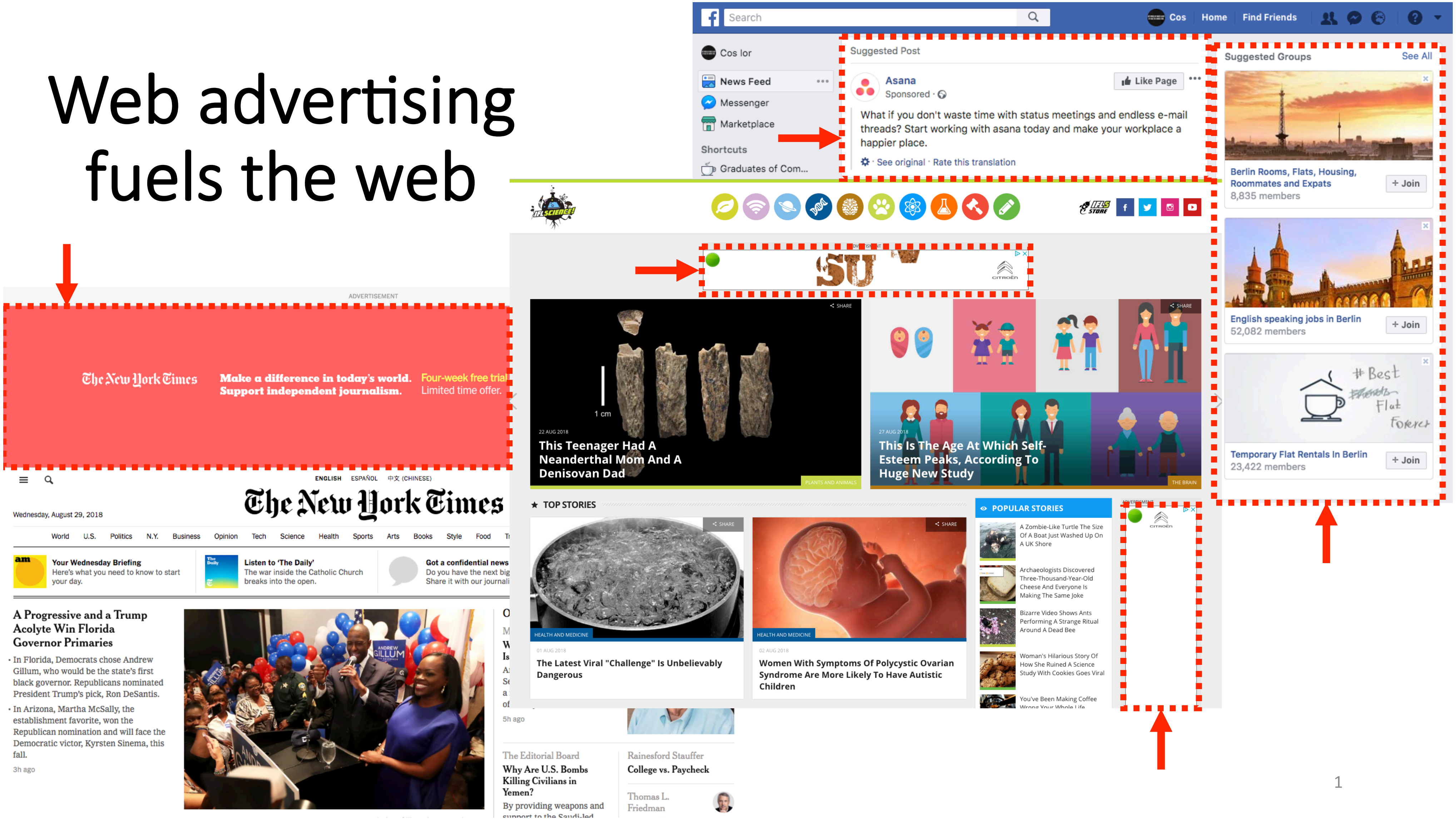
Tracing Cross Border Web Tracking

Costas Iordanou Georgios Smaragdakis
Ingmar Poese Nikolaos Laoutaris

Technische Universität Berlin
Universidad Carlos III de Madrid
BENOCS
DATA TRANSPARENCY LAB
erc European Research Council

Web advertising fuels the web



Search

Cos Home Find Friends

Cos Ior

News Feed

Messenger

Marketplace

Shortcuts

Graduates of Com...

Suggested Post

Asana

Sponsored ·

Like Page

What if you don't waste time with status meetings and endless e-mail threads? Start working with asana today and make your workplace a happier place.

See original · Rate this translation

Suggested Groups See All

Berlin Rooms, Flats, Housing, Roommates and Expats
8,835 members
+ Join

English speaking jobs in Berlin
52,082 members
+ Join

Temporary Flat Rentals In Berlin
23,422 members
+ Join

22 AUG 2018
This Teenager Had A Neanderthal Mom And A Denisovan Dad
PLANTS AND ANIMALS

27 AUG 2018
This Is The Age At Which Self-Esteem Peaks, According To Huge New Study
THE BRAIN

TOP STORIES

HEALTH AND MEDICINE
01 AUG 2018
The Latest Viral "Challenge" Is Unbelievably Dangerous

HEALTH AND MEDICINE
02 AUG 2018
Women With Symptoms Of Polycystic Ovarian Syndrome Are More Likely To Have Autistic Children

POPULAR STORIES

A Zombie-Like Turtle The Size Of A Boat Just Washed Up On A UK Shore

Archaeologists Discovered Three-Thousand-Year-Old Cheese And Everyone Is Making The Same Joke

Bizarre Video Shows Ants Performing A Strange Ritual Around A Dead Bee

Woman's Hilarious Story Of How She Ruined A Science Study With Cookies Goes Viral

You've Been Making Coffee Wrong Your Whole Life

ADVERTISEMENT

The New York Times Make a difference in today's world. Support independent journalism. Four-week free trial Limited time offer.

ENGLISH ESPAÑOL 中文 (CHINESE)

The New York Times

Wednesday, August 29, 2018

World U.S. Politics N.Y. Business Opinion Tech Science Health Sports Arts Books Style Food

Your Wednesday Briefing
Here's what you need to know to start your day.

Listen to 'The Daily'
The war inside the Catholic Church breaks into the open.

Got a confidential news
Do you have the next big
Share it with our journali

A Progressive and a Trump Acolyte Win Florida Governor Primaries

- In Florida, Democrats chose Andrew Gillum, who would be the state's first black governor. Republicans nominated President Trump's pick, Ron DeSantis.
- In Arizona, Martha McSally, the establishment favorite, won the Republican nomination and will face the Democratic victor, Kyrsten Sinema, this fall.

3h ago

5h ago

The Editorial Board
Why Are U.S. Bombs Killing Civilians in Yemen?
By providing weapons and support to the Saudi-led

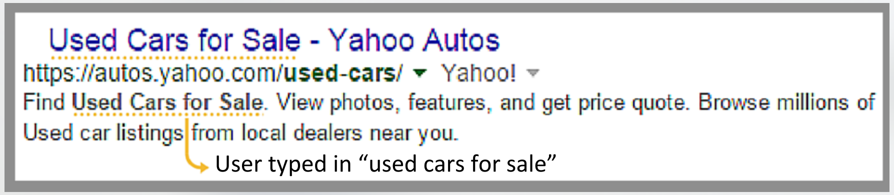
Rainesford Stauffer
College vs. Paycheck

Thomas L. Friedman

The rise of targeted ads

Why Targeted ads?

- Users get relevant ads
- Increase user engagement
- More efficient ad campaigns
- Higher ROI for the advertisers
- Better use of resources
- Etc.

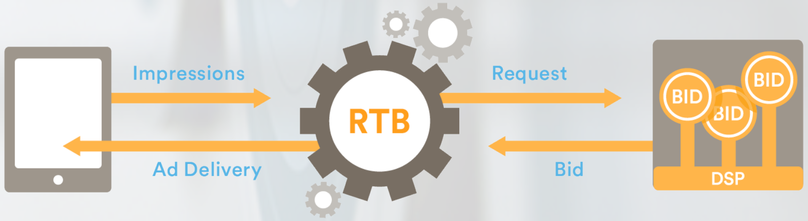


Used Cars for Sale - Yahoo Autos
https://autos.yahoo.com/used-cars/ ▾ Yahoo! ▾
Find Used Cars for Sale. View photos, features, and get price quote. Browse millions of Used car listings from local dealers near you.

User typed in “used cars for sale”

How it works?

- Tracking and profiling users
- Real time auctions of ads (RTB)
- Cookie synchronization
- Etc.



The reaction of users and regulators

Users

Browser extensions

ABP Adblock Plus

AdBlock

NoScript

GHOSTERY®

PRIVACY BADGER

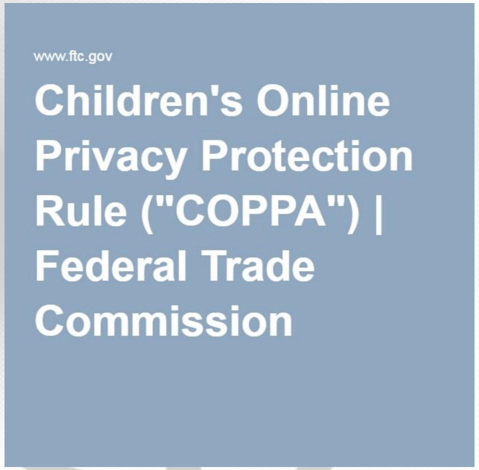
Browsers

brave

CLIQZ

TOR BROWSER

Regulators


www.ftc.gov

Children's Online Privacy Protection Rule ("COPPA") | Federal Trade Commission

Users and regulators reaction

Users

Browser extensions

ABP Adblock Plus
AdBlock
D.
NoScript
uO
GHOSTERY
PRIVACY BADGER

Browsers

brave
CLIQZ
TOR BROWSER

Regulators

www.ftc.gov

Children's Online Privacy Protection Rule ("COPPA") | Federal Trade Commission

General Data Protection Regulation - Details

One of the biggest changes with respect to privacy and regulation on the web in the last few years (Enforcement date: 25th May, 2018)

In general the new legislation:

1. tries to regulate how users' data are collected, processed and stored and
2. if they include any sensitive information about the user

General Data Protection Regulation - Details

One of the biggest changes with respect to privacy and regulation on the web in the last few years (Enforcement date: 25th May, 2018)

In general the new legislation:

1. tries to regulate how users' data are collected, processed and stored and
2. if they include any sensitive information about the user

Implementation − Per member state Data Protection Authority (DPA)

DPA: Responsible for complaints − investigations & enforcement

Investigation starting point − **Ad & Tracking** flows entry point servers location

RQ: How can we identify the physical locations of such servers?

Challenges

1. How to effectively **detect ad and tracking related domains in the wild**?

2. How to **ensure correct geolocation of infrastructure servers**?

Challenges

1. How to effectively **detect ad and tracking related domains in the wild**?

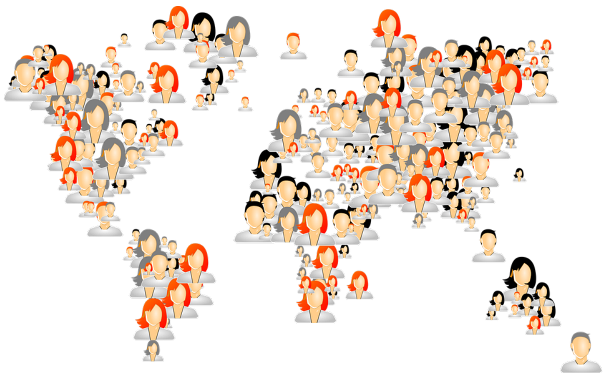
2. How to **ensure correct geolocation of infrastructure servers**?

3. How to ensure that **all possible ad and tracking servers are observed**?

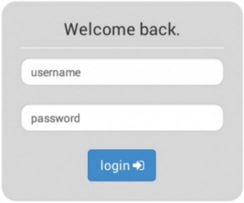
4. How to maintain a **balance between accuracy and scalability**?

Why real users instead of just Web crawling?

Real Users



User interaction

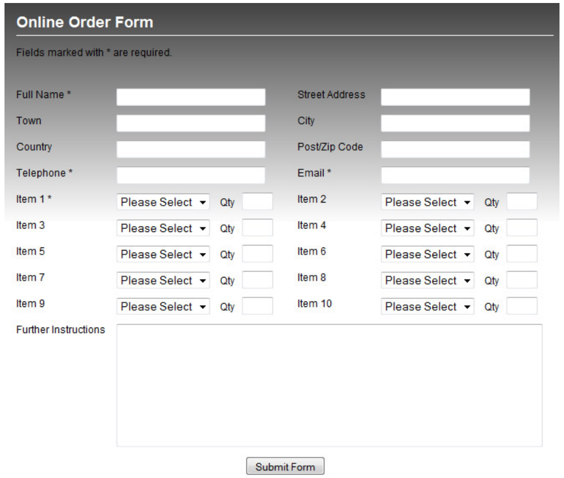


Welcome back.

username

password

login



Online Order Form

Fields marked with * are required.

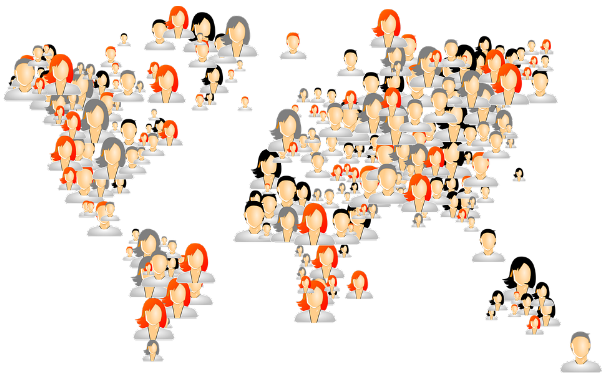
Full Name *				Street Address			
Town				City			
Country				Post/Zip Code			
Telephone *				Email *			
Item 1 *	Please Select	Qty		Item 2	Please Select	Qty	
Item 3	Please Select	Qty		Item 4	Please Select	Qty	
Item 5	Please Select	Qty		Item 6	Please Select	Qty	
Item 7	Please Select	Qty		Item 8	Please Select	Qty	
Item 9	Please Select	Qty		Item 10	Please Select	Qty	

Further Instructions

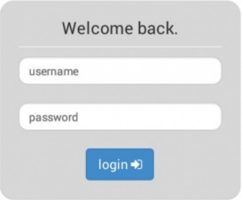
Submit Form

Why real users instead of just Web crawling?

Real Users



User interaction

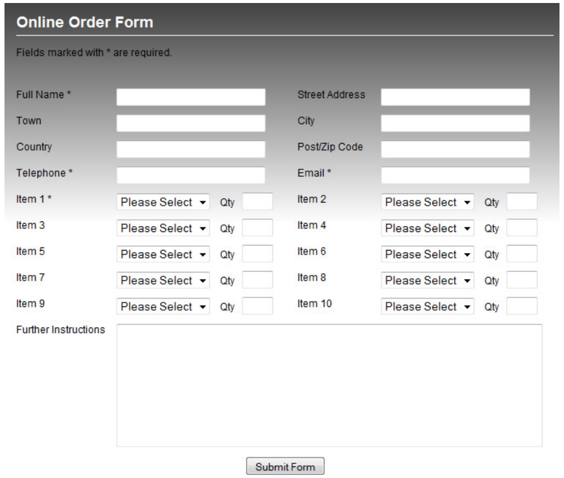


Welcome back.

username

password

login



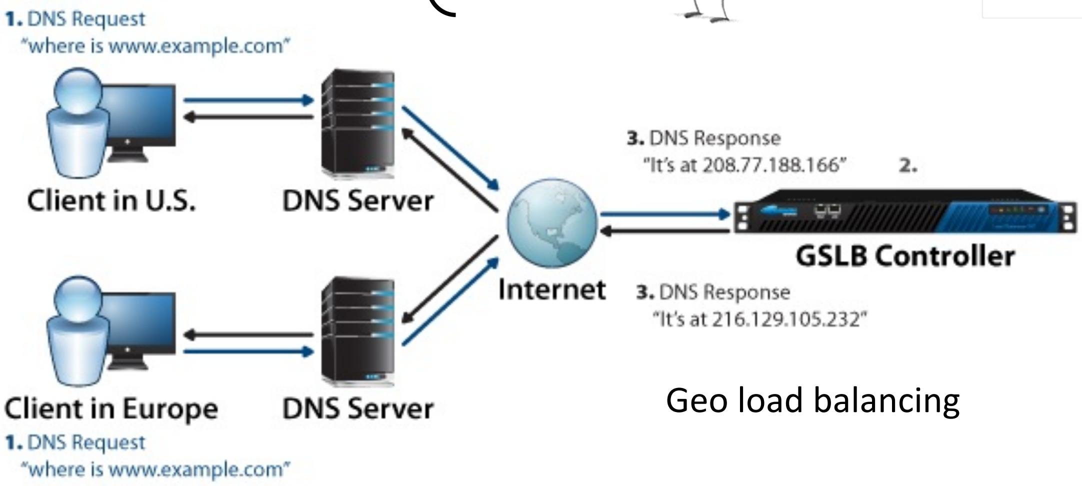
Online Order Form

Fields marked with * are required.

Full Name *			Street Address		
Town			City		
Country			Post/Zip Code		
Telephone *			Email *		
Item 1 *	Please Select	Qty	Item 2	Please Select	Qty
Item 3	Please Select	Qty	Item 4	Please Select	Qty
Item 5	Please Select	Qty	Item 6	Please Select	Qty
Item 7	Please Select	Qty	Item 8	Please Select	Qty
Item 9	Please Select	Qty	Item 10	Please Select	Qty

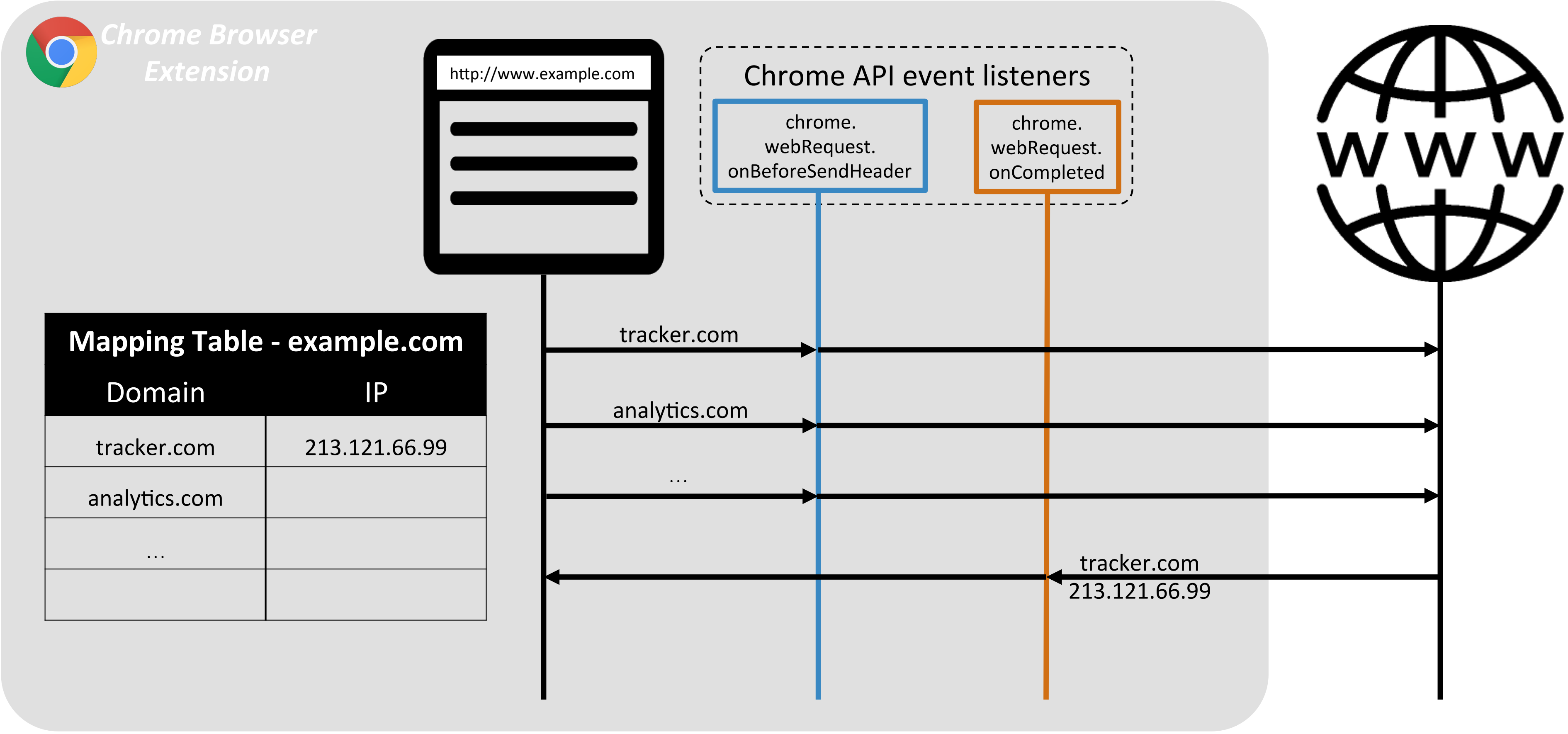
Further Instructions

Submit Form



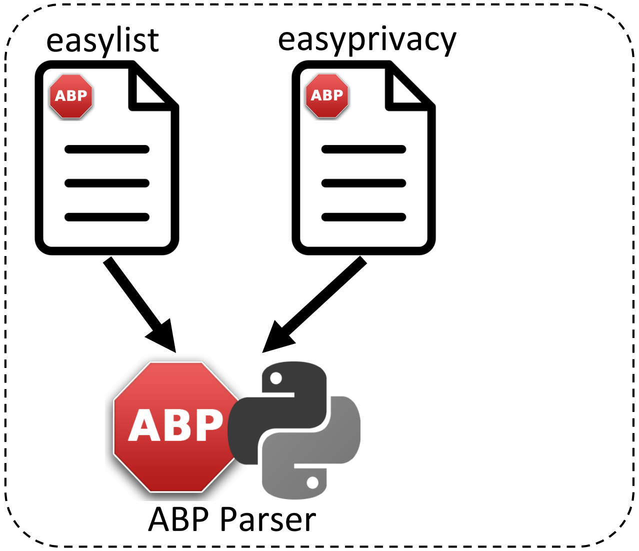
Geo load balancing

Mapping 3rd party domains to IPs



Mapping Table - example.com	
Domain	IP
tracker.com	213.121.66.99
analytics.com	
…	

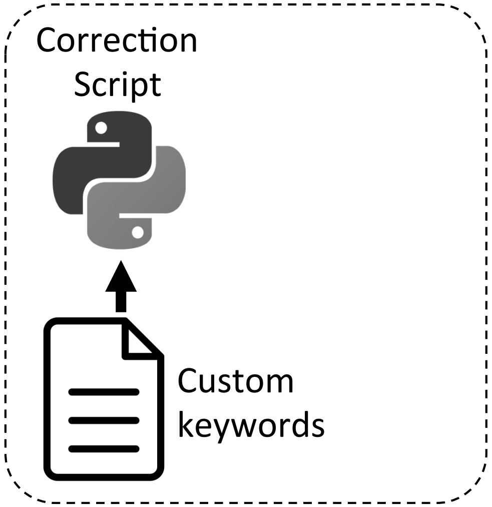
Identify Ad and Tracking related domains



easylist

easyprivacy

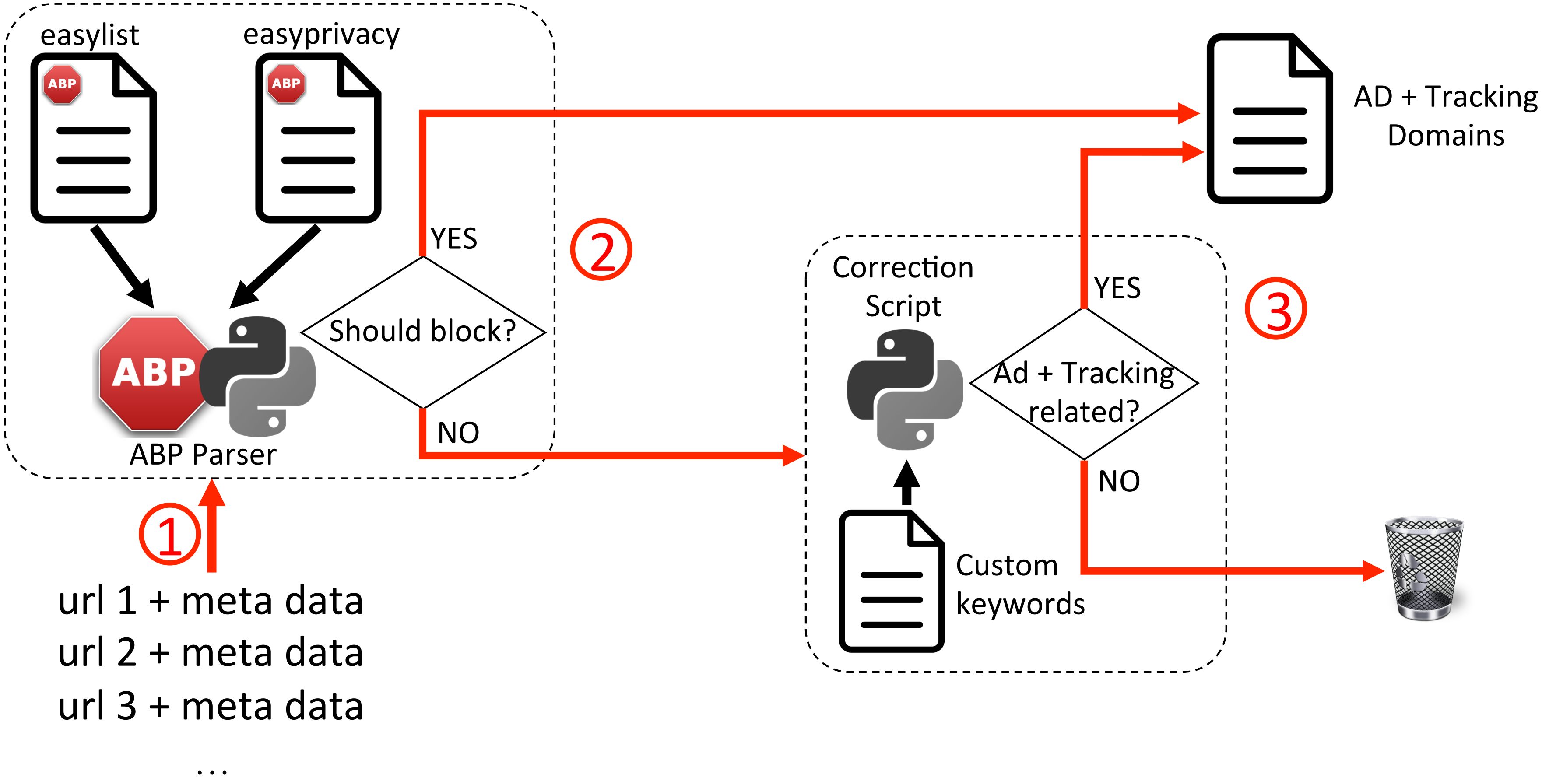
ABP Parser



Correction Script

Custom keywords

Identify Ad and Tracking related domains



Challenges

1. How to effectively **detect ad and tracking related domains in the wild**?

2. How to **ensure correct geolocation of infrastructure servers**?

3. How to ensure that **all possible ad and tracking servers are observed**?

4. How to maintain a **balance between accuracy and scalability**?

Accurate geo-location of server IPs

RIPE IPmap validation process - infrastructure servers IPs

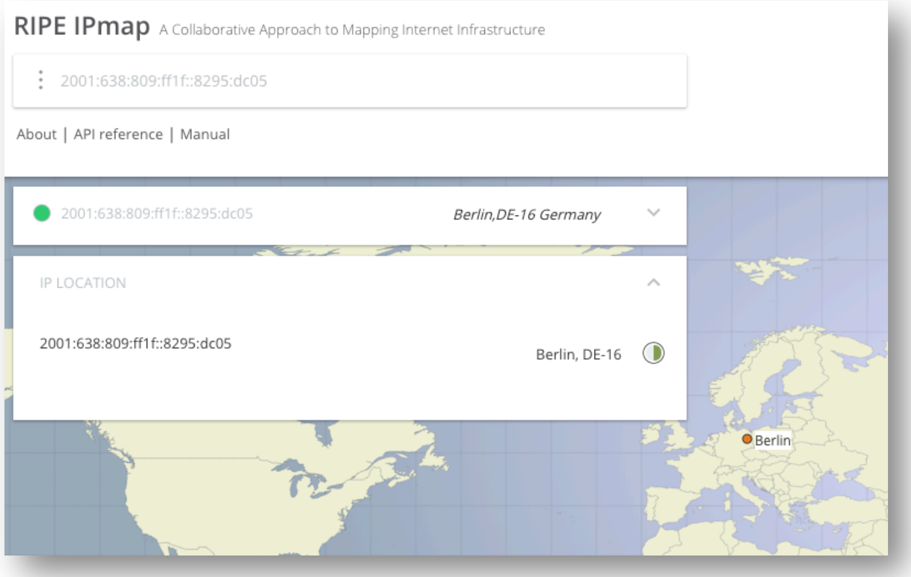
prefix	region	service
46.51.128.0/18	eu-west-1	AMAZON
46.51.216.0/21	ap-southeast-1	AMAZON
13.73.232.0/21	japaneast	AZURE
20.19.14.128/25	koreacentral	AZURE
…	…	…

Regions maps

eu-west-1: Ireland, Ireland
ap-southeast-1: Singapore, Singapore

RIPE IPmap

RIPE NCC
RIPE NETWORK COORDINATION CENTRE



RIPE IPmap A Collaborative Approach to Mapping Internet Infrastructure

2001:638:809:ff1f::8295:dc05

About | API reference | Manual

● 2001:638:809:ff1f::8295:dc05 Berlin,DE-16 Germany

IP LOCATION

2001:638:809:ff1f::8295:dc05 Berlin, DE-16

→ 99.6% match with the reported country

Challenges

1. How to effectively **detect ad and tracking related domains in the wild**?

2. How to **ensure correct geolocation of infrastructure servers**?

3. How to ensure that **all possible ad and tracking servers are observed**?

4. How to maintain a **balance between accuracy and scalability**?

Avoiding pitfalls…

- Identify **all domains behind each IP** (Reverse DNS query)

Query: https://freeapi.robtex.com/pdns/reverse/93.184.216.34

Response:

```
rrname:example.org,      rrdata:93.184.216.34, rrtype:A, time_first:1440526884, time_last:1535919774, count:18
rrname:www.example.org,  rrdata:93.184.216.34, rrtype:A, time_first:1440723354, time_last:1527899734, count:18
rrname:www.example.com,  rrdata:93.184.216.34, rrtype:A, time_first:1441108386, time_last:1535371292, count:18
rrname:www.example.net,  rrdata:93.184.216.34, rrtype:A, time_first:1436692690, time_last:1527900018, count:18
rrname:imrek.org,        rrdata:93.184.216.34, rrtype:A, time_first:1440827324, time_last:1508103356, count:18
rrname:example.net,      rrdata:93.184.216.34, rrtype:A, time_first:1440526998, time_last:1533895598, count:18
…
```

Avoiding pitfalls…

- Identify **all domains behind each IP** (Reverse DNS query)

Query: https://freeapi.robtex.com/pdns/reverse/93.184.216.34

Response:

```
rrname:example.org,      rrdata:93.184.216.34, rrtype:A, time_first:1440526884, time_last:1535919774, count:18
rrname:www.example.org,  rrdata:93.184.216.34, rrtype:A, time_first:1440723354, time_last:1527899734, count:18
rrname:www.example.com,  rrdata:93.184.216.34, rrtype:A, time_first:1441108386, time_last:1535371292, count:18
rrname:www.example.net,  rrdata:93.184.216.34, rrtype:A, time_first:1436692690, time_last:1527900018, count:18
rrname:imrek.org,        rrdata:93.184.216.34, rrtype:A, time_first:1440827324, time_last:1508103356, count:18
rrname:example.net,      rrdata:93.184.216.34, rrtype:A, time_first:1440526998, time_last:1533895598, count:18
…
```

- Identify **all IPs for each domain** (Forward DNS query)

Query: https://freeapi.robtex.com/pdns/forward/example.com

Response:

```
rrname:example.com, rrdata:2606:280::::::1946, rrtype:AAAA, time_first:1441278890, time_last:1535952170, count:18
rrname:example.com, rrdata:93.184.216.34,      rrtype:A,    time_first:1441278890, time_last:1535952170, count:18
rrname:example.com, rrdata:208.77.188.166,     rrtype:A,    time_first:1246678898, time_last:1246678898, count:1
```

Avoiding pitfalls…

- Identify **all c**

Query: https

Response:

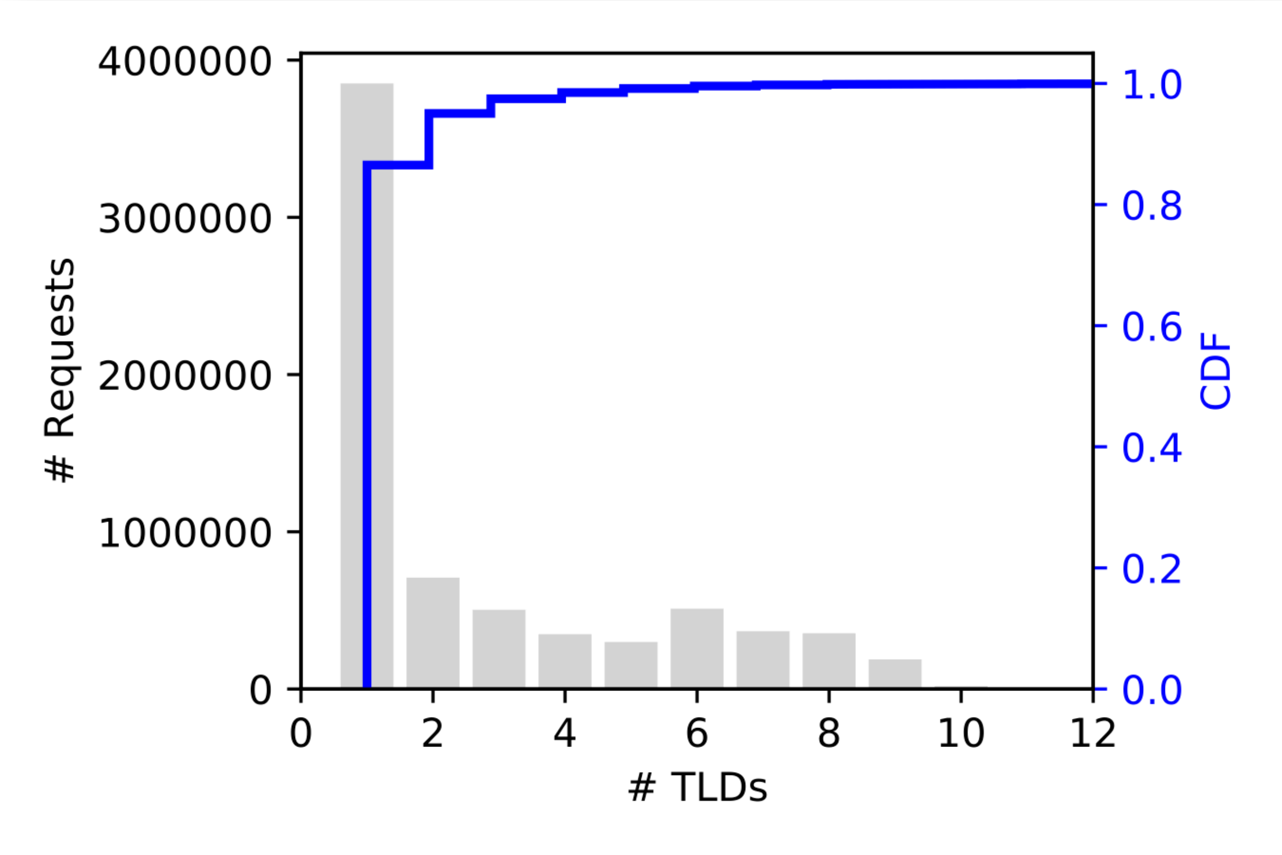
```
rrname:example.org                                                                     count:18
rrname:www.example                                                                     count:18
rrname:www.example                                                                     count:18
rrname:www.example                                                                     count:18
rrname:imrek.org,                                                                      count:18
rrname:example.net                                                                     count:18
…
```

- Identify **all I**

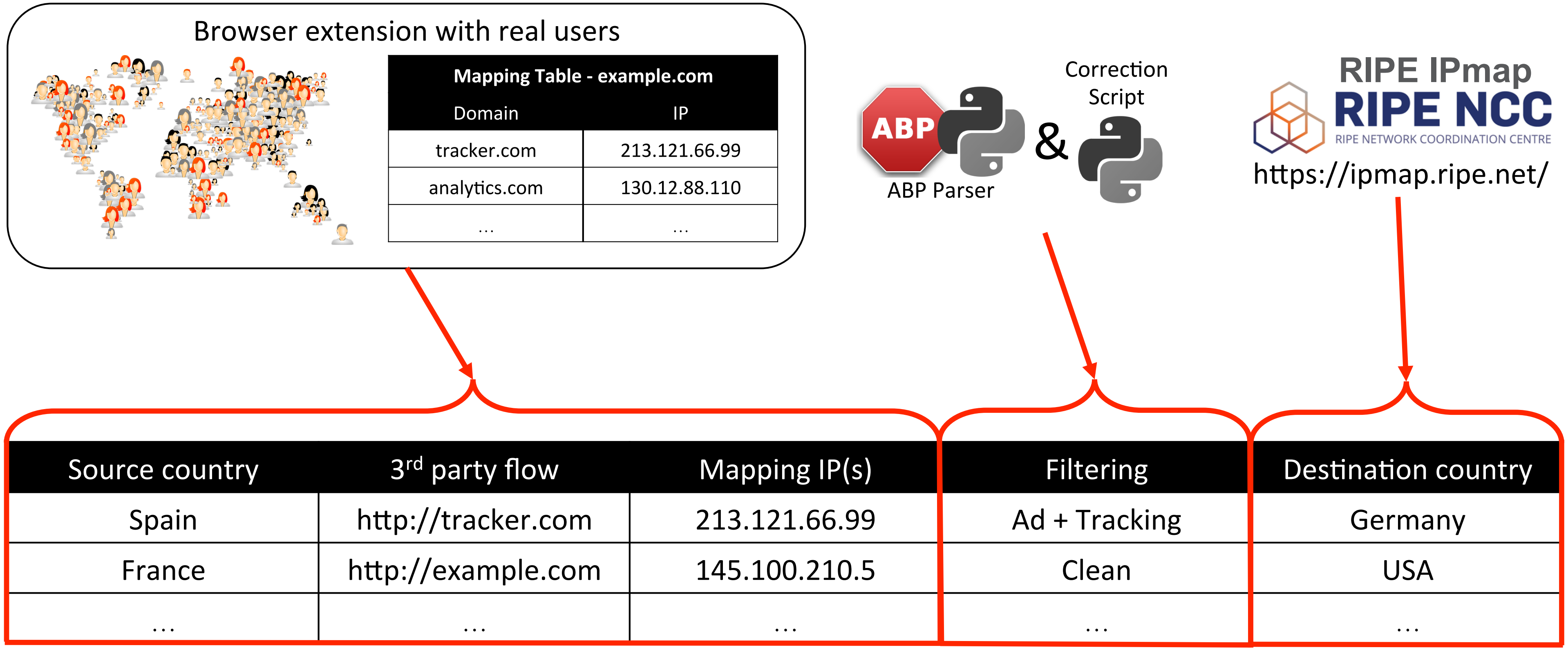
Query: https:

Response:

```
rrname:example.com                                                                   0,  count:2
rrname:example.com                                                                   0,  count:2
rrname:example.com                                                                   0,  count:18
rrname:example.com                                                                   0,  count:18
rrname:example.com,                                                                  8,  count:1
```



Joining everything together



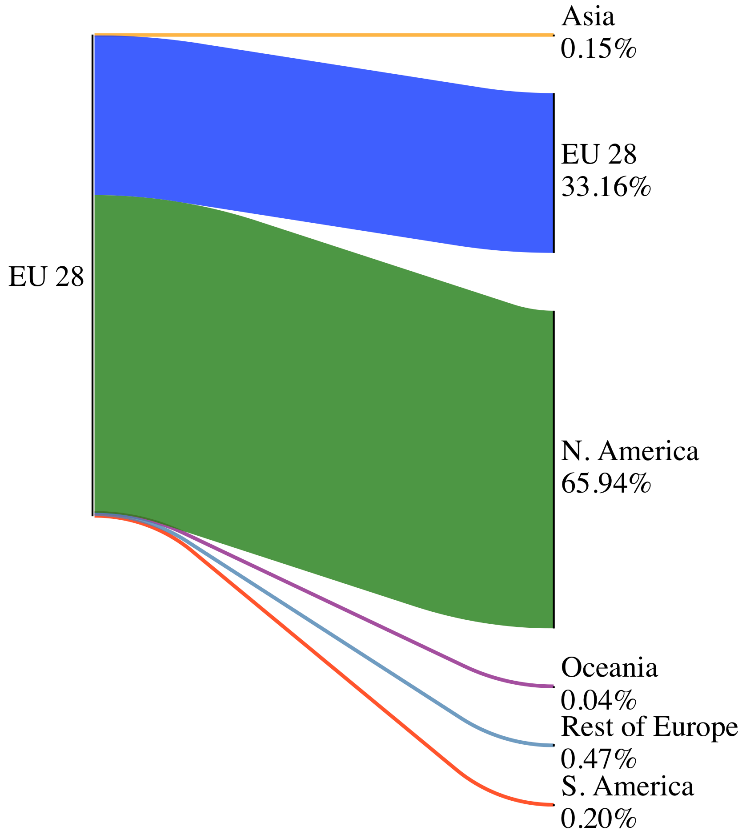
Browser extension with real users

Mapping Table - example.com	
Domain	IP
tracker.com	213.121.66.99
analytics.com	130.12.88.110
…	…

https://ipmap.ripe.net/

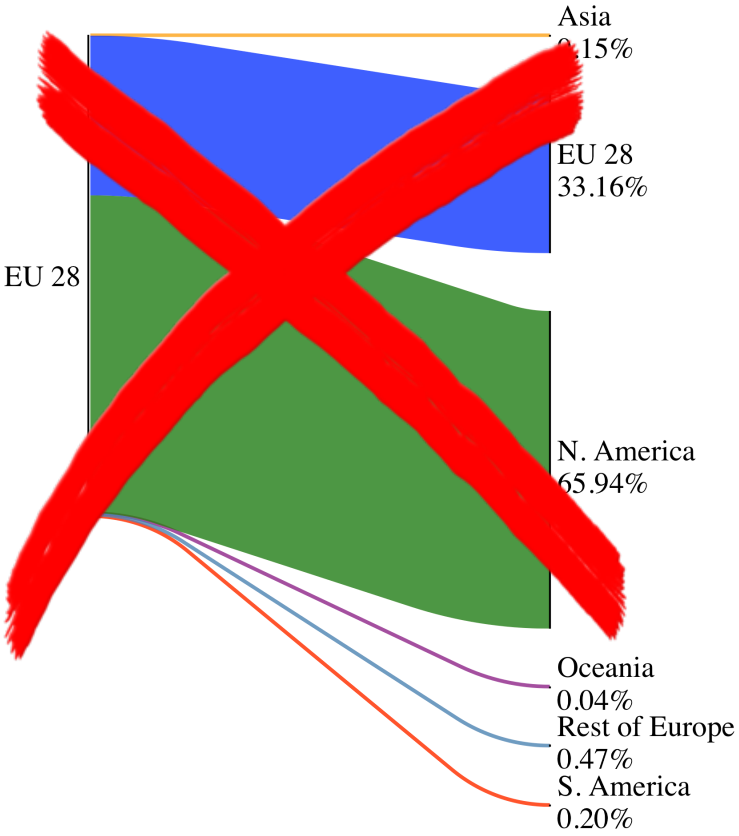
Source country	3rd party flow	Mapping IP(s)	Filtering	Destination country
Spain	http://tracker.com	213.121.66.99	Ad + Tracking	Germany
France	http://example.com	145.100.210.5	Clean	USA
…	…	…	…	…

Results - EU 28 member states confinement level

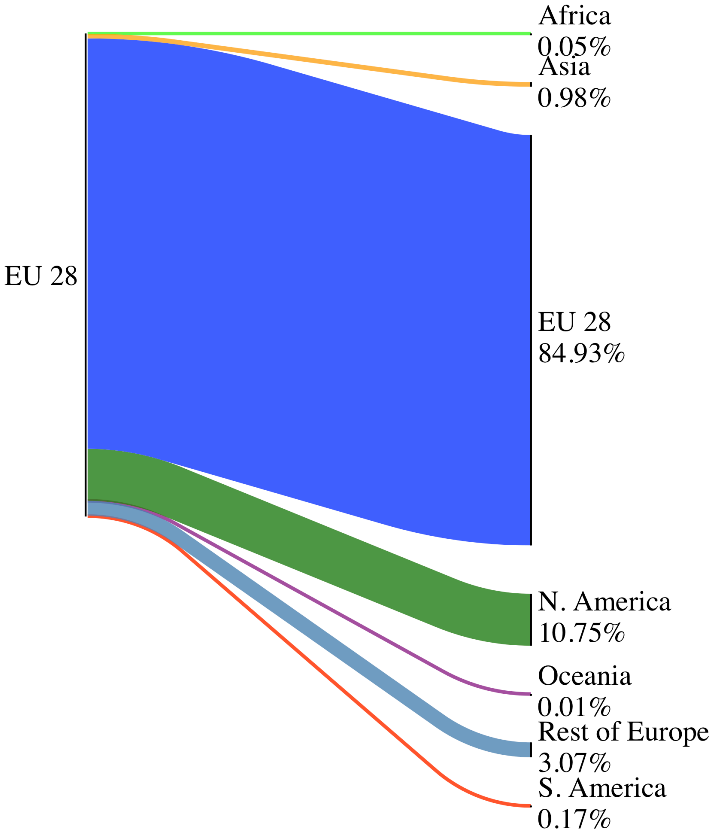


MaxMind geo-location

Results - EU 28 member states confinement level



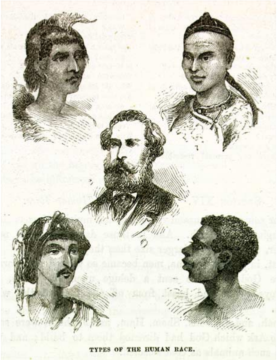
MaxMind geo-location



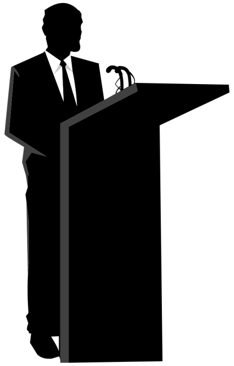
RIPE IPmap geo-location

What about sensitive websites?

Sensitive categories as defined by GDPR

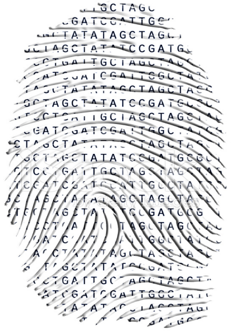


Race & Ethnicity

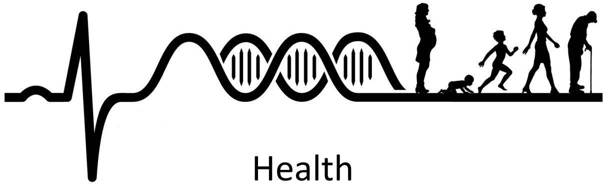


Political beliefs

Religion



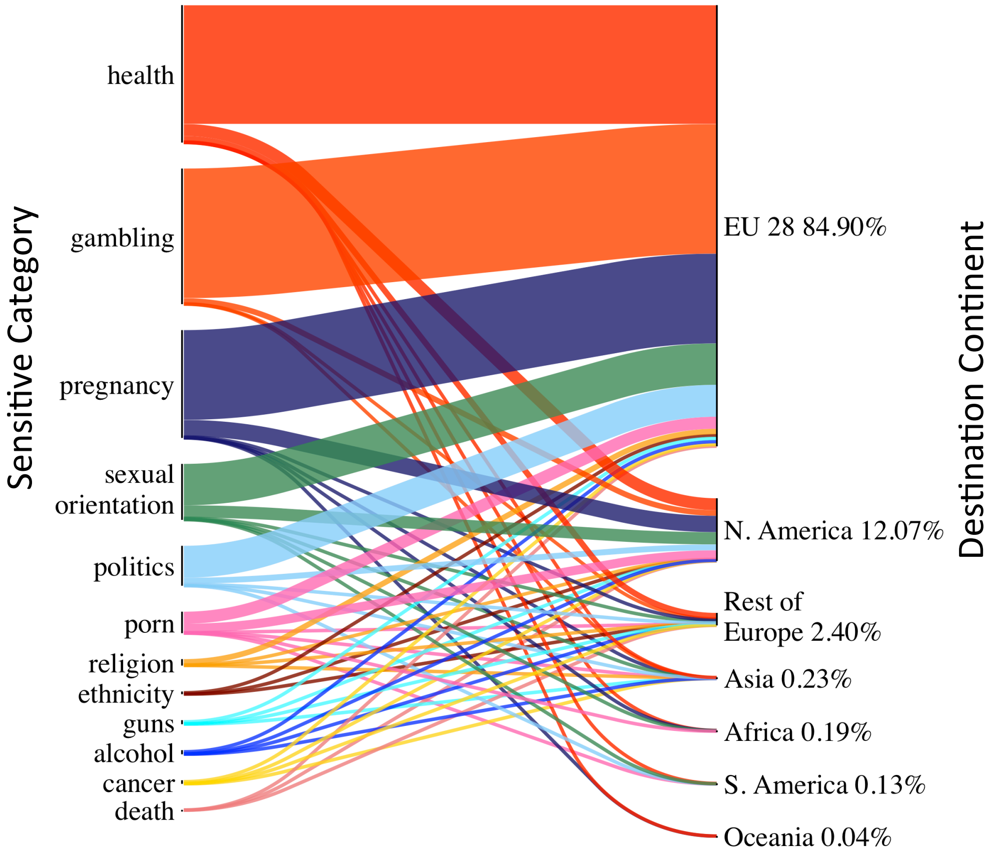
Genetic & biometric data



Health

Sexual Orientation

Results - Sensitive websites based on EU 28 users



Challenges

1. How to effectively **detect ad and tracking related domains in the wild**?

2. How to **ensure correct geolocation of infrastructure servers**?

3. How to ensure that **all possible ad and tracking servers are observed**?

4. How to maintain a **balance between accuracy and scalability**?

Scaling up – From real users to ISP flows

Scaling up – From real users to ISP flows

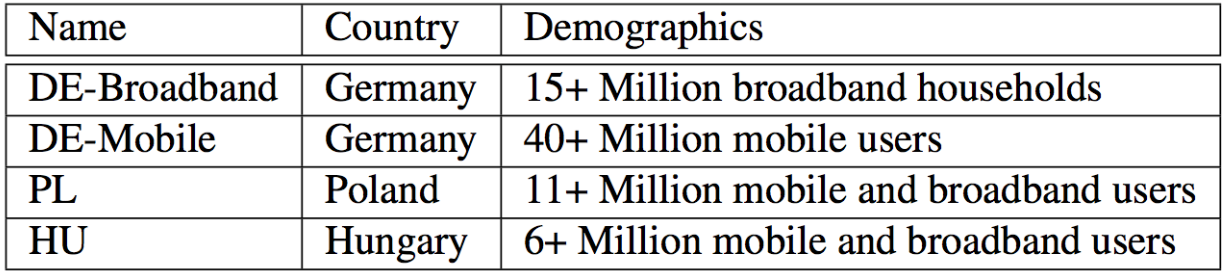
Datasets

List of Ad + Tracking IPs

< 28k IPs

+

ISPs Datasets



Name	Country	Demographics
DE-Broadband	Germany	15+ Million broadband households
DE-Mobile	Germany	40+ Million mobile users
PL	Poland	11+ Million mobile and broadband users
HU	Hungary	6+ Million mobile and broadband users

Scaling up – From real users to ISP flows

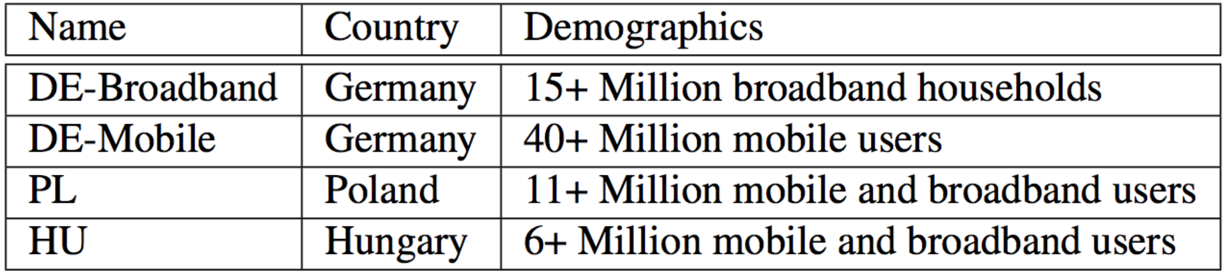
Datasets

List of Ad + Tracking IPs

< 28k IPs

+

ISPs Datasets



Name	Country	Demographics
DE-Broadband	Germany	15+ Million broadband households
DE-Mobile	Germany	40+ Million mobile users
PL	Poland	11+ Million mobile and broadband users
HU	Hungary	6+ Million mobile and broadband users

Four 24h daily snapshots

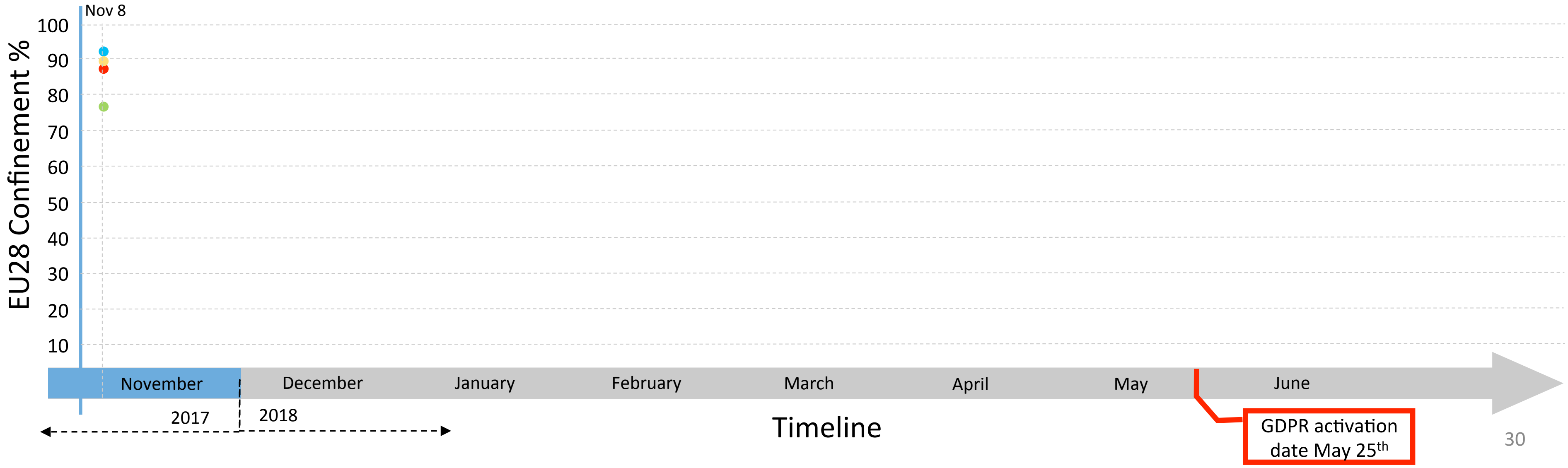
1. Wednesday Nov. 8, 2017
2. Wednesday Apr. 4, 2018
3. Wednesday May 16, 2018
4. Wednesday June 20, 2018

Scaling up – Continent level ISPs results

	● DE-Broadband				● DE-Mobile				● PL				● HU			
	Nov 8				Nov 8				Nov 8				Nov 8			
#Sampled Tracking Flows (in Millions)	1,057.0				70.4				13.8				43.3			
EU28	**88.5%**				**91.1%**				**77.5%**				**89.5%**			
North America	10%				6.9%				19.8%				10.2%			
Rest Europe	<1%				<1%				1.9%				<1%			
Asia	<1%				<1%				<1%				<1%			
Rest World	<1%				<1%				<1%				<1%			

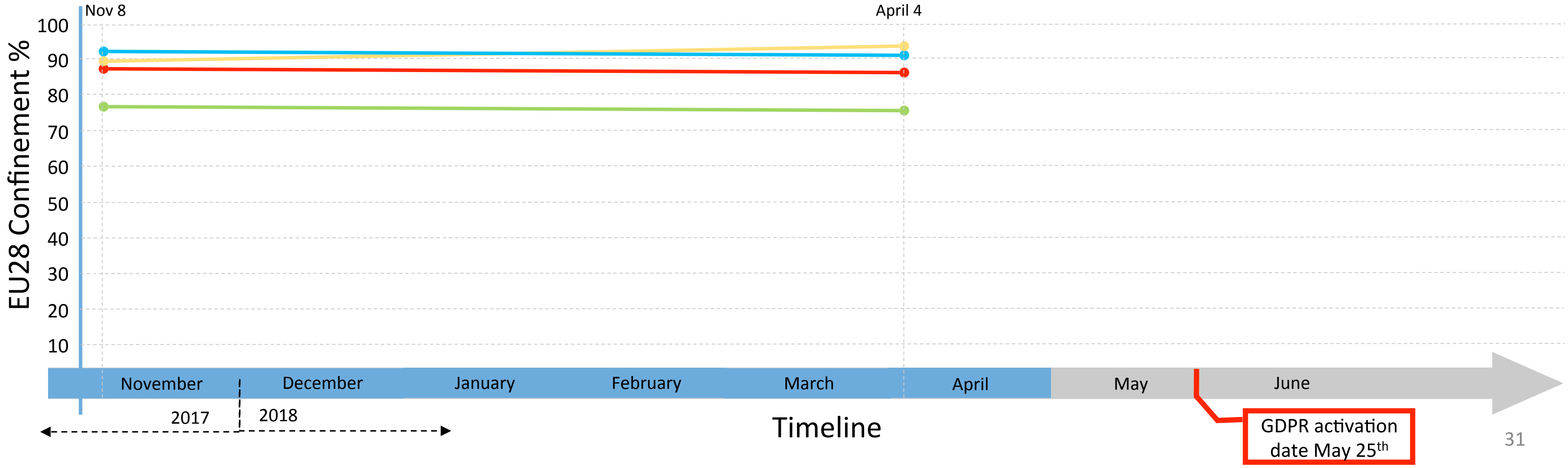
Scaling up – Continent level ISPs results

	● DE-Broadband				● DE-Mobile				● PL				● HU			
	Nov 8				Nov 8				Nov 8				Nov 8			
#Sampled Tracking Flows (in Millions)	1,057.0				70.4				13.8				43.3			
EU28	**88.5%**				**91.1%**				**77.5%**				**89.5%**			
North America	10%				6.9%				19.8%				10.2%			
Rest Europe	<1%				<1%				1.9%				<1%			
Asia	<1%				<1%				<1%				<1%			
Rest World	<1%				<1%				<1%				<1%			



Scaling up – Continent level ISPs results

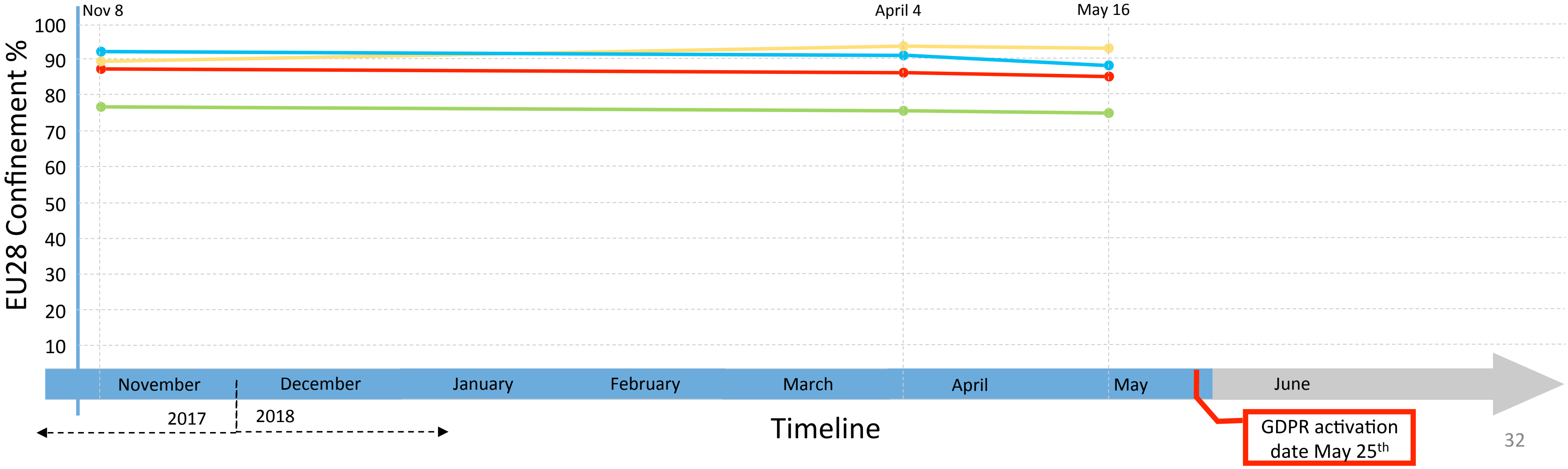
	● DE-Broadband				● DE-Mobile				● PL				● HU			
	Nov 8	April 4			Nov 8	April 4			Nov 8	April 4			Nov 8	April 4		
#Sampled Tracking Flows (in Millions)	1,057.0	1,200.8			70.4	77.4			13.8	13.8			43.3	50.2		
EU28	**88.5%**	**87.7%**			**91.1%**	**90.8%**			**77.5%**	**75.6%**			**89.5%**	**93.1%**		
North America	10%	9.3%			6.9%	6.6%			19.8%	21.5%			10.2%	6.3%		
Rest Europe	<1%	1.7%			<1%	2%			1.9%	1.9%			<1%	<1%		
Asia	<1%	<1%			<1%	<1%			<1%	<1%			<1%	<1%		
Rest World	<1%	<1%			<1%	<1%			<1%	<1%			<1%	<1%		



GDPR activation date May 25th

Scaling up – Continent level ISPs results

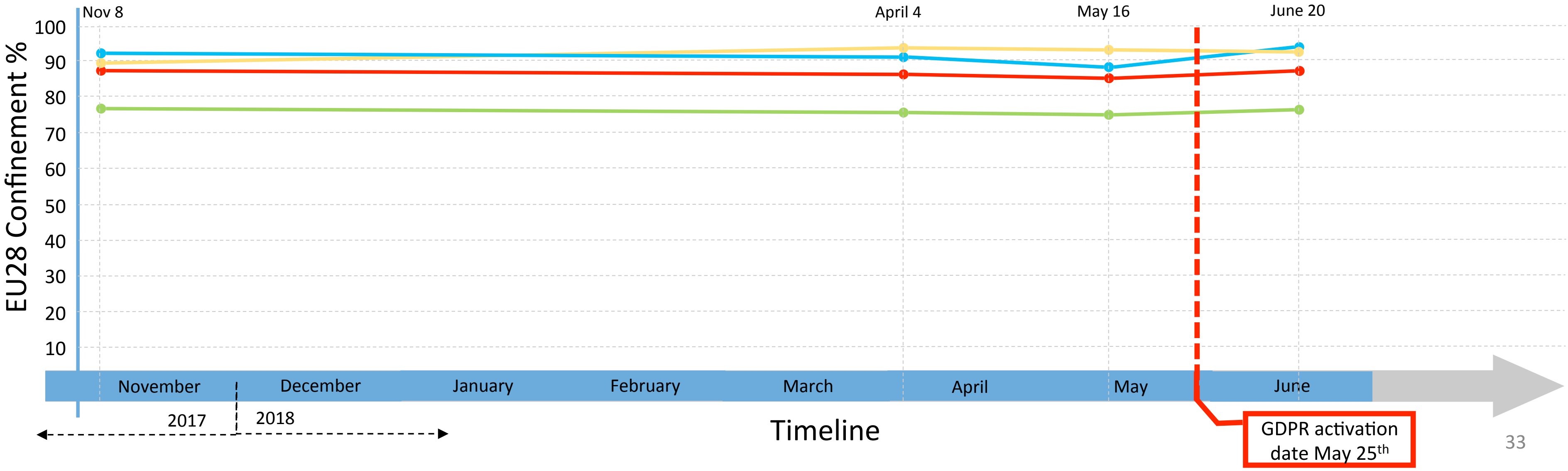
	● DE-Broadband				● DE-Mobile				● PL				● HU			
	Nov 8	April 4	May 16		Nov 8	April 4	May 16		Nov 8	April 4	May 16		Nov 8	April 4	May 16	
#Sampled Tracking Flows (in Millions)	1,057.0	1,200.8	1,105.3		70.4	77.4	70.8		13.8	13.8	12.4		43.3	50.2	39.3	
EU28	**88.5%**	**87.7%**	**86.5%**		**91.1%**	**90.8%**	**89.9%**		**77.5%**	**75.6%**	**74.7%**		**89.5%**	**93.1%**	**92.4%**	
North America	10%	9.3%	9.2%		6.9%	6.6%	6.4%		19.8%	21.5%	22%		10.2%	6.3%	7%	
Rest Europe	<1%	1.7%	2.9%		<1%	2%	3.1%		1.9%	1.9%	1.7%		<1%	<1%	<1%	
Asia	<1%	<1%	<1%		<1%	<1%	<1%		<1%	<1%	<1%		<1%	<1%	<1%	
Rest World	<1%	<1%	<1%		<1%	<1%	<1%		<1%	<1%	1.1%		<1%	<1%	<1%	



GDPR activation date May 25th

Scaling up – Continent level ISPs results

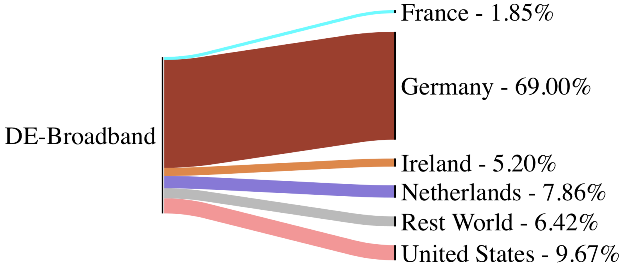
	● DE-Broadband				● DE-Mobile				● PL				● HU			
	Nov 8	April 4	May 16	June 20	Nov 8	April 4	May 16	June 20	Nov 8	April 4	May 16	June 20	Nov 8	April 4	May 16	June 20
#Sampled Tracking Flows (in Millions)	1,057.0	1,200.8	1,105.3	963.4	70.4	77.4	70.8	74.5	13.8	13.8	12.4	11.9	43.3	50.2	39.3	33.6
EU28	**88.5%**	**87.7%**	**86.5%**	**88.3%**	**91.1%**	**90.8%**	**89.9%**	**92.5%**	**77.5%**	**75.6%**	**74.7%**	**75%**	**89.5%**	**93.1%**	**92.4%**	**91.6%**
North America	10%	9.3%	9.2%	8.4%	6.9%	6.6%	6.4%	5.1%	19.8%	21.5%	22%	21.3%	10.2%	6.3%	7%	7.7%
Rest Europe	<1%	1.7%	2.9%	1.8%	<1%	2%	3.1%	1.3%	1.9%	1.9%	1.7%	3.4%	<1%	<1%	<1%	<1%
Asia	<1%	<1%	<1%	<1%	<1%	<1%	<1%	<1%	<1%	<1%	<1%	<1%	<1%	<1%	<1%	<1%
Rest World	<1%	<1%	<1%	<1%	<1%	<1%	<1%	<1%	<1%	<1%	1.1%	<1%	<1%	<1%	<1%	<1%

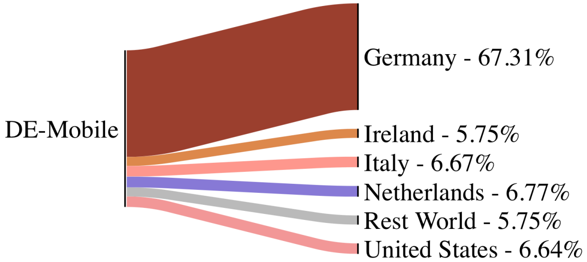


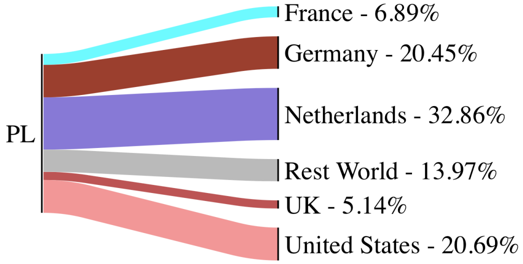
GDPR activation date May 25th

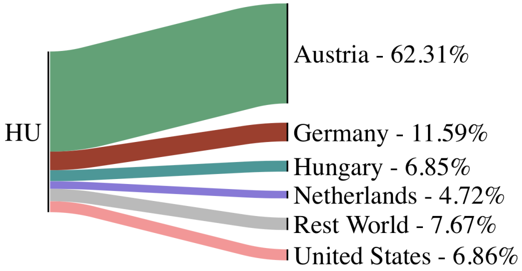
Country level confinements

ISPs dataset at April 4th









Can we further improve localization?

Two approaches:

1. Using DNS optimization

   Group server IPs (locations) based on:

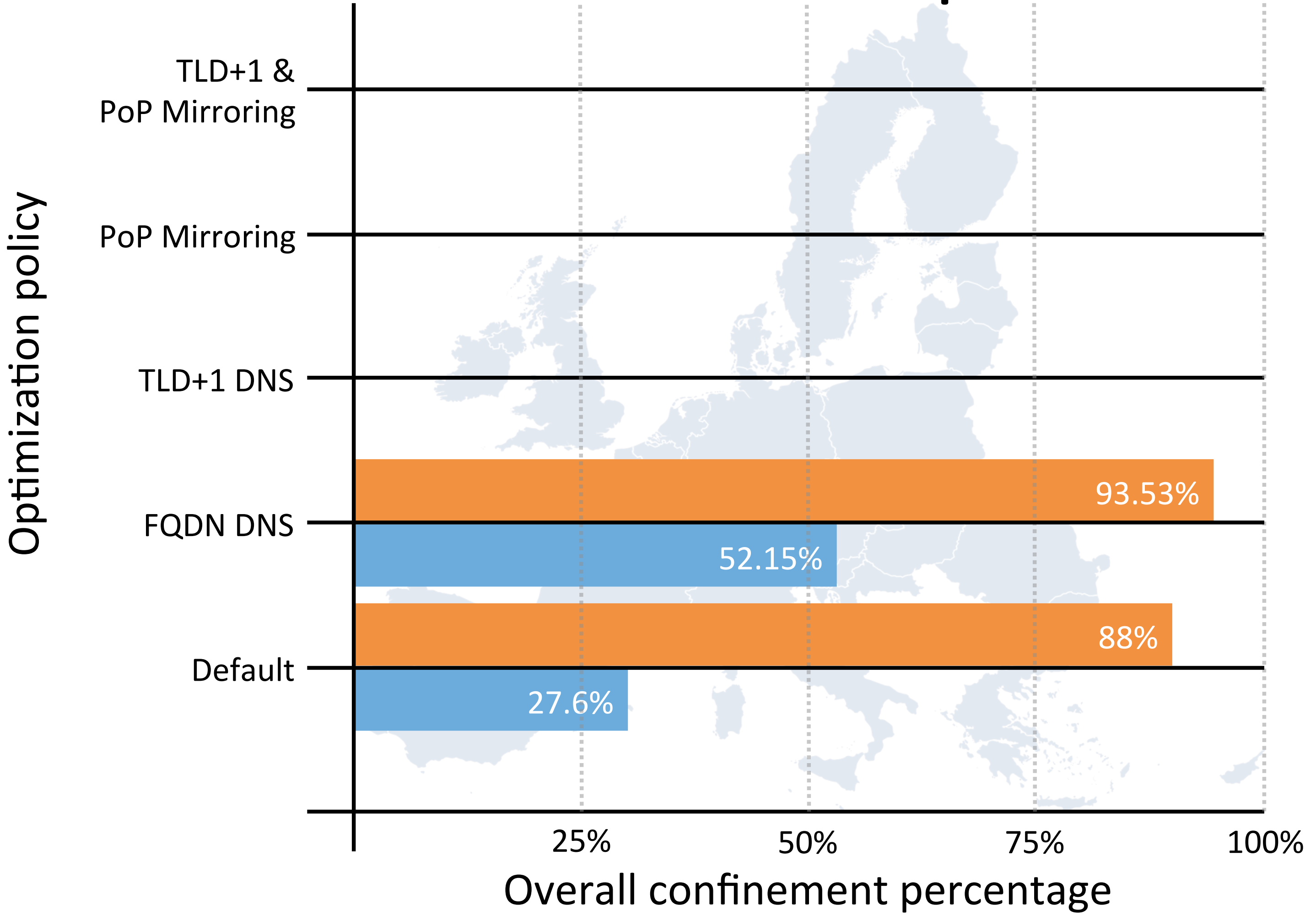
   a) Fully Qualified Domain Names (FQDN) *i.e., sub_d.tracker.com*

   b) Top Level Domain plus one (TLD+1) *i.e., tracker.com*

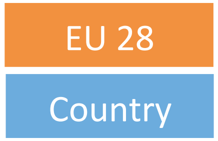
2. Using PoP Mirroring

   Deploy/migrate PoP servers based on cloud services datacenters availability

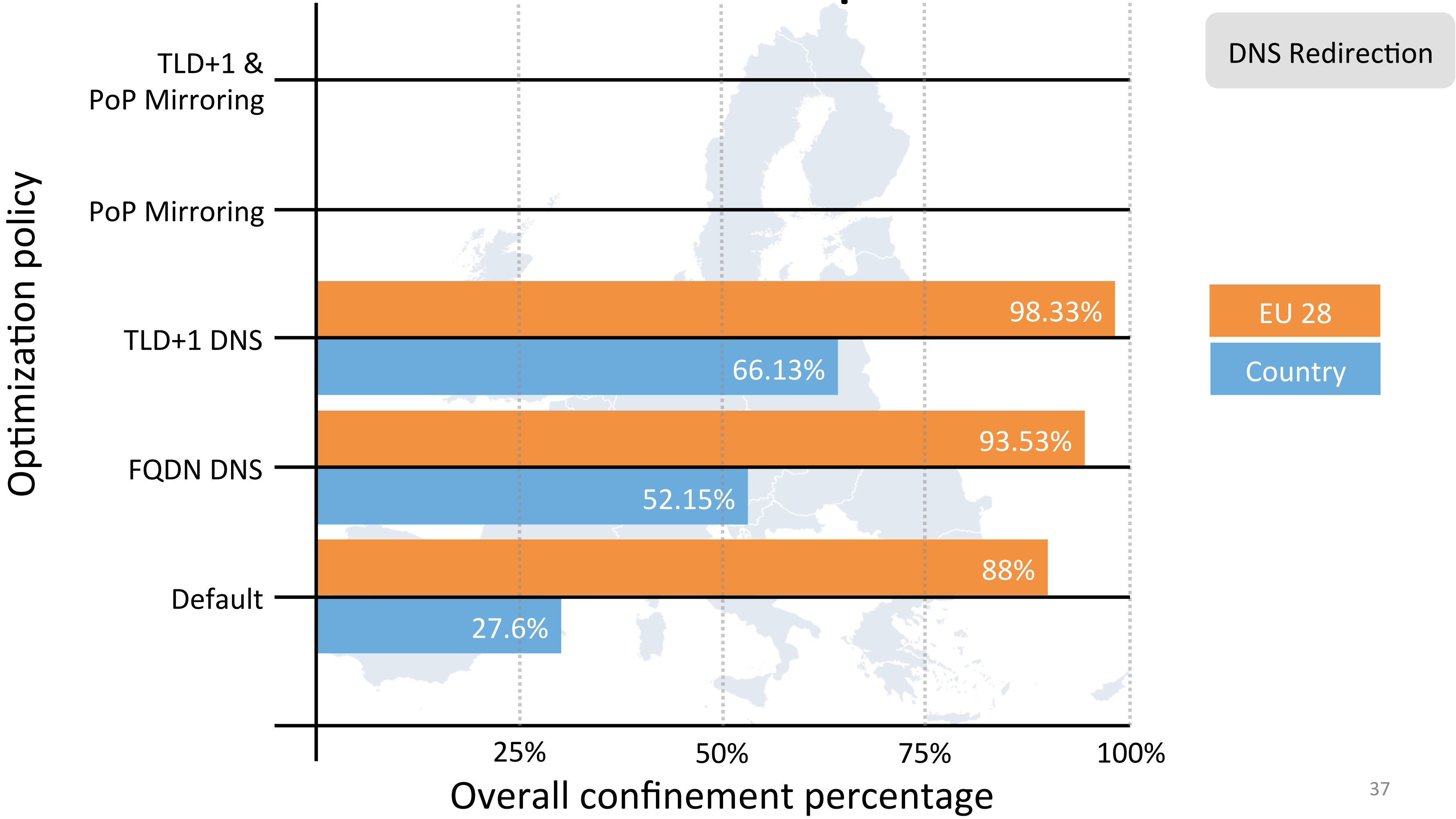
EU 28 localization improvement



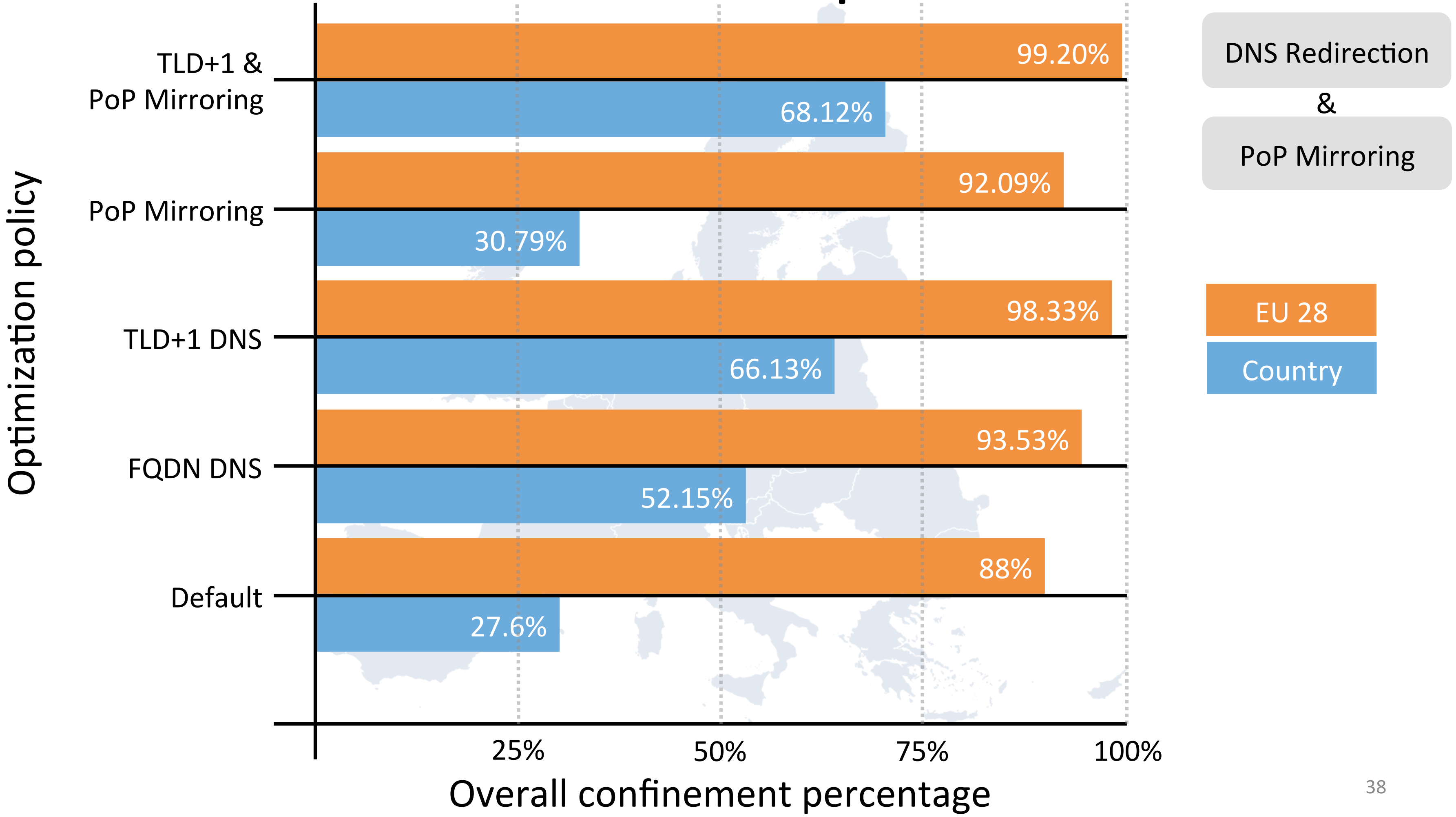
DNS Redirection



EU 28 localization improvement

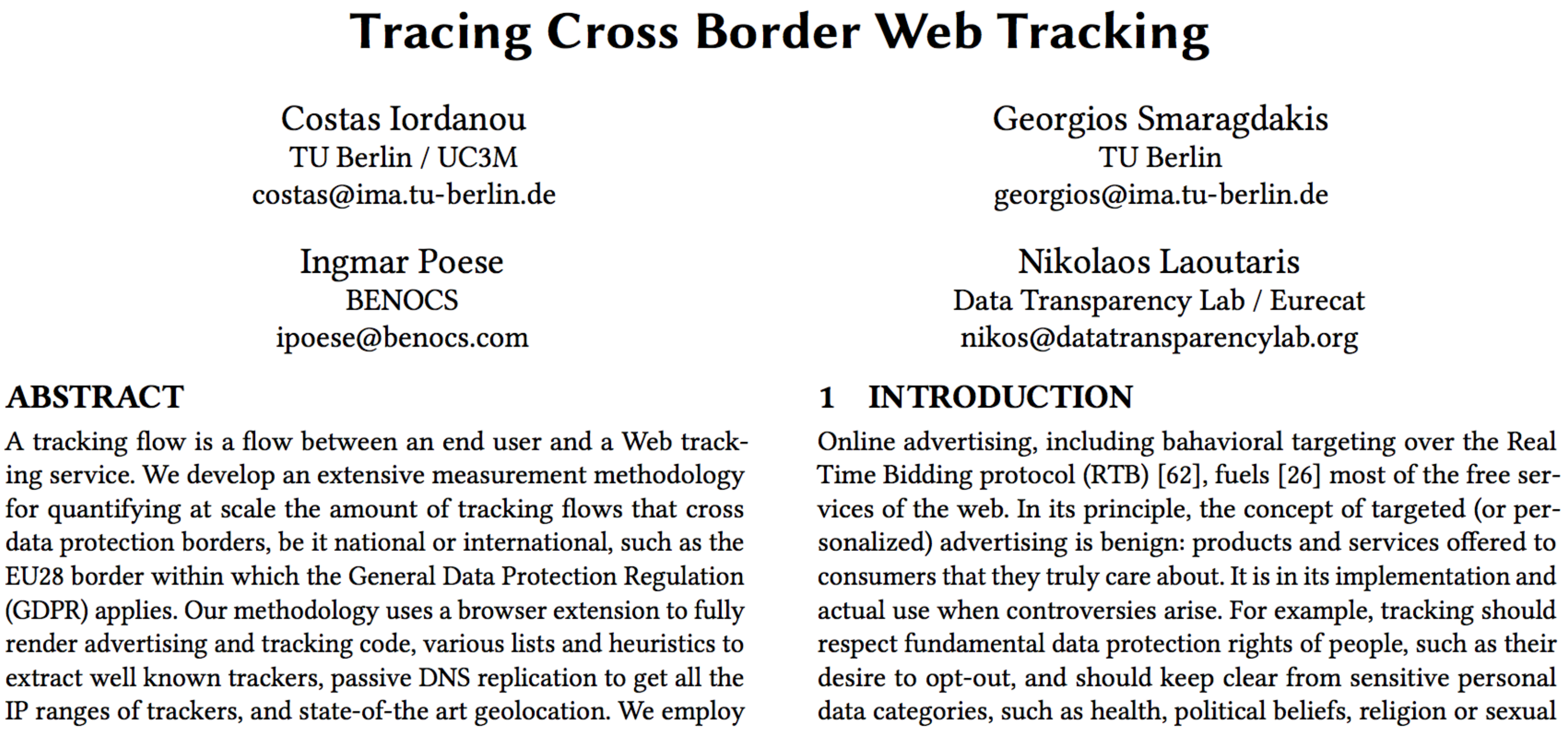


EU 28 localization improvement



In the paper

- Details on the methodology
- More results



Tracing Cross Border Web Tracking

Costas Iordanou
TU Berlin / UC3M
costas@ima.tu-berlin.de

Georgios Smaragdakis
TU Berlin
georgios@ima.tu-berlin.de

Ingmar Poese
BENOCS
ipoese@benocs.com

Nikolaos Laoutaris
Data Transparency Lab / Eurecat
nikos@datatransparencylab.org

ABSTRACT

A tracking flow is a flow between an end user and a Web tracking service. We develop an extensive measurement methodology for quantifying at scale the amount of tracking flows that cross data protection borders, be it national or international, such as the EU28 border within which the General Data Protection Regulation (GDPR) applies. Our methodology uses a browser extension to fully render advertising and tracking code, various lists and heuristics to extract well known trackers, passive DNS replication to get all the IP ranges of trackers, and state-of-the art geolocation. We employ

1 INTRODUCTION

Online advertising, including bahavioral targeting over the Real Time Bidding protocol (RTB) [62], fuels [26] most of the free services of the web. In its principle, the concept of targeted (or personalized) advertising is benign: products and services offered to consumers that they truly care about. It is in its implementation and actual use when controversies arise. For example, tracking should respect fundamental data protection rights of people, such as their desire to opt-out, and should keep clear from sensitive personal data categories, such as health, political beliefs, religion or sexual

Main takeaways

1. ≈90% of tracking flows from EU 28 terminates within EU 28
2. Incorrect geolocation approach can totally flip the results
3. Country level confinement is correlated with the IT infrastructure
4. DNS redirection & PoP Mirroring can improve confinement levels
5. ≈3% of the tracking flows are in sensitive categories

Tracing Cross Border Web Tracking

Costas Iordanou
email: costas@ima.tu-berlin.de

Technische Universität Berlin

Universidad Carlos III de Madrid

BENOCS

DATA TRANSPARENCY LAB

erc European Research Council