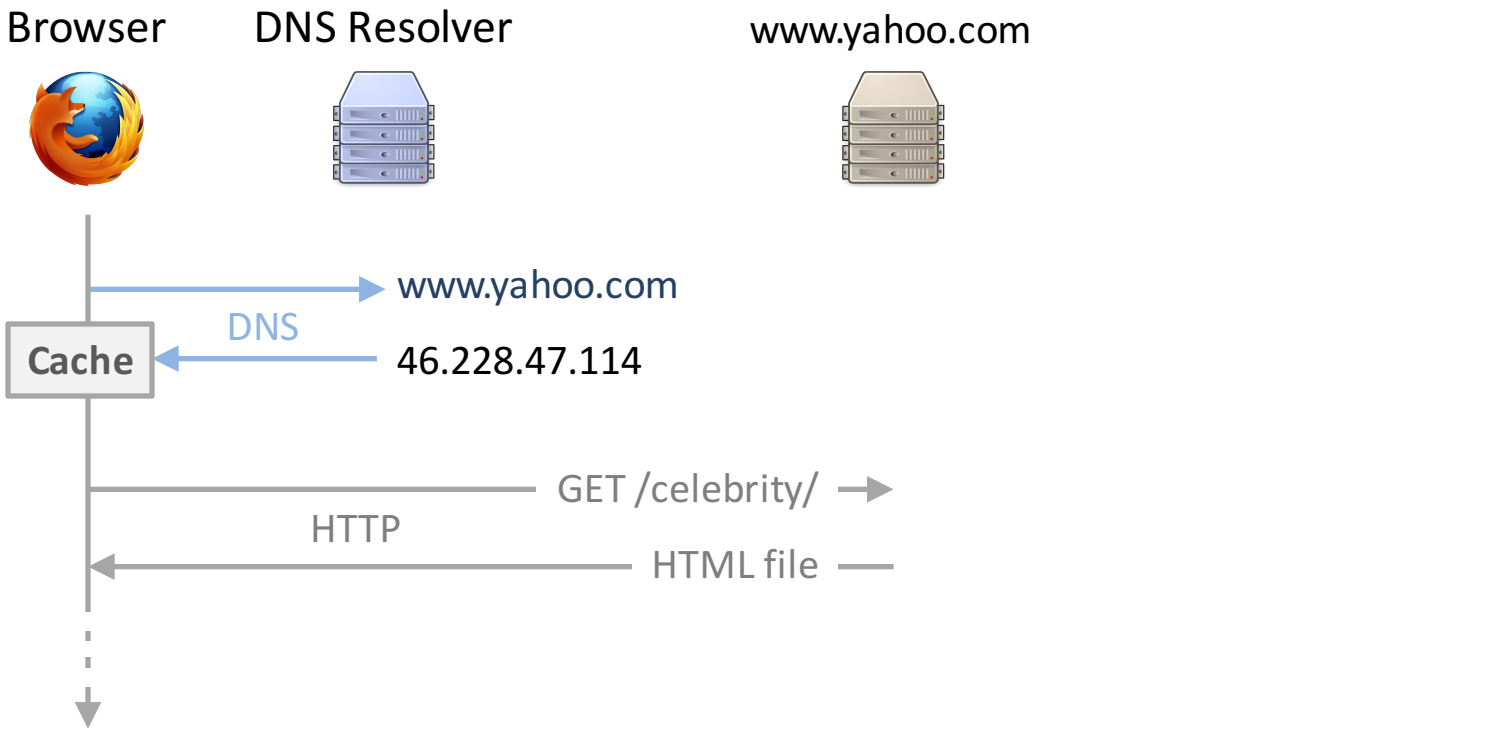
Potential threats to privacy

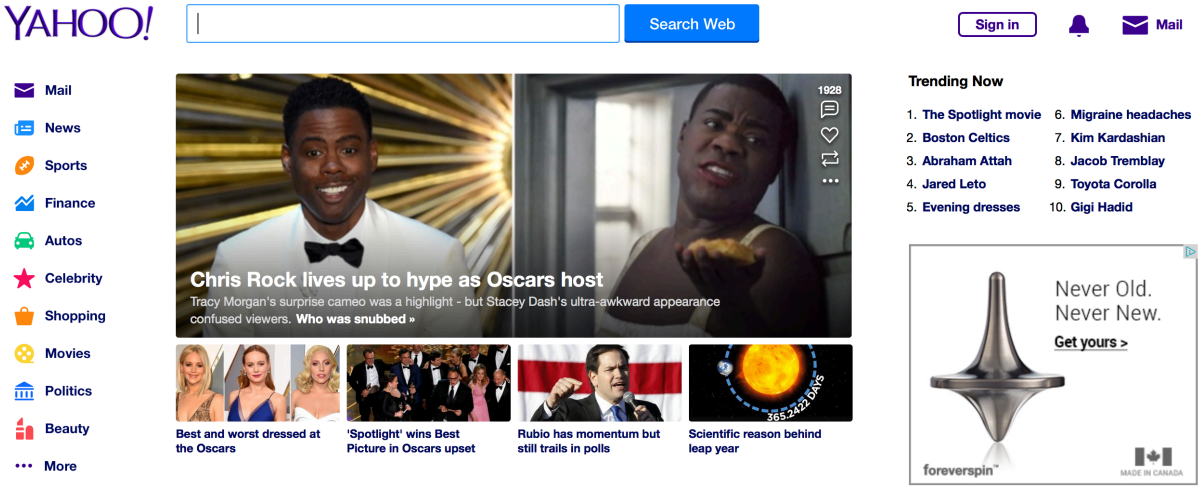
New ideas for protection



Dr. Dominik Herrmann

Slides available at
https://dhgo.to/dns-traffic-analysis


Browser
DNS Resolver
www.yahoo.com
www.yahoo.com
DNS
Cache
46.228.47.114
GET /celebrity/
HTTP
HTML file


YAHOO!
Search Web
Sign in
Mail
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
Never Old.
Never New.
Get yours >
foreverspin

Motivation of monitoring DNS

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

OpenDNS

tail

ead><title>Yahoo</title…

w.yahoo.com/celebrity/

oo.com

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

	2016-03-05	06:46:01.383 aus5.mozilla.org
	2016-03-05	**09:46:01.455 www.exploit-db.com**
?	2016-03-05	10:22:01.814 time.apple.com
	2016-03-05	10:22:01.950 b.config.skype.com
	2016-03-05	14:17:09.663 notify5.dropbox.com
	2016-03-05	14:17:10.411 ols.officeapps.live.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?

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.

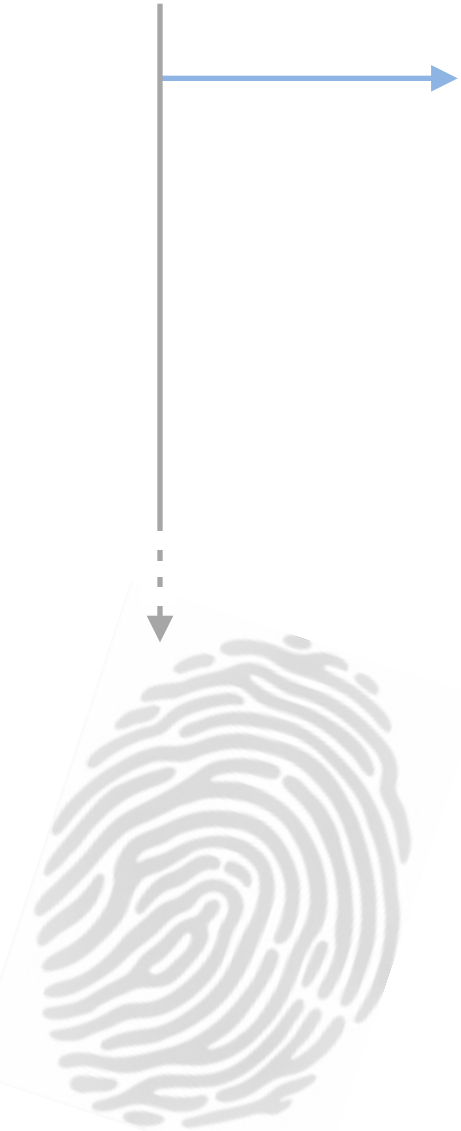
	Heuristic search: Δt > 5 sec	
09:41:20.242	ad4.adition.com	
09:41:21.770	ads.nuggad.com	
09:41:40.152	skypedata.akadns.net	
09:42:41.985	dl-debug.dropbox.com	
09:45:11.201	**google.com**	**true positive**
09:46:00.033	**www.heise.de**	**true positive**
09:46:00.133	dealbook.nytimes.com	
09:46:00.134	pressroom.yahoo.net	
09:46:00.169	www.yahoo.com	**true negative**

www.yahoo.com
cached for 1–5 min

09:46:30.812	**[visit Yahoo website]**	**false negative**

Browser DNS Resolver

51 domains resolved when Yahoo's home page is visited



www.yahoo.com	shopping.yahoo.com	search.yahoo.com
bs.serving-sys.com	**www.flickr.com**	sports.yahoo.com
pclick.yahoo.com	**www.tumblr.com**	**thinkprogress.org**
s.yimg.com	beap.gemini.yahoo.com	sync.adap.tv
sb.scorecardresearch…	finance.yahoo.com	sync.adaptv.advertisin…
crl-ds.ws.symantec.co…	ftw.usatoday.com	**www.cbsnews.com**
y.analytics.yahoo.com	geo-um.btrll.com	ads.yahoo.com
geo.query.yahoo.com	googleads.g.doublecli…	**www.chicagotribune.…**
csc.beap.bc.yahoo.com	match.adsrvr.org	**www.foxnews.com**
geo.yahoo.com	pagead2.googlesyndic…	**www.latimes.com**
comet.yahoo.com	help.yahoo.com	fonts.googleapis.com
answers.yahoo.com	info.yahoo.com	tpc.googlesyndication…
everything.yahoo.com	news.yahoo.com	cm.g.doubleclick.net
groups.yahoo.com	na.ads.yahoo.com	**www.npr.org**
login.yahoo.com	pr-bh.ybp.yahoo.com	**www.politico.com**
mail.yahoo.com	r.turn.com	**www.sbnation.com**
mobile.yahoo.com	rmx.pxl.ace.advertisin…	**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	inference of **whole (!) URL**
	ALEXA top 100 000 websites	**HEISE** 6283 news pages
$k = 1$	99 %	63 %
$k \leq 5$	99 %	76 %

Interesting problems:

- robustness
- threshold for match
- influence of cache

Browser DNS Resolver

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 correlated with domain name length

logging of flow records

(common practice)

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

Is DNS-based visited website verification still possible?

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

Measurements indicate:

domain lengths multiset is characteristic

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

drawing inferences from
DNS logs and flow records

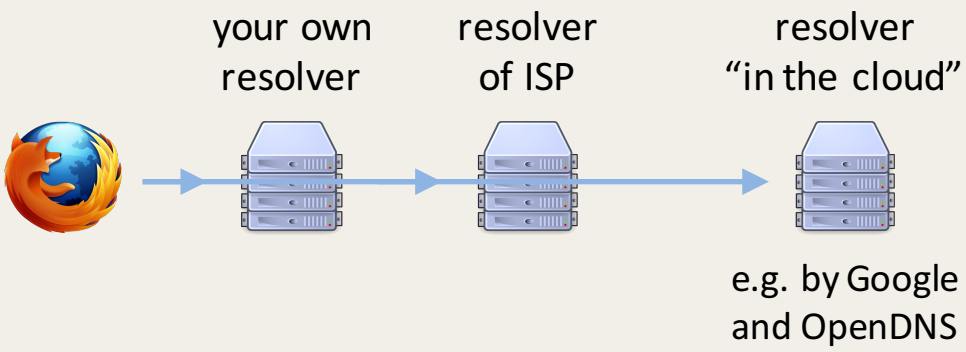
useful for forensics

real-world accuracy?

utility for law enforcement?

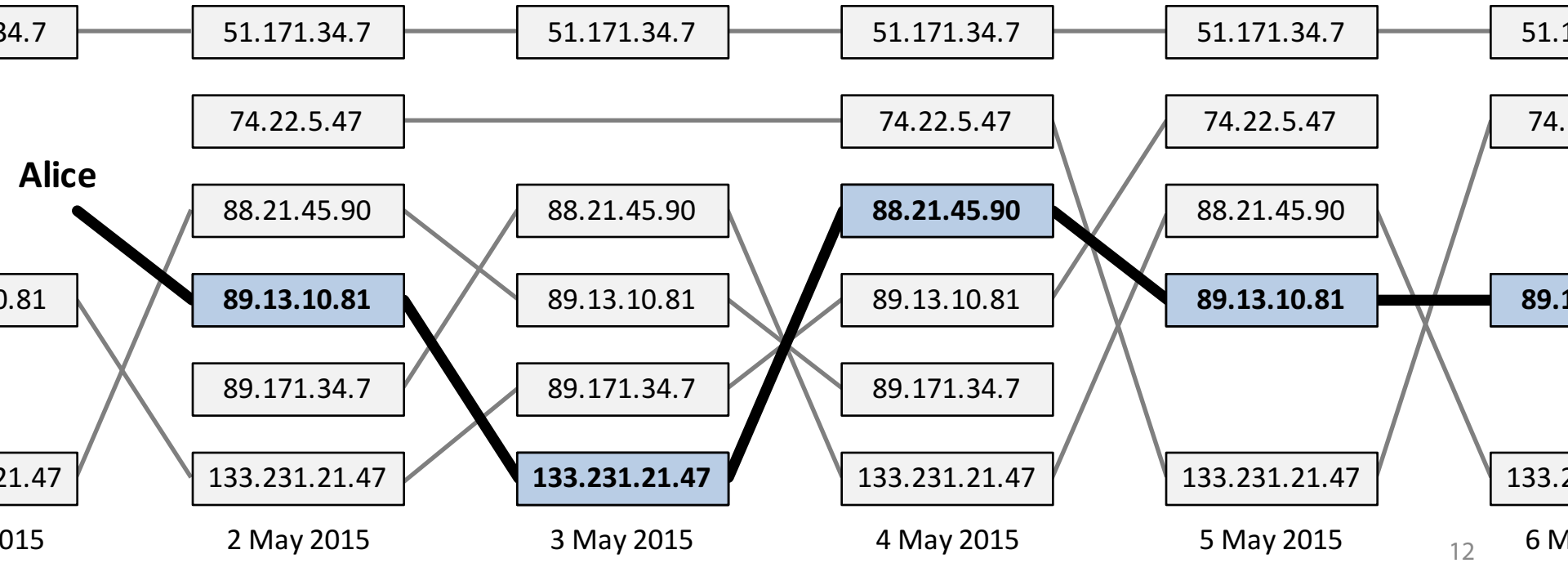
probative value of evidence?

privacy concerns



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

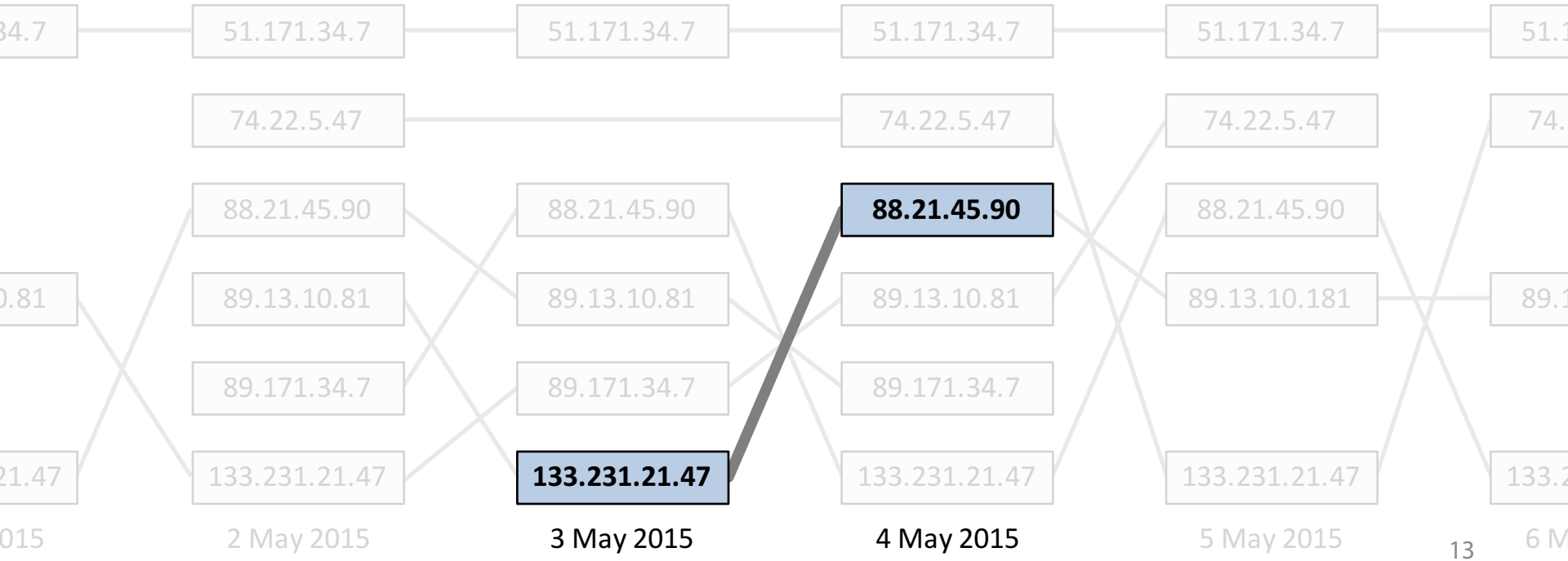
Challenge:
IP address changes frequently (daily)



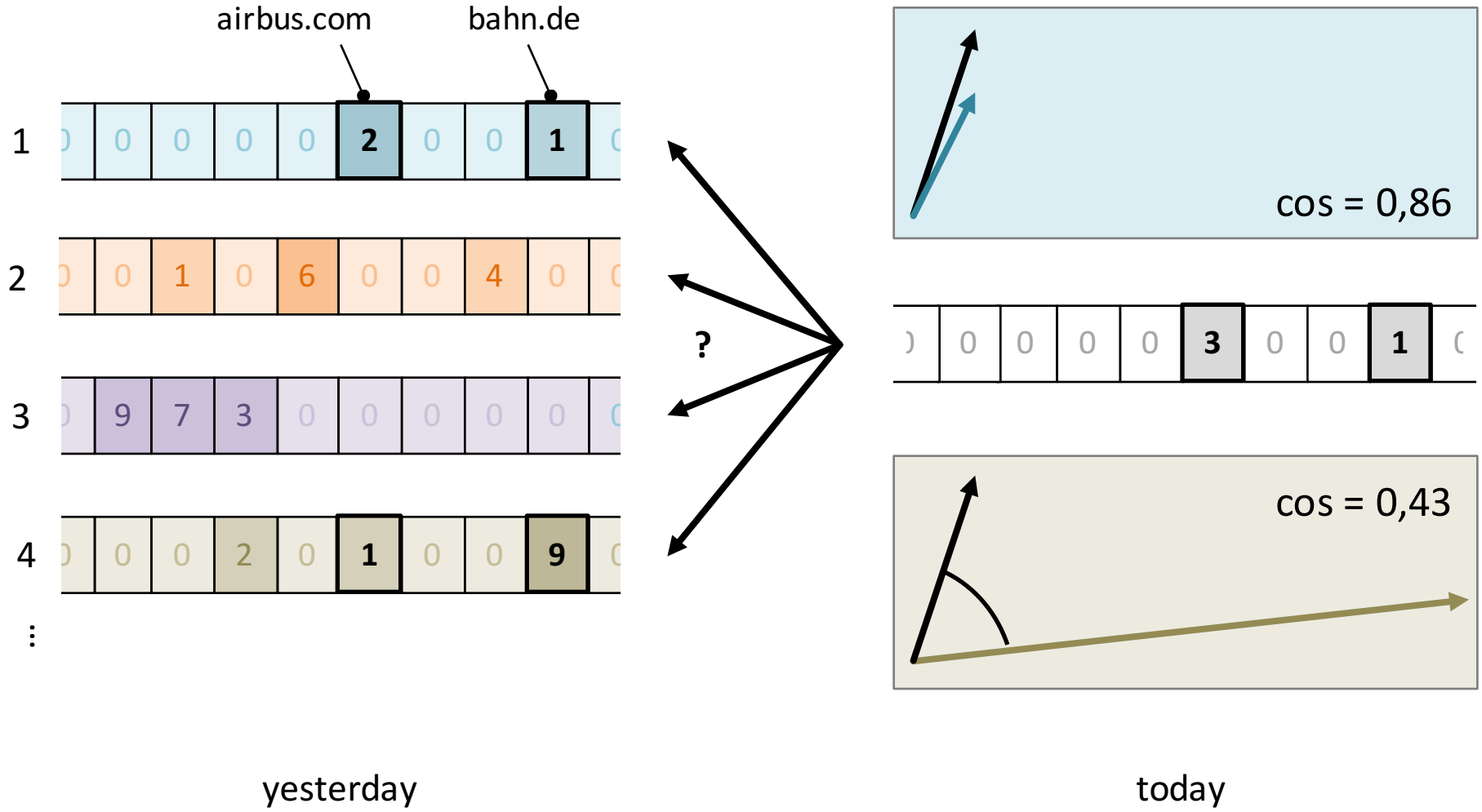
3 May 2015			4 May 2015
spiegel.de	4 x		1 x	spiegel.de
google.de	15 x		9 x	google.de
apple.com	1 x		0 x	apple.com
airbus.com	**3 x**		**6 x**	**airbus.com**
mpg.de	**2 x**		**3 x**	**mpg.de**

re-identification via resolved domains

Do users have distinct habits?



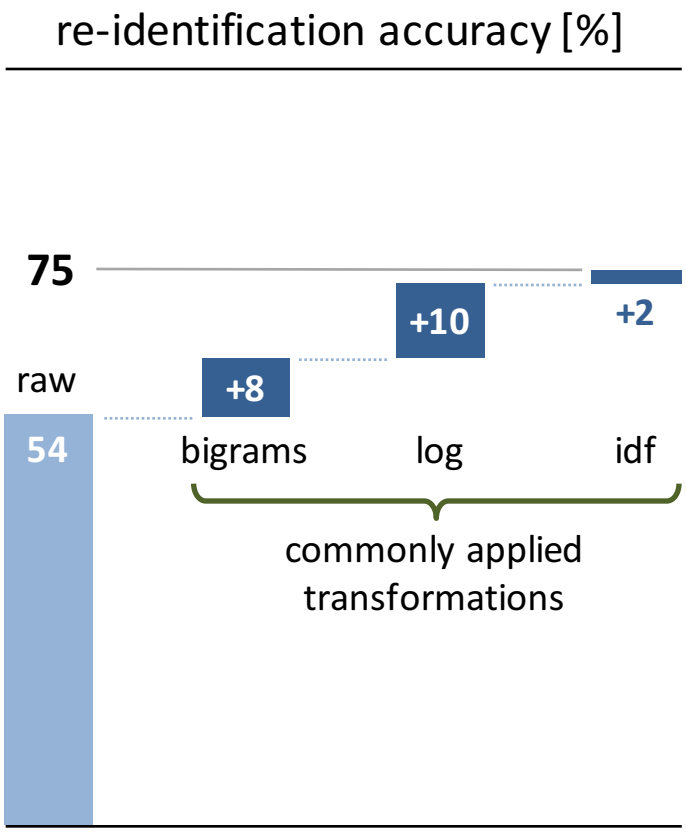
Sessions are modelled as **vectors** that are compared with **cosine similarity**
(nearest-neighbor classifier)

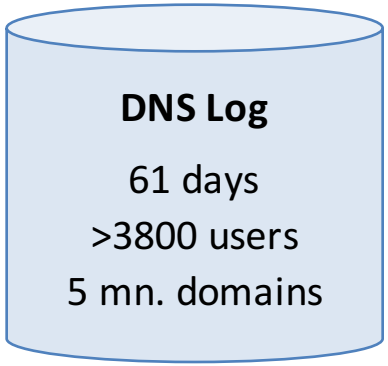


How accurate is behavior-based tracking in practice?

Experimental approach:

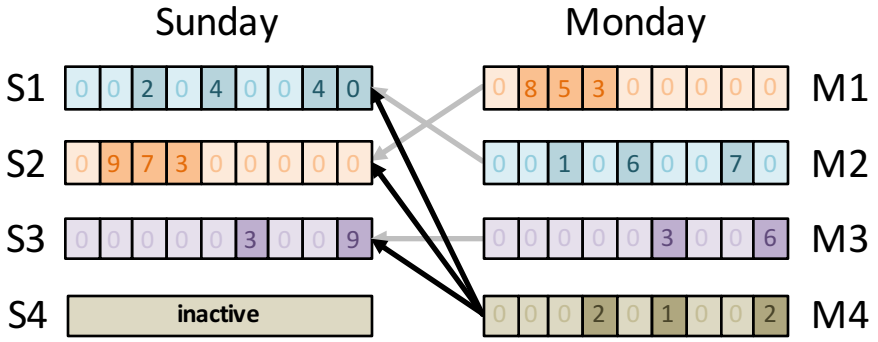
1. Obtain DNS log with realistic traffic
2. Track users day to day (24h sessions)
3. Determine overall accuracy



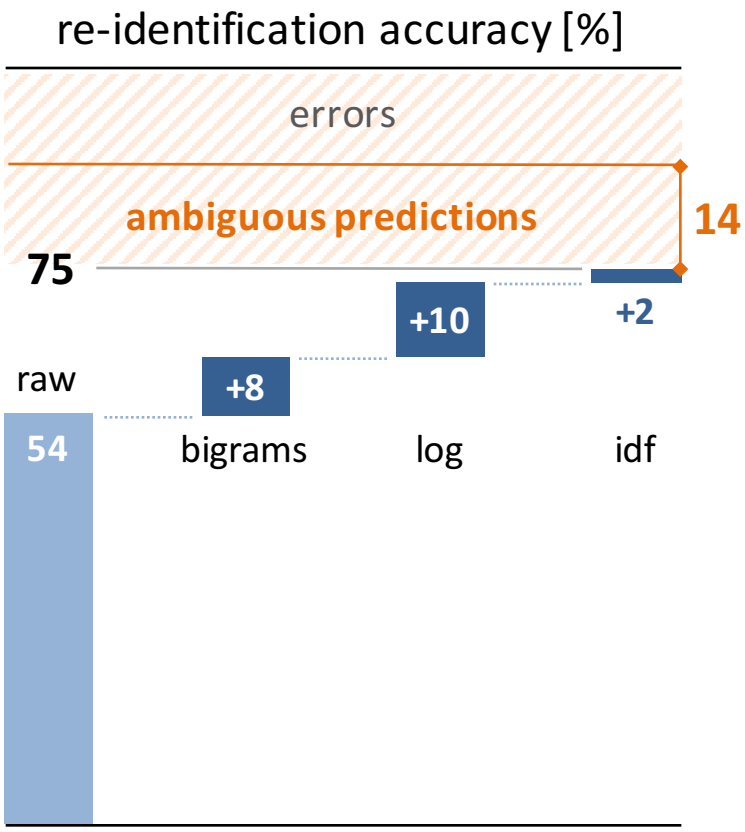


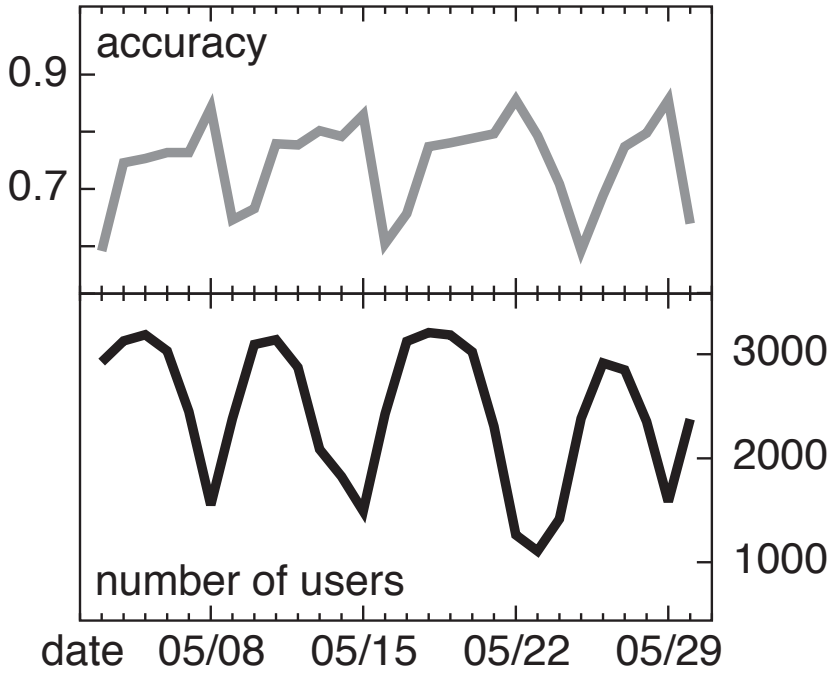
with »ground truth«
(pseudonymized)

How accurate is behavior-based tracking in practice?

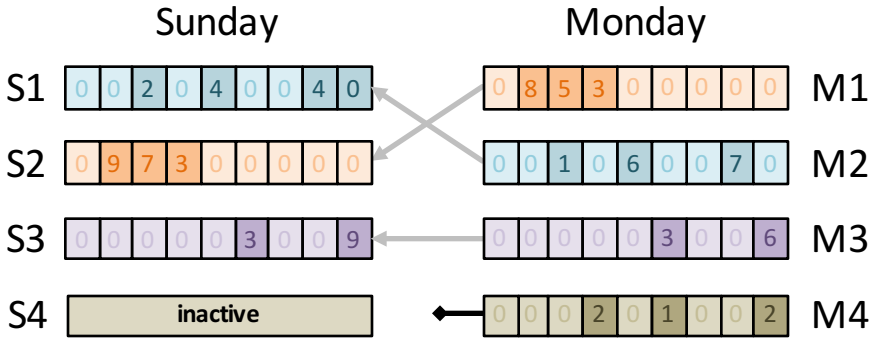


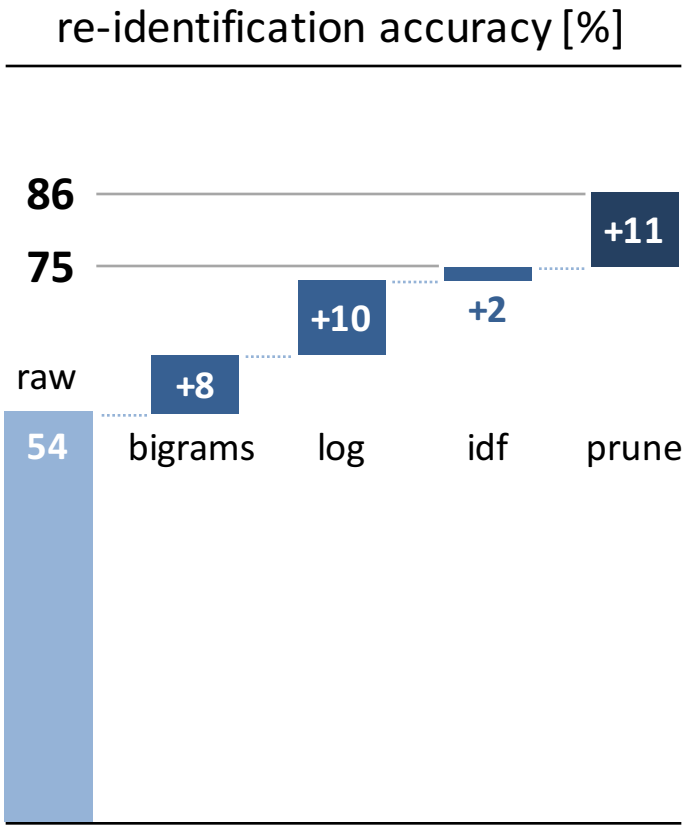
ambiguous prediction
… can be resolved

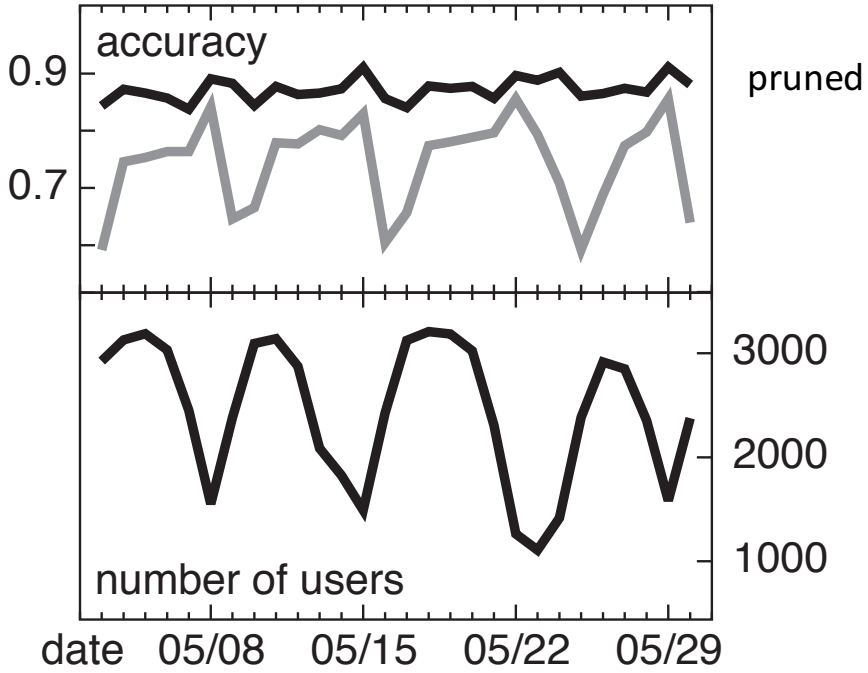




How accurate is behavior-based tracking in practice?







Application to network forensics:

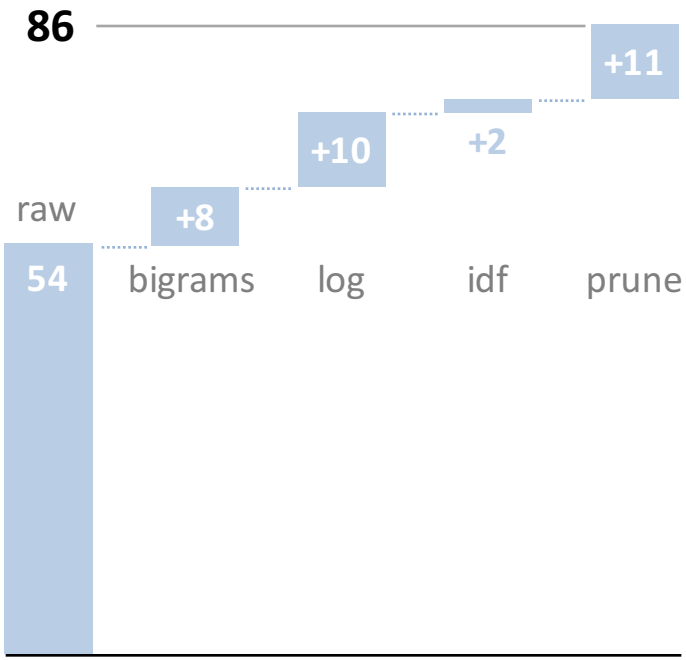
*How accurate is user re-identification with **flow records only**?*

Idea of ngram markers:

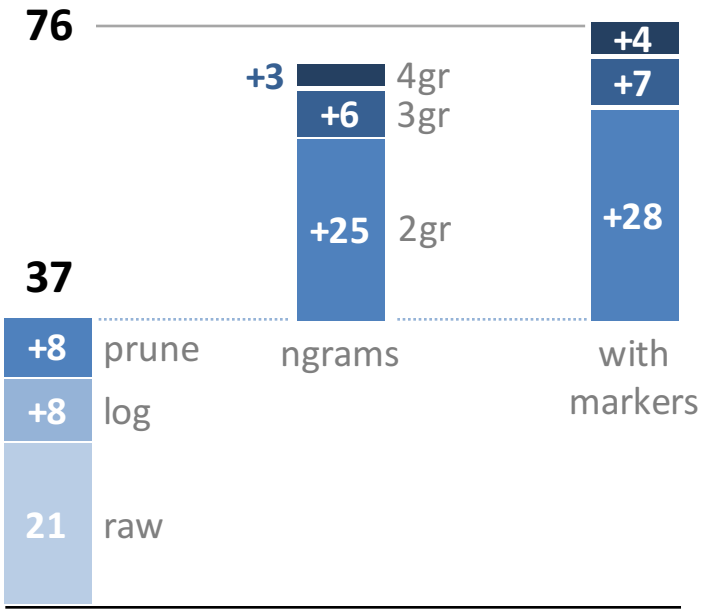
observed: 15 30 [pause ≥ 5 s] 18

bigrams: 15–30 30–P P–18

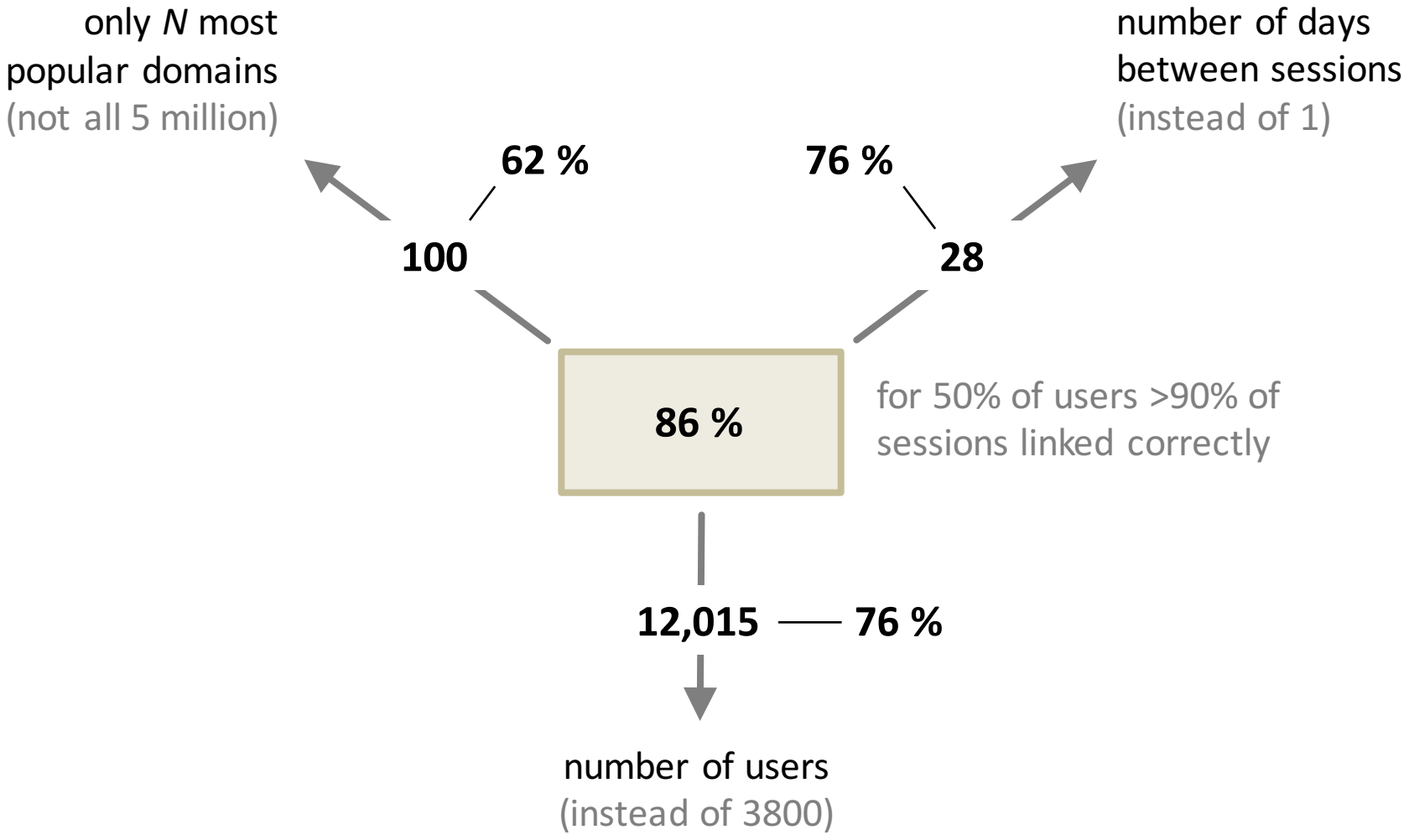
domain names
re-identification accuracy [%]



domain name lengths
re-identification accuracy [%]



Behavior-based re-identification is quite robust.



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?

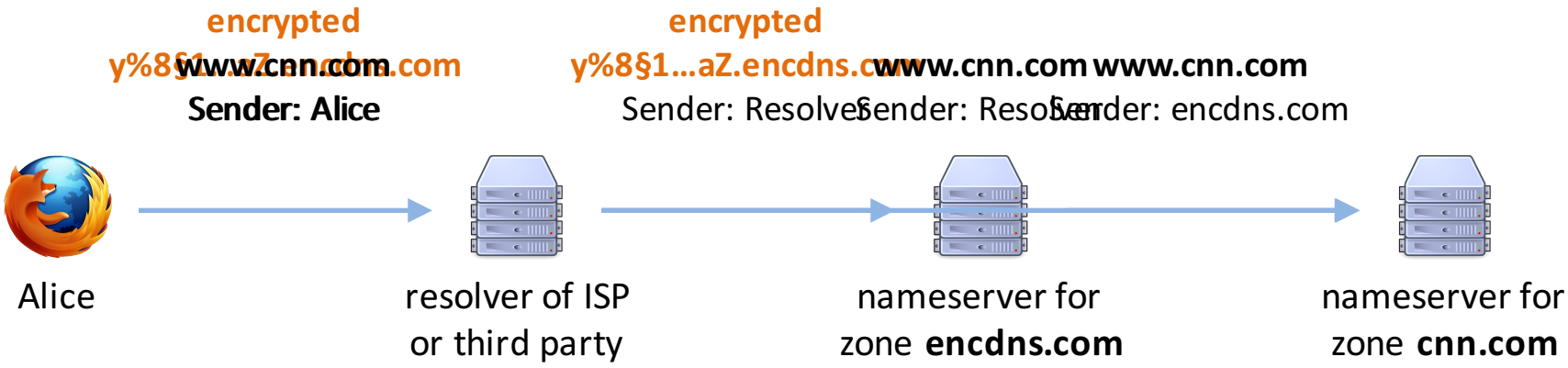
generic anonymization services (Tor) too slow

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

Challenge:
limited space (255 bytes)

cryptobox of Bernstein's NaCl library (Curve25519)

Measurements indicate:
fast and scalable (>6000 queries/sec)

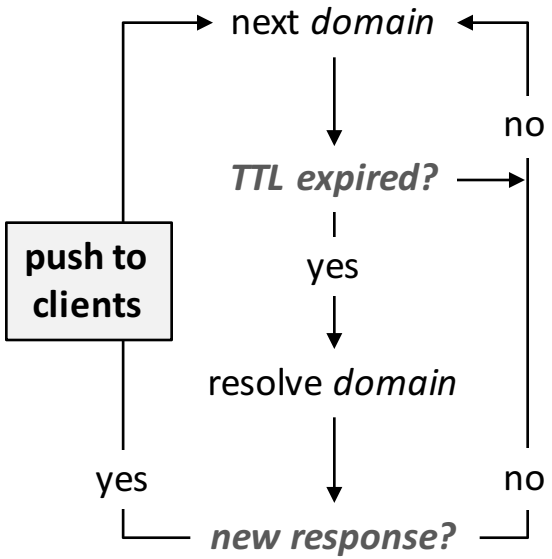


We can exploit **peculiarities of DNS** to improve performance and privacy.

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

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

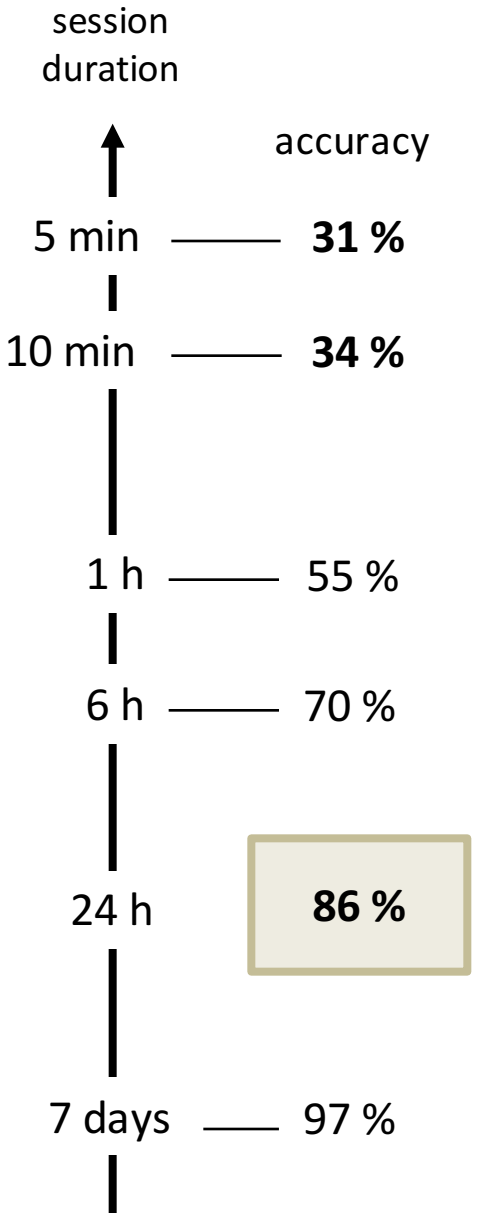


Traffic requirements (10,000 domains):
– resolving domains: 350 MB per day
– pushing updates: 0.8 KB/s per user

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

Protection against behavior-based tracking

… can be delegated to Internet Service Provider



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: **Traffic Analysis 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

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