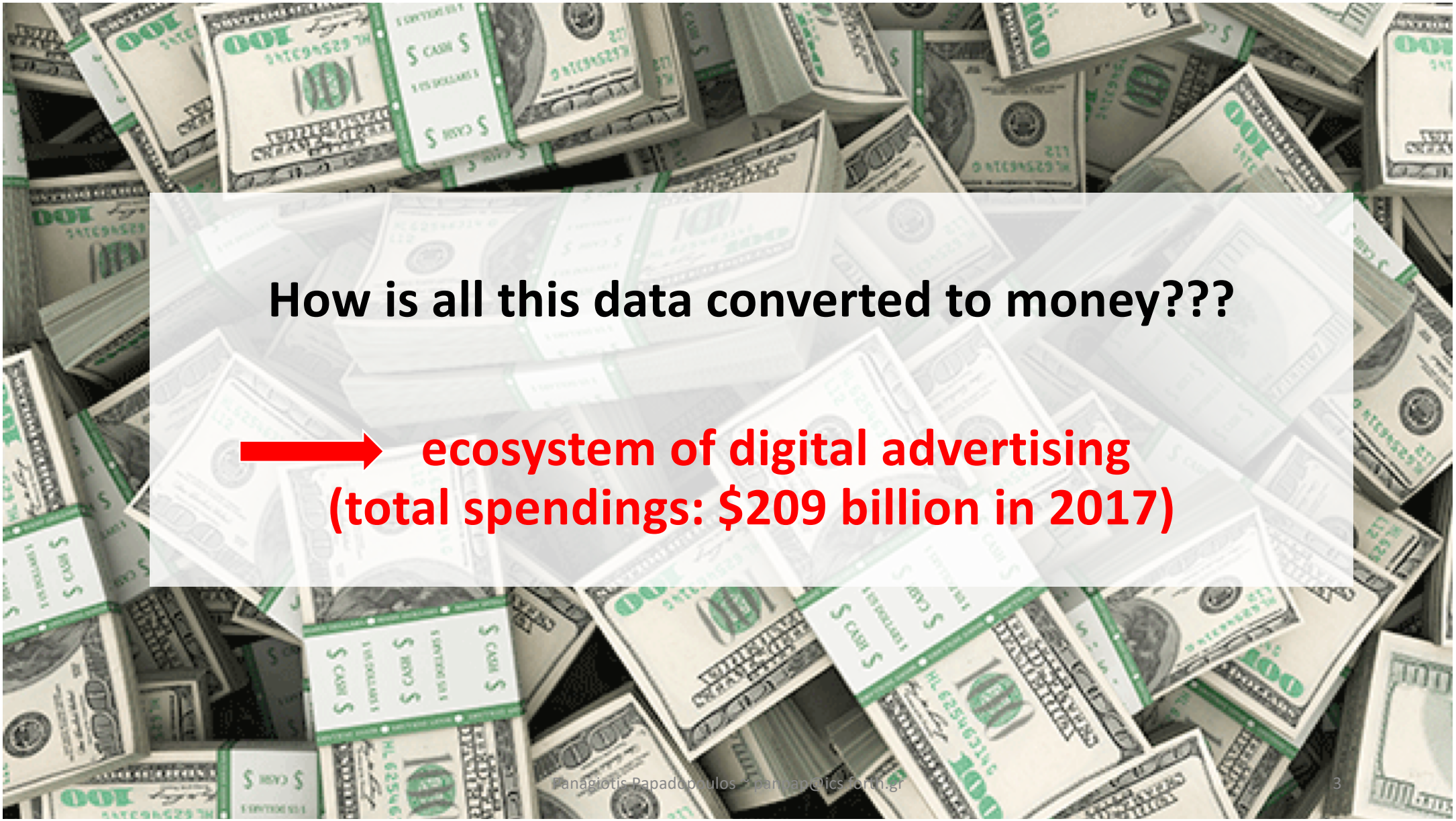
The Advertising Ecosystem and its Impact on the User Privacy

Panagiotis Papadopoulos
University of Crete

Data-driven economy

- The user data of an IT company
  → contribute to its overall market valuation

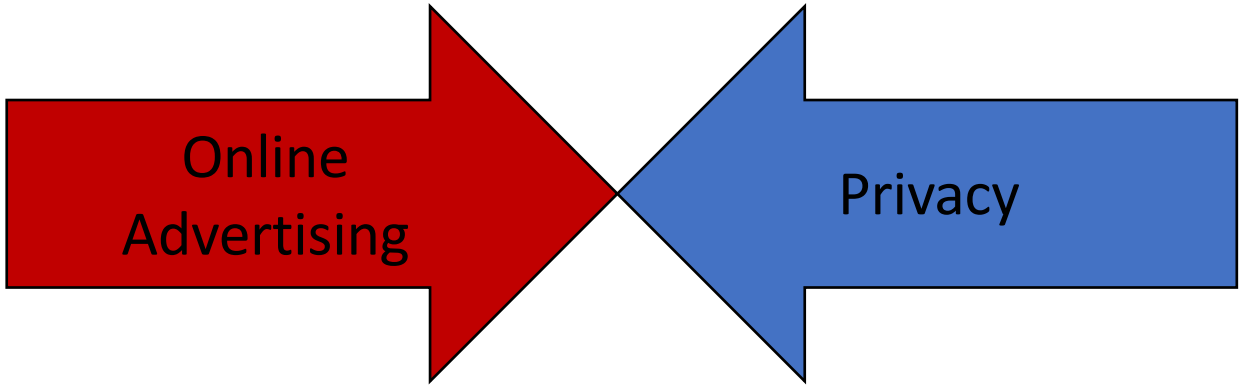
- Companies pursue more and more users personal data
  - By purchasing them
  - By providing free services (Google search, Facebook etc.)



How is all this data converted to money???

→ **ecosystem of digital advertising**
(total spendings: $209 billion in 2017)

Bidirectional effect between Advertising and User Privacy in the online world.



- Ad-ecosystem craves for personal data affecting the user's privacy
- The quality and quantity of these data affect the pricing dynamics of personalized ads.

Advertising ⇾⇽ Privacy

progressively moving towards a programmatic model

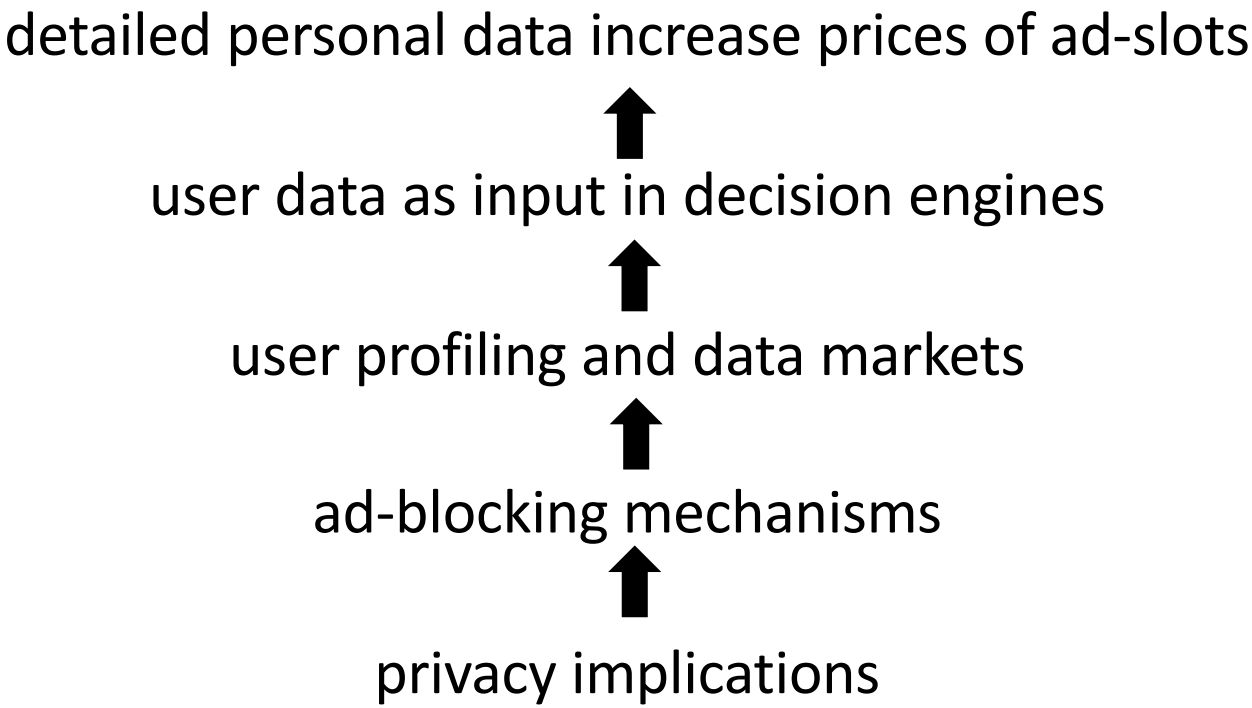
ads are matched to interests of individuals

elaborated user tracking

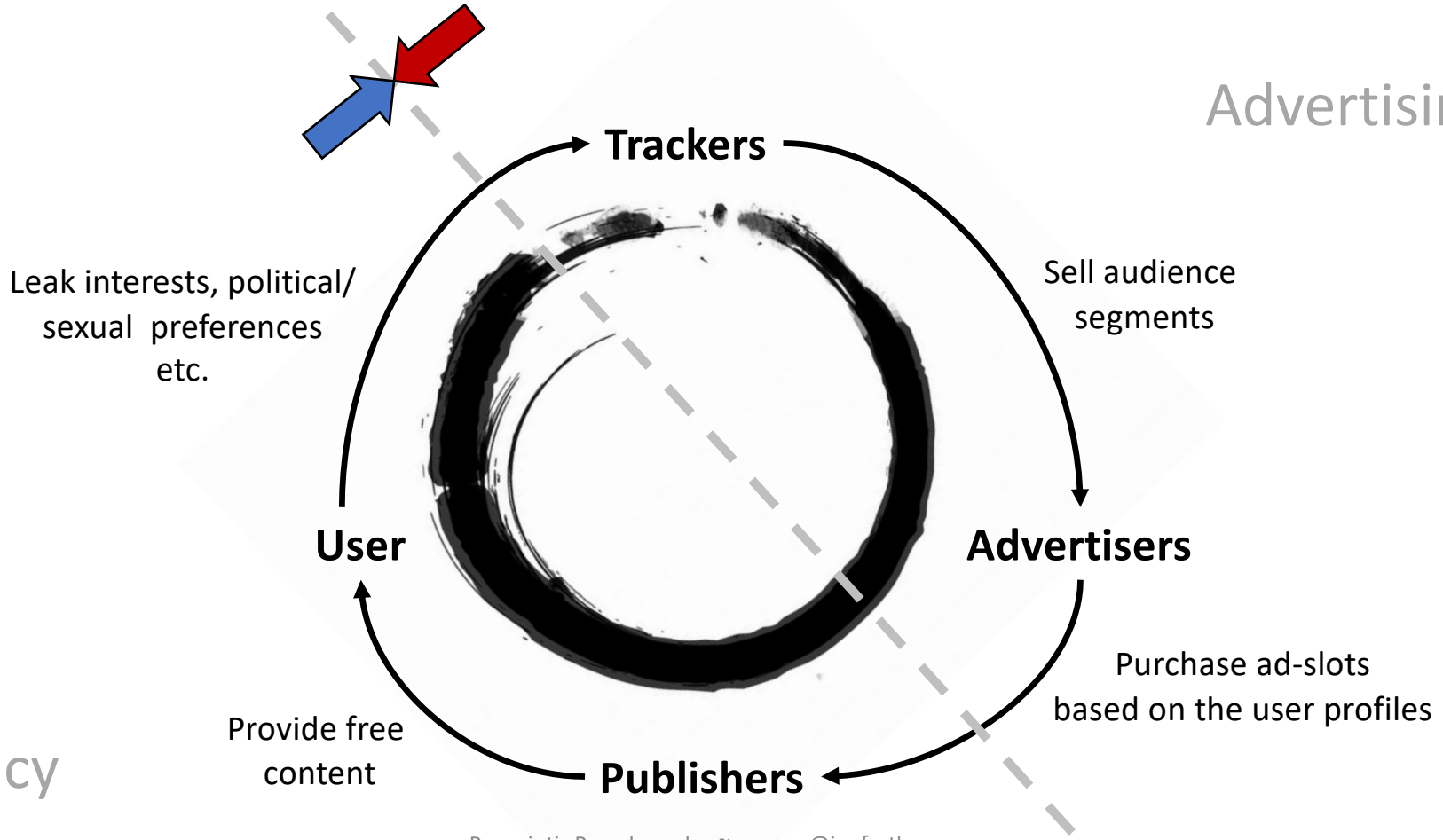
personal data collection

privacy implications

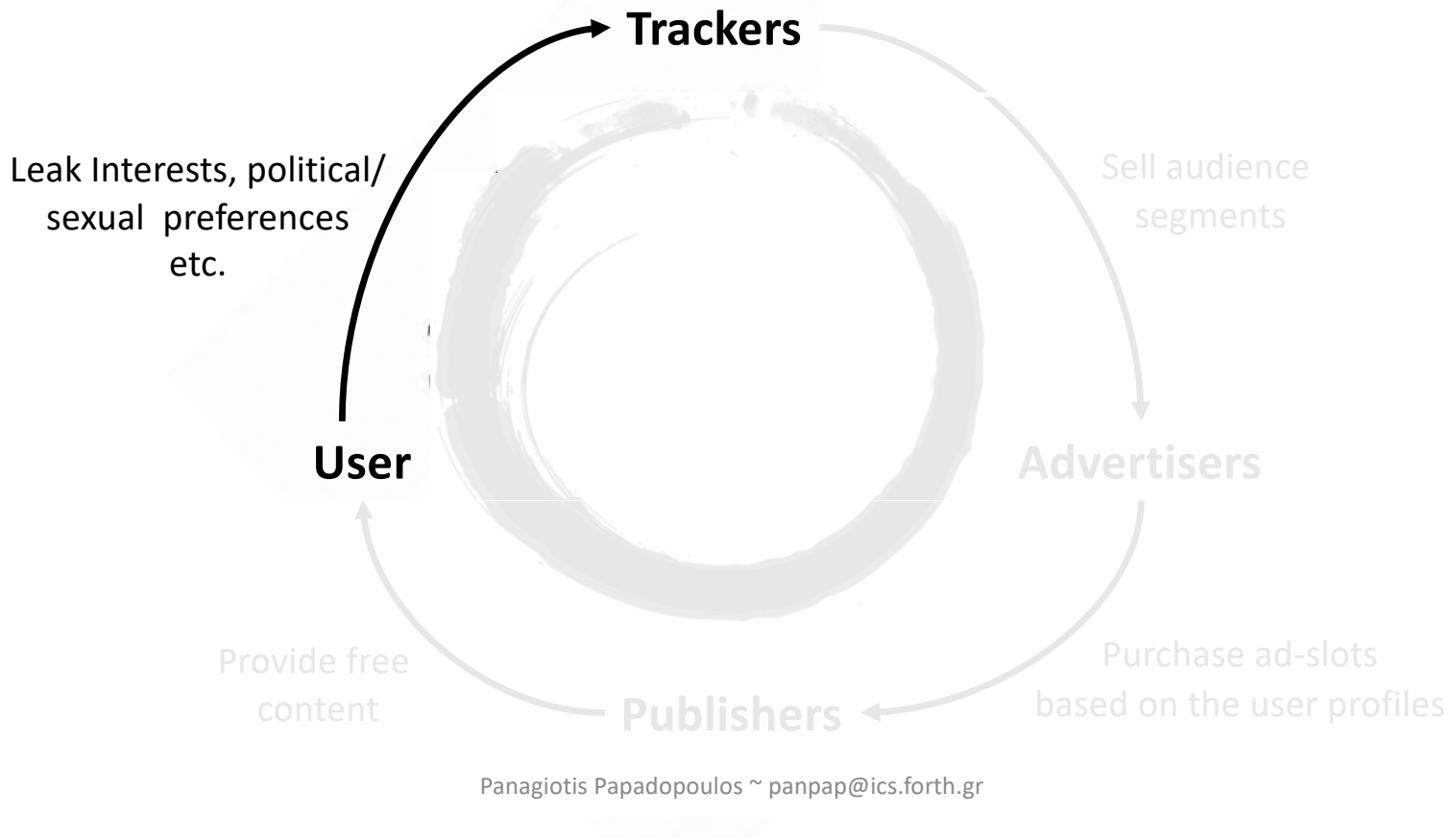
Privacy ⇾⇽ Advertising



The circle of our online life



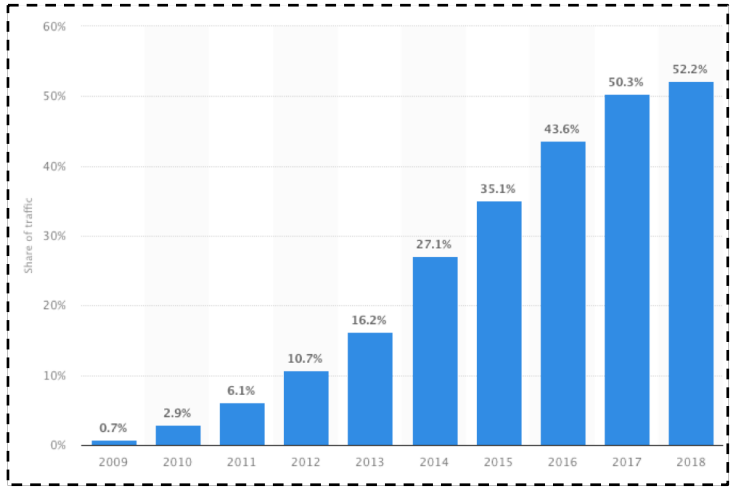
User Tracking



User's attention shifts to mobile devices

1. Mobile devices are *personal* devices. What kind of data are trackers able to exfiltrate from them?

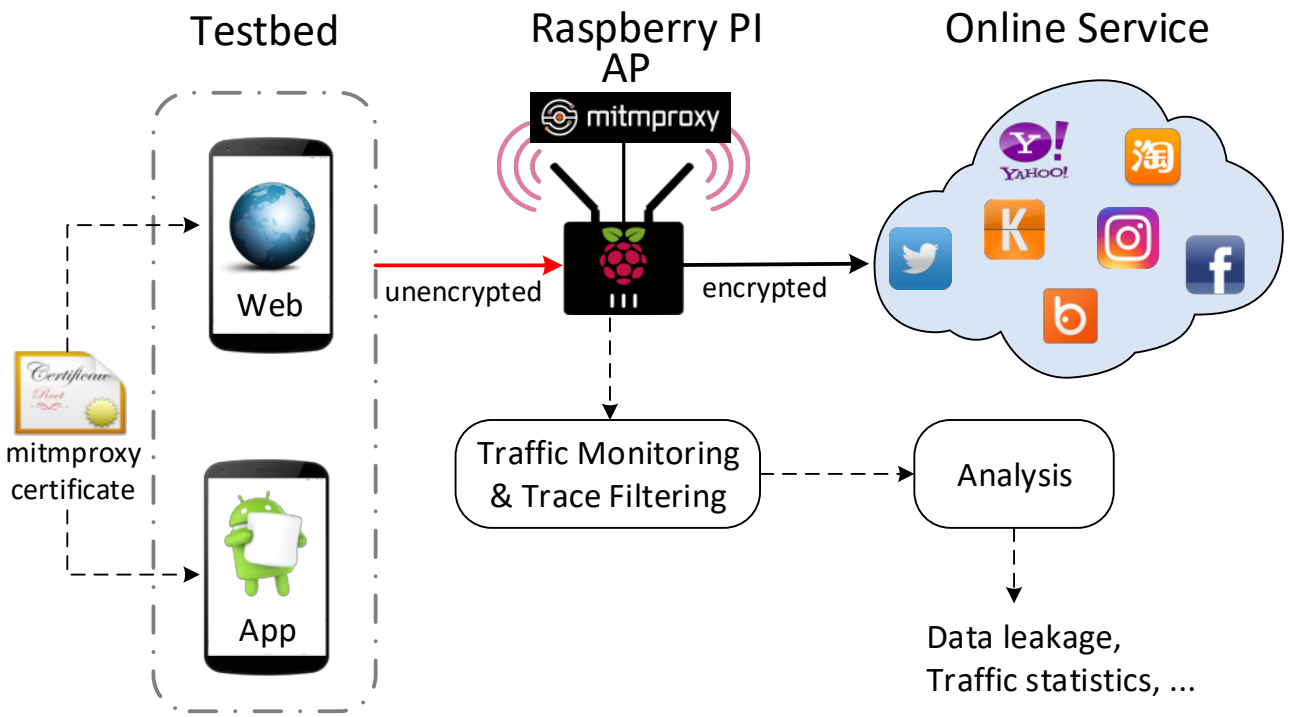
2. How many third-parties have access to these data?



Percentage of all global web pages served to mobile phones from 2009 to 2018, source: Statista Inc

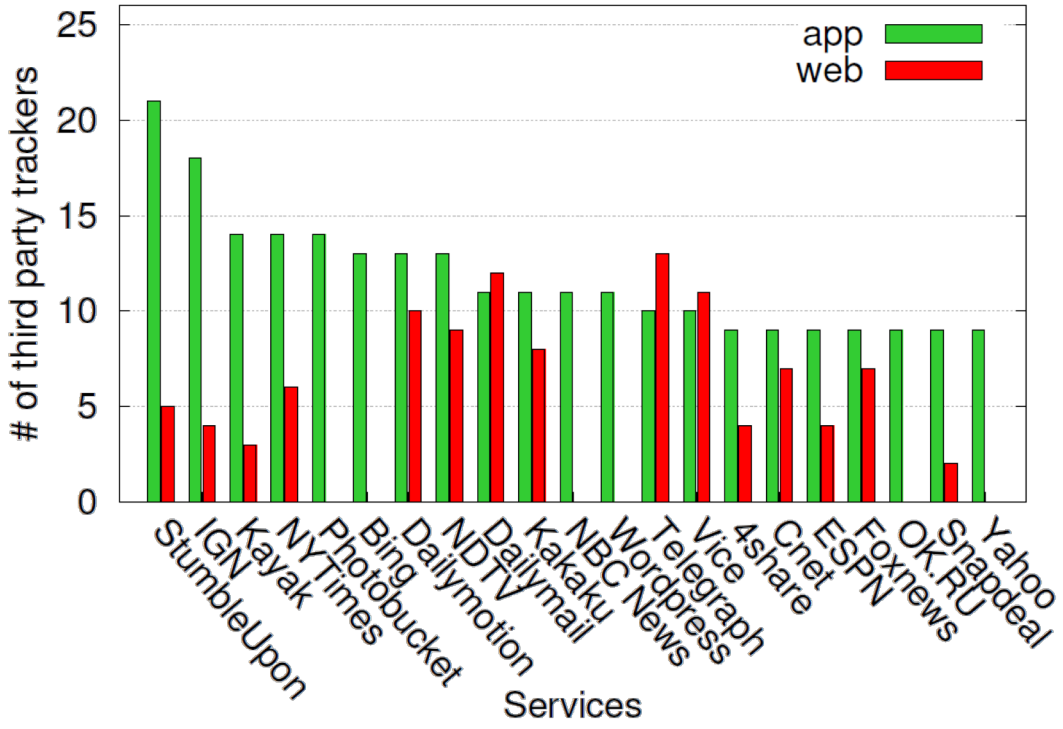
3. Which of the two options (apps or websites) facilitates the most privacy leaks?

Our approach: Privacy Leak Monitoring



- Top 120 Alexa services that allow access through web and app
- Capture traffic: Raspberry Pi → SSL-capable monitoring proxy
- Run each service for 20 mins:
  - through web (mobile browser)
  - through app
- Filter possible leaked identifiers using pattern matching + taint analysis

Third-party Trackers having access to you



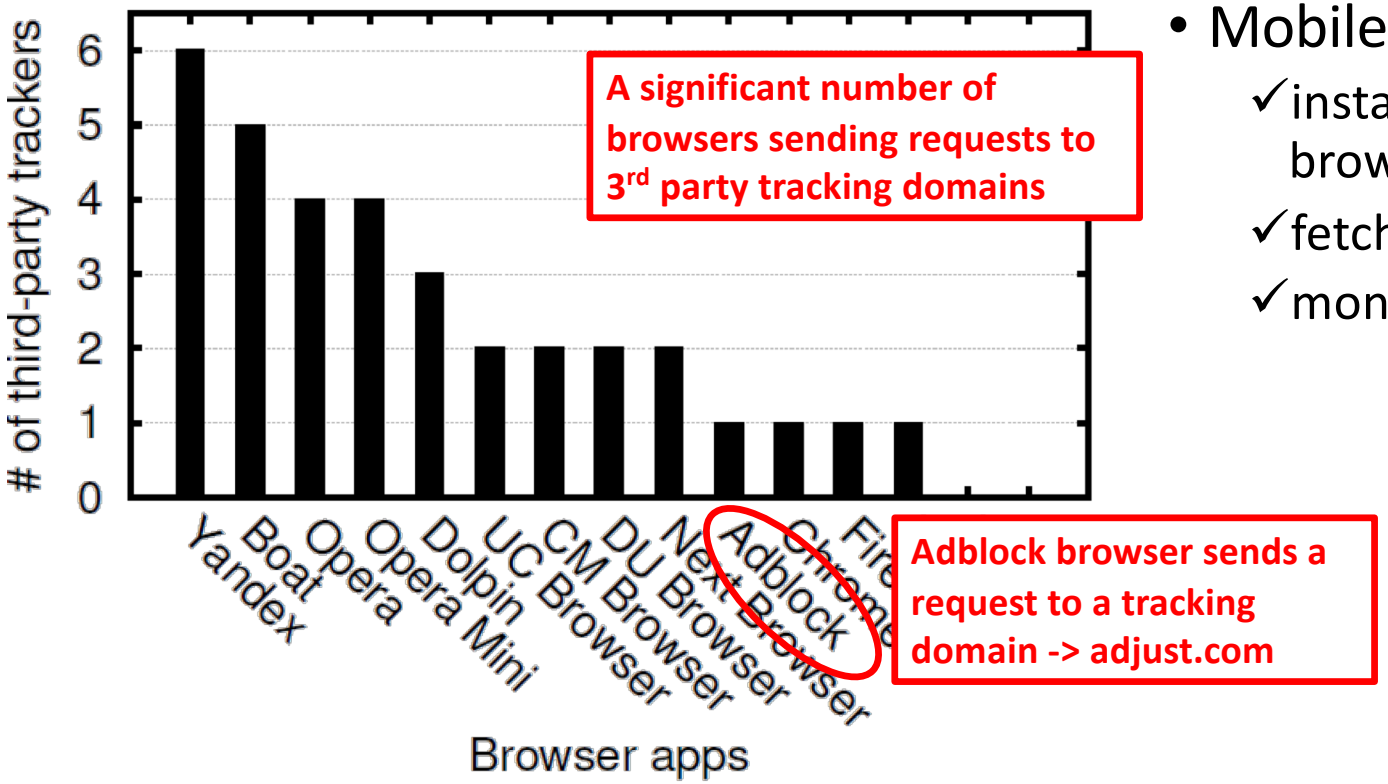
- Apps leak information to an average of **11.7** third-party tracking domains
- Websites leak information to an average of **5** third-party tracking domains

- **94%** of apps and **70%** of websites leak data **to at least one third-party tracker**

Type of Information Leaked: Apps vs Web

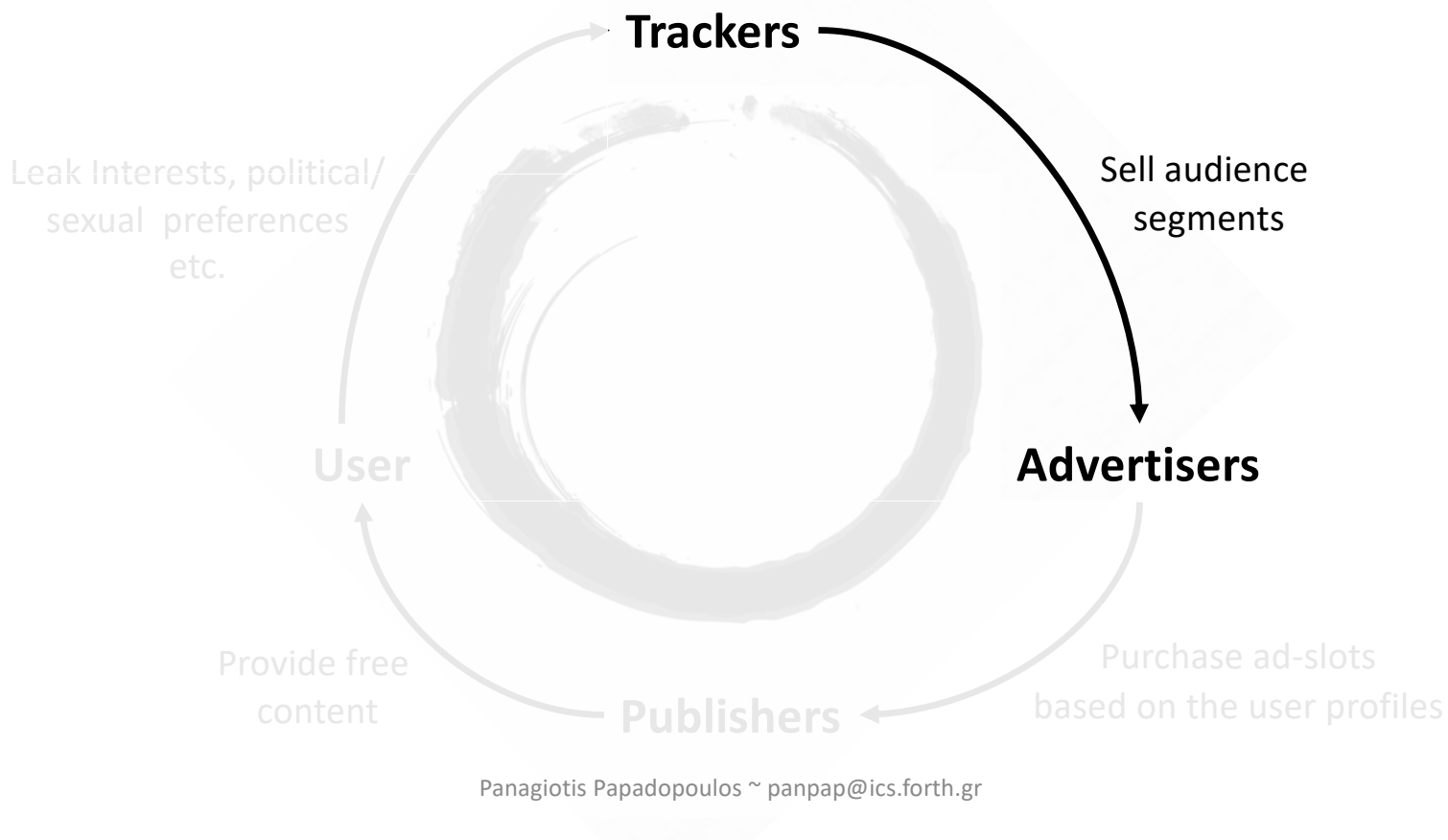
- 4.3% of the apps leak **nearby WiFi APs**
  - ➢ current geolocation and possible interpersonal relations of people being in the same location at the same time
  - ➢ 1 app leaked the entire list of **known APs** → **previous user's locations**
- 3.45% of the apps leak the list of **installed apps**
  - ➢ a tracker to easily infer important information like gender, age, preferences, interests etc.
- 85% of websites leak **GPS coordinates** (Vs 66.38% of apps)

Mobile Browsers leak, too…



- Mobile browsers are apps, too
  - ✓install 15 popular mobile browsers
  - ✓fetch google.com
  - ✓monitor traffic

User Data markets



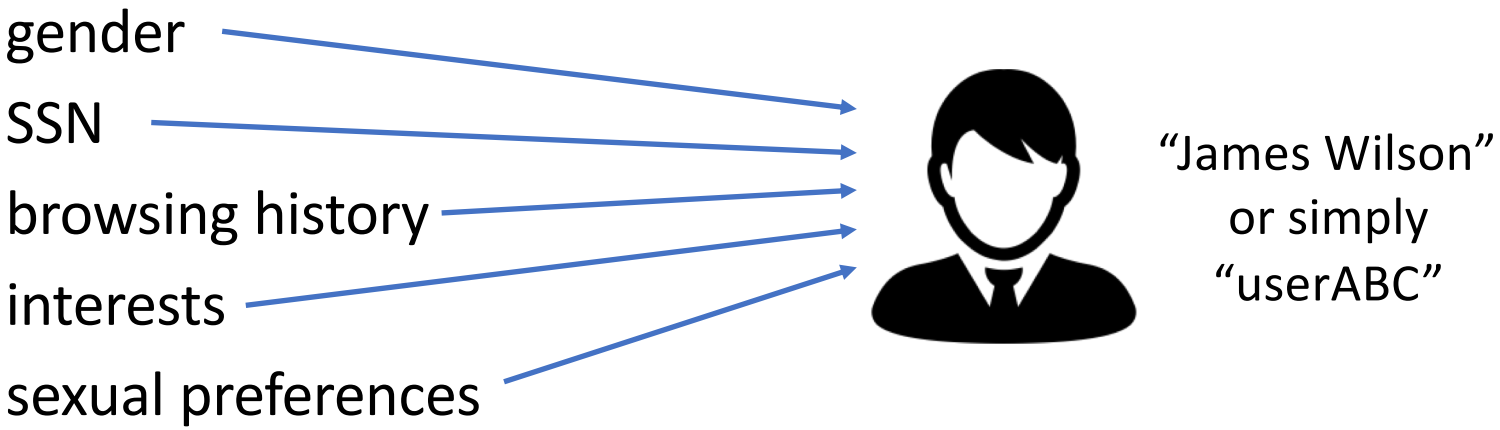
Data Markets

- Trackers and data brokers (e.g., Axciom, Experian) process collected user data and form user profiles

- These user profiles may contain information from both online and offline world
  - ✓ e.g., phone number, city/state, email address, bankruptcy, driving license, education information, SSN, employment details, information on marriage, divorce, property records, etc.
- Profiles are sold in data markets to advertisers and are used as input in programmatic auctions.

Collected data attribution

All this volume of collected data must be attributed to a single ID to make sense:



User Identification

- Cookies the most common user identification method nowadays
  - Upon first visit, domain1 drops a cookie on the user side naming her as "userABC"
  - Work as passports in offline world!
- domain-specific: cookies created by one third-party entity cannot be read by anyone else
  - domain2 cannot see "userABC"

Universal User Identification

So:

Broker sells data for “userABC” or “James Wilson”

- **but** the buyer (e.g., advertiser) does not know any such naming

Entity A knows James by the ID “userABC” and entity B knows the same user as “user123”

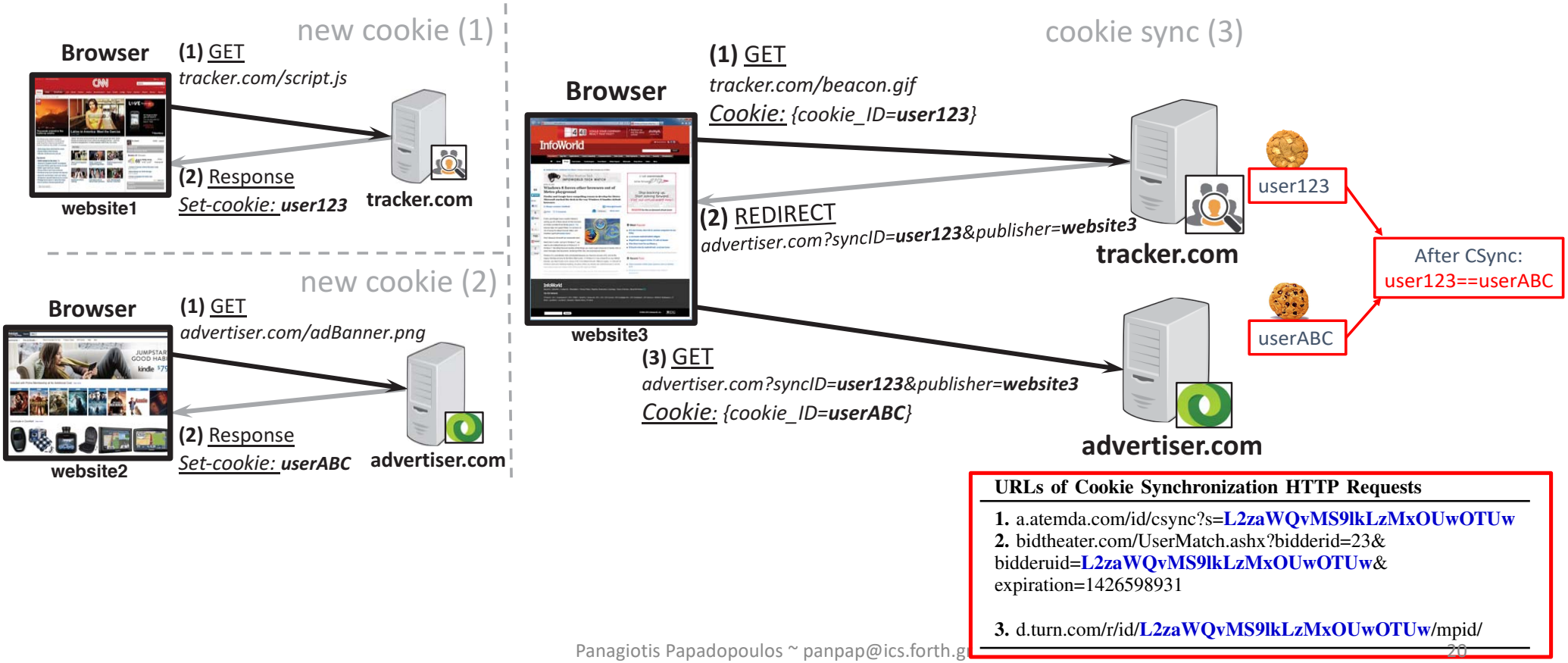
- How the seller and buyer finalize such data merges?
  - i.e., **userABC==user123**

Some universal user identification must appear!

Cookie Synchronization

to the rescue…

Cookie Synchronization



URLs of Cookie Synchronization HTTP Requests

1. a.atemda.com/id/csync?s=**L2zaWQvMS9lkLzMxOUwOTUw**
2. bidtheater.com/UserMatch.ashx?bidderid=23&
bidderuid=**L2zaWQvMS9lkLzMxOUwOTUw**&
expiration=1426598931

3. d.turn.com/r/id/**L2zaWQvMS9lkLzMxOUwOTUw**/mpid/

Example of a userID getting synced between different entities.

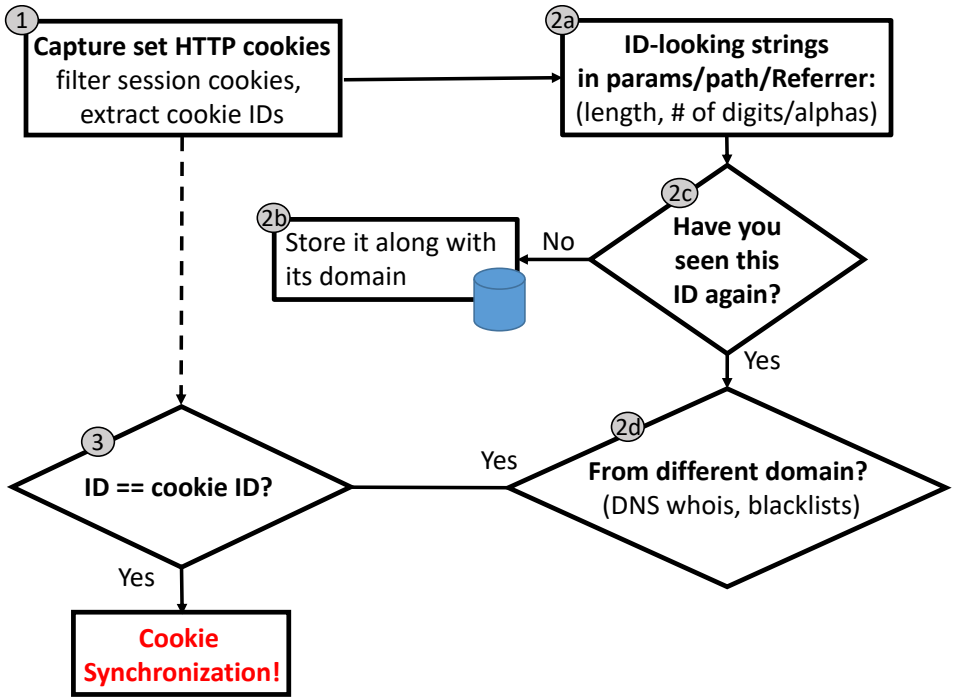
Privacy implications for users

advertiser.com learns that:

1. what it knew as “userABC” is also “user123”
   - Reduction of user aliases -> loss of anonymity
2. user has just visited website website3.com
3. server-to-server user data merges
   - merge data known for “userABC” and “user123” into a single profile
4. coupled with evercookie, or user fingerprinting, CSync allows re-identification of users even after they delete their cookies
   - It needs just one entity to respawn its cookie and through sync all other entities can re-link the two user profiles (before and after cookie erasure)
   - users are not able to abolish their assigned userIDs

Studying Cookie Synchronization in the wild

- Heuristic-based CSync detection
- 179M HTTP requests from 850 volunteering users across 2016
- web traffic redirection through a set of proxies

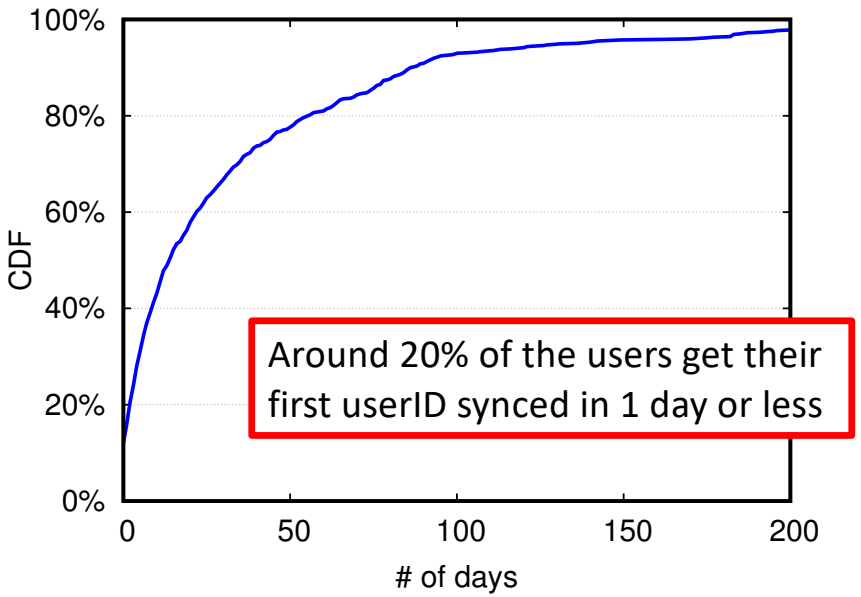


Heuristics-based Cookie Synchronization detection mechanism.

Analysis Results

- 97% of users with regular activity on the web (>10 HTTP requests per day) affected

- 63 cases of domains (after sync), set cookies using same userIDs
  - baidu.com sets cookie baiduid = {idA}
  - after a while different domains set their own cookie again using baiduid = {idA}
  - good job SOP… ☺



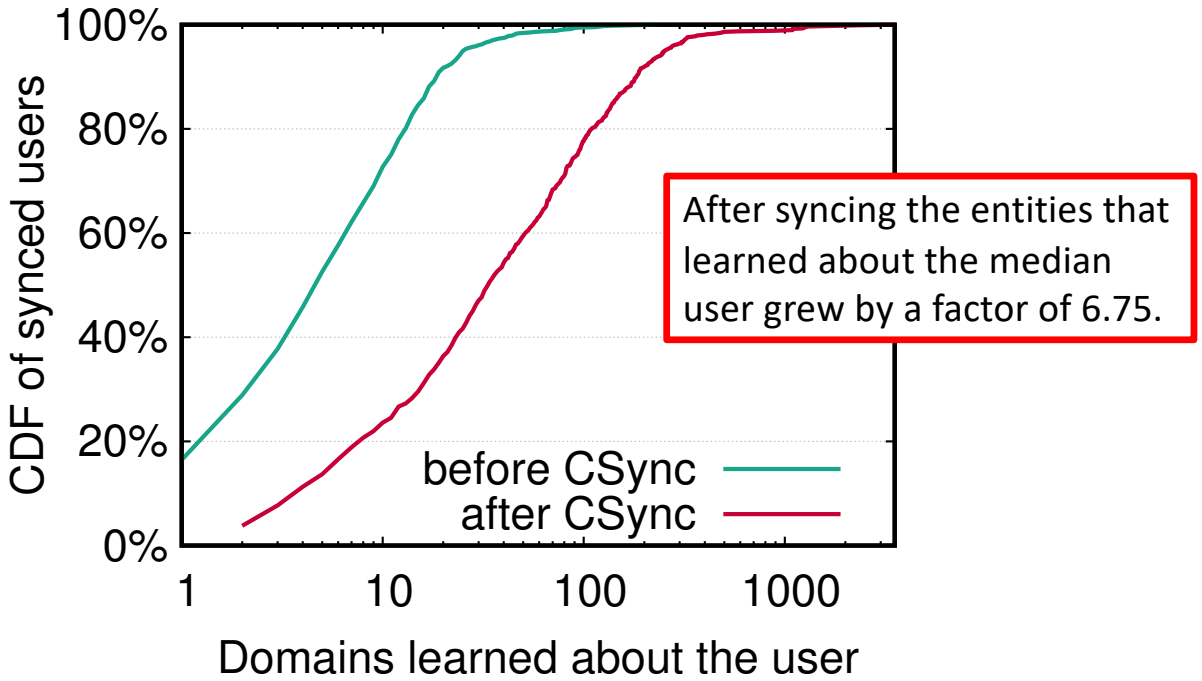
Distribution of the time it takes for the first CSync to appear per user.

Summaries of CSync results

ID Summary stored in cookie by adap.tv
"key=**valueclickinc**:value=708b532c-5128-4b00-a4f2-2b1fac03de81:expiresat=wed apr 01 15:03:42 pdt 2015,key=**mediamathinc**:value=60e05435-9357-4b00-8135-273a46820ef2:expiresat=thu mar 19 01:09:47 pst 2015,key=**turn**:value=268483050575917034 5:expiresat=fri mar 06 16:43:34 pst 2015,key=**rocketfuelinc**:value=639511 149771413484:expiresat=sun mar 29 15:43:36 pst 2015"

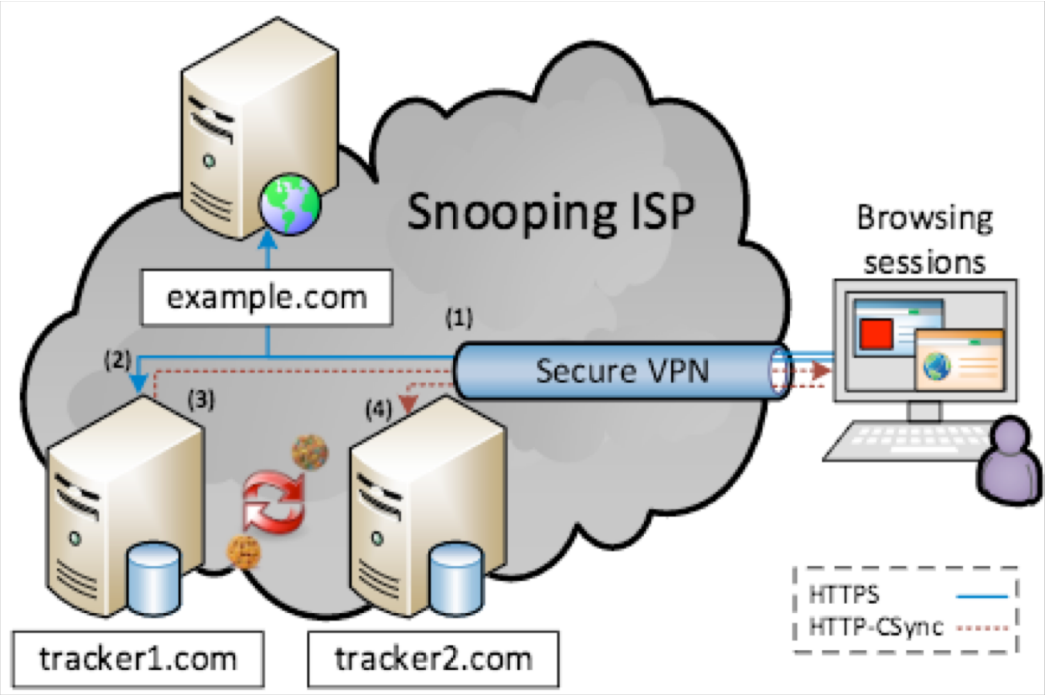
Example of an *ID Summary* stored in a cookie on the user's browser. It includes (previously synced) userIDs and expiration dates assigned by 4 different domains.

Diffusion of Privacy



Distribution of the number of entities learned at least one userIDs of the user (i) with and (ii) without the effect of Cookie Synchronization.

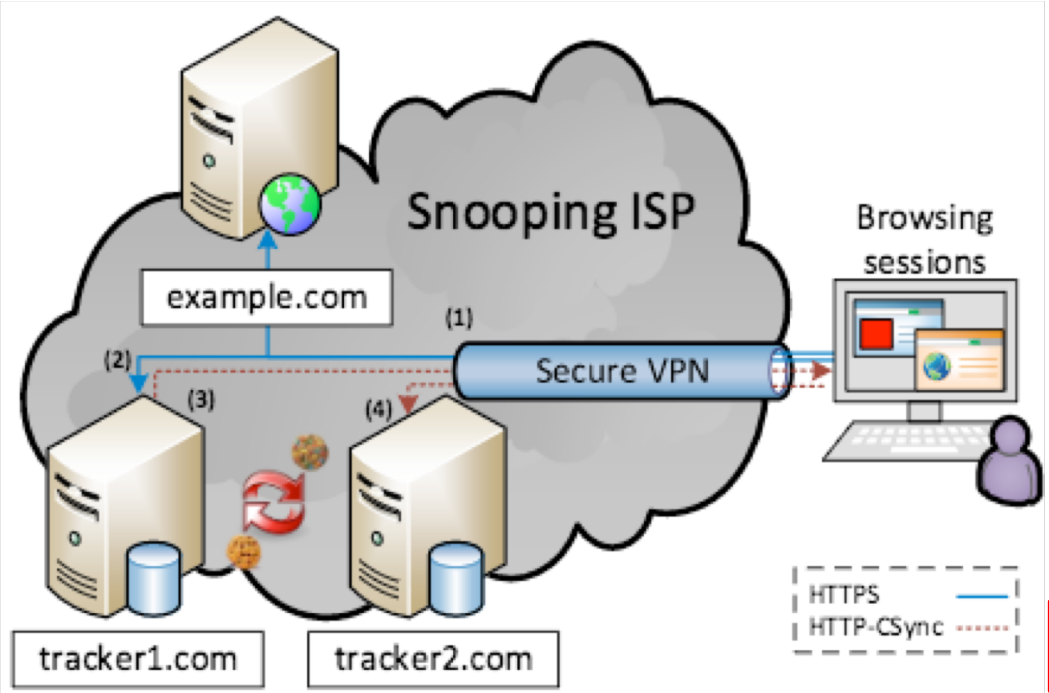
Spilling userIDs out of TLS (1/2)



1. User visits **https:**//example.com over VPN.

2. example.com is ad-supported collaborating with **https**://tracker1.com:
   - tracker1.com provides audience segments for personalized advertising
   - tracker1.com sets a cookie (*user123*) on the user-side

Spilling userIDs out of TLS (2/2)



3. tracker1.com redirects user to **http**://tracker2.com:
   - piggybacks its cookie in location URL (*user123*)
   - allows tracker2 to read (or set) its own cookie (*userABC*)

3xx Redirect request Headers
Location: tracker2.com?syncID=**user123**&partner=tracker1
Referrer: example.com

Response Headers
Set-Cookie: cookie_ID=**userABC**

(1) ID-spilling: **userABC==user123**

(2) browsing history leak: **user123** just visited example.com

Whenever ISP sees request to tracker2.com from *user123* it will reidentify the user who visited example.com

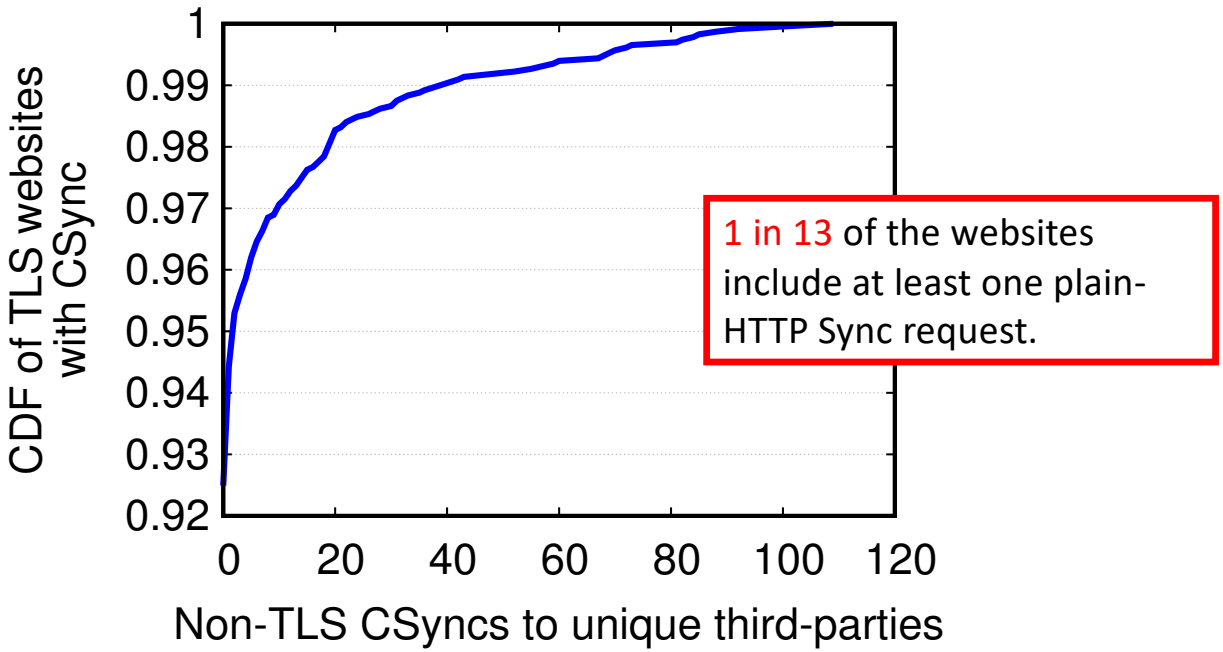
A real world leaking example

Role	Domain
Visited website:	https://financialexpress.com
Cookie setter: SetCookie:	https://tapad.com **D0821FA0-8A80-4D9E-BC85-C40EAC4E4FF5**
Cookie syncer:	http://delivery.swid.switchadhub.com/adserver/user_sync.php?SWID=cf43265166a9ccf5f6fd0472f23776fa&sKey=PM2&sVal=**D0821FA0-8A80-4D9E-BC85-C40EAC4E4FF5** referrer: **financialexpress.com** Get-cookie: {cf43265166a9ccf5f6fd0472f23776fa}
Cookie syncer:	http://tags.bluekai.com/site/3096?id=**D0821FA0-8A80-4D9E-BC85-C40EAC4E4FF5** referrer: **financialexpress.com** Get-cookie: {c57b29d1-f8e2-11e7-ac1b-0242ac110005}

The synced ID links together all consecutive set cookies

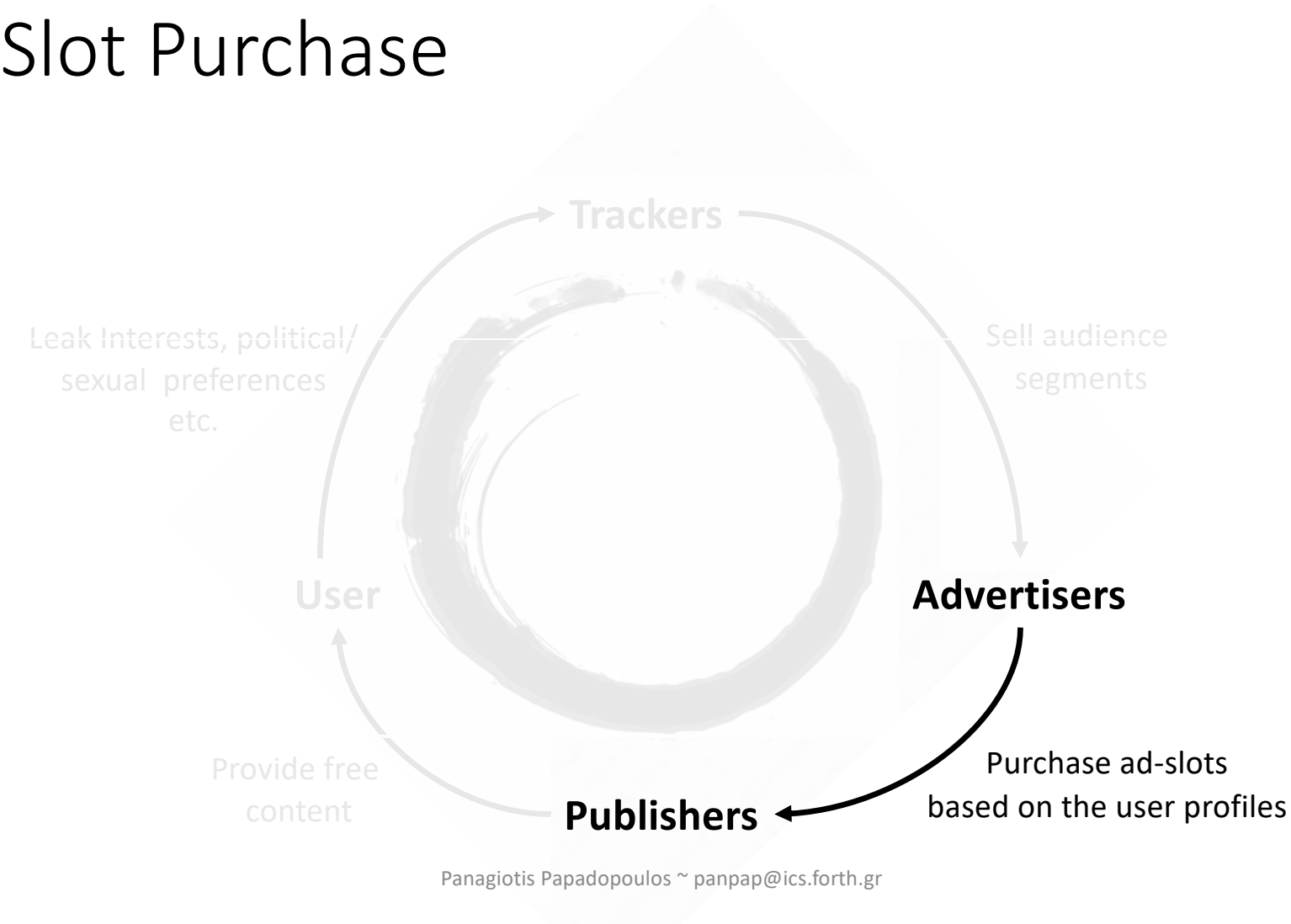
Track these 2 cookie IDs and you know who this user is

UserID spilling in Alexa top 12K

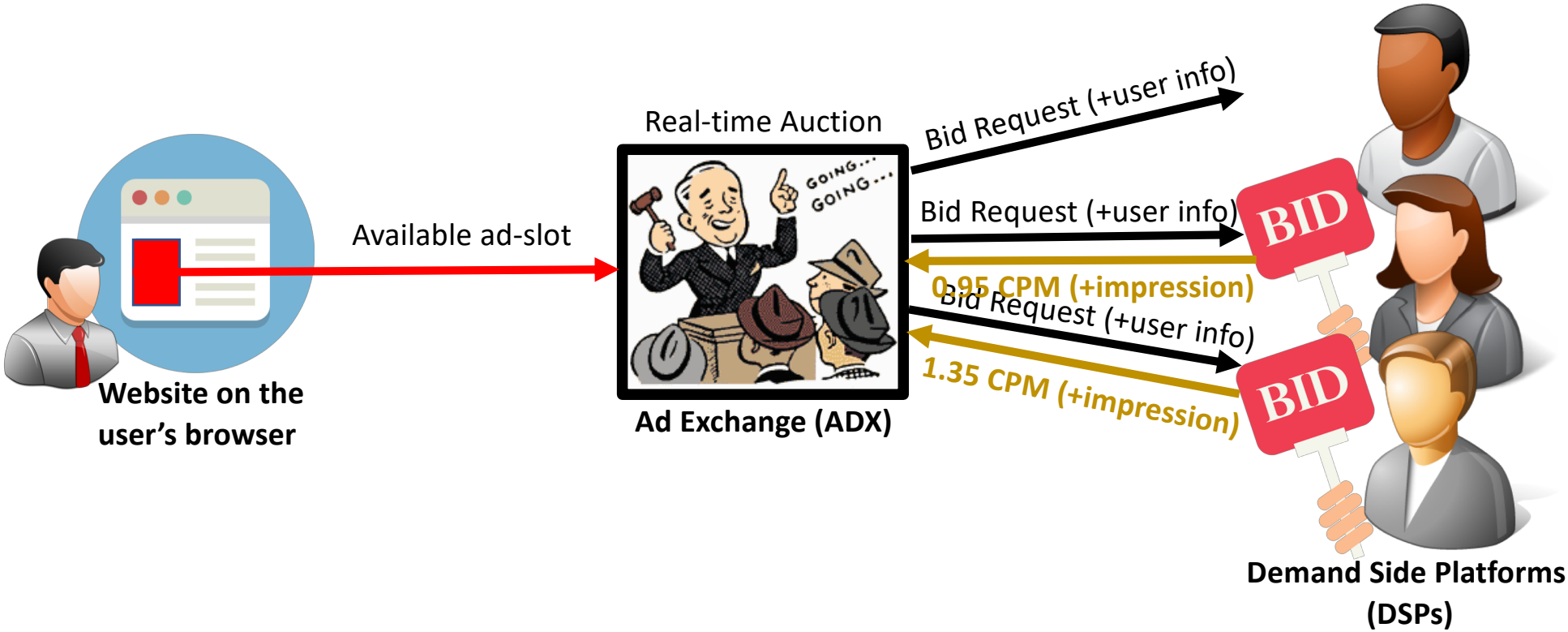


Distribution of non-TLS sync requests per TLS website.

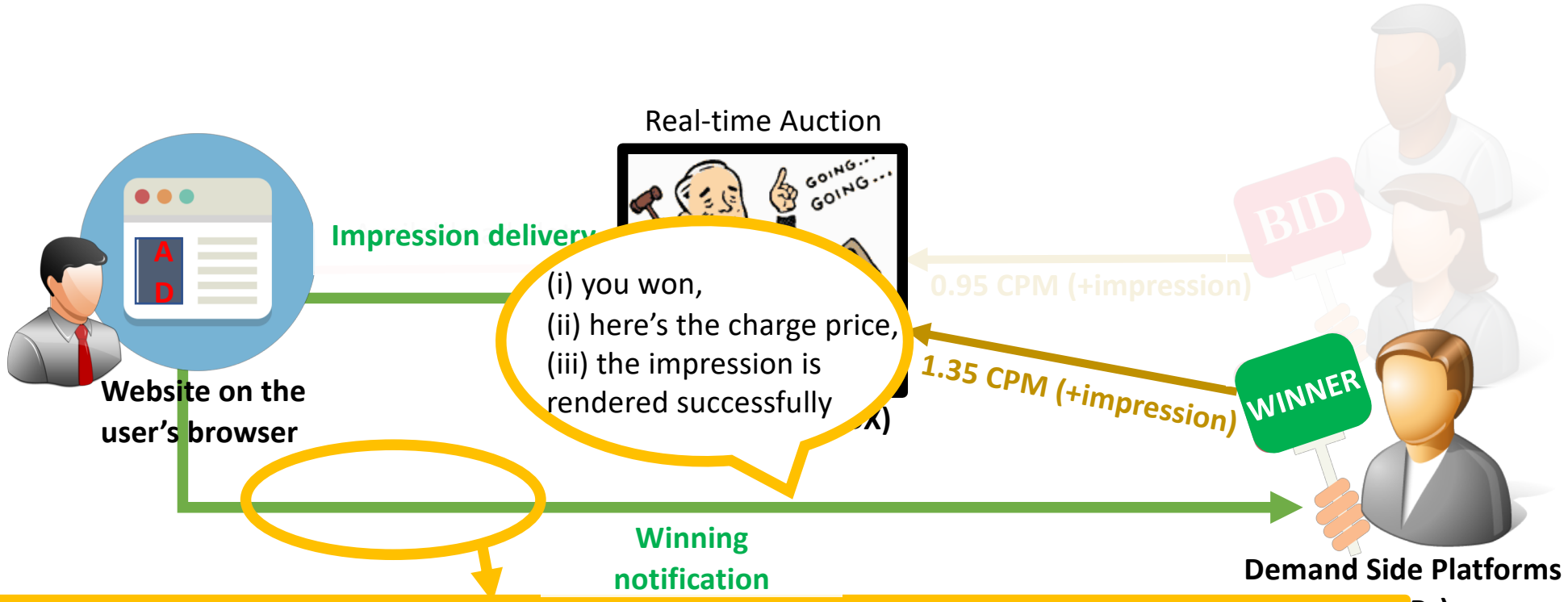
Ad Slot Purchase



Programmatic auctions of 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=..

What about some Transparency?

- How do the collected user data affect the pricing dynamics?
- Are all users valued equally by advertisers?
- What is the average money advertisers pay to reach the average user?

Measuring cost of advertisers per *individual*

1. Leverage Real-Time Bidding (RTB) protocol:
    1. 74% of programmatically purchased advertising
    2. $8.7 billion in 2016 only in US

2. Methodology to calculate **at real time** the overall value advertisers pay per ***individual*** user based on her leaked information.

3. Year-long dataset (2015) of 1600 volunteering mobile users + 2 real probing ad campaigns

Challenge: Encrypted prices on the rise

- Charge prices in nURLs tend to be encrypted
  - Encryption is a regular practice in desktop RTB auctions (~68%)
  - Lower but rapidly increasing in mobile RTB auctions (~30%)

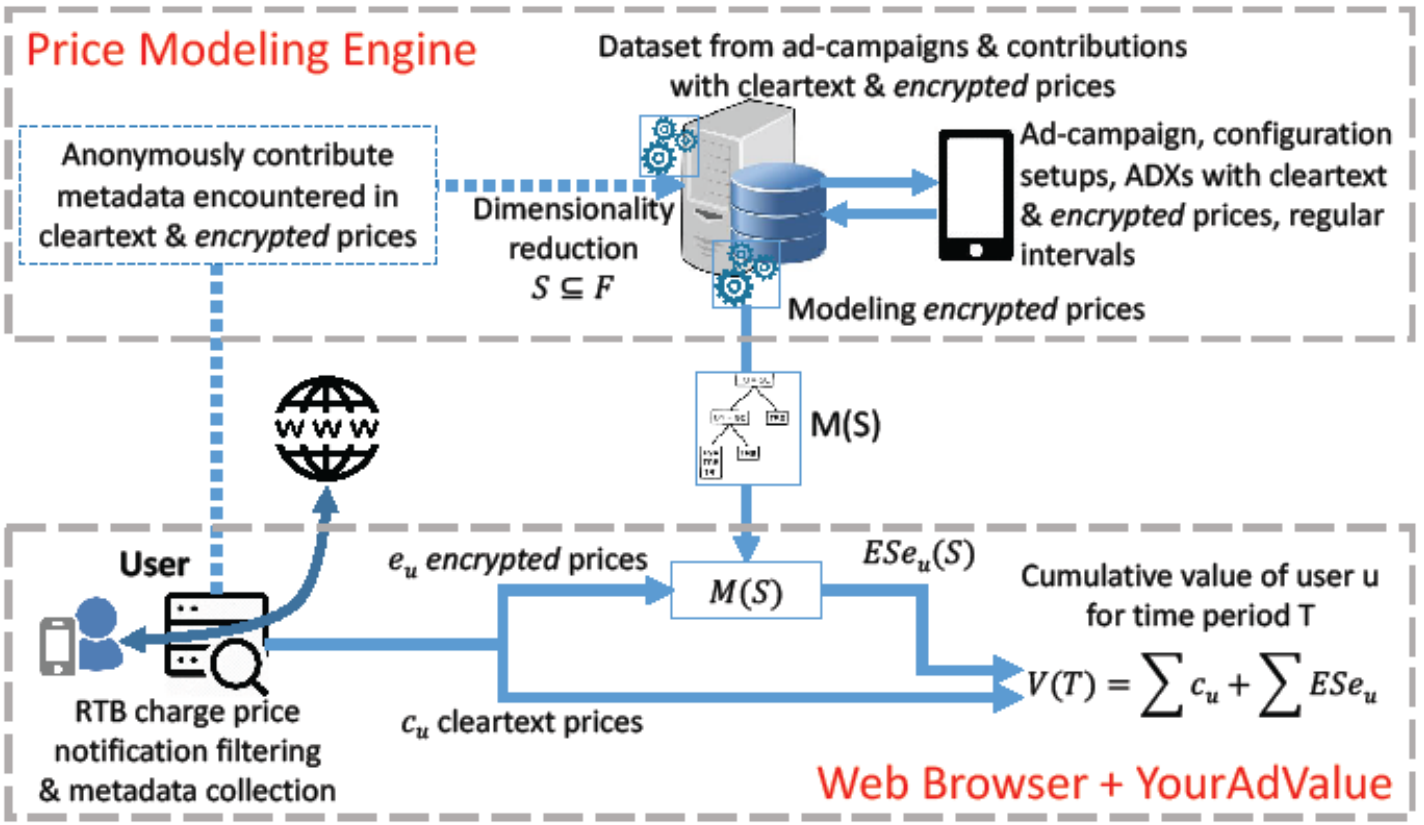
Cumulative value of user u
for time period T

$$V(T) = \sum c_u + \sum ESe_u$$

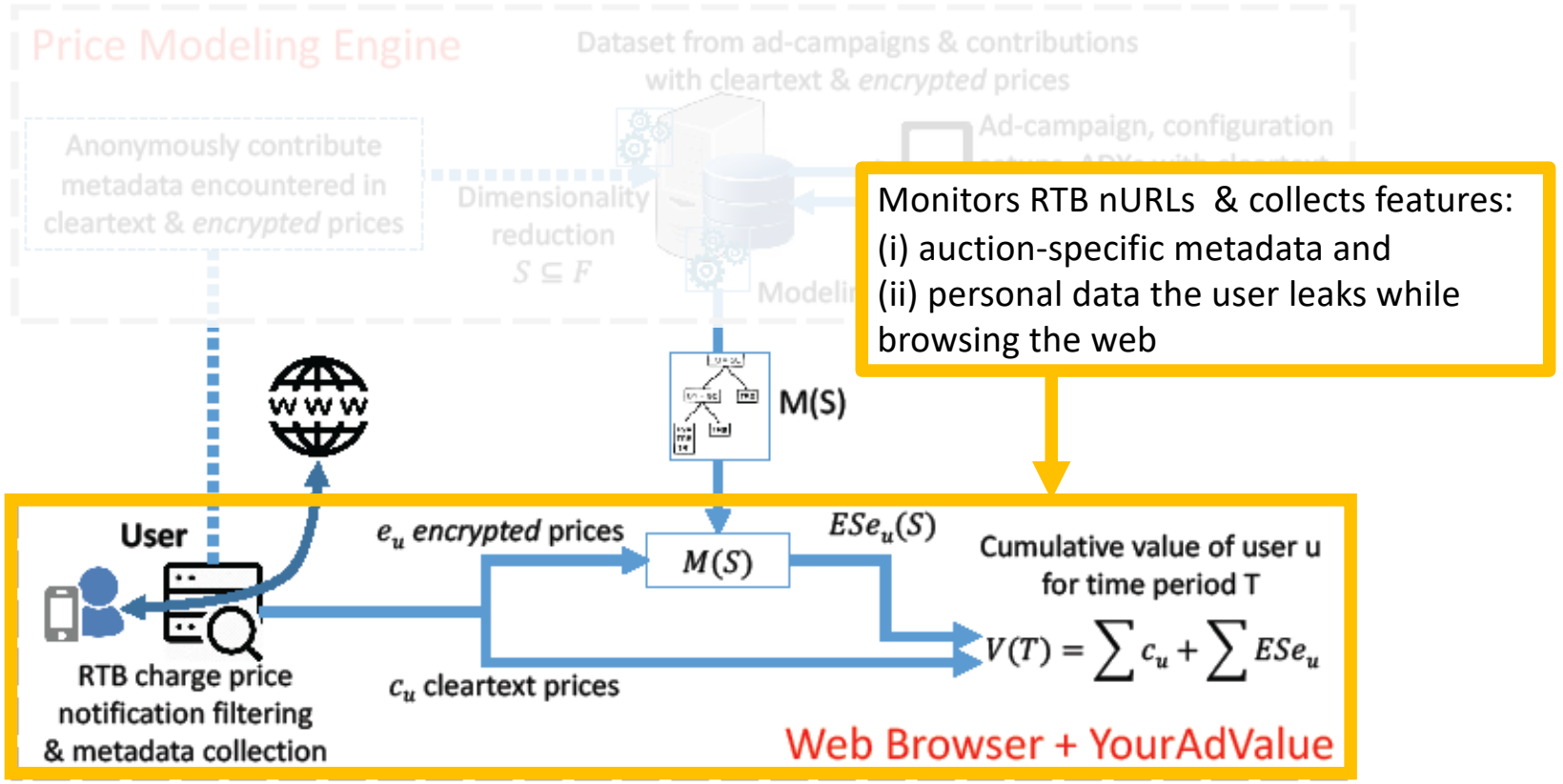
Encrypted

Previous work [Olejnik, 2013] assumes encrypted prices follow the same distribution as cleartext. **But is that so?**

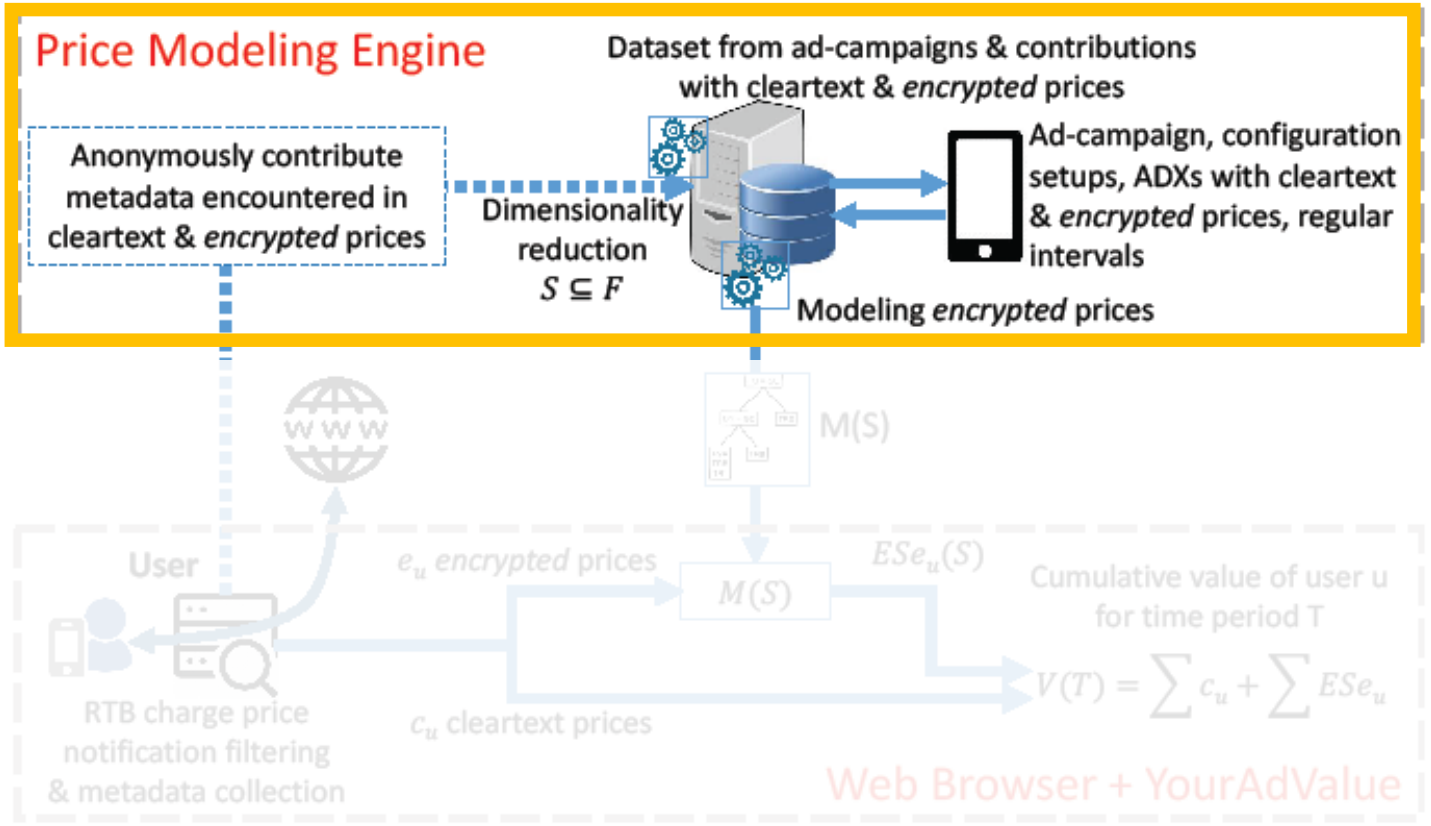
Methodology



YourAdValue browser extension



Price Modeling Engine (PME)



Features affecting charge prices

- user location (at city level) affects the median charge prices
- Android devices are more popular, but iOS-based devices draw higher prices
- during the day median charge prices are of similar range
  -> early morning hours - noon: more charge prices with increased values
- the user's interests severely affect the ad prices
  -> (e.g., users interested in "Business & Marketing" are more expensive than users interested in "Science")

Real probing ad-campaigns

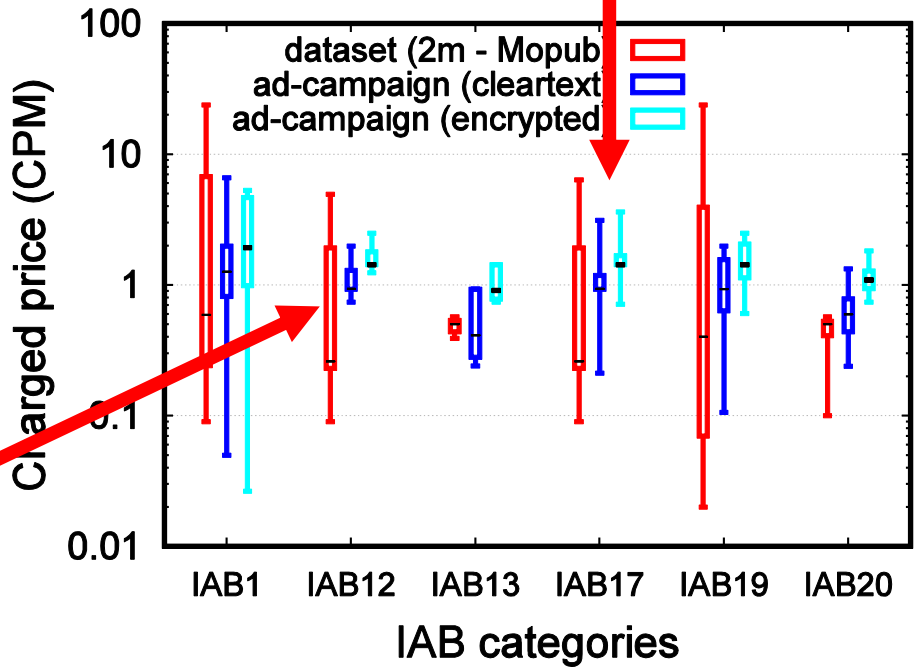
- 2 real probing ad-campaigns in 2016 (*A1*, *A2*): various experimental setups

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

Cost per IAB in cleartext and encrypted prices

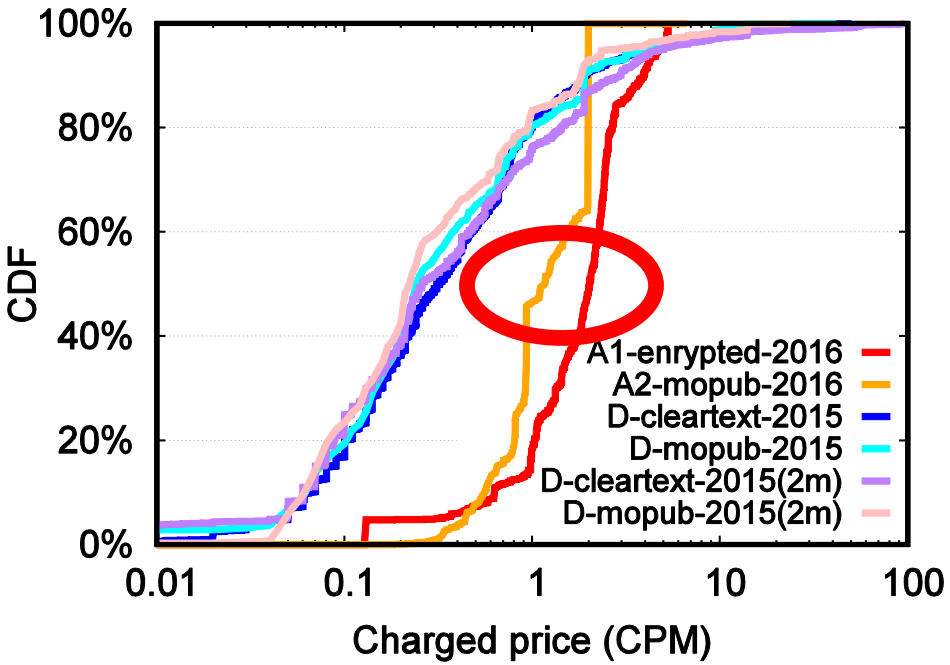
Median encrypted prices always higher than cleartext



Comparison of CPM costs for the different IAB categories in our dataset and the 2 probing ad-campaigns.

Time shift: More recent cleartext prices are higher than the ones last year

Encrypted Vs. Cleartext prices

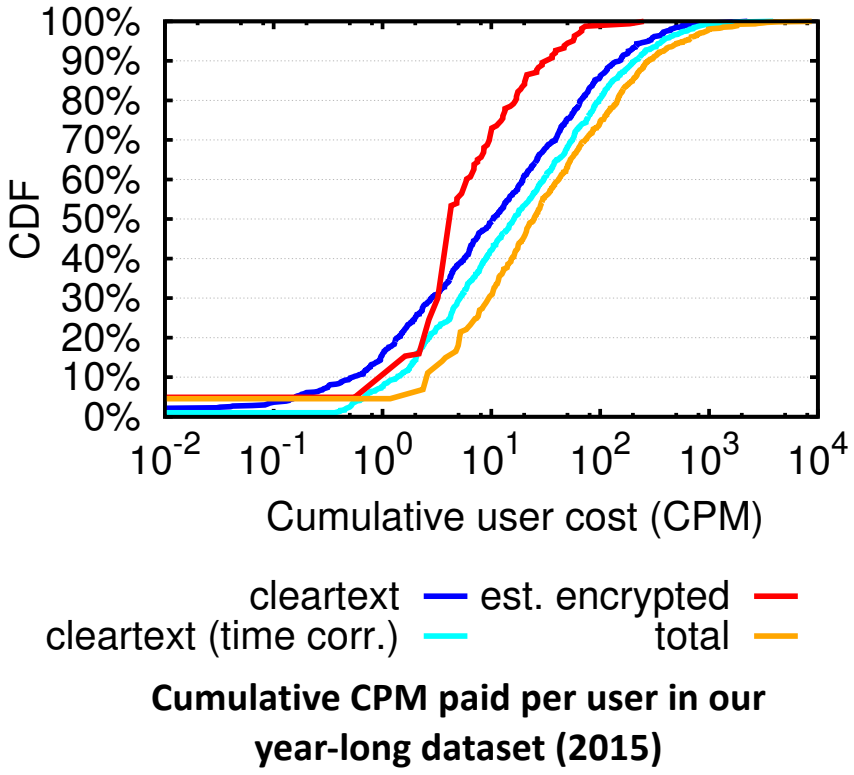


Comparison of price distributions between cleartext and encrypted, for different time periods and datasets (D vs. A1 and A2).

~~*"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)

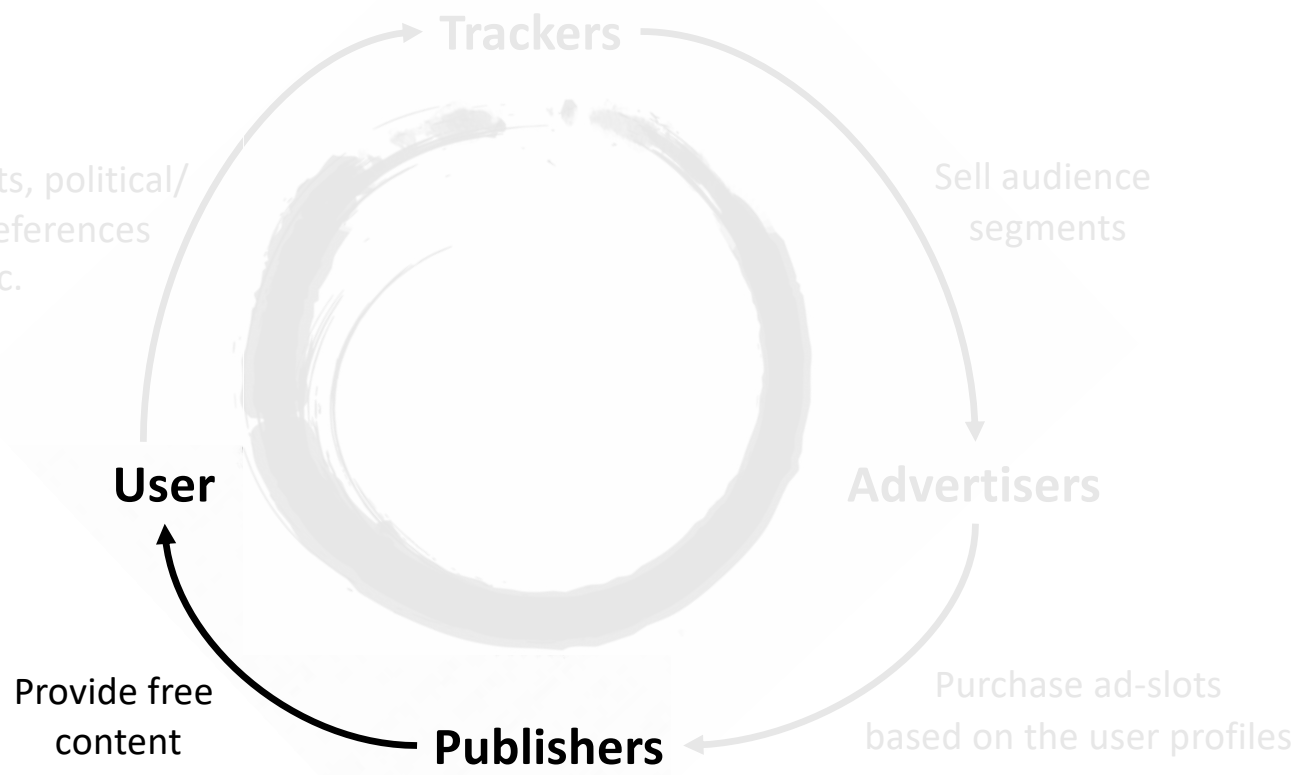
How much do advertisers pay to reach you?



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 (euros) (73% of the users cost < 100 CPM)
- 2% of users cost 10-100× more to the ad-ecosystem than the average user!

Free content it is?



In Digital Advertising

1. Advertiser pays to deliver ad
2. Publisher gets revenue and provides space
3. User **also pays** to get ads inside the received content

Free content it is **NOT**

1. After examining the costs of mobile advertising for advertisers, analyze the hidden costs on the users using the same dataset
2. compare them for the same user profiles
3. how fairly they are shared among the two sides?

Goal: Transparency in the costs of digital advertising for both advertiser and user

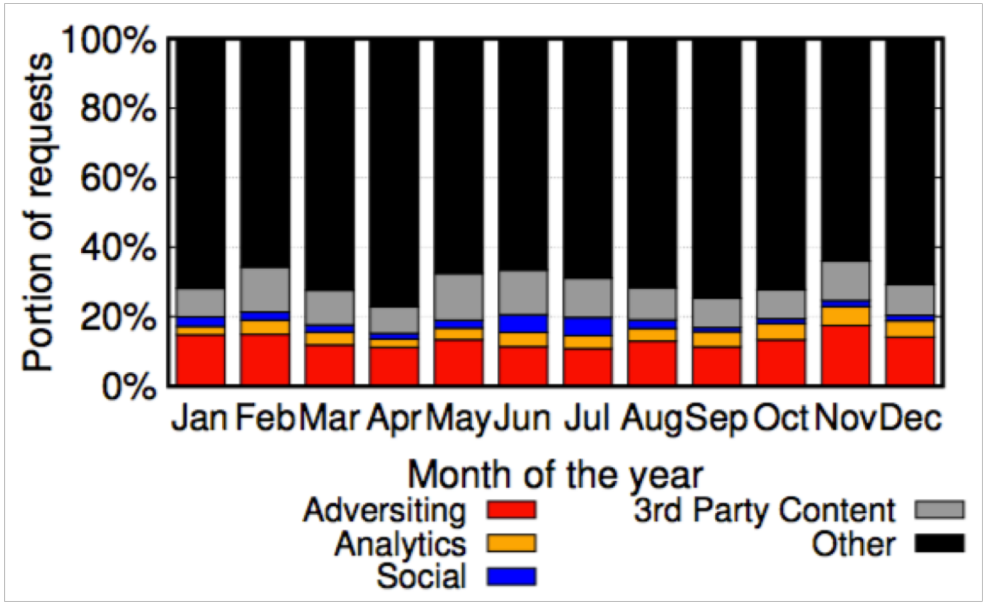
The view of the user (1/4)

Portion of HTTP requests per content category for average user.

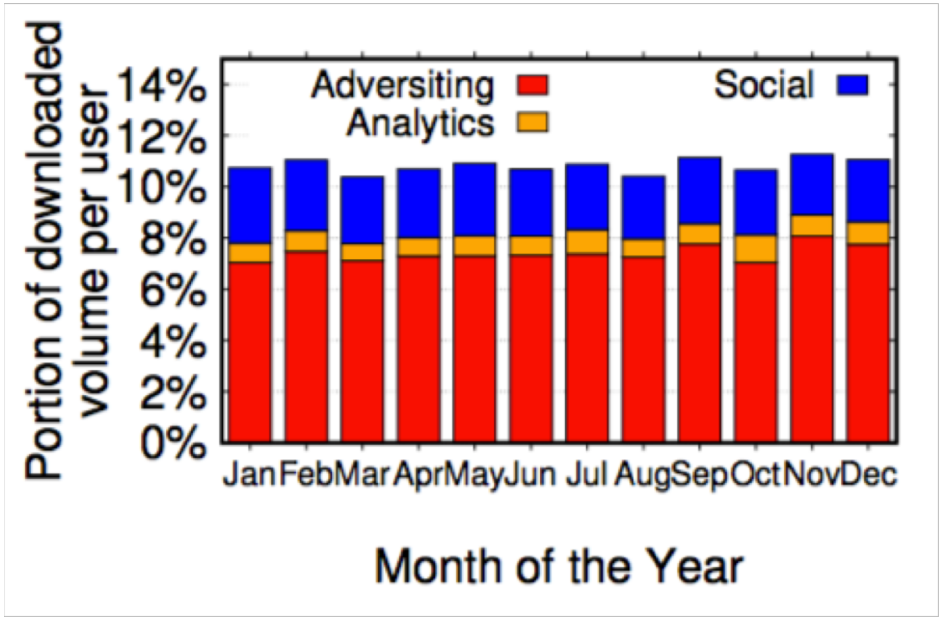
- Categories:
  - **Advertising**,
  - **Analytics**,
  - **Social** (Facebook,Twitter plugins),
  - **3rd Party Content** (CDNs, widgets),
  - **Other**

Actual Content

- 23% of HTTP requests associated with content user is **not** actually interested in.



The view of the user (2/4)



How much of the downloaded volume is related to ads?

- Across the year:
  8.2% of total transferred bytes Ad and Analytics related

- Ad volume increased:
  - ➢from 5.6%* (2012) to 7.3% (2016)

*Narseo Vallina-Rodriguez, Jay Shah, Alessandro Finamore, Yan Grunenberger, Konstantina Papagiannaki, Hamed Haddadi, and Jon Crowcroft. Breaking for Commercials: Characterizing Mobile Advertising. IMC'12

The view of the user (3/4)

- power consumption of network component:
  - 7.98% related to ad-traffic

- mobile device's battery:
  - 10 hours of ad-free browsing
  - 9.2 hours of ad-supported browsing.

Based on the proposed model of Chanmin Yoon, Dongwon Kim, Wonwoo Jung, Chulkoo Kang, and Hojung Cha.
AppScope: Application Energy Metering Framework for Android Smartphone Using Kernel Activity Monitoring, USENIX ATC'12

The view of the user (3/4)

Anonymity Loss: userIDs synced

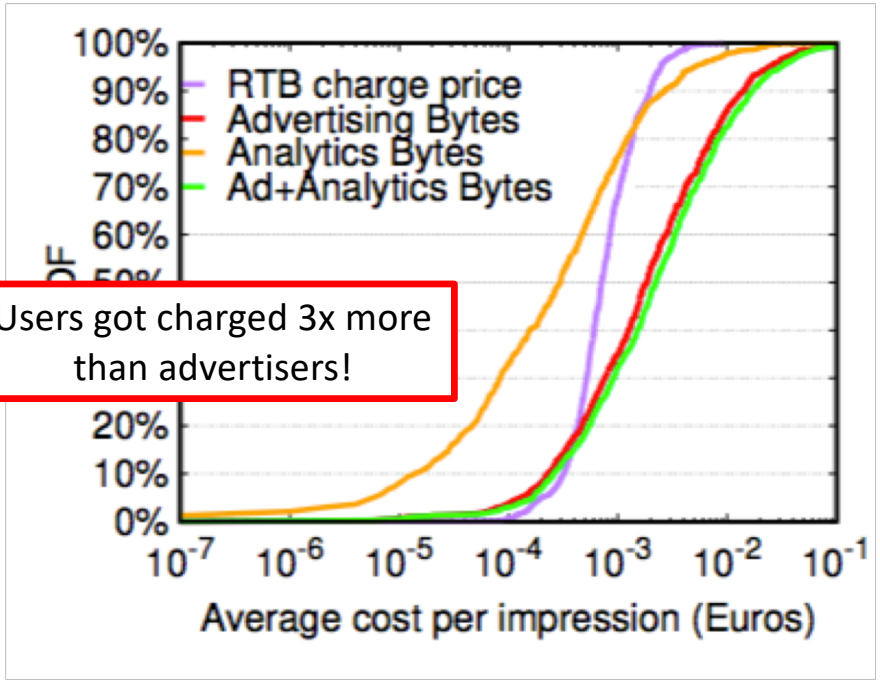
✓median user exposed to 1 CSync every 140 HTTP requests
- ➢(every 3-4 website visits)

How much do tracking entities know about a user?
- Median user:
  - ➢As many as 22 tracking entities learned each up to 20% of all her userIDs
  - ➢3 tracking entities learned each up to 40% of her userIDs

Comparison of both views

Per ad impression:

- median advertiser paid 0.00071 Euros
- median user paid 0.0022 Euros (in downloaded ad-bytes)*
- median user got userIDs leaked to 3.4 different entities



*Considering average cost per byte in Europe

Ad-blockers kill free internet: Any alternative web monetization models?

- Subscriptions, pay for the received content

- User Compensation: e.g., Basic Attention Token (BAT)
  - ➢users are rewarded for their attention

- **Increasingly popular** -> CPU borrowing: user-side cryptomining

<u>Websites dataset</u>

- 100K ad supported
- 100K mining supported
- 27 different mining libraries

<u>Monitoring Module</u>

- CPU & Memory Utilization
- Power Consumption
- System Temperature
- Parallel process interference

Web-mining Vs Ads Analysis

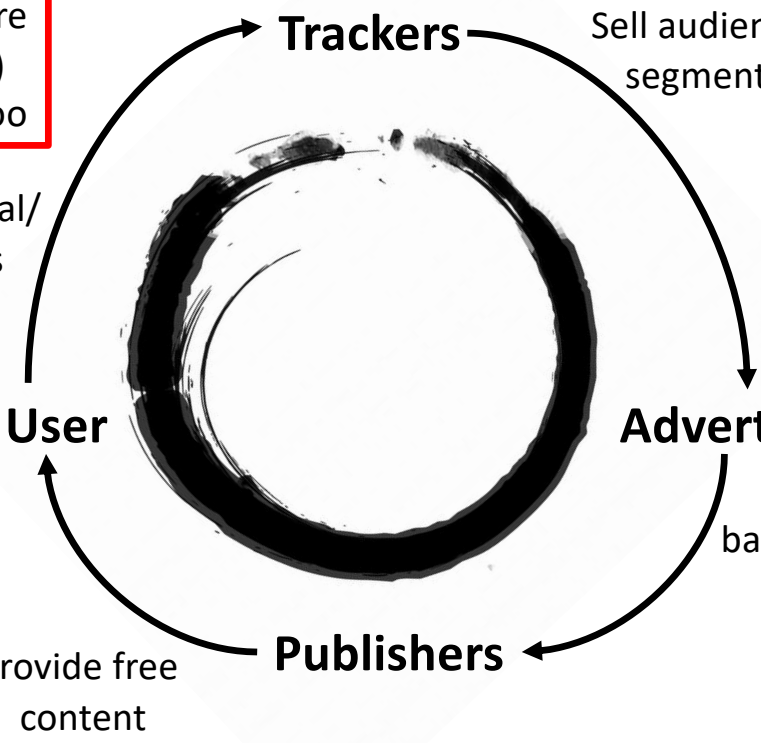
- Zero user data needed by miner -> privacy preserving
- Profitable for website visits > 5.53 min
- median mining website utilizes up to 59x more the visitor's CPU
- a visit to a mining website consumes on average 2.08x more energy
- a visitor's system operates in up to 52.8% higher temperatures
- web-miners affect parallel running processes: may degrade 57% their performance.
  - ➢scalability issues for multiple open mining tabs

- **Hybrid approach:** begin with ads, continue with mining when tab switches on the background

Conclusion

1. WWW'17

- Apps leak significantly more (SSIDs, installed apps etc.)
- Web browser apps leak, too



2. EuroSec'18,
3. arXiv:1805.10505 (submitted in NDSS)

- CSync is the catalyst for any user data sharing
- Reduces anonymity of web users
- Increases the number of trackers learning about the user by 6.75x.

4. IMC'17

- In mobile, median user costs 25 CPM per year (euros).
- User Interests, demographics affect pricing dynamics

5. WWW'18
6. arXiv:1806.01994 (submitted in INFOCOM)

Mobile Users get charged 3x more than advertisers