Measuring the advertising value of real users & the geographical footprint of web tracking

Nikolaos Laoutaris

Chief Scientist DTL
Director of Data Science Eurecat

Data Transparency: Concerns and Prospects

By NIKOLAOS LAOUTARIS
Data Transaprency Lab, Barcelona 08019, Spain

I. INTRODUCTION

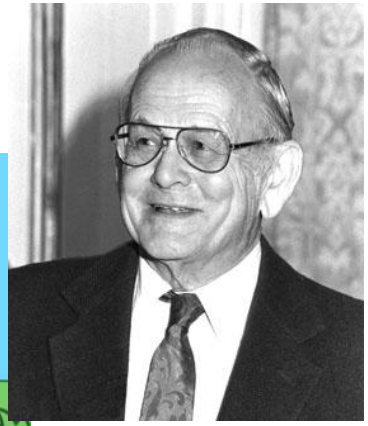
The question of "how far" technologies and business models of the web should go into collecting personal data of unassuming, or at best moderately informed data protection problems, as well as to point toward concerns and challenges to be addressed in order for this to materialize. Most of the discussion applies to the use of personal data by marketers on the fixed and mobile web, but some parts may also be relevant to other online and offline use cases and/or types of data (e.g., off-web health and financial data).

For years, the practice of collecting data on individuals at unprecedented scale was a not an issue for most people, for the simple reason that the public, and even governments, were just unaware of its magnitude, precision, and detail. The last few years, however, attitudes have started to change and the topic of privacy is increasingly appear-

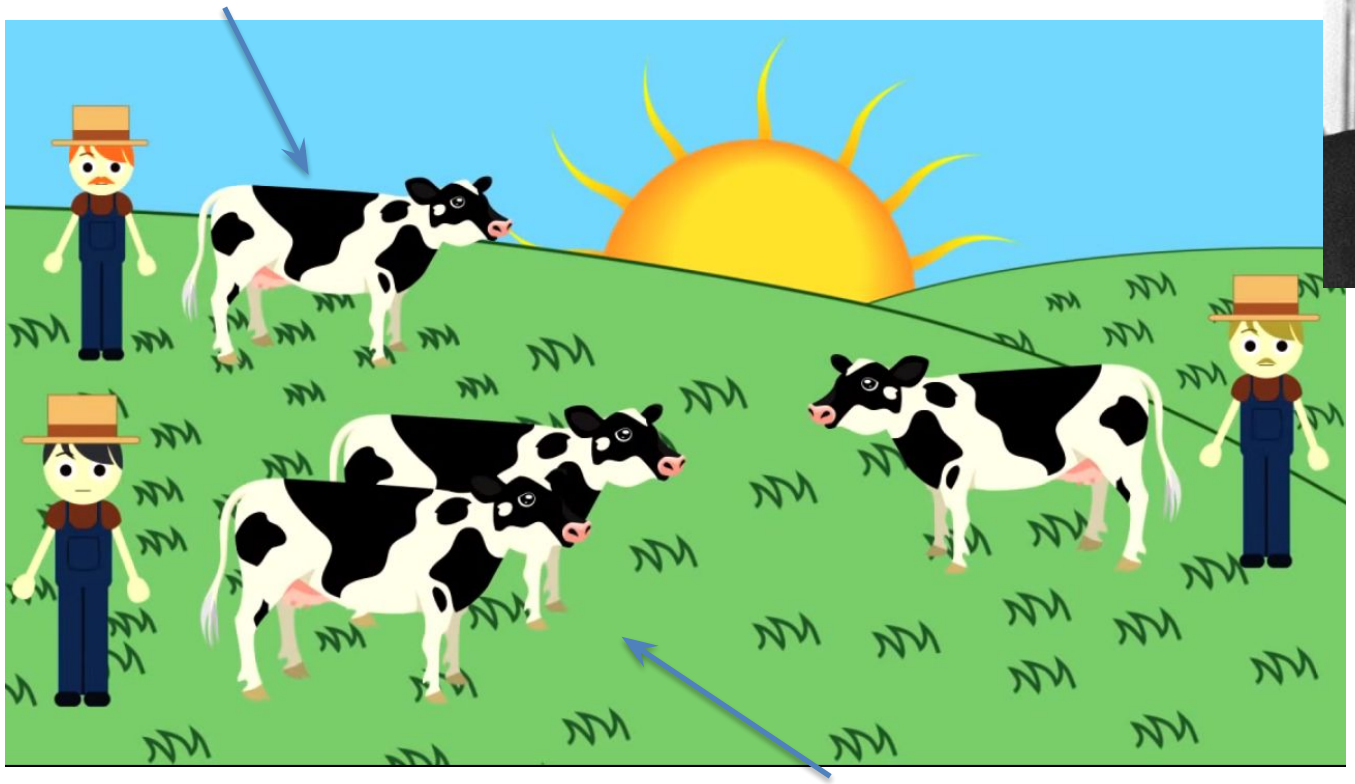
The utility-vs.-privacy tradeoff

Tragedy of the commons around privacy

Garrett Hardin, 1968

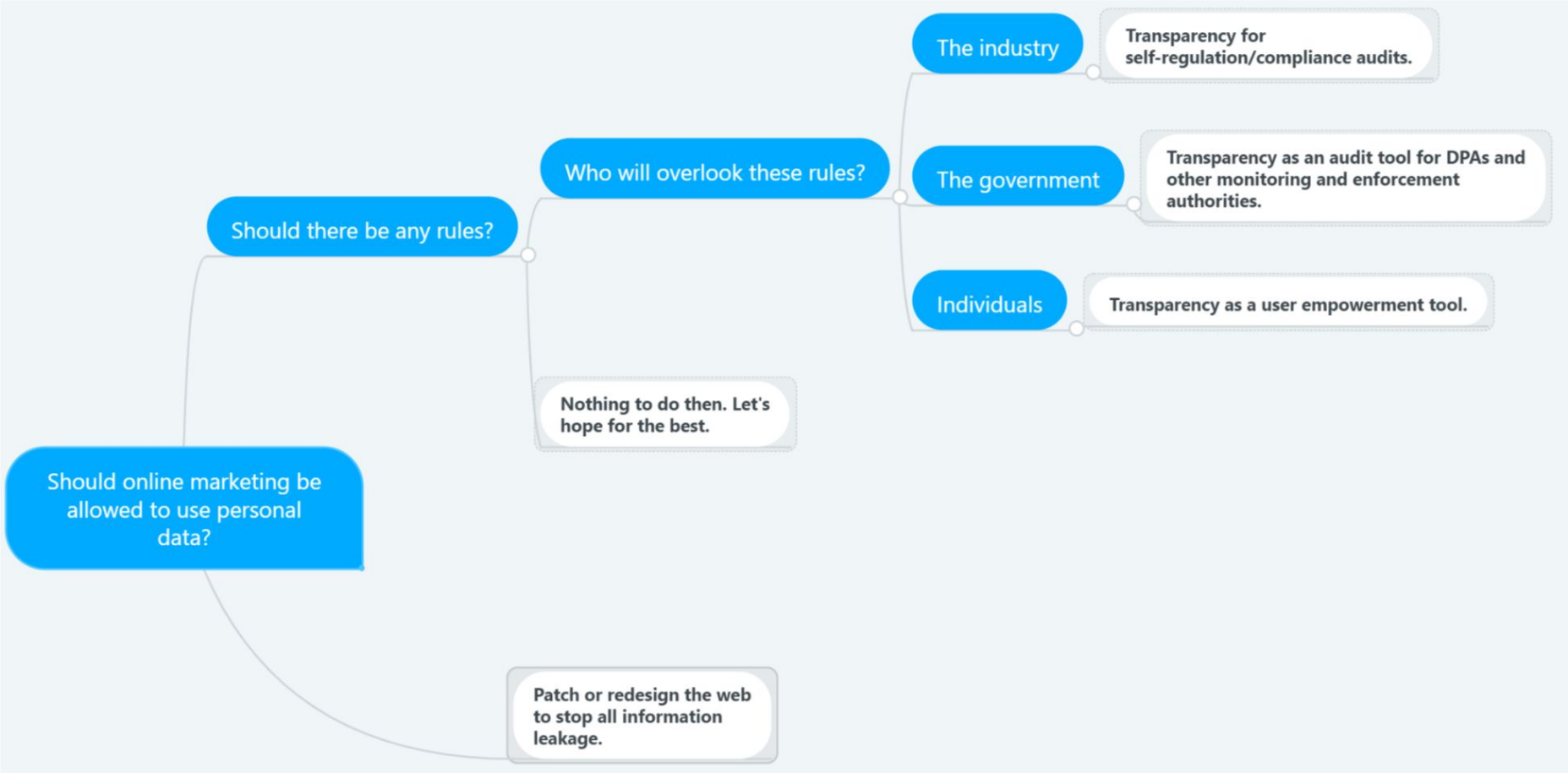


Internet company in
Web Economy … crossing privacy red lines



The “commons”: consumer trust on the web and it’s business models

Transparency & Data Protection



Transparency Software
Creation of the first DTL tool

$heriff

Detecting Price Discrimination

1. Select price

Availability: In Stock

Price: ~~$200.50~~ $189.00

Ex Tax: $90.00

Price in reward points: 400

10 or more $105.40

2. Check it

189.00

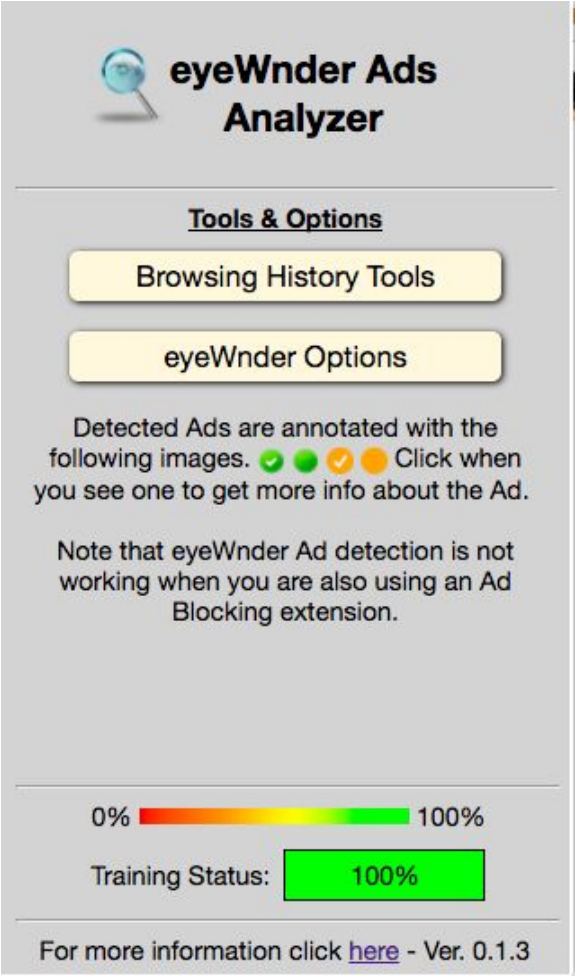
Check it

3. Examine differences

S, Safari, Spain	$189.00
Firefox, Spain	$189.00
m, Liège	$165.99
São Paulo	$189.00
d, Tampere	$189.00
any, Berlin	$201.50

Jakub Mikians
UPC (now Amazon)

Kostas Iordanou
Telefonica-UC3M



Detects & annotates all rendered ads

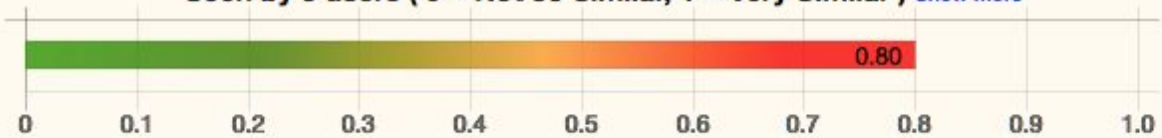
eyeWnder Ad Results - Ad ID: 7dd66cfb-38a3-45a0-9cfc-d64d0f5b195d

Check your browsing history using the eyeWnder analysis tool: Analysis Tool

Selected Advertisement

User Feedback: Do you think it was a targeted ad?

Users Demographics Similarity

How similar are you to others that have seen this ad?	**Seen by 3 users (0 = Not so Similar, 1 = Very Similar)** Show more 

Advertisement Information Table

This ad takes you to:	https://subscribe.wired.com/subscribe/wired/103100?source=WIR_Footer_IntlTargeting_Apr16
This ad is about:	People & Society > Men's Interests Books & Literature > Magazines News > Technology News

Analyze Advertisement

Get Intermediaries

This talk is about

A) how much advertisers pay to show you an ad

B) geographical aspects of tracking

If you are not paying for it, you are the product: How much do advertisers pay to reach you?

Panagiotis Papadopoulos
FORTH-ICS, Greece
panpap@ics.forth.gr

Nicolas Kourtellis
Telefonica Research, Spain
nicolas.kourtellis@telefonica.com

Pablo Rodriguez Rodriguez
Telefonica Alpha, Spain
pablo.rodriguezrodriguez@telefonica.com

Nikolaos Laoutaris
Data Transparency Lab, Spain
nikos@datatransparencylab.org

ABSTRACT

Online advertising is progressively moving towards a programmatic model in which ads are matched to actual interests of individuals collected as they browse the web. Letting the huge debate around privacy aside, a very important question in this area, for which little is known, is: *How much do advertisers pay to reach an individual?*

In this study, we develop a first of its kind methodology for computing exactly that – the price paid for a web user by the ad ecosystem – and we do that in real time. Our approach is based on tapping on the Real Time Bidding (RTB) protocol to collect cleartext and encrypted prices for winning bids paid by advertisers in order to place targeted ads. Our main technical contribution is a method for tallying winning bids even when they are encrypted. We achieve this by training a model using as ground truth prices obtained by running our own "probe" ad-campaigns. We design our methodology through a browser extension and a back-end server that provides it with fresh models for encrypted bids. We validate our methodology using a one year long trace of 1600 mobile users and demonstrate that it can estimate a user's advertising worth with more than 82% accuracy.

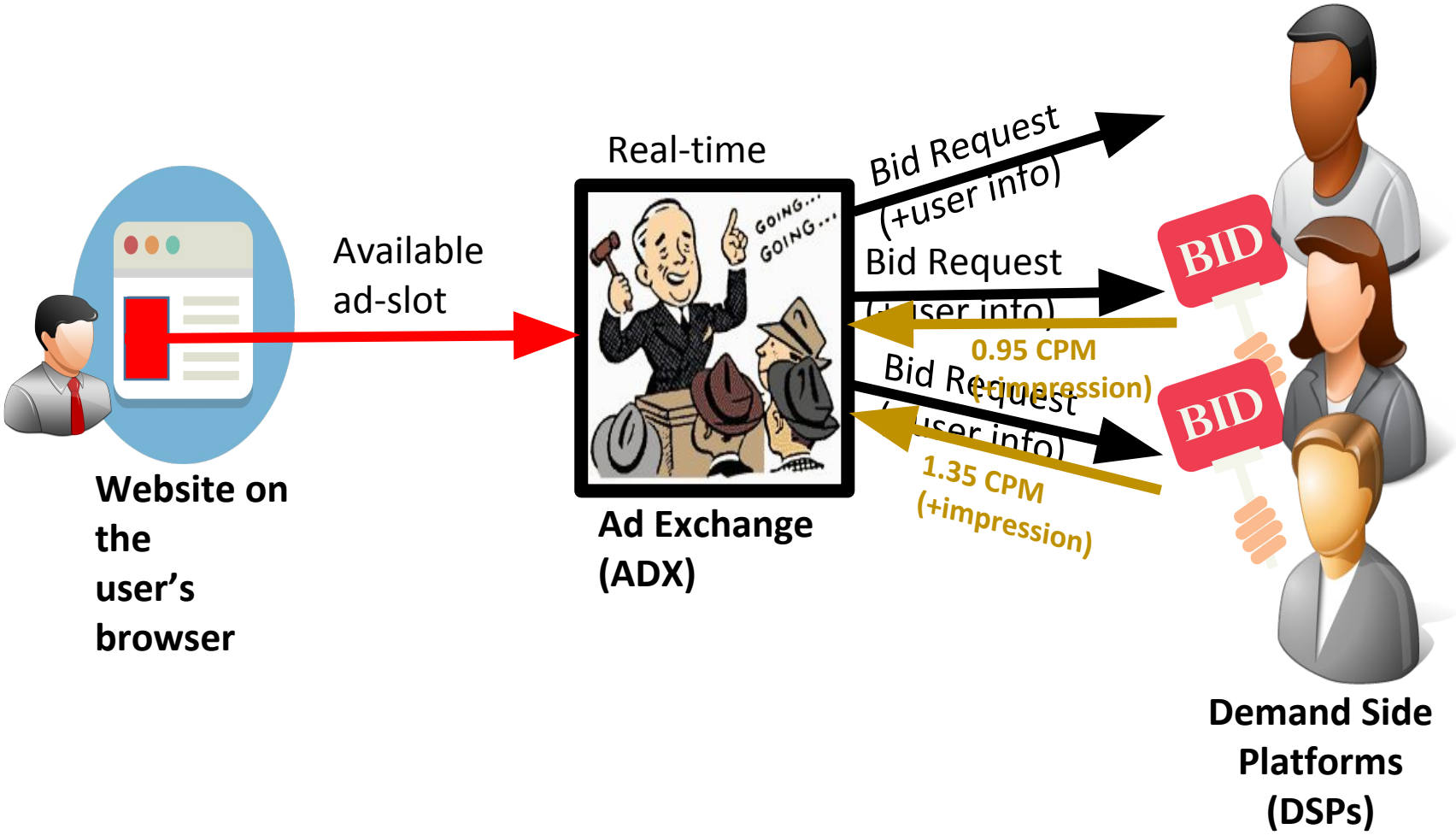
of auctions is the Real-Time Bidding (RTB) protocol for transacting digital display ads in real time. RTB has been growing with an annual rate of 128% [80], and currently accounts for 74% of programmatically purchased advertising. In US alone it created a revenue of $8.7 billion in 2016 [8].

Consequently, the collection of user personal data has become more aggressive and sometimes even intrusive [29, 33], raising a huge public debate around the tradeoffs between (i) innovation in advertising and marketing, and (ii) basic civil rights regarding privacy and personal data protection [51, 55]. These increasing privacy concerns, drew the attention of a significant body of research, which studied users' privacy loss in conjunction to existing user tracking techniques [1, 17, 21, 52, 60], and proposed various defence mechanisms to users [59, 64, 65]. Still, there is an outstanding question that remains unaddressed by the related work in the area. This question concerns transparency and is the following: *Based on the exposed user personal data, how much do advertisers pay to reach an individual?*

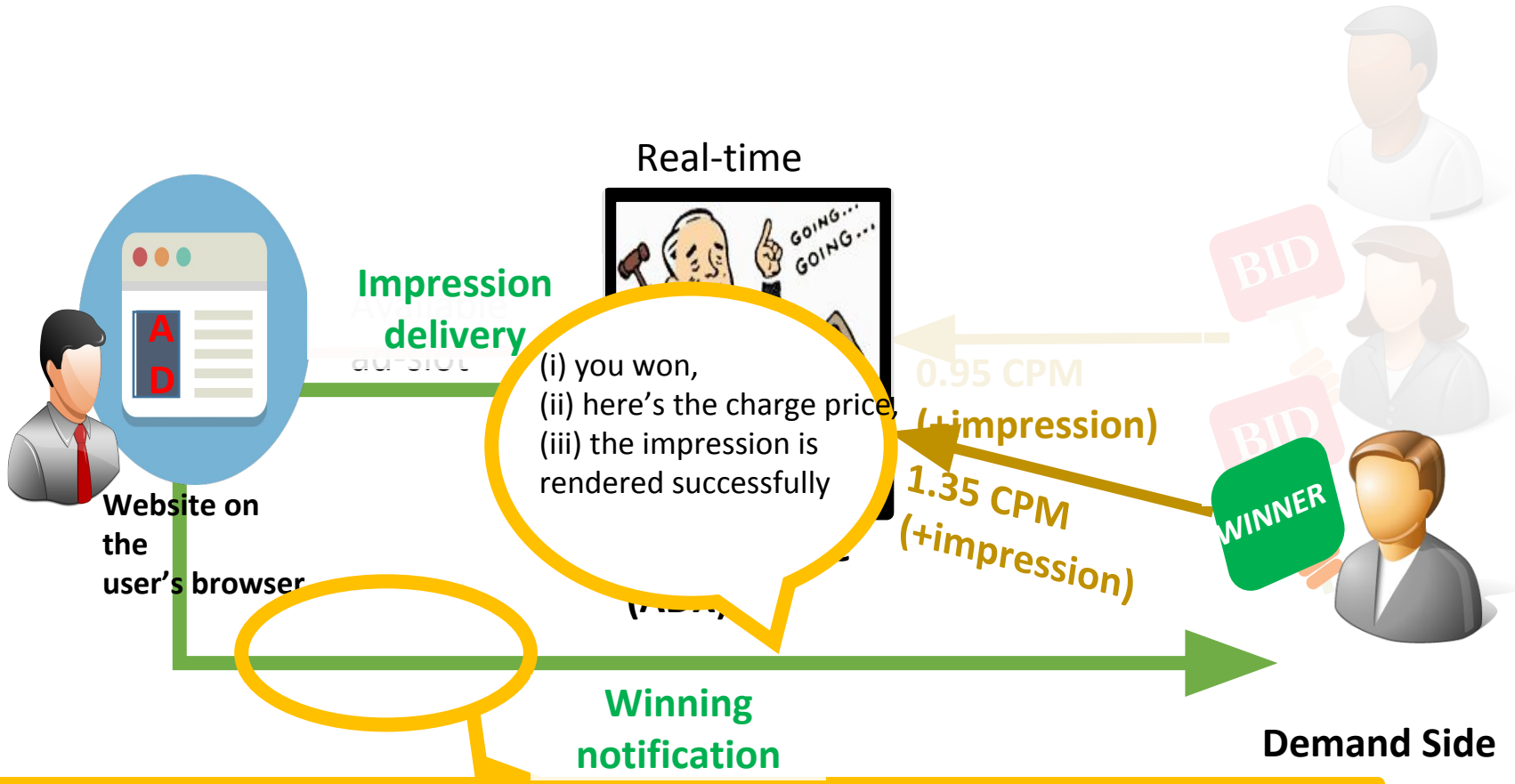
Despite the importance of this question, it is surprising how little is known about it. There exist several reports about the *average* revenue per user (ARPU) from online advertising [13, 30, 67], but

How much do advertisers actually pay to reach you?

Programmatic auctions over RTB



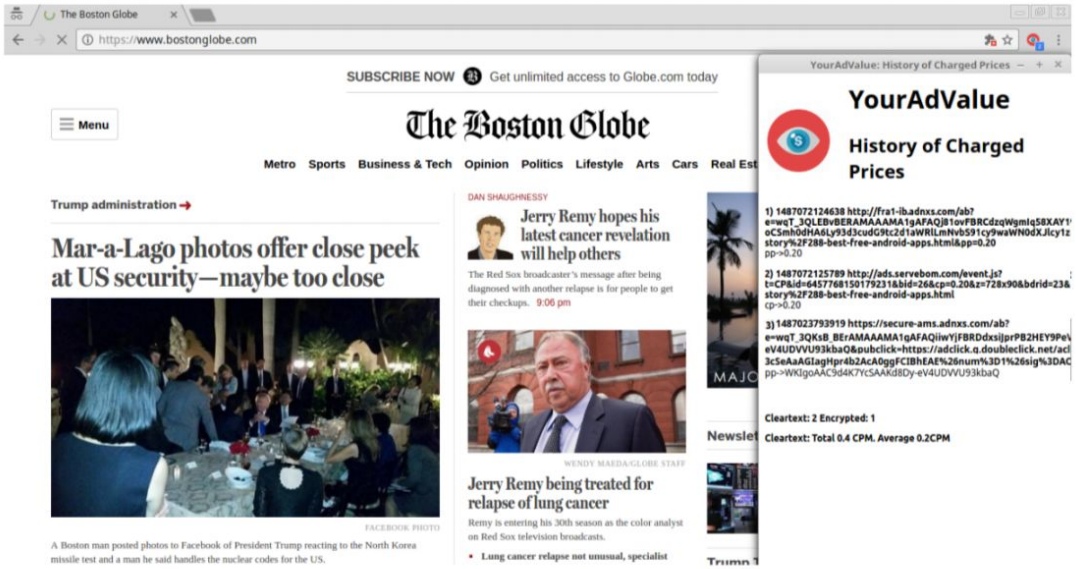
RTB price notification channel



nURL example:
cpp.imp.mpx.mopub.com/imp?ad_domain=amazon.es&ads_creative_id=ID&bidder_id=ID&..&bidder_name=..&
charge_price=0.95&country=ESP&currency=EUR&latency=0.116&mopub_id=ID&pub name=..

Tapping RTB for user valuation

A plugin for your browser that combines your online activity with Facebook's Public APIs to estimate your advertising value



The same service as FDVT but for RTB based programmatic advertising

DEMO VIDEO: https://youtu.be/QPfc-gXGdjI

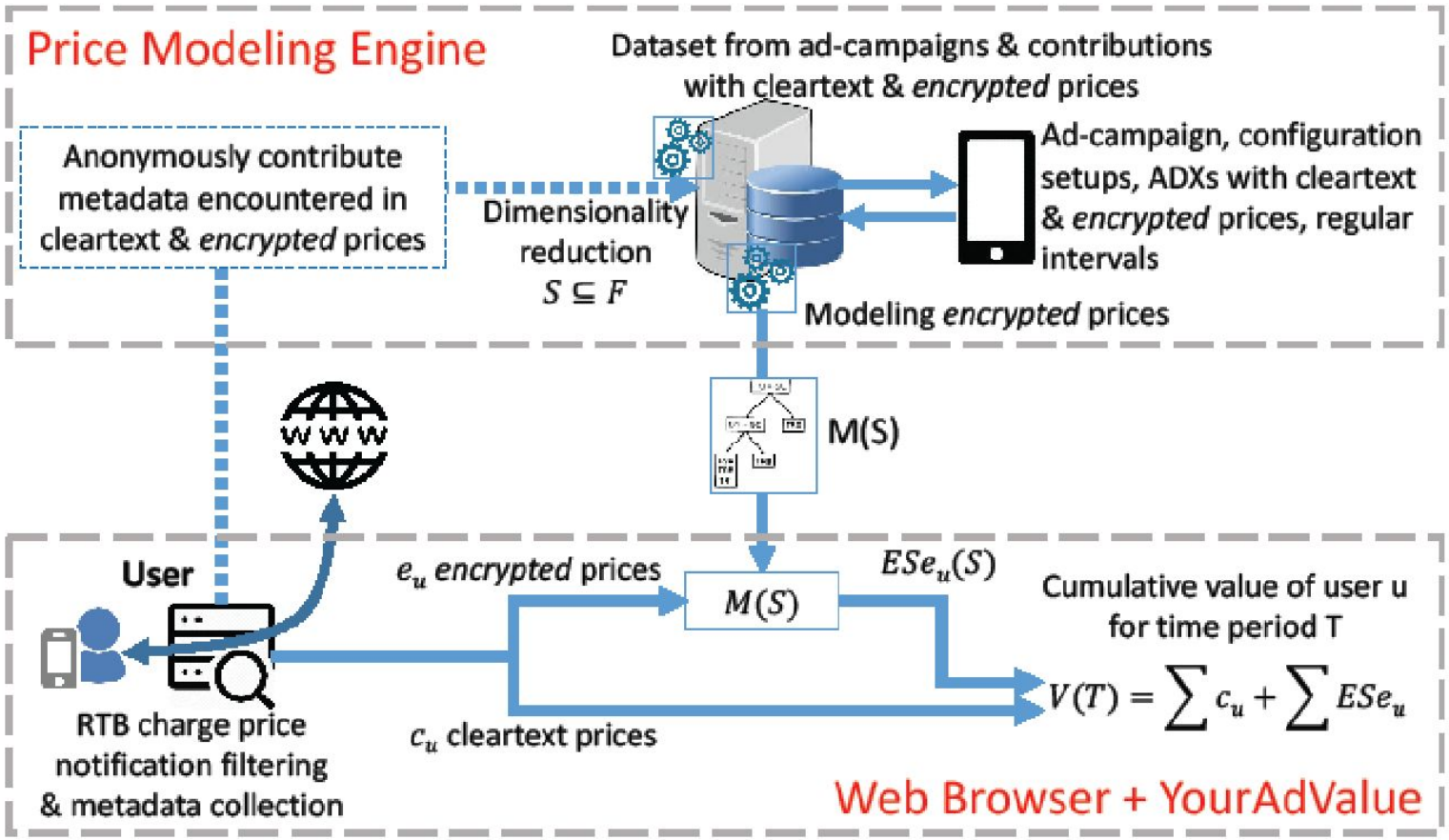
Methodology

1. Winning bid important features from unencrypted RTB

2. “probe” ad campaigns for encrypted RTB

3. “probe” campaign winning bids as ground truth for training a classifier for encrypted RTB

Methodology



Evaluating our approach

- Offline year-long (2015) dataset *D* with mobile traffic from 1600 real users

- Weblog Ads Analyzer:
  - ➢ filter RTB traffic and
  - ➢ extract features (auction's metadata and user data) from nURLs

Metric	D
Time period	12 months
Impressions	78,560
IAB category of publishers	18
RTB publishers	~5.6k/month

Real probing ad-campaigns

- 2 real probing ad-campaigns in 2016 (*A1*, *A2*): various experimental setups

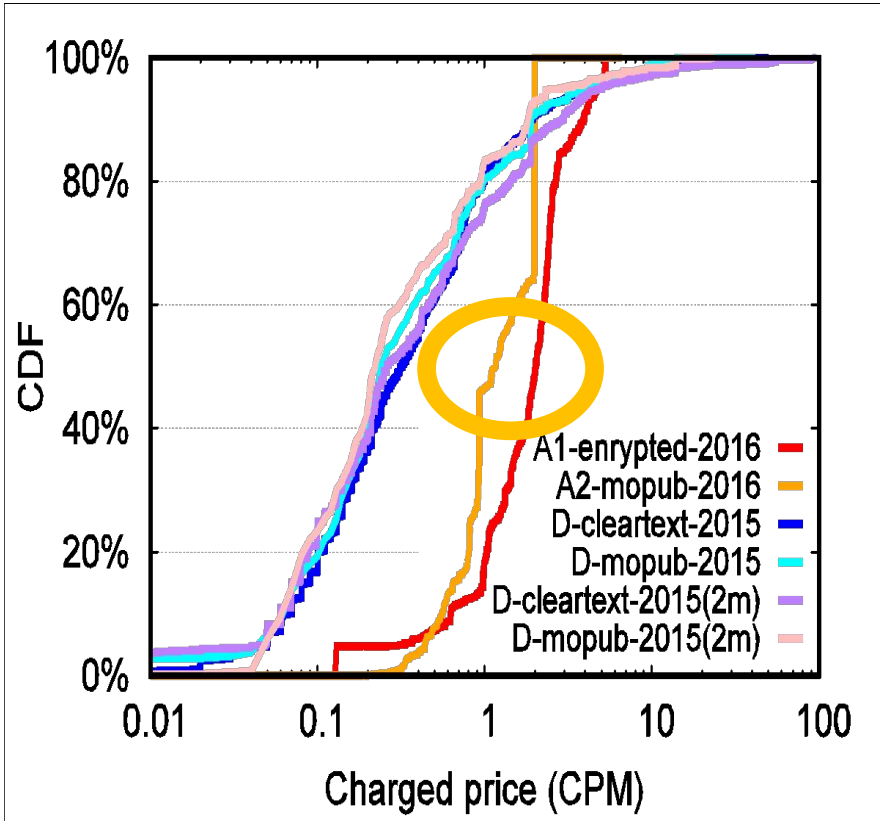
Filter name	Range of values (type)
Cities	Madrid, Barcelona, Valencia, Seville
Time of day	12am-9am, 9am-6pm, 6pm-12am
Day of week	Weekday, Weekend
Type of device	Smartphone, Tablet
Type of OS	iOS, Android
Ad-format (smartphone)	320x50, 300x250, 320x480 or 480x320
Ad-format (tablet)	728x90, 300x250, 768x1024 or 1024x768
Ad-exchange	MoPub, OpenX, Rubicon, DoubleClick, PulsePoint
Content category of publisher	all IABs possible

Metric	D	A1 (enc)	A2 (clr)
Time period	12 months	13 days	8 days
Impressions	78,560	632,667	318,964
IAB category of publishers	18	16	7
RTB publishers	~5.6k/month	~0.2k	~0.3k

Features that affect prices

- application/web-browser
- device type
- user location
- time of day
- day of week
- ad format
- publisher IAB categories (“Business & Marketing” more expensive than “Science”)
- ad exchange

Encrypted Vs. Cleartext prices

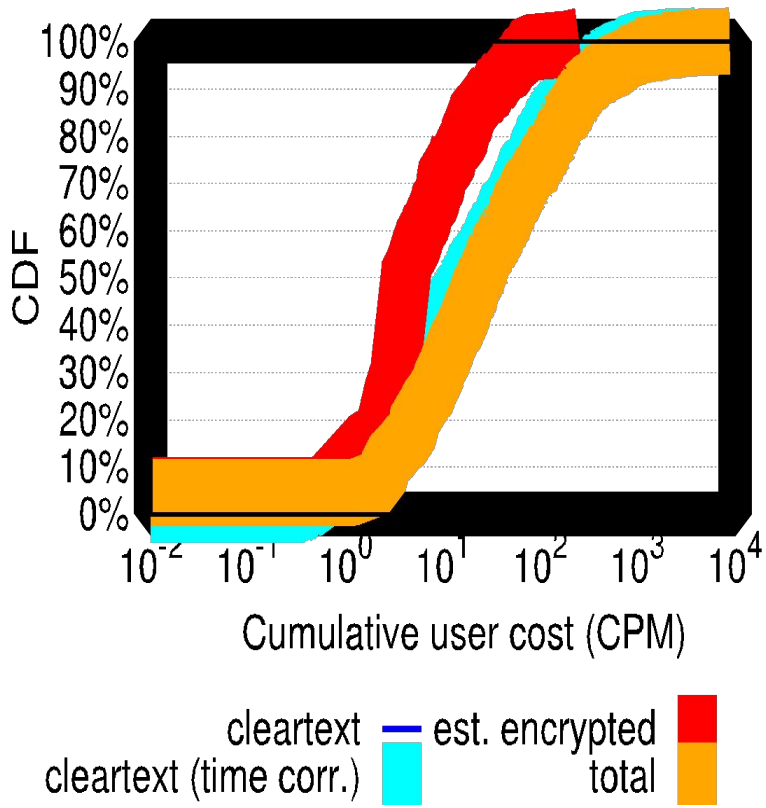


Comparison of price distributions between cleartext and encrypted, for different time periods and datasets (D vs. A1 and A2).

~~*"It's safe to assume that encrypted prices follow the same distribution with cleartext prices."*~~

price distribution of encrypted prices (A1):
→ distinctly different
→ about 1.7x higher median value than cleartext prices (A2)

How much do advertisers pay to reach you?



Cumulative CPM paid per user in our year-long dataset (2015)

- Cumulative cost from encrypted prices: cannot surpass cleartext (still dominant).
- some users more costly than others
- median user costs 25 CPM (73% of the users cost < 100 CPM)
- 2% of users cost 10-100× more to the ad-ecosystem than the average user!

Tracing Cross Border Web Tracking

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

Georgios Smaragdakis
TU Berlin
georgios@inet.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 our methodology on a dataset from 350 real users of the browser extension over a period of more than four months, and then generalize our results by analyzing billions of web tracking flows from more than 60 million broadband and mobile users from 4 large European ISPs. We show that the majority of tracking flows cross national borders in Europe but, unlike popular belief, are pretty well confined within the larger GDPR jurisdiction. Simple DNS redirection and PoP mirroring can increase national confinement while sealing almost all tracking flows within Europe. Last, we show that cross boarder tracking is prevalent even in sensitive and hence protected data categories and groups including health, sexual orientation, minors, and others.

1 INTRODUCTION

Online advertising, including bahavioral targeting over the Real Time Bidding protocol (RTB) [60], fuels [25] 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 orientation. One of the most important changes on how to process and store personal data is the European Union General Data Protection Regulation (GDPR) [4]. GDPR offers protection to European citizens across a wide range of privacy threats, including tracking on sensitive categories such as those mentioned above. Now that Europe's new data protection law is in place (implementation date of the GDPR across the European Union was on May 25, 2018; the regulation entered into force on May 24, 2016), the next challenge becomes implementing it in practice. GDPR has provisions that include steep fines reaching up to 4% of worldwide turnover or 20 million euros, whichever is higher, for any company found in violation. Monitoring the effectiveness of the law, investigating complaints, and prosecu

[Distinguished paper award]

Methodology

Browser extension with real users

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

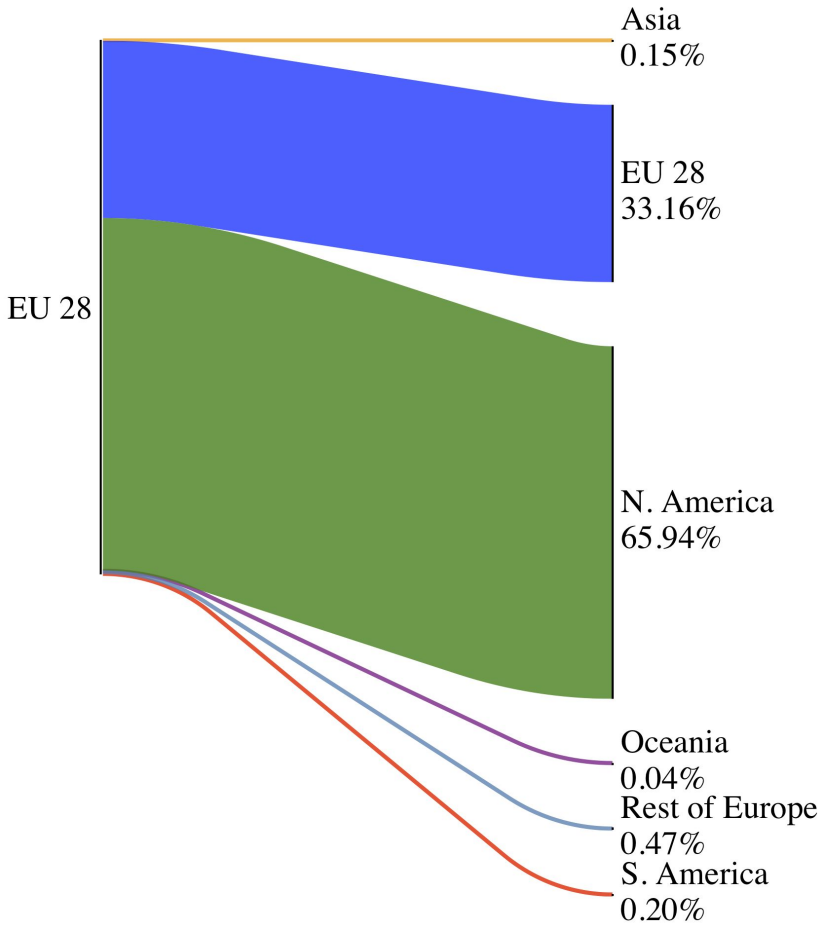
ABP Parser

Correction Script

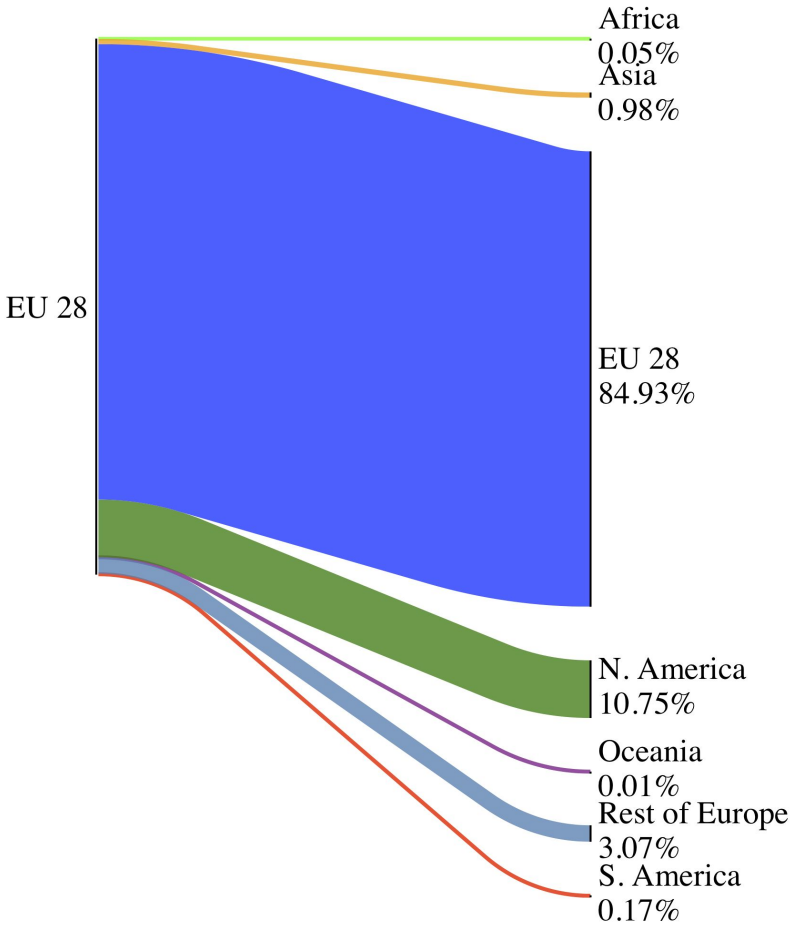
RIPE NCC
RIPE NETWORK COORDINATION CENTRE
https://ipmap.ripe.net/

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 confinement level



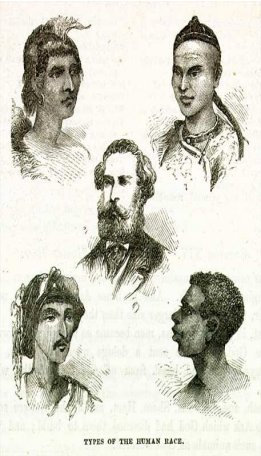
MaxMind geo-location



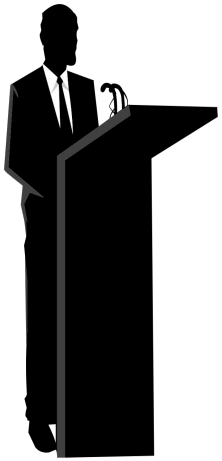
RIPE IPmap geo-location

What about sensitive websites?

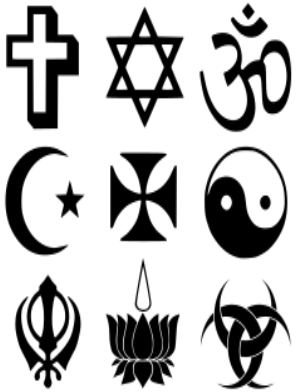
Sensitive categories as defined by GDPR



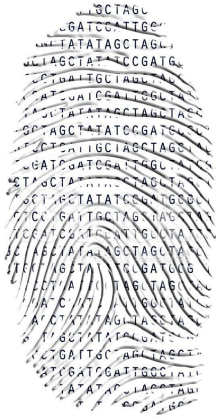
Race & Ethnicity



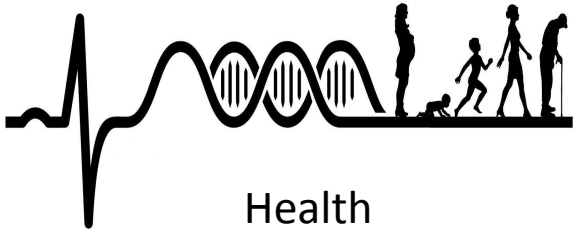
Political beliefs



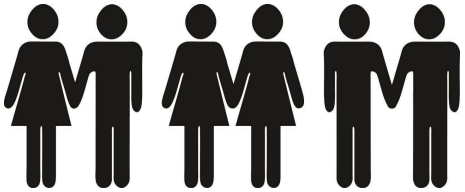
Religon



Genetic & biometric data

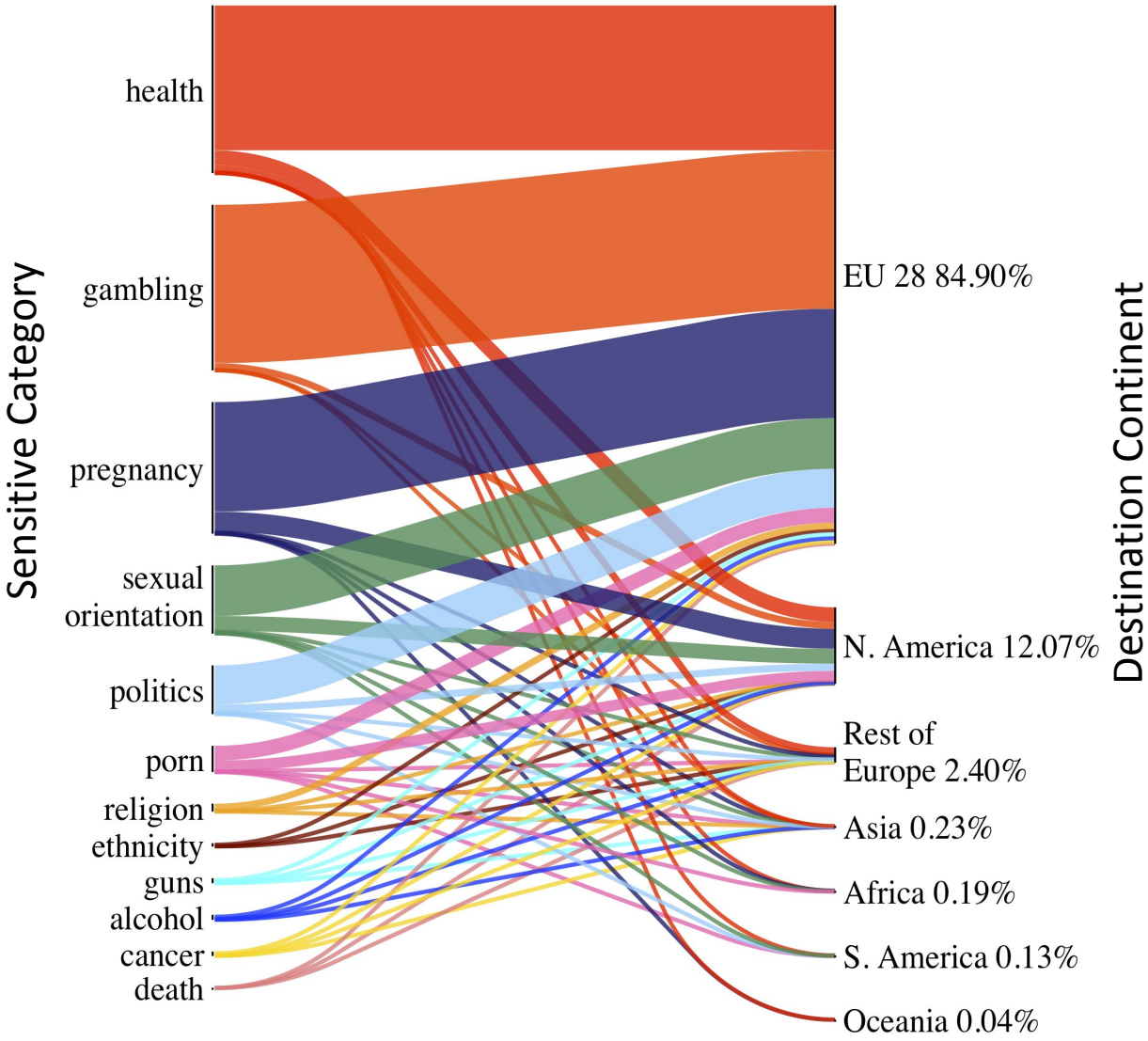


Health



Sexual Orientation

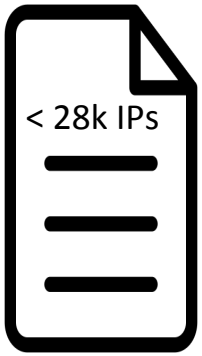
Sensitive websites based on EU 28 users



Scaling up – From real users to ISP flows

Datasets

List of Ad + Tracking IPs



+

ISPs Datasets

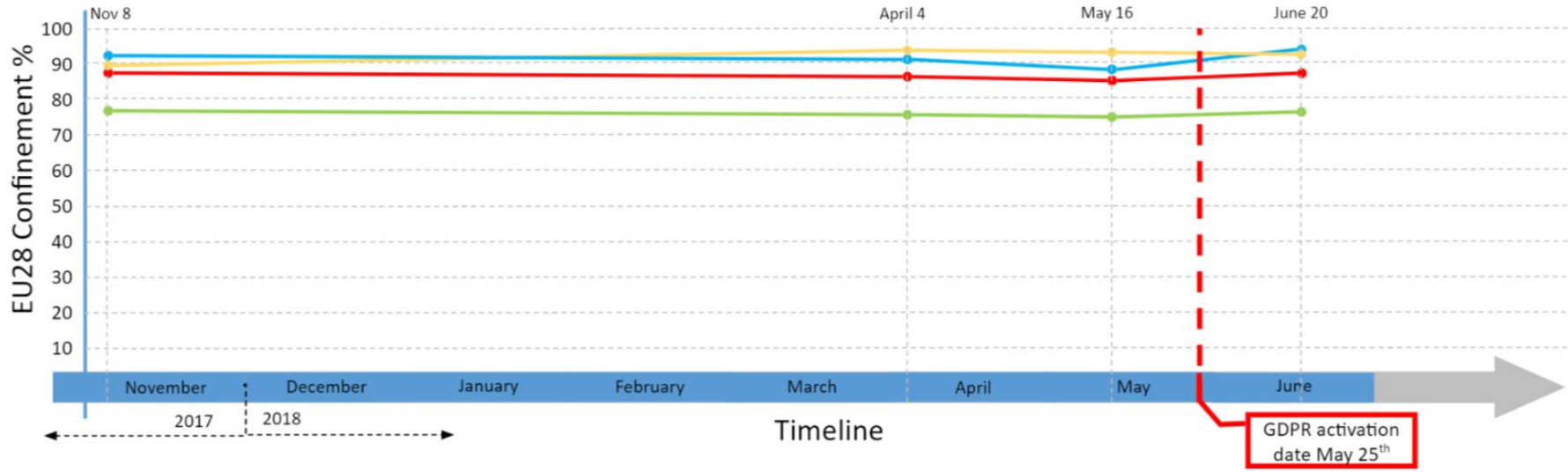
Name	Country	Demographics
DE-Broadband	Germany	15+ Million broadband households
DE-Mobile	Germany	40+ Million mobile users
PL	Poland	11+ Million mobile and broadband users
HU	Hungary	6+ Million mobile and broadband users

<u>Four 24h daily snapshots</u>

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

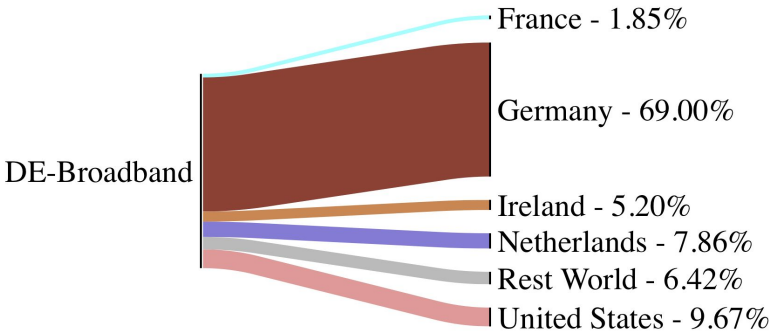
Scaling up – Continent level ISPs results

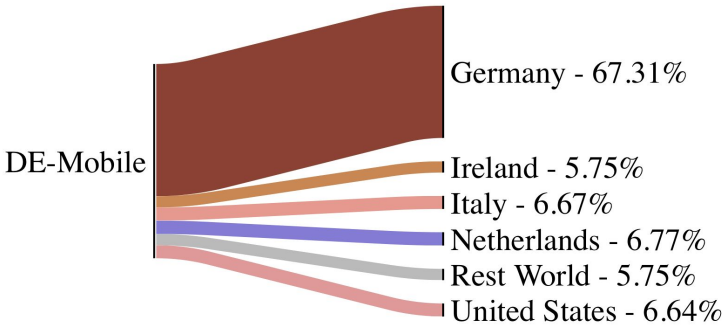
	DE-Broadband				DE-Mobile				PL				HU			
	Nov 8	April 4	May 16	June 20	Nov 8	April 4	May 16	June 20	Nov 8	April 4	May 16	June 20	Nov 8	April 4	May 16	June 20
#Sampled Tracking Flows (in Millions)	1,057.0	1,200.8	1,105.3	963.4	70.4	77.4	70.8	74.5	13.8	13.8	12.4	11.9	43.3	50.2	39.3	33.6
EU28	**88.5%**	**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%

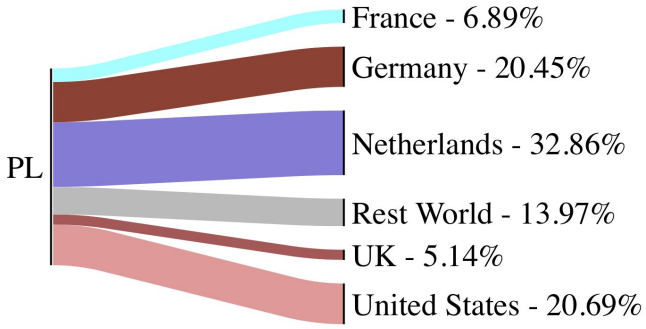


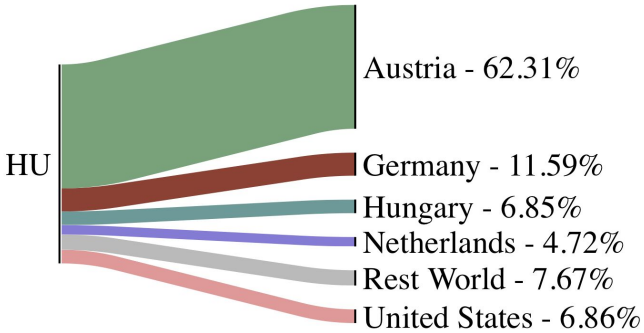
Country level confinements

ISPs dataset at April 4th









Can we further improve localization?

Two approaches:

1. Using DNS optimization

   Group server IPs (locations) based on:

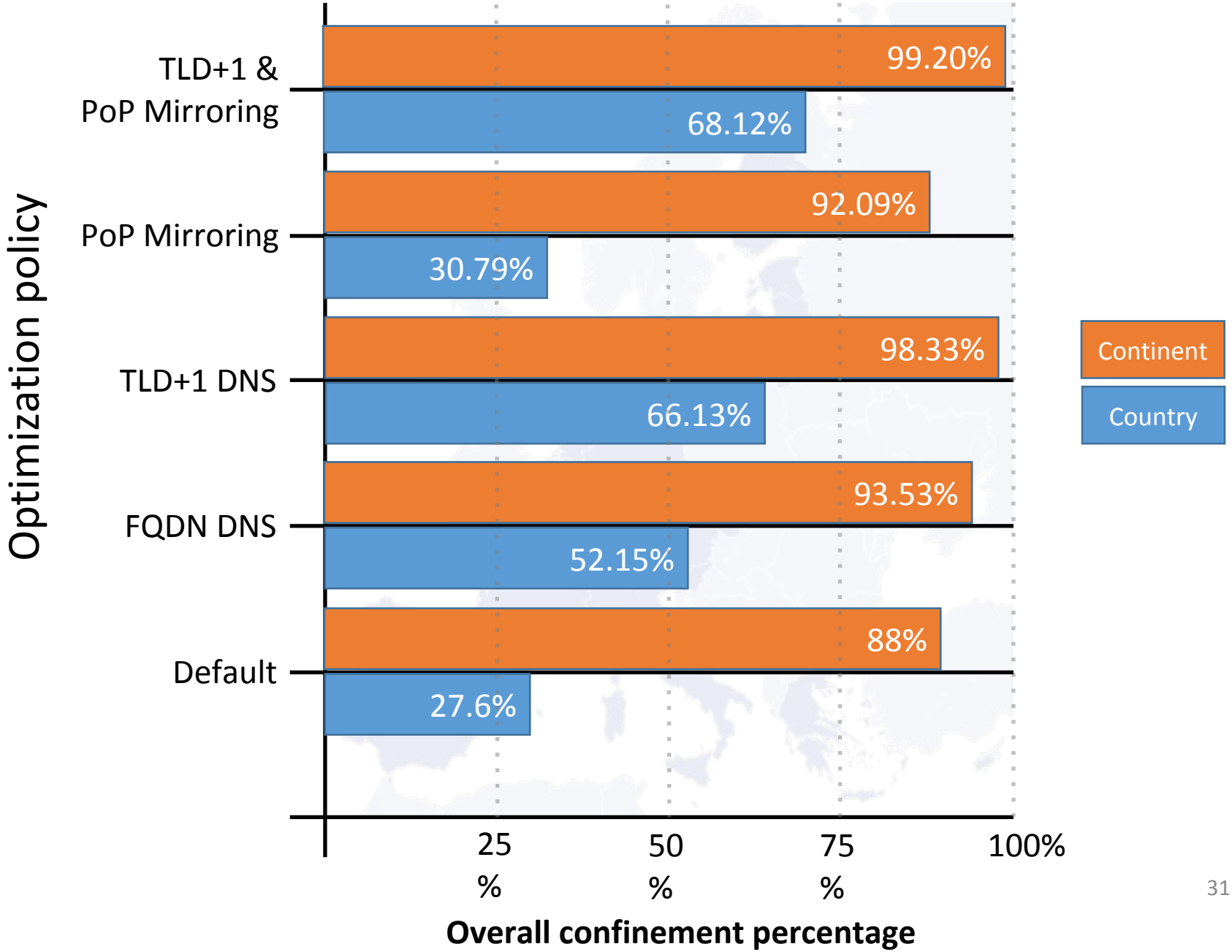
   a) Fully Qualified Domain Names (FQDN) *i.e., sub_d.tracker.com*

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

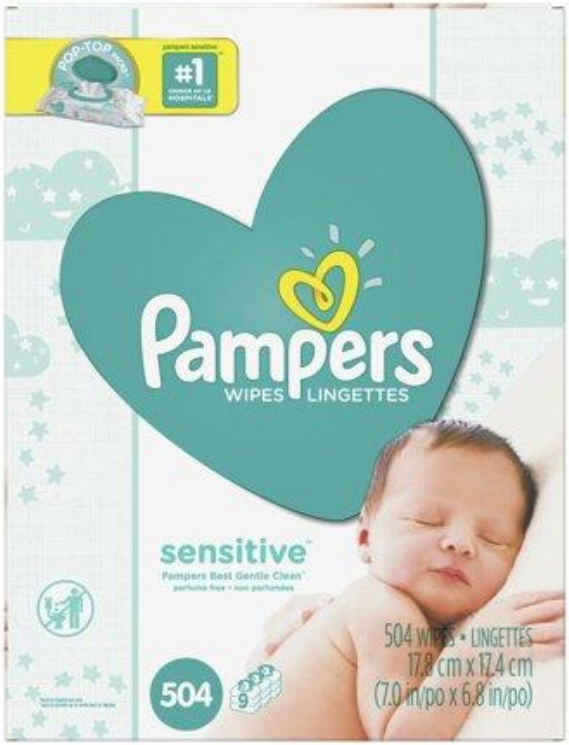
2. Using PoP Mirroring

   Deploy/migrate PoP servers based on cloud services datacenters availability

EU 28 localization improvement



Ongoing work



Who's tracking sensitive domains?

How can we tell if a domain is sensitive?

people don’t care about privacy

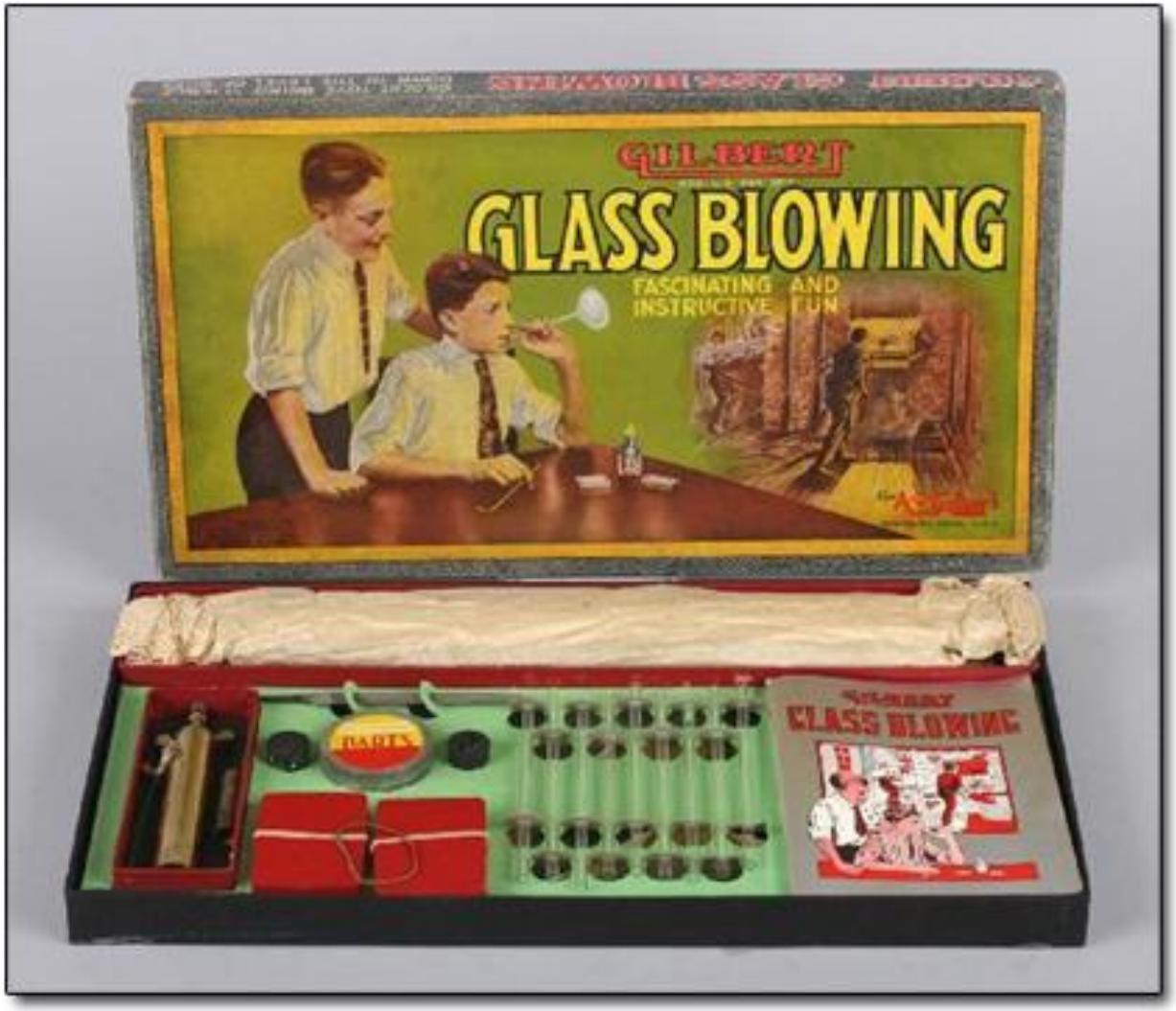
(some say)

some other things people didn’t care about

smoke

flight security

kids playing with melted glass



kids playing with melted iron

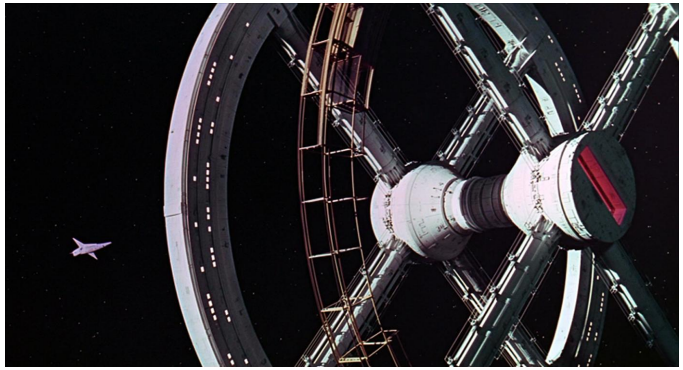
kids playing with power tools

still available at ebay

kerosene train

haven’t located one yet

societies evolve



Won’t be long before we look back and shake our head

Thank you!