The Past, Present and Future of Online Tracking

Dr Michael Veale
Faculty of Laws, University College London
In the early 90s, the Web was ‘stateless’ — it had no memory of its visitors.

Cookies were invented to solve this problem: they are simply text placed on your browser by a web server that a server can look at later.

Syntax of the Set-Cookie HTTP Response Header:
```plaintext
Set-Cookie: NAME=OPAQUE\_STRING \n    [; expires= ] \n    [; path=] \n    [; domain=] \n    [; secure]
```

Syntax of the Cookie HTTP Request Header:
```plaintext
Cookie: NAME=OPAQUE\_STRING *[; NAME=OPAQUE\_STRING]
```

First proposal for state management on the web (Apr. 18, 1995)
In the early days of the Web, all content on a webpage came from the same server.

An early, popular browser, Netscape Navigator, introduced the function of rendering two webpages in a single browsing window in 1996 (frames).

This created a problem: could the second website access the cookies the first had laid?
The same origin policy

• The solution — the Same Origin Policy.
• Cookies only accessible by servers that share features (particularly the domain) of the one that laid them.
• A user visiting ucl.ac.uk should expect only ucl.ac.uk cookies to be read — not kcl.ac.uk cookies.
• Didn’t fix the problem for long:
  • Websites started calling many distinct servers. Used to be 1, now 100s — because a website would instruct your computer to query many domains.
  • These many domains collaborate to share information about users’ Web usage and more — called Cookie Syncing.

• Google calls home with unique identifiers for at least 28% of all web page loads, while Facebook does the same for approximately 15%. The proportion is significantly higher in certain sectors, such as news, compared to others, such as banking.*

• Collaboration between trackers means that even under conservative estimates, 53 firms observe more than 91% of users’ browsing behaviour.**

Apps are pretty bad

• Users have more ability to control the Web, through browsers. For apps, they have almost none.

• One recent study identified 2,121 separate advertising tracking services in apps in the Android ecosystem, which can be grouped by ownership into approximately 292 parent organisations.*

• Another study found that 88.4% of apps contained a tracker owned by Alphabet (Google), 42.6% by Facebook, 33.9% by Twitter, 26.3% by Verizon and 22.2% by Microsoft. 30% of News apps, 28% of Family apps, and 25% of Gaming & Entertainment apps contain trackers from more than ten distinct tracker companies.**

---

The US military, Customs and Border Protection, the Secret Service and Homeland Security buy location data from adtech firms as a way of avoiding obtaining warrants.

- ‘X-Mode’ trackers in apps: e.g. a Quar’an/prayer time app downloaded by 98m Muslims around the world; a Craigslist searching app and a spirit level app.
- Bidstream data from real-time bidding (Venntel, Babel Street)
- Many vendors found selling data that can specifically be used to locate women who have visited abortion clinics (Kochava, SafeGraph, Placer.ai)

This data isn’t just used for advertising.
This infrastructure isn’t just used for advertising

- NSA and GCHQ utilise the uniquely identifying Google “PREF” cookies to single out a user’s computer and allow it to be remotely exploited using hacking tools developed by these state actors.
- DoubleClick cookies to reidentify Tor Browser users
- Major Belgacom hack achieved by GCHQ and CSEC through MUTANT BROTH, a system bulk-storing a range of cookies laid on popular websites, including those from Google and Facebook, in order to both identify users and build up a pattern of their daily habits and routines.

But when they are used for ads, it’s pretty bad too.
Real-Time Bidding
From about 2010, automated auctions for your eyeballs.

Data sent to bidders each time this happens

- **Site**
  - URL of the site being visited
  - Site category or topic

- **Device**
  - Operating system
  - Browser software and version
  - Device manufacturer, model
  - Mobile provider
  - Screen dimensions

- **User**
  - Unique identifiers set by vendor and/or buyer.
  - Advertising exchange’s cookie ID.
  - A demand-side platform’s user identifier
  - Year of Birth
  - Gender
  - Interests
  - Metadata reporting on consent provided

- **Geography**
  - Longitude and latitude
  - Postal/ZIP code
Bid requests go to hundreds or thousands of companies; little oversight.

**Vectaury** in France — small company, with only 3.5m€ annual turnover — retained 68m bid request records (and fined by the French data regulator, CNIL) in 2018.

Their website even claimed that they discarded 70% of all data, and only kept any of it for 12 months meaning that this small company was possibly sent 1/4 billion bid requests in just a single year.
## Leading RTB exchanges, daily bid request estimates

<table>
<thead>
<tr>
<th>Exchange</th>
<th>Daily Bid Requests</th>
</tr>
</thead>
<tbody>
<tr>
<td>Index Exchange</td>
<td>50 billion</td>
</tr>
<tr>
<td>OpenX</td>
<td>60 billion</td>
</tr>
<tr>
<td>Rubicon Project</td>
<td>Unknown. Claims to reach 1 billion people’s devices.</td>
</tr>
<tr>
<td>PubMatic</td>
<td>70 billion</td>
</tr>
<tr>
<td>Oath/AOL</td>
<td>90 billion</td>
</tr>
<tr>
<td>AppNexus</td>
<td>131 billion</td>
</tr>
<tr>
<td>Smaato</td>
<td>214 billion</td>
</tr>
<tr>
<td>Google DoubleClick</td>
<td>Unknown. DoubleClick is the dominant exchange.</td>
</tr>
</tbody>
</table>

---

6. 500 Billion / 29.6 - 38.8 billion impressions per day. Using AppNexus 1:1:1.5 ratio, this is 214 auctions per day. 500 impressions figure cited in “Optimize your mobile strategy”, Smaato, (URL: https://www.smaato.com/).
7. “Transacting at a peak of 11.4 billion daily impressions, our marketplace handles more traffic each day than Visa, Nasdaq, and the NYSE combined” at https://www.appnexus.com/sell. Note that in 2017, AppNexus said in “AppNexus Scales with DriveScale”, 2017, (URL: http://go.drivescale.com/ee/461-2838-800/images/DEY_Gase_Study_AppNexus-final_v1.pdf) that 10.7 billion “impressions transacted” came as a result of running 123 billion auctions. The impressions transacted to auctions ratio appears to be roughly 1:1:1.5. Therefore, the 11.4 daily impressions reported in 2018 equates to 134 billion auctions per day.
• Some ad infrastructures facilitate messages to be targeted to the level of specific individuals.

• Studies on Facebook’s ad infrastructure have modelled that this “nanotargeting” is possible using either the 4 rarest interests of an individual or the 22 random interests from the interests set Facebook assigns — both options make users unique on Facebook with a 90% probability (González-Cabañas et al., 2021).

• Various “war stories” of this occurring, from intimate partner abuse to Labour Party HQ targeting Jeremy Corbyn.
‘Cookie Banners’
We asked five menstrual health apps for our data, and here's what we found...

We asked five menstruation apps to give us access to our data. We got a dizzying dive into the most intimate information about us.
• Personal data in data protection law: information relating to an identified or identifiable natural person.

• Data protection law (e.g. the GDPR) requires a ‘lawful basis’ for all personal data processing

  • **Not a consent-first law**: If you are doing something aligned with the user, not using sensitive categories of data like ethnicity or health, you typically won’t need consent.

• However, ePrivacy Directive: consent for storing or retrieving data for terminal devices (history of rootkits, tracking) if not necessary for the requested service.
Article 4
Definitions

11. ‘consent’ of the data subject means any freely given, specific, informed and unambiguous indication of the data subject’s wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her;

Article 7
Conditions for consent

1. Where processing is based on consent, the controller shall be able to demonstrate that the data subject has consented to processing of his or her personal data.

2. If the data subject's consent is given in the context of a written declaration which also concerns other matters, the request for consent shall be presented in a manner which is clearly distinguishable from the other matters, in an intelligible and easily accessible form, using clear and plain language. Any part of such a declaration which constitutes an infringement of this Regulation shall not be binding.

3. The data subject shall have the right to withdraw his or her consent at any time. The withdrawal of consent shall not affect the lawfulness of processing based on consent before its withdrawal. Prior to giving consent, the data subject shall be informed thereof. It shall be as easy to withdraw as to give consent.

4. When assessing whether consent is freely given, utmost account shall be taken of whether, inter alia, the performance of a contract, including the provision of a service, is conditional on consent to the processing of personal data that is not necessary for the performance of that contract.
Recital 32

Consent should be given by a clear affirmative act establishing a freely given, specific, informed and unambiguous indication of the data subject's agreement to the processing of personal data relating to him or her, such as by a written statement, including by electronic means, or an oral statement. This could include ticking a box when visiting an internet website, choosing technical settings for information society services or another statement or conduct which clearly indicates in this context the data subject's acceptance of the proposed processing of his or her personal data. Silence, pre-ticked boxes or inactivity should not therefore constitute consent. Consent should cover all processing activities carried out for the same purpose or purposes. When the processing has multiple purposes, consent should be given for all of them. If the data subject's consent is to be given following a request by electronic means, the request must be clear, concise and not unnecessarily disruptive to the use of the service for which it is provided.
With hundreds of trackers... how?

Consent management platforms emerge
Legal entrepreneurship of an unsavoury kind

Figure 1. The three components of the QuantCast CMP on https://sourceforge.net as of September 2019.
Many vendors: but are they compliant with the law?
• Together with Aarhus University and MIT, we investigated whether these interfaces were providing valid consent under EU law.

• Built a bespoke web scraper and fed it the top 10K UK websites in 2019. We coded it to be able to analyse the top 5 CMPs to see how they were configured.

Empirical, computational legal analysis to find out

List of URLs to check

Software analyses pop-ups

Returns data on compliance

<table>
<thead>
<tr>
<th>CMP</th>
<th>Pop-up/open contest options</th>
<th>Reconcile terms</th>
<th>Personal options</th>
<th>Minimum compliance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cookiebot</td>
<td>450</td>
<td>767</td>
<td>64 (75.9%)</td>
<td>2 (54.6%)</td>
</tr>
<tr>
<td>ConsentFlow</td>
<td>6477</td>
<td>5231</td>
<td>67 (80.7%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>OneTrust</td>
<td>47718</td>
<td>1387</td>
<td>68 (85.4%)</td>
<td>3 (13%)</td>
</tr>
<tr>
<td>Quтелlet</td>
<td>3754</td>
<td>120147</td>
<td>50 (33.3%)</td>
<td>73 (25.2%)</td>
</tr>
<tr>
<td>TrustArc</td>
<td>4276</td>
<td>2845</td>
<td>61 (75.9%)</td>
<td>2 (25%)</td>
</tr>
<tr>
<td>All</td>
<td>459/233</td>
<td>46/234</td>
<td>382 (56.2%)</td>
<td>80 (13%)</td>
</tr>
</tbody>
</table>

Table 1: Key statistics on accepted CMPs.
And what did we find?

Turned case law into three legal tests

1. No optional boxes preselected

2. Reject all as easy as Accept all

3. Consent is explicit
Future of Targeting
Privacy Enhancing Technologies

Example: Multi-party computation

$$f(a, b, c)$$

multi-party computation among servers

secret sharing
On-device targeting
Let’s do it at my place instead

- Can we target users in as before, but without data leaving devices?
- Train shared models with privacy enhancing data analysis so tracking data never leaves the phone.
- Use other technologies such as ‘zero-knowledge proofs’ — or just locked down platforms — to check individuals are profiling themselves sufficiently, seeing the adverts.

Emerging moves

- Google’s Privacy Sandbox
- Investigation by UK competition and markets authority
- Others: Apple’s hires in AdTech; Meta and Mozilla’s proposals in the IETF.
Questions for the future of online advertising

- **Interaction between input data and confidential computing**: what is the theoretical, ethical, and legal basis not to use e.g. blood pulse, eye-tracking, etc data when the output is confidential?

- **Is your device betraying you?** Rights for people facilitating computing, rather than just rights related to how data about you is used. Links to research ethics.