SalingerPrivacy

Cookies and Other Online Identifiers

Research Paper for the Office of the Australian Information Commissioner

15 June 2020

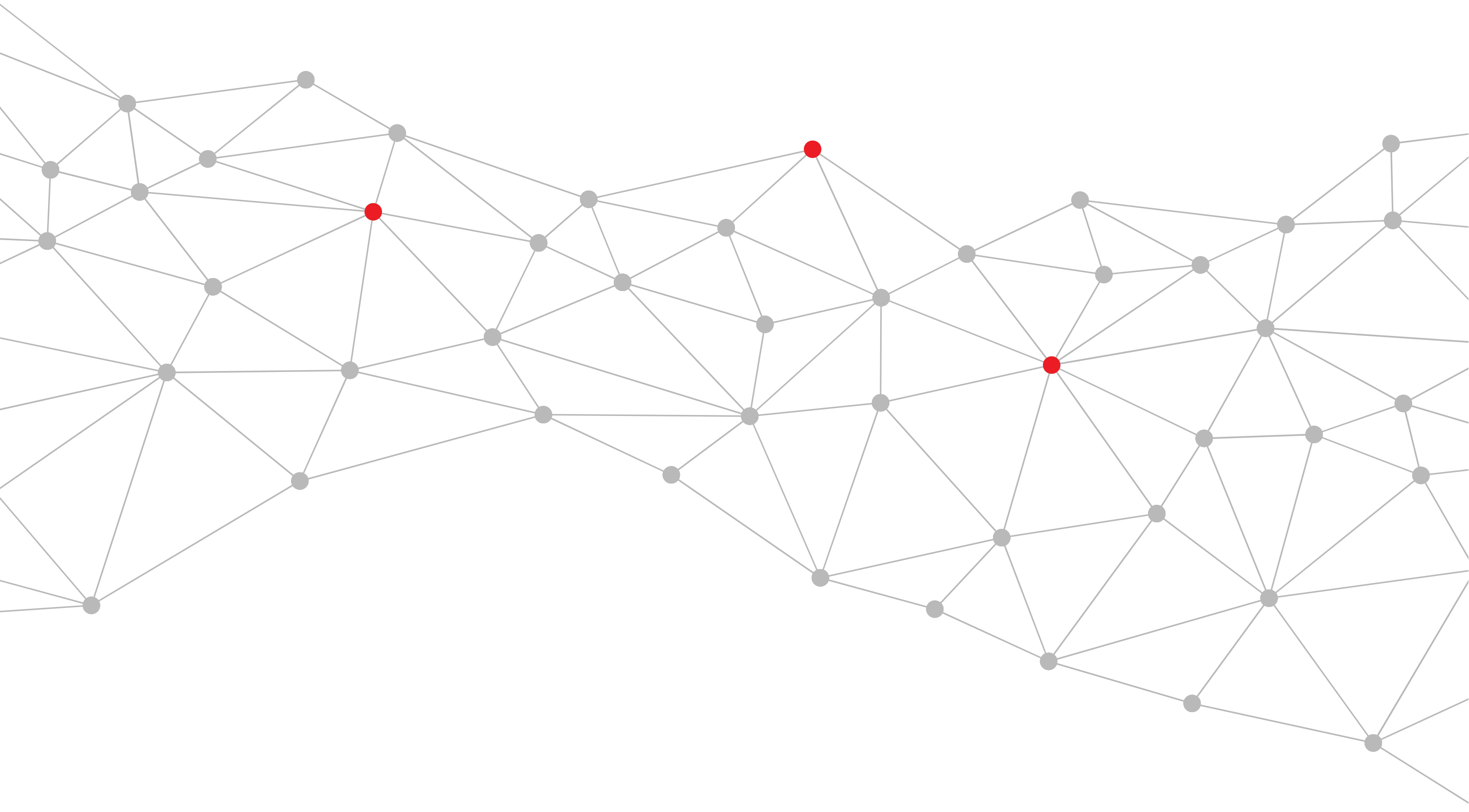


Table of Contents

Executive Summary

This Research Paper was commissioned by the Office of the Australian Information Commissioner, to inform a review of the Privacy Act. This Research Paper builds upon our previous Research Paper which recommended updating the definition of ‘personal information’ in the Privacy Act, as well as considering additional options for reform and regulatory responses, in relation to cookies and online identifiers in particular.

From the digital breadcrumbs we leave behind in the form of geolocation data shed from our mobile devices, to the patterns of behaviour we exhibit online as we browse, click, comment, shop, share and ‘like’, we can be tracked. Tracked, then profiled, and finally targeted.

Led by – but by no means exclusive to – the online behavioural advertising industry, entire ecosystems have been built around the use of online identifiers, which are deployed to track the behaviour of consumers and citizens across different sites and services, and over time.

Online identifiers can include:

- files embedded on devices
- device identifiers
- device fingerprinting, and
- digital platform user IDs used across multiple sites.

The purpose of online identifiers is simple: they aim to track users across multiple pages, visits, websites and even devices. What varies is the intent and purpose behind their use. Some uses are essential for the functioning of the internet, such as session identifiers that allow e-commerce sites to operate, or authentication identifiers which allow users to login to online services. However, in addition to those essential functions, cookies and other online identifiers have been repurposed to facilitate longitudinal data collection, profiling, and targeting, constituting a significant incursion into the private lives of users.

By linking an individual or a device to behaviour such as searches, queries, posts, browsing sites and purchases, and even offline behaviour or purchases, the party doing the tracking can start to profile individuals, drawing inferences about their interests and preferences, behaviour and budget, and divide them into segments accordingly. The individual user of the device can then be targeted to receive a particular ad, offered personalised content or recommendations, sent political messaging, or subjected to an automated decision such as differential pricing.

Chapter 1 of this Research Paper offers an introduction to online identifiers, and the current regulatory context in Australia. Chapter 2 provides an in-depth review of various forms of online identifiers, and explains how online identifiers allow many types of communications – not just advertisements – to be targeted at particular individuals.

Chapter 3 demonstrates that the privacy harms facilitated by the unfair and/or intrusive use of online identifiers to track, profile and micro-target individuals online - and even offline - include social and market exclusion, price discrimination resulting in economic inequality, prejudice and discrimination, manipulation leading to negative physical and mental health impacts, and manipulation of voting intentions. These activities each hold the potential to impact on individuals' autonomy, by narrowing or altering their market or life choices.

In relation to the use of cookies and online identifiers to collect personal information about individuals, there appears to be a significant disconnect between community expectations in Australia, and common business practices. Research conducted for the OAIC in 2017 found that Australians feel uncomfortable about online tracking, with 79% uncomfortable about targeted advertising based on their online activities, and 83% uncomfortable with social networking companies keeping databases of information about their online activities. A survey conducted by Roy Morgan in 2018 found that almost 90% of Australians say it is unacceptable for social media and search companies to use their personal data in order to tailor ads and offers to consumers.

Robust data protection regulation is necessary to achieve both consumer protection outcomes, and consistency of the playing field for industry. It will therefore be critical to ensure that the Privacy Act remains fit for its purpose of enabling effective regulation of personal information handling, in line with community and business expectations.

Chapter 4 reviews regulatory responses to the risks posed by online identifiers from around the world, and highlights that a number of privacy statutes explicitly refer to online identifiers within their definitions of personal information or personal data.

Whether or not any particular piece of data meets the definition of 'personal information' is a threshold legal issue for the operation of privacy law in Australia: the definition of 'personal information' determines the boundaries of what is regulated, and what is protected. Yet through its Digital Platforms Inquiry, the ACCC found that the current definition of 'personal information' suffers from a lack of certainty around its coverage of technical data, including online identifiers.

In recommending that the Privacy Act should be amended to explicitly include online identifiers, the ACCC is not alone. There is increasingly global recognition that online identifiers pose privacy risks (by facilitating privacy harms), and require more consistent and robust regulation.

However we caution against a regulatory response which relies primarily on the regulation of any particular type of identifier or technology. As Chapter 2 demonstrates, cookies are neither inherently good, nor inherently bad. It is not possible to prohibit the use of online identifiers altogether, since their use is essential in authentication, session management, security management, and network routing. Similarly, personalisation of messaging to users is neither inherently good, nor inherently bad.

Further, the privacy risks stem from the data collected and generated from the use of online identifiers, rather than from identifiers themselves. Whilst it is important to regulate the use of identifiers to stem intrusive forms of data collection, it is also important to not focus solely on identifiers. In particular, if a dataset is assembled through the use of online identifiers but later shared without the identifiers, the privacy risks are not necessarily reduced.

We therefore caution against a regulatory response which overly restricts all practices, rather than focussing on the level of harm posed by any particular practice. For this reason, we have not recommended following the model currently proposed by the UK's ICO, the effect of which would appear to be to make all online targeted advertising and messaging unlawful in the absence of a proactive consent from the user.

Further, as Chapter 4 demonstrates, the European approach to regulating cookies and other tracking technology via the ePrivacy Directive has produced legislation which is already out-of-date. The current reform process is mired in debates over what is or is not a 'strictly necessary' cookie, rather than debating what is fair or intrusive. Meanwhile that debate is increasingly moot; Chapter 2 provides a number of examples, such as Login Management and reCAPTCHA, in which 'strictly necessary' identifiers have been re-purposed for tracking.

In our view, an overarching 'fairness' regulatory framework, such as that used in Canada, is preferable, because rather than focusing on any particular type of technology, it focuses on intent and outcomes, as well as whether there are meaningful opportunities for individuals to understand and control what is happening to their data.

We suggest that the appropriate response is to regulate privacy-intrusive behaviours or practices, rather than any particular technology. As such, we recommend explicitly bringing online identifiers within the Privacy Act's scope, but then allowing the Australian Privacy Principles to do the heavy lifting, in terms of determining what use cases will be considered lawful and fair, and what will not.

We also suggest that any proposed reform must be mindful of the need for global consistency, which is beneficial for consumers, regulated entities and regulators alike; but must also ensure that the definition is 'fit for purpose' for Australian conditions now and into the future.

Our recommendations in Chapter 6 were drafted with the following objectives:

- Resolving the lack of clarity around coverage of technical data, in line with the ACCC's recommendations
- Enabling consistency with the GDPR where suitable
- Maintaining the technological neutrality of the Privacy Act
- Making minimal regulatory change for maximum effectiveness
- Updating the Privacy Act to ensure it remains fit for purpose in protecting against multiple forms of privacy harms in digital environments, while
- Avoiding regulatory over-reach into technologies, practices or behaviours which do *not* pose risks of privacy harms.

Our recommendations encompass:

- reforming the Privacy Act, in relation to:
    - the definitions of personal information and de-identified data
    - the definition of consent
    - introducing an overarching fairness requirement on all APPs regulating collection, use and disclosure of personal information
    - repealing APP 7, the direct marketing principle, and
    - strengthening APP 6, in relation to the secondary use of personal information
- developing Codes and Guidelines under the Privacy Act, in relation to:
    - specifying 'no-go zones' under the overarching fairness requirement, such as targeted marketing to children
    - developing guidelines to outline the features that online identifiers *should* have (namely, observable, resettable and blockable); and
- other policy responses the OAIC could take, to promote best privacy practices with respect to the fair use of online identifiers.

Finally, in Appendix A, we have mapped out examples of how implementation of our recommendations would likely impact on some of the practices outlined in this report.

Anna Johnston
Principal | Salinger Privacy

15 June 2020

Glossary and acronyms

ACCC	Australian Competition and Consumer Commission
API	Application Programming Interface
APPs	Australian Privacy Principles, found in the Privacy Act
CCPA	California Consumer Privacy Act
COPPA	Children's Online Privacy Protection Act 1998 (USA)
DPI	Digital Platforms Inquiry
EDPB	European Data Protection Board
ePD	Directive on Privacy and Electronic Communications 2009 (EU), aka the ePrivacy Directive
ePR	ePrivacy Regulation; will replace the ePD, currently in draft form
EU	European Union
GDPR	General Data Protection Regulation 2016 (EU)
ICO	Information Commissioner's Office (UK)
OAIC	Office of the Australian Information Commissioner
OPCC	Office of the Privacy Commissioner of Canada
PECR	Privacy and Electronic Communications Regulations (UK)
PIPEDA	Personal Information Protection and Electronic Documents Act 2000 (Canada)
Privacy Act	Privacy Act 1988 (Australia)

Chapter 1

Why online identifiers matter

Digital breadcrumbs

From the digital breadcrumbs we leave behind in the form of geolocation data shed from our mobile devices, to the patterns of behaviour we exhibit online as we browse, click, comment, shop, share and 'like', we can be tracked. Tracked, then profiled, and finally targeted.

Led by – but by no means exclusive to – the online behavioural advertising industry, entire ecosystems have been built around the use of online identifiers, which are deployed to track the behaviour of consumers and citizens across different sites and services, and over time.

The objective of online behavioural advertising is, like any advertising, to predict purchasing interests, and drive purchasing decisions. Online, however, the repercussions are much greater, because of the degree to which advertising – and indeed, even what non-advertising content users are shown – has become 'personalised'. Personalisation means precise decisions are made at an individual level about who sees what, and equally what will be withheld from whom.

Online identifiers can include:

- files embedded on devices (e.g. cookies, web beacons and tracking pixels)
- device identifiers (e.g. an IMEI mobile phone identification number, or a MAC address for devices which connect to networks or other devices using Wi-Fi or Bluetooth)
- device fingerprinting (e.g. information about the IP address, operating system, language settings and browser being used on a device), and
- digital platform user IDs used across multiple sites (e.g. using an existing Facebook or Google account to 'log in' to a third party site).

By linking an individual or a device to behaviour such as searches, queries, posts, browsing sites and purchases, and even offline behaviour or purchases, the party doing the tracking can start to profile individuals, drawing inferences about their interests and preferences, behaviour and budget, and divide them into segments accordingly. The individual user of the device can then be targeted to receive a particular ad, offered personalised content or recommendations, sent political messaging, or subjected to an automated decision such as differential pricing.

How online identifiers are used

The purpose of online identifiers is simple: they aim to track users across multiple pages, visits, websites and even devices. What varies is the intent and purpose behind their use. Some uses are essential for the functioning of the internet, such as session identifiers that allow e-commerce sites to operate, or authentication identifiers which allow users to login to online services. However, in addition to those essential functions, online identifiers have been repurposed to facilitate longitudinal data collection, profiling, and targeting, constituting a significant incursion into the private lives of users.

The technologies for storing and establishing online identifiers is ever-changing. What started with cookies has moved onto cache identifiers, network identifiers, device and browser fingerprinting, and even user behaviour fingerprinting. These advances have diminished the user's ability to control (or even be aware of) when online identifiers are being used, and therefore when they are being profiled and tracked.

The UK Information Commissioner defines micro-targeting as "targeting techniques that use data analytics to identify the specific interests of individuals, create more relevant or personalised messaging targeting those individuals, predict the impact of that messaging, and then deliver that messaging directly to them".[1]

Online identifiers have three roles to play in tracking, profiling and delivering messages to individuals; first to facilitate longitudinal data collection, second to aid the linking of disparate datasets, and third to allow the targeting of the user with the results of the analysis of the collected data.

The first step, data collection, is entirely dependent on online identifiers. They provide a way of tracking the same user not only across one site, but also across multiple sites. It is this latter capability that facilities profiling, since it allows a picture of how a user moves across the web to be recorded, along with the actions taken on the various sites they visit. This detailed picture of the digital-self forms the dataset on which machine learning and predicative analytics can be run.

Often a single data collection will not provide sufficient insight or coverage to be useful. When this occurs, there is a need to combine multiple disparate datasets – the second step. Sometimes this is done in real-time via processes like Cookie Syncing, other times it occurs in the background through data brokers which collect and collate multiple data sources for resale.[2] When the data set is sufficiently large, the identifier may not even be required for linking, as the data itself is so detailed that it forms an identifier in and of itself.

At the end of the second step is a detailed dataset ready for analysis, using cutting edge machine learning or predictive analytics to determine behaviours, susceptibility to

[1] UK ICO, *Democracy disrupted? Personal information and political influence*, July 2018; available at https://ico.org.uk/media/2259369/democracy-disrupted-110718.pdf

[2] See https://www.vice.com/en_us/article/bjpx3w/what-are-data-brokers-and-how-to-stop-my-private-data-collection

manipulation and likely triggers for actions. In its benign form this analysis can predict the right advertisement to show the user, but it can also be used to manipulate mood,[3] opportunity,[4] price[5,6] and political opinions,[7] amongst many other aspects of an individual's personality and character.

Once the predictive model has been built, or the susceptibility to manipulation determined, the online identifier comes back into play. In the third step, the online identifier is used to target the results of the analysis back to the intended user. Without it there would a wealth of analysis with no determinable targets. As such, the online identifier bookends sophisticated analytics that facilitates precise manipulation of individual behaviour. It is therefore both an enabler and executor in the modern data driven world.

Some might argue that such profiling and targeting is also an essential part of the modern World Wide Web, which is dominated by the provision of free services supported by advertising. However, such advertising does not necessarily have to be based on profiling the user; many believe contextual advertising is sufficient.[8,9]

In fact, the advertising market is dominated by Google and Facebook, which together receive over 60% of all online advertising spending in the US;[10] Amazon is a distant third with just over 7%. Both Google and Facebook use behavioural profiling and targeting in their advertising businesses, and represent the two dominant players in the market. Google's advertising revenue for 2019 was estimated to be US$134.81 billion, accounting for 70.9% of its global revenue,[11] while Facebook had advertising revenue of US$69.66 billion, accounting for 98.5% of its global revenue.[12] As such, it is clear behavioural advertising is delivering returns for Google and Facebook, although whether it is in the interests of the users and advertisers is facing increased scrutiny.[13] One aspect that is clear is that both companies are generating vast amounts of revenue from the data they are collecting and processing about users.

Advertising is just one use case of the technology. Many large businesses today will have their own analytics teams, collecting their own data from their interactions with their customers. Such data is not limited to just online data; it is often linked with offline transaction details and loyalty card records. Additional data is purchased through data brokers to augment and expand the picture that can be built of a business's customers. Those businesses will then target customers through online platforms, advertisements and offline promotions. They may target their existing customers but on new platforms the

[3] See Kramer, Adam DI, Jamie E. Guillory, and Jeffrey T. Hancock. "Experimental evidence of massive-scale emotional contagion through social networks." Proceedings of the National Academy of Sciences 111.24 (2014): 8788-8790. Available at https://www.pnas.org/content/111/24/8788.full

[4] See https://www.theguardian.com/technology/2015/jul/08/women-less-likely-ads-high-paid-jobs-google-study

[5] See https://www.forbes.com/sites/neilhowe/2017/11/17/a-special-price-just-for-you/#2c791c6c90b3

[6] See https://www.wsj.com/articles/SB10001424052702304458604577488822667325882

[7] See https://www.theguardian.com/news/series/cambridge-analytica-files

[8] See [100] on Page 29

[9] See https://digiday.com/media/gumgumtest-new-york-times-gdpr-cut-off-ad-exchanges-europe-ad-revenue/

[10] See https://www.emarketer.com/content/facebook-google-duopoly-won-t-crack-this-year

[11] See https://www.statista.com/statistics/266249/advertising-revenue-of-google/

[12] See https://www.statista.com/statistics/271258/facebooks-advertising-revenue-worldwide/

[13] Australian Competition and Consumer Commission, Digital Platforms Inquiry: Final Report, June 2019; available at https://www.accc.gov.au/publications/digital-platforms-inquiry-final-report

customer may not expect (e.g. by providing an existing mailing list to a platform such as Facebook); or they may target 'lookalike' audiences (i.e. people with similar interests, behaviours or characteristics to existing/past customers).[14]

Online identifiers can ultimately be used to directly impact individuals, via:

- online behavioural advertising (as opposed to broadcast or contextual advertising)
- online behavioural targeting (such as targeting individuals to receive particular political messaging)
- business decisions about the precise prices or products shown to (or hidden from) a consumer
- personalised content (examples include Facebook's news feeds, and Netflix or Amazon making recommendations about other movies or books the user might like), and
- online tracking in the offline world (such as creating a biometric identifier to conduct 'facial detection' (without necessarily performing facial recognition), in order to track an individual as they move through a space such as an airport or a shopping centre).

It would be wrong to suggest that all behavioural analytics is bad, or detrimental to the user or their privacy. The same techniques can deliver a more personal online experience, which better delivers the information someone is interested in, for example, via relevant product offerings,[15] making sure products are locally available prior to a customer ordering them[16] or recommendations for films,[17] amongst many others. This presents one of the challenges in regulating the area: the same technologies, data, and methods can be used to both serve and manipulate the user. The distinguishing factor between these different outcomes is the intent of the organisation deploying the capability.

As such, there are no easy technical solutions, nor simple legislative solutions. Formulating an appropriate regulatory response first requires establishing fundamental principles, and ensuring the end user has the capability and power to control their interactions and data, so they can continue to enjoy the benefits of the online environment, while avoiding exploitation.

[14] UK ICO, *Draft direct marketing code of practice*, v1.0 for public consultation, January 2020, pp.90-91; available at https://ico.org.uk/media/about-the-ico/consultations/2616882/direct-marketing-code-draft-guidance.pdf

[15] See https://www.amazon.science/the-history-of-amazons-recommendation-algorithm

[16] See https://press.aboutamazon.com/news-releases/news-release-details/aws-announces-general-availability-amazon-forecast

[17] See https://mobilesyrup.com/2017/08/22/80-percent-netflix-shows-discovered-recommendation/

The regulatory context

In 2019 the Australian Competition and Consumer Commission (ACCC) released the final report arising from its Digital Platforms Inquiry (DPI). The ACCC found that:

> "the Privacy Act needs reform in order to ensure consumers are adequately informed, empowered and protected, as to how their data is being used and collected. This will increase trust in the digital economy and spur competition between businesses on the basis of privacy."[18]

In particular, the ACCC noted existing problems with legal uncertainty about the scope of the definition of 'personal information' under the Privacy Act in terms of its coverage of what the ACCC termed 'technical data'; and the need to offer Australian consumers more effective data protection standards to match those found in other jurisdictions, particularly Europe. The ACCC concluded:

> "there are significant benefits in updating the definition of 'personal information' so that it covers the realities of how data is collected on individuals in the digital economy and to bring the Australian privacy regime into greater alignment with standards set by overseas data protection regulations."[19]

Recommendation 16(a) of the DPI Final Report was to update the definition of 'personal information', "to clarify that it captures technical data such as IP addresses, device identifiers, location data, and any other online identifiers that may be used identify an individual".

In December 2019 the Australian Government released its response to the DPI, including an implementation roadmap. The three responsible Ministers announced the Government's immediate commitment to a number of the DPI's recommendations, including to:

> "ensure privacy settings empower consumers, protect their data and best serve the Australian economy … through further strengthening of Privacy Act protections, subject to consultation and design of specific measures as well as conducting a review of the Privacy Act".[20]

The Government response and roadmap indicated the immediate commencement of consultation on Recommendation 16(a) in relation to the definition of personal information, as well as to shortly begin a broader review of the Privacy Act.

[18] Australian Competition and Consumer Commission, *Digital Platforms Inquiry: Final Report*, June 2019, p.3; available at https://www.accc.gov.au/publications/digital-platforms-inquiry-final-report

[19] Australian Competition and Consumer Commission, *Digital Platforms Inquiry: Final Report*, June 2019, p.461; available at https://www.accc.gov.au/publications/digital-platforms-inquiry-final-report

[20] Australian Government, *Regulating in the digital age: Government Response and Implementation Roadmap for the Digital Platforms Inquiry*, December 2019, p.3; available at https://treasury.gov.au/sites/default/files/2019-12/Government-Response-p2019-41708.pdf

Salinger Privacy was commissioned earlier this year by the Office of the Australian Information Commissioner (OAIC) to produce a Research Paper to inform the consultation on Recommendation 16(a). That first Research Paper, produced in February 2020, focused on the definition of 'personal information'.

This Research Paper, focussing on cookies and online identifiers, was then commissioned by the OAIC to inform the broader review of the Privacy Act. This Research Paper builds upon the previous Research Paper, by providing additional context for our earlier recommendations for updating the definition of 'personal information' in the Privacy Act, as well as considering additional options for reform and regulatory responses, in relation to cookies and online identifiers in particular.

Chapter 2

Types of online identifiers

Introduction to cookies

Cookies are an essential part of how the World Wide Web (the web) works. Without them much of the functionality taken for granted in modern web services would cease to work. However, their usage has moved beyond just essential functionality to become a primary tool for tracking and profiling devices and users.

The dual purpose of cookies creates challenges when evaluating their privacy impact. Determining whether a cookie is essential to the functionality of the website or being used for the purposes of tracking requires an understanding of the underlying technology. However, it is important to note that a purely technological focused evaluation is not sufficient. It is not possible to classify cookies as "safe" or "unsafe" based purely on their technical attributes. A holistic approach is required that encompasses evaluating the party that is creating the cookie, what it contains, its persistence, its stated purposes, and its potential future usage. This section will provide an overview of the technology underlying cookies, explaining why they are necessary, before moving on to examining the different ways they are set and used.

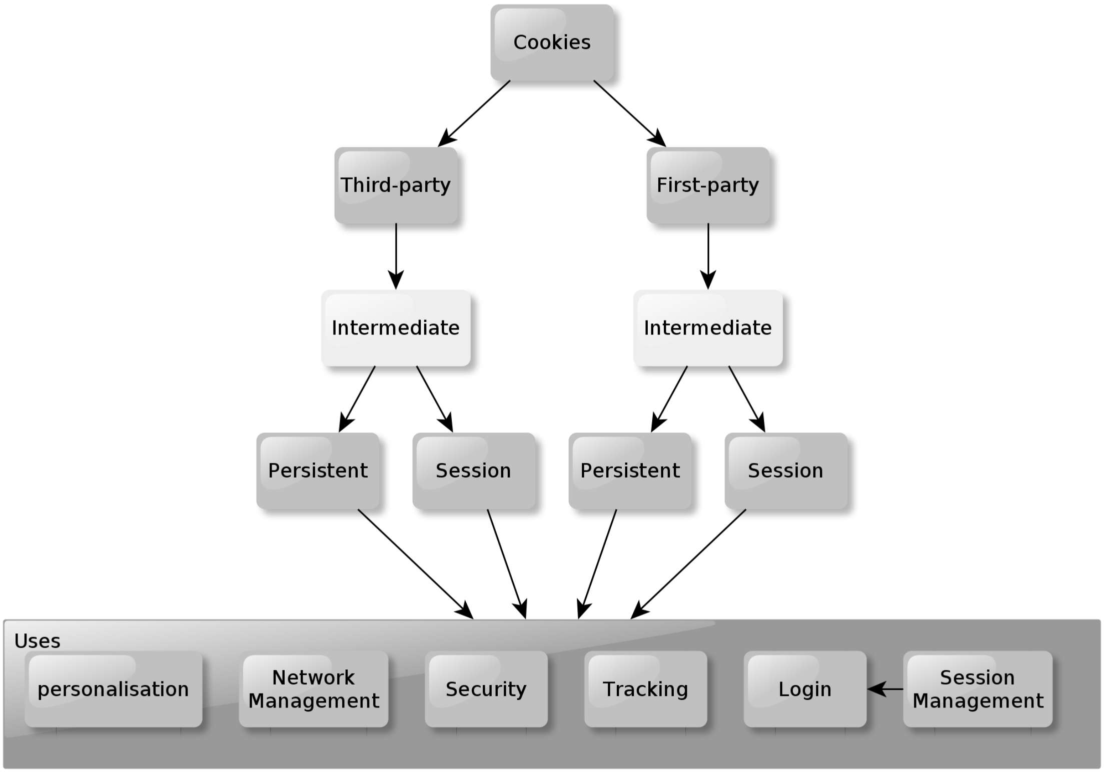


Figure 1: Cookie types and uses

What are cookies?

Cookies are small text files that are stored on a user's device via their web browser. The range of devices that store cookies has increased as people have adopted an always-connected lifestyle, through use of smartphones, tablets, smartwatches, and Internet of Things (IoT) devices. Cookies were originally intended as a way of keeping track of a user as they interacted with a website.[21] They can store small amounts of information, but most frequently are used to store a unique identifier.

Why are cookies needed?

In order to understand why such tracking is essential it is necessary to first look at the protocol that the web is built on. The web uses a protocol called the Hyper Text Transfer Protocol (HTTP). It is a fast, simple protocol that allows one device to request a resource from another device, typically a user makes a request in their web browser, on their device, for a resource on a remote web server. Each request is for a single resource and is independent of all other requests. As such, HTTP is considered to be *stateless*, which means each request is executed without any knowledge of previous requests. Such an approach works well for exchanging static information, but does not work well for modern web services, which are often more complicated and interactive, involving multiple requests related to each other, for example, e-commerce, online banking, or social media. Such interactions are known as *stateful*, in essence, they allow a series of requests and responses to be grouped together as a single session, with past requests and actions impacting on future responses. For example, adding items to an online shopping basket is *stateful*, since the shopping basket consists of all previously added items in that session.

Cookies were designed to provide *stateful* interactions on top of the *stateless* HTTP protocol, thereby allowing the same user to be tracked across multiple independent HTTP requests.

How are cookies set?

To set a cookie the web server, which is hosting the webpage the user wishes to view, includes the cookie value in its response to the request made by the user.[22] The browser stores the cookie data in a text file that is bound to the web host that set it, so only that host can access the contents. The host of a website can be found by looking at the Uniform Resource Locator (URL) more commonly referred to as a web address.

For example, https://www.oaic.gov.au/privacy/your-privacy-rights/ is made up of the following components: scheme://host:port/path?query

[21] See https://montulli.blogspot.com/2013/05/the-reasoning-behind-web-cookies.html

[22] Cookies can also be set by client-side JavaScript, however, in the context of third-party cookies this isn't possible.

Scheme	https (secure variant of HTTP)
Host	www.oaic.gov.au
Port	443 (default for https)
Path	/privacy/your-privacy-rights/
Query	[not set]

If the OAIC server sets a cookie, it will be associated with the host oaic.gov.au or a more specific portion of it, for example subdomains or Path. Crucially, it is not possible to set a cookie for a different host or access a cookie set by another host.

Once set, all future requests by the user, from that web browser, to the same host will include the contents of the cookie in the request. As such, all future requests, until the cookie is deleted, will be associated with the same user.

First-party, Third-party, and Intermediary Cookies

First-party and Third-party Cookies

The cookies described above can be considered as first-party cookies, they are set by, and retrieved by, the primary server that is being interacted with. This is illustrated in Figure 2, showing how the server sets the cookie value of "ABC1234" and that value is subsequently sent with future requests for page resources and even other pages on the same server.

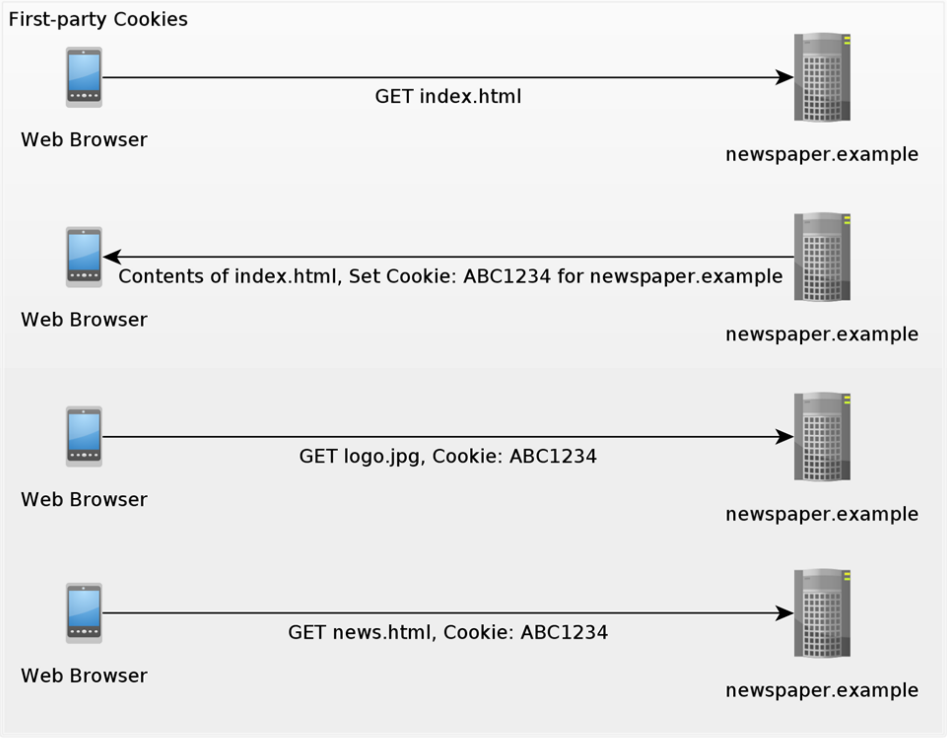


Figure 2: Example of setting and retrieving first-party cookies

In contrast, third-party cookies are set by hosts other than the main host being interacted with. Such usage is possible because a single web page may be made up of resources from multiple different hosts; the images may come from one host, the JavaScript from another, and the fonts from yet another. When requests are made to those third-party hosts for resources, the browser will send to the server any cookies already set for that third-party host, and will allow the third-party to set further cookies in the response.

Figure 3 illustrates the process of setting third-party cookies, in which the *index.html* page on *newspaper.example*[23] references JavaScript on a third party ad server, *ad-server.example*. When the *index.html* is loaded the browser will make a request to *ad-server.example* for the JavaScript and in doing so permit *ad-server.example* to set a cookie in the browser. *newspaper.example* includes the JavaScript from *ad-server.example* on its website to serve targeted advertisements to its users. *newspaper.example*'s earnings from such advertisements is partly, or possibly even wholly, dependent on visitors to its website clicking on the displayed advertisements. As such, *newspaper.example* is motivated to assist *ad-server.example* in delivering the most targeted advertisements possible, based on the assumption that targeted advertisements are more likely to be clicked on.

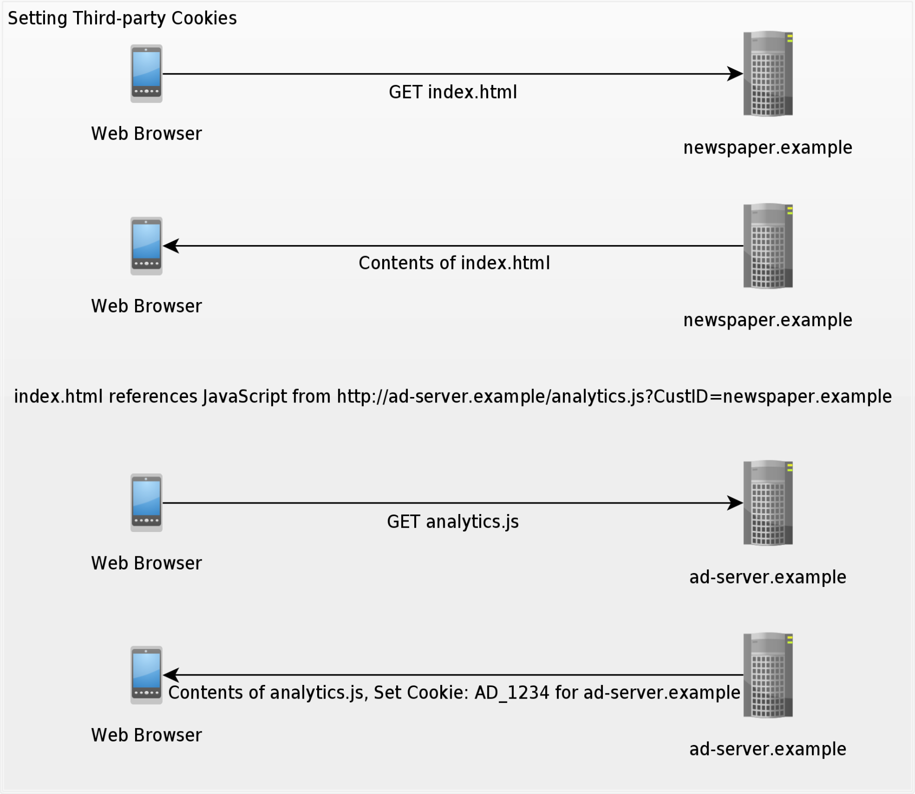

Setting Third-party Cookies
Web Browser
GET index.html
newspaper.example
Web Browser
Contents of index.html
newspaper.example
index.html references JavaScript from http://ad-server.example/analytics.js?CustID=newspaper.example
Web Browser
GET analytics.js
ad-server.example
Web Browser
Contents of analytics.js, Set Cookie: AD_1234 for ad-server.example
ad-server.example

Figure 3: Setting third-party cookies

[23] *.example* is used in place of *.com* throughout the report to ensure no references are made to active websites. *.example* is a reserved Top Level Domain that can be used for such purposes.

This functionality has been exploited by advertising and analytics companies to facilitate cross domain tracking. Cross-domain tracking is when a user can be tracked across different domains, or websites. For example, if a user visits a newspaper website followed by a shopping website, and they can be identified via an online identifier as the same user on both websites, the identifier allows cross-domain tracking to take place. Such tracking provides valuable data on a user's browsing history and allows for tracking the effectiveness of advertisements and more detailed profiling.

Figure 4 illustrates what happens if two different websites (*newspaper.example* and *bigshop.example*) both reference JavaScript from *ad-server.example*. When the user browses *bigshop.example* after browsing *newspaper.example*, the same cookie value set by *ad-server.example* on *newspaper.example* will be sent in the request. Note that the request for the JavaScript can include an identifier for the website being visited, for example, "CustID=bigshop.example". This allows *ad-server.example* to track the same user across both *newspaper.example* and *bigshop.example*. The scope of the information that is collected by the third-party is at the discretion of the first-party and depends on the commercial relationship between the two parties. It should be noted that such tracking is only possible because of the cooperation of the first-party websites, in this instance *newspaper.example* and *bigshop.example*.

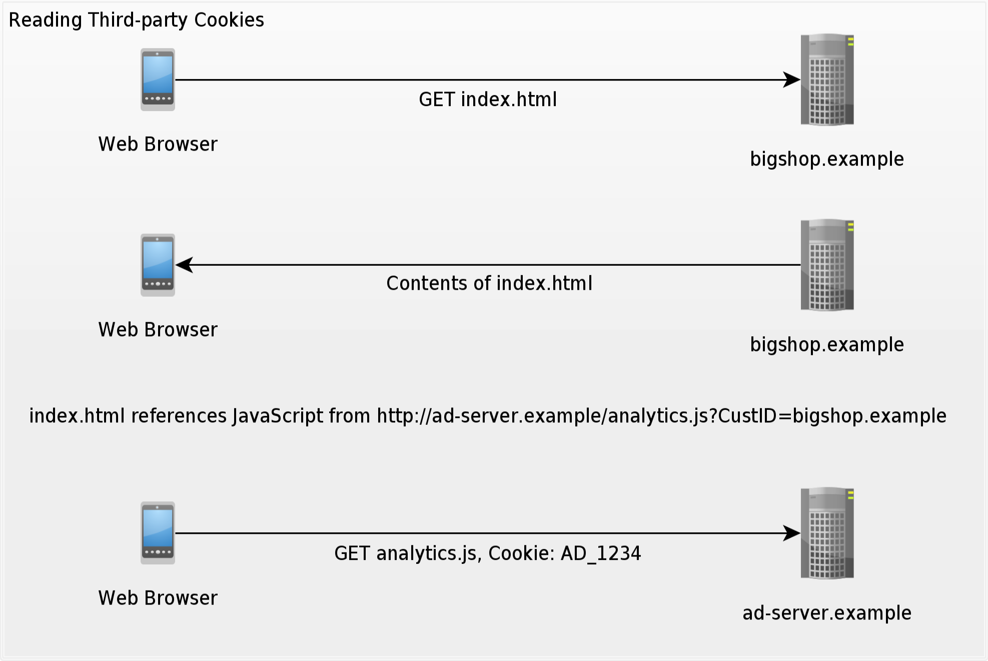

Reading Third-party Cookies
Web Browser
GET index.html
bigshop.example
Web Browser
Contents of index.html
bigshop.example
index.html references JavaScript from http://ad-server.example/analytics.js?CustID=bigshop.example
Web Browser
GET analytics.js, Cookie: AD_1234
ad-server.example

Figure 4: Retrieving third-party cookies

ad-server.example knows that the user visited both sites and can use that information in its profile of the user. *bigshop.example* and *newspaper.example* do not necessarily have direct access to this information, unless *ad-server.example* shares it with them. The reason *bigshop.example* and *newspaper.example* cooperate with the data collection is that it allows more targeted advertising to be shown, and *bigshop.example* and *newspaper.example* stand to benefit if more users click on the adverts shown. Additionally, they may have data sharing agreements in place that allow targeting by both organisations directly, in which case, profile data can be shared in the background between each pair of parties involved, or even all three parties together.

The risks associated with third-party cookies were known back when cookies were first conceived in 1994.[24] There was even consideration for the blocking of third-party cookies. However, the risk of permitting them was considered to be the lesser of two evils. Lou Montulli, the inventor of the cookie, justified the decision to not block third-party cookies in two ways:

> "Any company that had the ability to track users across a large section of the web would need to be a large publicly visible company. Cookies could be seen by users so a tracking company can't hide from the public. In this way the public has a natural feedback mechanism to constrain those that would seek to track them.

If 3rd party cookies were disabled ad companies would use another mechanism to accomplish the same thing, and that mechanism would not have the same level of visibility and control as cookies. We would be trading out one problem for another".[25]

Point one has turned out to be true in some regards, in that many people are aware of the tracking that takes place, and the companies undertaken such tracking are large publicly listed organisations. However, the scale of the tracking, and the capability of the users to defend against, or boycott companies that engage in such tracking, appears increasingly debatable. This is a particular problem where dominant players occupy a market, for example Google, Facebook, Amazon, who dominate the search and advertising, social media, and online shopping markets respectively. Where there is no viable alternative, the capability of the user to exercise the natural feedback mechanism is limited.

The relevance of point 2 is gaining prominence as browser makers address concerns with third-party cookies by blocking them by default.[26] As such, the risks from alternative tracking mechanisms is likely to increase. An overview of some of those possible alternatives can be found below at *Non-cookie based tracking and identifiers*.

By default, whether a cookie is a first-party or third-party cookie is not fixed.[27] As such, a cookie that is initially set as a first-party cookie becomes a third-party cookie if another website accesses resources from the host that originally set the cookie. This is further used as a method of tracking logged-in users across the internet. For example, if a user logs into a social media account the cookie will be set as a first-party cookie during the login process. However, if that user later accesses a website that has resources from the same

[24] See https://montulli.blogspot.com/2013/05/the-reasoning-behind-web-cookies.html

[25] See https://montulli.blogspot.com/2013/05/the-reasoning-behind-web-cookies.html

[26] See https://webkit.org/blog/10218/full-third-party-cookie-blocking-and-more/

[27] It is possible for security reasons to lock a cookie to being first-party to prevent against an attack known as Cross-Site Request Forgery. Such restrictions are normally found on login cookies for secure services like e-commerce and banking. Further examination of this is beyond the scope of this report. However, it should be noted, the restriction is at the discretion of the party setting the cookie, not the user.

social media company, for example, a Like or Tweet button, that first-party cookie will become a third-party cookie in the context of the other site.

Even with third-party cookies blocked, most browsers continue to send first-party cookies in a third-party context. This has led to the use of redirect bounces to set first-party cookies, which are subsequently used for third-party tracking. For example, a user may request *newspaper.example*, but they are immediately redirected to *ad-server.example/?redir= newspaper.example*. When the page on *ad-server.example* loads it sets a first-party cookie and then immediately redirects back to *newspaper.example*. Once set, the *ad-server.example* cookie will be sent to any requests on any page that references a resource on *ad-server.example*, even with third-party cookie blocking enabled. This is type of tracking is sometimes referred to as a first-party bounce tracker.[28]

newspaper.example cooperates in this redirecting because it benefits from earning more money from showing targeted adverts, or potentially more detailed profiling of their users for further analytics. There is a mutual benefit to greater data acquisition for both the website containing the tracking scripts, and the trackers themselves. For example, *newspaper.example* may receive detailed analytics about their users based on the data collected by *ad-server.example*, providing a broader picture than could be collected by *newspaper.example* alone.

The techniques for performing the redirects can be quite sophisticated. For example, JavaScript can be written to monitor when the page loses focus, indicating the user is moving to a different tab or window. By listening for this event the advertising script can trigger the redirect bounce in the background, leaving the user completely unaware of it.

There are techniques for blocking such cookies, including Firefox's First Party Isolation,[29] although such techniques are not currently enabled by default because they have led to the breaking of some popular web services, including federated login – where a user logs into one site via a third-party, for example, Facebook Login or Google Login.[30,31] Safari also provides some blocking via its Intelligent Tracking Protection.[32]

Intermediary Cookies

Conventionally only first and third-party cookies are considered, but increasingly there is another kind of cookie being set. This report has termed them *intermediary cookies*, on the basis they are set by a party acting as a network intermediary for the host. Crucially, they appear as first-party cookies, in that they relate to the original host, but are read and retrieved by the intermediary, sometimes without the host even having visibility of them. This is possible because of something called a TLS Proxy, more commonly found in Content Delivery Networks, Web Application Firewalls, and DDoS protection providers. In essence,

[28] See https://webkit.org/blog/8311/intelligent-tracking-prevention-2-0/
[29] See https://www.ghacks.net/2020/04/17/mozilla-adds-dynamic-first-party-isolation-option-to-firefox-77/
[30] See https://www.ctrl.blog/entry/firefox-fpi.html
[31] See https://blog.mozilla.org/data/2018/01/26/improving-privacy-without-breaking-the-web/
[32] See https://webkit.org/blog/9521/intelligent-tracking-prevention-2-3/

the TLS Proxy intercepts all requests made for resources on the host, inspects them for security risks, and then forwards to the host only the requests deemed to be safe.

It is necessary for the TLS Proxy to keep track of users for a number of reasons, including security and networking routing. This is achieved by setting a cookie. To set this cookie the TLS Proxy injects JavaScript and Cookie setting headers into the responses as they pass through from the origin server. As such, the cookie appears to be a first-party cookie from the host, when in fact it is coming from the intermediary. Such cookies are not restricted to purely network and security, they are also used for some analytics, although assertions are made that cookie IDs are not synced across sites or included in profile information.[33] However, as third-party cookies start to be blocked the potential value in these intermediaries could increase.

Cookie Syncing

The restrictions placed on cookies to only allow them to be read by the host that set them makes sharing Cookie IDs intentionally difficult. Cookie ID Syncing is a process that attempts to bypass this restriction, allowing two different parties to share their respective online identifiers, facilitating subsequent sharing and linking of their respective profile data. It is a popular approach used by third-party advertisers to boost the value of their data through exchanging profile data and to allow co-ordinated targeting of a user. It is primarily used with third-party cookies, but the techniques could be used to sync non-cookie based identifiers in the future, which could become more important once first-party data collection becomes the norm following third-party cookie blocking.

There are various approaches to cookie syncing,[34,35] many are automated using Machine Learning and AI to link different datasets. However, the easiest approach is if the IDs can be explicitly linked by the platforms themselves.

A common approach is to use pixel beacons (see discussion below at *Web Beacons/Pixel Tags* for further details on what they are). A simple example would be two organisations; *ad-server.example* and *ad-partner.example*, both having cookies set in a user's browser and wanting to synchronise their IDs, so they can exchange the data they both hold on that user. The user navigates to *newspaper.example*, which contains a third-party ad from *ad-server.example*. That ad triggers a request for a pixel beacon from *ad-server.example/beacon.gif*. As normal, the request will contain the ID from the cookie associated with *ad-server.example*. *ad-server.example* responds with a redirect to *ad-partner.example* that includes the userId from *ad-server.example* as a URL parameter. For example, *ad-server.example/beacon.gif?userId=abc123*. The browser will follow that redirect and send the cookie associated with *ad-partner.example* in the request. In doing so, *ad-partner.example* can extract the userId set by *ad-server.example* from the URL and map it to its own userID from its cookie. The mapping can be mutual if *ad-partner.example*

[33] See https://support.cloudflare.com/hc/en-us/articles/200170156-Understanding-the-Cloudflare-Cookies#12345682
[34] See https://docs.adobe.com/content/help/en/id-service/using/intro/match-rates.html
[35] See https://developers.google.com/authorized-buyers/rtb/cookie-guide

redirects in the same way back to *ad-server.example* with its userID in the URL. At the end of the process both *ad-server.example* and *ad-partner.example* have mapped their IDs to each other and can now perform a server-to-server exchange of the data they hold.

Cookie syncing is an extremely common occurrence.[36] A web census study in 2016 found that the majority of the most popular third-parties "sync cookies with at least one other party: 45 of the top 50, 85 of the top 100, 157 of the top 200, and 460 of the top 1,000".[37] The impact this has on users was analysed in a further study in 2018, which found that:

> "...97% of the regular web users are exposed to Csync [cookie synchronisation]: most of them within the first week of their browsing, and the median userID gets leaked, on average, to 3.5 different domains".[38]

Despite the prevalence of cookie syncing it remains under researched, with a general lack of transparency making monitoring and analysis difficult to establish.

Session vs Persistent Cookies

As well as being restricted by host, cookies can have a lifetime set, either via a MaxAge attribute or an expiry time; such cookies are known as persistent cookies. They will be stored by the browser and will continue to be available until they expire or are deleted. As such, they will survive machine and browser restarts.

When no lifetime is set the cookie is considered to be a session cookie. Such cookies are supposed to only last the lifetime of the browser session, in other words, when the browser is closed session cookies should be discarded. As such, the lifetime of a session cookie is determined by the lifetime of the browser session.

Dependence on the lifetime of a browser session can lead to results that may not be expected by the user. In an age of tablets, phones, and laptops that are rarely rebooted, and apps that remain open in the background, a browser session could last days, weeks, or even months. It is important to note, that the session is associated with the browser, not the tab. Closing a tab in the browser does not close the session. Even those using Private Browsing or Incognito mode can get caught out by this behaviour. Such modes only clear their contents when ALL Private Browsing tabs are closed. As such, if someone logs into their social media account in a Private Browsing window and keeps it open for extended period of time the benefits of Private Browsing can be diminished.

Further issues occur when the browser is set to restore tabs at start-up. When selected, the browser session is saved during shutdown and restored at restart. What may be

[36] See https://www.iabuk.com/opinions/closer-look-third-party-cookie-matching
[37] Steven Englehardt and Arvind Narayanan. "Online tracking: A 1-million-site measurement and analysis." Proceedings of the 2016 ACM SIGSAC conference on computer and communications security. 2016. Extended Version See https://webtransparency.cs.princeton.edu/webcensus/
[38] Panagiotis Papadopoulos, Nicolas Kourtellis, Evangelos P. Markatos, "Cookie Synchronization: Everything You Always Wanted to Know But Were Afraid to Ask", available at https://arxiv.org/abs/1805.10505

unexpected is that the entire session is restored, not just the data associated with tabs being restored. As such, the session could last weeks, months, years, or even for the lifetime of the machine. The restore feature is a popular function within browsers; concerns about the potential privacy and security issues associated with restoring of session data have been known about for over a decade.[39,40] It remains unknown whether end users are aware of the full consequence of enabling the feature.

Cookie usage

As has already been mentioned, there are a number of uses for cookies, many of which are legitimate uses, some of which are essential to the functioning of the web. The challenge is in correctly categorising cookies and their respective purpose and intent. The purposes of the cookie are not mutually exclusive, for example, a network management cookie may also be used for analytics. Likewise, a cookie's usage can be repurposed to provide additional tracking capabilities at a later date.

Personalisation

Cookies can be used to store personalisation data, for example, language, colour options, preferences, etc. An increasingly common example in the European Union is storing cookie preferences. When the user provides or denies consent for the storing of cookies, that very preference is often stored in a cookie. On the face of it such usage is benign. However, when the set of preferences themselves can form an identifier, a preference cookie can be repurposed for tracking. These cookies are normally persistent cookies with a long lifetime.

Session Management

As was discussed above at *Why are cookies needed?*, the original reason for cookies was to create sessions that could track a single user's interaction with a website. This form of tracking is essential to how the web works, and blocking it would break many popular web services. Most dynamic websites will create sessions for users when they first visit the website, irrespective of whether the user has an account with the website. This allows showing of personalised content, for example, on a news website other stories of interest will be shown based on the stories viewed during the current session. Likewise, with e-commerce, the ability to add items to a shopping basket before having to register or login.

Such cookies are extremely common. Provided that the data collected during the session is not linked to a longer term identifier, and the session cookies are regularly cleared from the browser, the privacy risk from such cookies remains low.

[39] See https://bugzilla.mozilla.org/show_bug.cgi?id=345345
[40] See https://bugs.chromium.org/p/chromium/issues/detail?id=128513

Login Management

An extension of session management is login management, or authentication management. In this approach a cookie is used to store an authentication token that permits the user to take actions as an authenticated, logged-in, user. This is an essential part of the web, since users would not want to re-enter their username and password on every request. It was normally considered good practice to set login management cookies to expire at the same time as the login session expired, i.e. when the user is logged out automatically after 20 minutes. It is not strictly necessary from a security point of view, since the login should be expired by the server in the any case.

However, increasingly, longer life authentication tokens are being used to facilitate third-party tracking. This exploits the ability to receive third-party cookies that were originally set via a first-party interaction. As such, logging into a social media site will result in a long-life authentication token being set, which will then be tracked via social media buttons and widgets on other websites to enable widespread, cross-domain tracking. This is a good example of how an essential cookie is repurposed for tracking purposes. For example, Google's authentication cookies don't expire for two years.

Network Management

Network management cookies are primarily used to assist correct routing of traffic. They are sometimes referred to as load balancing cookies, although there are broader uses than just load balancing. As was described above at *Why are cookies needed?*, HTTP is stateless and therefore requests are treated independently. This creates a problem when an organisation has a large web presence and runs multiple servers across different locations and regions. It can be necessary for both performance and functionality to have requests from the same client be routed to the same server. To achieve this, when the client first connects, or passes through the TLS Proxy/Content Delivery Network (CDN), a cookie is set defining which server the client will connect to. When future requests are sent they will include the cookie and allow the CDN or load balancer to route the request to the same server.

This allows for more efficient usage of caches and simplifies the process of updating data stored on the server. In some instances it is essential for the operation of a protocol. For example, WebSockets create a two-way link to the server to facilitate services like push notifications when new emails or messages arrive. WebSockets are stateful connections between the server and the client, and as such, they require all requests to be routed to the same server.

Even though such cookies are inherently sessional in nature, they can often be set as persistent cookies with short lifetimes. This allows the expiration of the routing to occur at the same time the user's session on the server expires.

If used correctly, such cookies should not contain information that could identify the user, it should only encode the target server to which the traffic is being routed.[41] If that is the case there is no risk to user privacy as the contents is neither unique nor about the user. However, some services will use an identifier in the cookie instead of a direct mapping, and then have a look-up table on the load balancer. In these situations the cookie can act as an identifier and therefore presents a privacy risk.

Security

There are number of uses for cookies to help protect websites from attack (see discussion above at *Intermediary Cookies*). Broadly speaking the cookie is used to establish reputation and therefore trustworthiness of the user. As such, the cookies often contain an identifier that is used to profile behaviour over a period of time, possibly only across a session, or possibly over a longer period. One such example is monitoring the number of requests coming from a single user. If the user exceeds a threshold they may be temporarily denied access or required to complete a challenge in order to continue to access the service.[42] This is done to prevent overwhelming the server from attacks like Distribute Denial of Service[43] attacks, and automated bot-net attacks. As such, they do perform an essential service in protecting websites.

However, the lines between monitoring for security and profiling can become blurred. A good example is ReCAPTCHA,[44] something most users will be familiar with, having probably been asked to decipher garbled text or identify elements in a photo before. They are designed to test the request is coming from a human and not an automated bot. However, the service is run by Google and in order to operate requires allowing third-party cookies, including the cookie named NID, which google itself classifies as being used for advertising.[45]

ReCAPTCHA v3 is potentially worse, since it removes the challenge part and instead requires the website to include the ReCAPTCHA v3 script on every page. The ReCAPTCHA service monitors user behaviour to determine what “normal” behaviour on the website is and calculates a per user score. Google looks for its own cookies, and if found provides a lower risk score on the basis that a request is more likely to be from a human if there is an associated Google account. As for what other data is collected it is not specified by Google in either the Privacy Policy[46] or Terms of Service.[47]

Security cookies are often persistent cookies, and frequently third-party cookies, potentially set via first-party redirects. Whilst they clearly have a purpose in protecting websites from

[41] See https://developers.cloudflare.com/load-balancing/understand-basics/session-affinity/
[42] See https://www.cloudflare.com/en-gb/learning/ddos/ddos-mitigation/?utm_referrer=https://www.cloudflare.com/en-gb/learning/ddos/what-is-a-ddos-attack/
[43] See https://www.cloudflare.com/en-gb/learning/ddos/what-is-a-ddos-attack/
[44] See https://www.google.com/recaptcha/intro/v3.html
[45] See https://policies.google.com/technologies/types?hl=en-US
[46] See https://policies.google.com/privacy?hl=en
[47] See https://policies.google.com/terms?hl=en

cyber attacks, their dual purpose for profiling and close ties to advertising cookies makes them a potential high risk category of cookie.

Tracking

Tracking cookies present the greatest privacy concern given their objective is to track users across multiple domains. They are a staple of online analytics and marketing. Such cookies will typically contain a randomly generated identifier. That identifier is the index to a detailed profile built by the issuer to better discern, segment, and target the user. Such cookies are almost universally persistent cookies to facilitate the long term tracking, often with lifetimes in years. Tracking is dominated by a few big players, including Google, Facebook, and Twitter, each of which have the advantage of being able to easily set first-party cookies through normal usage, thereby bypassing restrictions on third-party cookie setting.

Non-cookie based tracking and identifiers

Awareness of the privacy invasiveness of cookies has led to increased blocking of some kinds of cookies. The approaches taken have varied, from partial blocks[48], AI-based intelligent blocking,[49] to minimal default blocking in Google Chrome.[50] However, in January 2020, Google announced that its browser Chrome would also start to block third-party cookies,[51] although not until 2022. It should be noted that Google's intention is to replace third-party cookies with a new technology they have termed privacy sandboxes.[52,53]

Google's Privacy Sandbox proposal is still in its infancy, but at a high-level it aims to shift the storage and processing of user data from the third-party into the browser itself. By using various advanced technologies, including machine learning and possibly differential privacy (see further discussion below at *Advances in Technology*), the aim is to allow profiling and targeting, but without allowing advertisers to collect data about, or track, individuals.

There are currently a number of Application Programming Interfaces (APIs) being proposed and defined. The following provides a brief summary of each:

- Trust Tokens API – used to establish trust in a user by allowing anonymous credentials to be set and shared.
- Click Through Conversion Measurement Event-Level API – used to measure when a user clicks on advertisements without tracking the individual user.
- Aggregated Reporting API – stores data locally until sufficient data exists across multiple browsers and users to allow privacy preserving aggregation on a server.

[48] See https://blog.mozilla.org/blog/2019/09/03/todays-firefox-blocks-third-party-tracking-cookies-and-cryptomining-by-default/
[49] See https://webkit.org/blog/9521/intelligent-tracking-prevention-2-3/
[50] See https://blog.cryptographyengineering.com/2018/09/23/why-im-leaving-chrome/
[51] See https://blog.chromium.org/2020/01/building-more-private-web-path-towards.html
[52] See https://www.blog.google/products/chrome/building-a-more-private-web/
[53] See https://blog.chromium.org/2019/08/potential-uses-for-privacy-sandbox.html

Useful for conducting trends analysis on large groups.

- Federated Learning of Cohorts (FloC) – groups users together under shared interests based on local browsing history. Advertisers can profile and target the group as a whole instead of the individuals that make up the group.
- Two Uncorrelated Requests, Then Locally-Executed Decision On Victory (TURTLE-DOV) – the browser holds the information about what an advertiser believes the individual is interested in. This information, combined with contextual information, is used to target an advert, with the decision-making taking place in the browser itself.
- WebID – Improved federated login, which is where a user can login into one site by using their username and password for another site, for example, Facebook Login.

The proposals are still at a very early stage and are subject to change. In principle, they may offer some advantages, particularly in regard to keeping the data in the browser. However, they are dependent on the effective blocking of alternative tracking technologies.[54] Furthermore, there needs to be further consideration of whether the techniques proposed can be disabled by users, for example, removing or clearing groups that have been assigned through the FloC API. The proposals are highly technical and whether they can be implemented in a way that is privacy preserving and not open to abuse remains unknown. As such, Google's announcement is not an abandonment of tracking, but a change in how it will be conducted.

Increased blocking of third-party cookies has led to development of alternative methods for recording and retrieving online identifiers. It should be expected that such developments will continue,[55] and may even accelerate as widespread third-party cookie blocking becomes a reality. As was warned by Lou Montulli,[56] the inventor of the cookie, it is possible that the alternative approaches developed will be harder to detect and prevent.

The following techniques are a summary of some of the most well-known or recent developments. This list should not be considered to be exhaustive; if anything the most advanced techniques in use today are probably not widely known about, since secrecy of the approaches is necessary to prevent advertising and tracking blockers from finding and blocking them.

Alternative approaches can be broken down into two categories: *stateful* and *stateless*. *Stateful* tracking attempts to set an explicit identifier somewhere in the browser that can be later be retrieved and therefore allow the tracking of the same user between requests and possible sessions. *Stateless* tracking by contrast does not attempt to set an explicit identifier, instead relying on fingerprinting of the device, the user, or the user's behaviour to form an identifier. In the case of device fingerprinting, such identifiers could allow tracking across different browsers. Where user or behaviour fingerprinting is performed it could allow tracking across devices. Furthermore, such identifiers are not limited by privacy protection

[54] See https://www.chromium.org/Home/chromium-privacy/privacy-sandbox
[55] See https://www.idimension.com/2018/08/defeat-itp-2-0-with-gtm-and-conversion-linker-sitewide-tagging/
[56] See https://montulli.blogspot.com/2013/05/the-reasoning-behind-web-cookies.html

measures like private browsing, potentially allowing tracking even in contexts a user may believe to be private.[57]

Stateful Tracking Technologies

Web Beacons/Pixel Tags

Web beacons have already been mentioned above at *Cookie Syncing*. Originally web beacons were single pixel images coloured to be either the same colour as the background or entirely transparent. As already covered, they would be requested from a third-party domain during the loading of the page, and in doing so would reveal to that third-party any cookies stored in the browser for the tracking host, as well as metadata about the user and network request, and potentially additional information included in the URL by the first-party, as demonstrated by their use in cookie syncing.

Whilst the beacons were originally invisible, in recent years they have broadened to include visible elements and widgets, in particular social media buttons. When a website includes such buttons on their page, the button itself acts as a web beacon, sending valuable analytics data to the social media platforms. The effectiveness of web beacons is currently dependent on the sending of third-party cookies to allow the beacon to link the request to a previously set identifier. If third-party cookies are blocked in their entirety[58] beacon usage will need to change in order to continue to be effective.

Local Shared Object (Flash cookies)

Similar to conventional cookies except they are stored outside of the normal cookie storage. More commonly known as Flash Cookies, since the cookies are set via the Adobe Flash plug-in. The plug-in maintains its own cookie storage outside the web browser. Originally such cookies were not removed during clearing of browsing data, however, since 2011 an interface (API) has been provided to allow browsers to delete Local Shared Objects as well.

Flash Cookies have fallen out of favour in recent years for a number of reasons, including unpopularity with consumers,[59] provision of easier ways to delete them,[60] declining support for Flash in web browsers.[61] However, they are still regularly mentioned in privacy and cookie policies. It is currently unclear whether third-party blocking of cookies includes flash cookies, however, the end of Flash support in 2020[62] should resolve the issue in any case.

[57] See https://www.forbes.com/sites/daveywinder/2020/06/03/google-chrome-privacy-lawsuit-could-you-get-a-5000-payout-incognito-mode-class-action/#40385f3b1485
[58] This would include blocking first-party cookies in a third-party context
[59] See https://www.nytimes.com/2010/09/21/technology/21cookie.html
[60] See https://bugzilla.mozilla.org/show_bug.cgi?id=625495
[61] See https://bugzilla.mozilla.org/show_bug.cgi?id=1519434
[62] See https://theblog.adobe.com/adobe-flash-update/

ISP UIDH (Super cookies)

Internet Service Provider (ISP) Unique Identifier Headers (UIDH), sometimes referred to as super cookies are injected into HTTP requests by the ISP.[63] Due to HTTP connections not being secure it is possible for the ISP to modify and even add additional headers to the requests passing over its network. This has been used to embed a unique identifier, or in the case of Optus[64] and Telstra,[65,66] the phone number of the subscriber.

In the instance of Optus and Telstra the purpose was for Direct Carrier Billing, which allows service providers to charge a customer via their mobile phone bill for services such as downloads of mobile games or ringtones. The use of HTTP header injection came to light when customers discovered they had been subscribed to services without having provided their phone number to the service provider.[67] This was possible because the phone number was included in the HTTP header by the ISP, allowing the charge to be made after just one or two clicks, and seemingly without explicit notification to the customer. The issue of Direct Carrier Billing has been a controversial one in Australia, with the ACCC taking action against number of ISPs, including Optus,[68] Vodafone,[69] and Telstra,[70] which has led to a number of the providers ending the billing service. However, the injection of the identifier in the HTTP header has not been specifically addressed, nor prohibited in future alternative scenarios, for example targeted advertising.

The ISP typically sells access to such injections by signing partnership deals with third-parties to enable injection into requests for their websites. The identifier is not stored on the client device and there is no way to delete or prevent it being injected. Using TLS (HTTPS) helps to some degree but can be circumvented by the receiving website by triggering an insecure HTTP request in order to receive the identifier. Once received this can form the starting point for a zombie cookie (see further discussion below at *Zombie Cookies*), since the same identifier could then be stored in conventional web cookies. To avoid such injection a user would need to use a Virtual Private Network (VPN) provider to encrypt traffic before it travels across the ISP network.[71]

Browser Cache

Web browsers include a cache to hold resources downloaded during the viewing of a website. For example, if a user visits a page with an image in it, the downloaded image will be stored in the browsers cache. If the user returns to the same page, or the resource is referenced on another page it will be loaded from the cache instead of having to be downloaded again. The website administrator is able to set how long resources should be

[63] See https://www.eff.org/deeplinks/2014/11/verizon-x-uidh
[64] See https://www.itnews.com.au/news/optus-admits-handing-user-phone-numbers-to-websites-405656
[65] See https://www.abc.net.au/news/2015-06-26/optus-and-telstra-sharing-mobile-phone-numbers/6575424?nw=0
[66] See https://www.itnews.com.au/news/optus-admits-handing-user-phone-numbers-to-websites-405656
[67] See https://forums.whirlpool.net.au/archive/2418905
[68] See https://www.accc.gov.au/media-release/optus-penalised-10-million-for-misleading-customers-over-digital-purchases
[69] See https://www.accc.gov.au/media-release/vodafone-to-compensate-customers-over-direct-carrier-billing-charges
[70] See https://www.accc.gov.au/media-release/telstra-refunds-93m-to-72000-customers-0
[71] See https://nordvpn.com/blog/super-cookies-going-global/

cached for, up to maximum recommended time of 1 year, although some browsers will accept longer cache times.

The cache mechanism can be used to set third-party identifiers in a manner that is even more powerful than cookies. For example, if the page on *newspaper.example* contains tracking JavaScript it can create a reference to a further JavaScript file: *ad-server.example/userId.js*. If that file does not exist in the browser cache a request will be made to the server. When *ad-server.example* receives the request it generates a new random ID and returns it as a variable in the JavaScript file, with a cache time of 1 year. The tracking script can then read this variable and include it as URL parameter or set it as first-party cookie on the site. If the user revisits *newspaper.example* at a later date, or any other site that includes the same tracking script, the browser will load the cached script instead of requesting it from the server, and the same ID will be propagated across domains and sessions.

As a result of the browser cache being a shared cache, the same host restrictions do not apply. There are restrictions on not loading cached content between HTTP (not secure) and HTTPS (secure) pages, but there are no restrictions on which site may access a resource in the cache. It is difficult to block such usage of the cache without unduly effecting the browsing experience. Browser caches can and are cleared, but it often requires an active choice by the user, or for the user to be using private browsing which uses a separate cache that is cleared when all private browsing tabs are closed.

ETag and Last-Modified Dates

An extension of the Browser Cache technique is to use ETag or Last-Modified dates to encode identifiers. Both attributes are used by the browser cache to determine whether the cached resource is current or needs to be updated form the server. When a resource is first loaded the server can return either an ETag header, an (almost) arbitrary length[72] value that is intended to uniquely identify the version of a resource, or a Last-Modified header indicating the time the resource was last modified, along with a cache time. When the cache time expires, if the resource is requested again, rather than immediately downloading the resource, the browser will revalidate the resource by sending the existing ETag or Last-Modified date to the server to check if the resource has actually been updated or if it can continue to use the existing cached copy.

As such, the ETag is conceptually similar to a cookie.[73] It allows the server to set an arbitrary value on the client that will automatically be returned to the server when a request for that resource is made. Much like the browser cache technique, it is cross-domain accessible, and by setting a cache time of 0, all requests for the resource will be revalidated and send the ETag to the server. It is difficult to detect when ETags are being used for tracking, as the identifiers are normally arbitrary, for example, a hash of the contents of the file. Disabling ETags would have a significant impact on caching and negatively impact on the user

[72] No strict limit is placed on the length of the ETag in the header, but it is generally expected that the entire HTTP header will not exceed 8kb.

[73] See https://www.chromium.org/Home/chromium-security/client-identification-mechanisms#TOC-Cache-metadata:-

experience. ETag identifiers were first found in the wild in a 2011 study that found Hulu.com was using the KISSmetrics service, which was setting ETag identifiers.[74] A subsequent lawsuit on the matter was filed and subsequently settled.[75]

However, the approach of using ETags is still viable and is even described by some as a way of tracking in a post cookie environment.[76] If this is not already being extensively used as a mechanism for storing identifiers it is likely to gain attention when third-party cookies are blocked. In many regards it is more powerful than third-party cookies, and therefore presents a high risk to privacy.

Local Storage

Local storage encompasses all forms of data storage on the browser. Some have already been covered in this report, notably cookies and Local Storage Objects. However, there are a range of alternative mechanisms for storing data, including HTML localStorage, Service Workers, IndexDB. Each provide slightly different functionality, but at the core are techniques for persisting data in the browser of the user for future reference. The type of data varies by application, but could include an identifier, an offline copy of an email or message, saved game data, or any other application data. The ability to store arbitrary data in the browser presents a risk of tracking via the storage of identifiers. Even when restrictions are placed on cross-domain access to storage, for example, preventing *newspaper.example* from accessing the local storage set by *ad-server.example*, there are workarounds, for example iFrames and postMessages, which in effect allow sharing of information or identifiers between domains.[77] The same technique can be used to share cookie data as well.

Device Identifiers

Many devices including laptops, desktops, mobile phones, and tablets will contain a series of device identifiers. These identifiers fall into two categories: resettable and non-resettable. Examples of resettable IDs would be the device wide advertising identifier, or an app generated ID. Examples of non-resettable IDs would be the IMEI (International Mobile Equipment Identity) or DeviceID.[78] Up until Android 10 it was possible to access these IDs via a common permission requested by many apps, which was needed to monitor the phone state, for example, whether a call was incoming. Android 10 enforces much stronger restrictions on accessing non-resettable IDs, preventing normal apps from accessing them altogether.[79]

[74] Ayenson, Mika D and Wambach, Dietrich James and Soltani, Ashkan and Good, Nathan and Hoofnagle, Chris Jay, “Flash Cookies and Privacy II: Now with HTML5 and ETag Respawning” (July 29, 2011). Available at SSRN: https://ssrn.com/abstract=1898390 or http://dx.doi.org/10.2139/ssrn.1898390

[75] See https://www.wired.com/2012/10/kissmetrics-tracking/

[76] See https://ensolvers.com/blog/user-behavior-tracking-using-etags/

[77] See https://www.simoahava.com/analytics/cookieless-tracking-cross-site-iframes/

[78] See https://developer.android.com/training/articles/user-data-ids

[79] See https://developer.android.com/about/versions/10/privacy/changes

The concern with collecting non-resettable identifiers led to the introduction of an advertising identifier on both Android and iOS.[80] Such identifiers are randomly generated unique identifiers that facilitate the profiling and delivery of targeted advertising on the device. The system-wide nature of the advertising identifier makes it a powerful tracking identifier. In many regards it can be considered similar to a third-party cookie, except it is not possible to block it. There appears to be an inconsistent approach from Apple and Google with regard to third-party cookies and advertising identifiers. Greater control over the former is being introduced, but users continue to have little control in the case of the latter, and no indication that improvements will be forthcoming.

Whilst the advertising identifier is user resettable, most users may not be aware of how to reset it,[81,82] and even if they do, there is no ability to deny apps access to it, or disable it entirely. This has led to a recent claim it is not consistent with the GDPR.[83]

Whilst such identifiers are most common on mobile devices, Microsoft Windows also includes an advertising identifier that apps can use to track users on desktops and laptops.[84]

Digital Platform Identifiers

Digital Platform Identifiers are online identifiers associated with a user account on a digital platform, for example, Facebook. These are identifiers that are stored centrally, and retrieved when a user signs in or is detectable by the platform. These identifiers act as a link that cuts across technologies and devices. They are used extensively in cookies and beacon pixels to track users across the web. They are also used in apps on devices and via federated login.[85]

Digital platform identifiers are particularly powerful because they link supposedly anonymous identifiers to a known user, and as such the associated data stored on the platform itself. When the digital platforms are also dominant players in the advertising market this type of identifier creates an increased privacy risk, due to the depth and breadth of data collected on an individual in a single location. There are three phases in which digital platform identifiers might operate.

Pre-Login

During this phase any identifiers are not linked to a known individual, and are equivalent to random identifiers stored in a cookie. If the platform, *mybook.example*, operates as a tracker or advertising platform it will have set cookies within the browser whilst the user is browsing across the web. At this point the profile is only attached to the random identifier, and if the user were to clear their cookies the profile would cease, and a new profile would

[80] See https://support.apple.com/en-gb/HT205223
[81] See https://www.wikihow.tech/Reset-Your-Advertising-ID-on-Android
[82] See https://support.apple.com/en-us/HT205223
[83] See https://noyb.eu/en/complaint-filed-against-google-tracking-id
[84] See https://support.microsoft.com/en-us/help/4459081/windows-10-general-privacy-settings
[85] Federated login allows someone to login into a service you credentials from another platform, for example, Facebook Login: https://developers.facebook.com/docs/facebook-login/security/

be created. If the user does not clear those cookies a detailed profile will be constructed over time.

Post-Login Same Device

If the user subsequently logs into their account on *mybook.example* in their browser, *mybook.example* can immediately link the previously constructed profile to the user. From this point onward any future browsing will also be associated with the user, not just a random identifier. Even if the user clears their cookies, the profile remains attached to their account on the platform. Furthermore, each subsequent login will re-establish the link between any current identifiers and the user. In effect, this allows linking of what are supposed to be separate profiles, and maps the random identifier to an individual. The ability to re-establish the link between a random identifier and a user at each login makes it difficult for the user to enforce separation and protect their privacy. In effect, they would need to completely isolate their access to *mybook.example* to avoid unintentional profile linking, requiring the clearing of cookies and all other identifiers before and after each login.

Such isolation is not necessarily even possible if the user is using the *mybook.example* app on their phone or tablet. The ability to access non-resettable identifiers makes it impossible to enforce a divide, potentially allowing mobile app behaviour to be linked to their account without any reasonable way to prevent it in the future, short of buying a new device. As such, a single login event on a device could incur a significant ongoing privacy cost.

Post-Login Different Device

Furthermore, if the same user logs into *mybook.example* on a different device, for example, a work laptop, their phone, or tablet, *mybook.example* can now link the identifiers on that device, and any associated profiles held, with the user as well. This is one of the common ways in which cross-device tracking and profiling can be performed using conventional identifiers. Similar to the consequence of logging into the platform on the same device, the true privacy cost of enabling such linking may not be apparent to the user when they are logging into their account.

The dominance of Google and Facebook also raises concerns with regard to what has been termed the “walled-garden”. In effect, both Google and Facebook have a very accurate and reliable way of identifying and tracking their users across their own platforms. Because both platforms act as the ‘stepping-off’ point for their users, the platform’s own online identifiers give unprecedented ability to track browsing habits from within their own ecosystem. For example, clicking a link on Facebook is trackable by Facebook itself. If users primarily access content they find or see on such platforms the potential for tracking is high, and difficult to block.

Stateless Tracking

Browser or device fingerprinting

Browser or device fingerprinting uses the available attributes and properties of the browser or device to construct an identifiable fingerprint. Providing this fingerprint remains static over a reasonable period of time it can act as an identifier. It is important to note that with such fingerprinting the individual attributes are rarely identifying on their own, in fact, they could be considered to be relatively common. However, when combined into a set, it is the combination of those common values that can be unique. In the case of browser fingerprinting some common attributes that will be used to construct the fingerprint are:

- Screen resolution
- Colour depth
- Installed extensions
- Installed Fonts
- System information (platform, operating system)
- System language
- HTTP Header information

A more extensive list is available from the EFF Panopticlick project[86] as well as an online test of uniqueness of the browser.

This type of fingerprinting is common, in particular for service likes TLS Proxies/Web Application Firewalls, ReCAPTCHA. In recent years browser makers have reduced the information available via these attributes to prevent unique fingerprints being generated, and greater public awareness has seen a fall in usage. The Web Census discovered that HTML Canvas fingerprinting had fallen out of favour with prominent trackers, but was used on more sites, although by smaller players in the tracking market who may not have such a large public profile. Additionally, the tracking was increasingly used in the context of tracking users for security reasons.[87] However, the same study discovered three previously unseen tracking techniques: AudioContext Fingerprinting, WebRTC Local IP Discovery, and Canvas-Font Fingerprinting.

WebRTC Local IP Discovery is an example of network fingerprinting, in which attributes of the network connection, in this particular instance, local network information that should not have been accessible, are used to create a unique signature. The types of information collected could include device name, MAC Address, IP Address, routers, open ports, configured proxies, amongst many more. Many of these values are not permanent and can change, but due to the configuration in modern home networks, the values tend to remain static for a long time, thereby facilitating their use for tracking. This provides a way for the

[86] See https://panopticlick.eff.org/about

[87] See Steven Englehardt and Arvind Narayanan. "Online tracking: A 1-million-site measurement and analysis." Proceedings of the 2016 ACM SIGSAC conference on computer and communications security. 2016. Extended Version See https://webtransparency.cs.princeton.edu/webcensus/

website to acquire network metadata that it should not be able to access. (See further details below at *Metadata Collection and Usage*.)

To give an idea of the sophistication of the techniques being deployed, the AudioContext Fingerprinting used an Audio API, which was new at time, to generate an audio signal via an oscillator, which was then processed to generate a hash. Due to subtle differences in how the systems generated and processed the signal a device fingerprint was generated. It is not known how effective such fingerprinting is, since that would require a large-scale study to determine uniqueness. However, despite the novelty of the techniques at the time, various combination of those techniques were discovered on a range of major brand websites, including hotels.com, expedia.com, travelocity.com,[88] ancestory.com and dell.com[89] to name just a few.

Browser and device fingerprinting is an evolving area, with browser and device makers playing catch-up with the developments coming from tracking companies. One of the advantages that tracking organisations have is that they can try out new techniques on a large audience to determine accuracy and uniqueness. In effect, they are able to experiment and test their developments on real-world users. The theme of experimenting on users is explored further below at *Behavioural Analytics*.

Biometric Identifiers

Biometric Identifiers covers any use of measurements of the body or parts of the body to construct an identifier. One of the most common forms is facial recognition. The usage of facial recognition is not new. Facebook deployed automatic facial recognition on its US sites back in 2010, although it subsequently made it an opt-in feature and settled a lawsuit with regard to its original incarnation.[90] A more recent example comes from ClearView AI which has built a facial recognition database from publicly available social media photos.[91]

Whilst facial recognition has gained the most attention, it is not the only form of biometric identifier being investigated. In 2019, it was revealed that the Pentagon had been developing a system for gait tracking via mobile phone sensors.[92] China had already deployed gait tracking in 2018, although their approach was via analysis of recorded video.[93]

The use of biometric identifiers is inherently problematic due to the inability of the user to block or reset the attributes. As such, they should be considered as a high-risk online identifier.

[88] See https://webtransparency.cs.princeton.edu/webcensus/audio_fp_scripts.html
[89] See https://webtransparency.cs.princeton.edu/webcensus/font_scripts.html
[90] See https://www.bbc.co.uk/news/technology-51309186
[91] See https://www.vox.com/recode/2020/2/11/21131991/clearview-ai-facial-recognition-database-law-enforcement
[92] See https://www.washingtonpost.com/news/powerpost/paloma/the-cybersecurity-202/2019/02/26/the-cybersecurity-202-your-phone-could-soon-recognize-you-based-on-how-you-move-or-walk/5c744b9b1b326b71858c6c39/
[93] See https://www.abc.net.au/news/2018-11-06/chinese-gait-recognition-tech-ids-people-by-how-they-walk/10469974

Zombie Cookies

It is common to find trackers using multiple different tracking technologies to record the same identifier or a linkable identifier. This has led to the creation of what has been termed zombie cookies. These are cookies that despite being deleted reappear afterwards. The way it is achieved is through storing the identifier in multiple locations, hoping one survives any reset or clearing action. If it does, the original cookie value can be reset on the next visit. For example, a user might clear their cookies, but not clear their cache. If the identifier had been stored in an ETag, the advertiser can recreate the same cookie on the next request. Digital Platform Identifiers can play a part in this as well, acting as alternative tracking technique. Even if a user deletes their cookies, clears their cache, resets their advertising ID, the next time they log into a digital platform new cookies can be set, new entries made in the cache, and the new advertising ID linked to the underlying platform identifier.

Conclusion

Not only does the user have the challenge of trying to reset all possibly linked identifiers at the same time, which increasingly looks infeasible, but it will be futile against digital platform identifiers. As such, the power the user has to control such tracking appears limited against the biggest platforms. Even against smaller non-digital platform trackers, it remains debatable whether the average user has the capability to defend against the tracking technologies being deployed against them.

Types of information collected

It is clear that tracking organisations are going to great lengths to maintain online identifiers of individuals and devices. Which raises the question as to why it is so important to establish such identifiers? Fundamentally, it is about maximising the longitudinal data collection that can be performed. The longer the period of data, and the greater the resolution, the more valuable the data and the better the predictive modelling that can be performed on it.

Once an identifier has been established a vast array of information can be collected. Initially the data could be contextual information about the current interaction, for example, pages on the website being visited, the search terms being entered, timing of access, and metadata of the requests. Over time a longitudinal record of those interactions will be built, including all interactions with the service. The scale and scope of that data will vary depending on the nature of the service provider.

Whilst it can be difficult to discern the exact data that is being collected, it is possible to establish a baseline by examining openly available standards that collect similar information, and to examine the types of data available to users who download their own data from digital platforms. In the case of the former, one indicator is to look at the standard for 3D-Secure,[94] the payment industries security tool for evaluating online transaction. The standard contains a description of the parameters sought.[95] This of particular interest because it shows the scale of the data being collected. In the case of an Android device up to 151 data points about the device are collected per transaction. This list includes non-resettable identifiers, IMEI, DeviceID, as well as the advertising id and user specific identifiers like phone number. Additionally, a raft of typical device fingerprinting parameters like screen resolution, language, and system parameters, and more sensitive information like longitude and latitude are also collected.

The justification for the collection of such data is to provide transaction security. However, constraints under which the collected data may be used, user awareness of the scale of the data collection, and how such data might be shared are far from clear. Most of the data points do not require additional app permissions to access. Assuming those data points are useful in profiling and fingerprinting, it is reasonable to expect other apps and tracking organisations to be collecting similar data points.

Further indications of the type of data being collected can be found by examining what data is available to users who choose to download their own data from one of the digital platforms.[96] The nature of the data collected will vary according to the privacy settings the user has selected. However, taking Google as an example, there is scope to collect the following:[97]

- Google search history, including maps
- Chrome browsing history
- History of website interactions that include Google services or trackers
- Location history (typically obtained via either maps app or Android device)
- Voice and audio recordings of all triggers of Google Assistants on all registered devices, phones, tablets, smart speakers
- Google Play store history, including searches and installs
- Usage of apps on Android devices (not just Google apps), including timing information
- Received and dismissed Google notifications, including Google News notifications
- YouTube History
- Contact information
- Google Pay Transaction history (including time, amount, location (GPS)

[94] See https://www.emvco.com/emv-technologies/3d-secure/

[95] See https://www.emvco.com/wp-content/plugins/pmpro-customizations/oy-getfile.php?u=/wp-content/uploads/documents/EMVCo_3DS_SDKDeviceInfo_1_4_102019.pdf

[96] See https://support.google.com/accounts/answer/162744?hl=en

[97] See https://www.wired.com/story/google-tracks-you-privacy/

The breadth of the data collected provides a detailed overview of not just the online activities, but also offline transactions paid for using Google Pay, and detailed record of the movements of the device.

Merging online and offline data

Of particular concern is the location information, which is considered to be high value, due to the power it has to reveal user behaviour. For example, it can establish where someone lives, where they work, where they eat lunch. If such data is collected from a large group of people, it helps build a detailed social graph, including both people they know and are linked to online, and those who they are not digitally linked to. This can be achieved by analysing devices that are in the same location at the same time. When augmented with auxiliary data, like key dates, it is possible to start discerning more detailed relationships. For example, where someone is and who they meet on Father's Day, or their birthday, for example, to discern family relationships.

It is not just a social graph that can be built from location data, if that location information can be extended to indoor tracking it becomes commercially valuable, to both the venue and for advertising. Wi-Fi and Bluetooth tracking are the two primary technologies for indoor tracking. The spread of Free Wi-Fi has facilitated an increase in the collection of indoor location data.[98] Such services, particularly when linked to social media logins, can be used to tie online profiles to offline behaviours. Additionally, Bluetooth Beacons have increasingly been deployed to facilitate accurate indoor localisation. Both Google and Facebook have in the past provided beacons to advertising partners to deploy in their physical stores. These beacons work by transmitting a fixed known ID value. The location of the beacon is known and recorded when it is installed. Apps on a user's phone listen for these IDs and when found can send them back to the provider to get an estimation of the location of the device. In doing so the device gets an accurate indoor location value, the platform is able to track the user's location and movements, and the venue is able to target advertisements.

The use of Beacons also facilitates linking online advertising to offline transactions,[99] which is becoming increasingly important in the online advertising space, indicated by Google's purchase of MasterCard transaction data in the US to better tie offline purchases with online advertisements.[100, 101]

Contextual vs Behavioural Information

It is a challenge to categorically define what information constitutes contextual information, and what information constitutes behavioural information. When viewed as a single interaction much of the above data could be considered to be contextual, for example, location. However, when viewed over a period of time, it can form a detailed picture of

[98] See https://skyfii.io/guest-wifi/
[99] See https://www.wordstream.com/blog/ws/2018/10/04/beacon-technology
[100] See https://www.bbc.co.uk/news/technology-45368040
[101] See https://support.google.com/google-ads/answer/6190164?hl=en-GB

behaviour. Crucially, such data can show reactions to stimuli and therefore allow the modelling of reactions, mood, and behavioural triggers. As such, entering a shop goes from being contextual, to being behavioural if it can be tied to the showing of a particular advert.

A clearer distinction can be found between contextual targeting and behavioural targeting. In the case of contextual targeting the context stems from the content the ad will be displayed within. (For example, a newspaper travel story about Thailand might show adverts for a resort in Phuket or airfares to Bangkok.) If performed correctly, there is no need to track the individual over time, since the targeting of the ad is based on the context of the ad placement, not the context of the target. Conversely, behavioural advertising is targeting the individual, and is dependent on maintaining online identifiers to build up a detailed profile of behaviour in order to be able to target an appropriate advert. The context of the ad is not the content of the page, but the target of the advert.

The distinction between the two approaches is best demonstrated in the differences between Google search and DuckDuckGo search. Google collects vast quantities of data about individuals to tailor adverts to them. In contrast, DuckDuckGo does not collect any information about its users. It relies on placing adverts based only on the search term entered. As described by the Founder and CEO through written testimony at a US Senate Judiciary hearing: "Every time you search on DuckDuckGo, it's like you are searching on our site for the first time. We do not even have the concept of a search history."[102] In contrast Google's page on targeted advertising highlights the ability to target affinity audiences, "With affinity audiences, you can reach people based on a holistic picture of their lifestyles, passions and habits".[103] Building that holistic picture requires a detailed longitudinal record of the individual's behaviour.

How are online identifiers used?

Online identifiers play two key roles, firstly to build the profile, and second, to target the results. The first step, building the profile, requires establishing online identifiers to record behaviour over time. It is unlikely any single identifier will be exhaustive,[104] instead there will be a need to link multiple datasets and therefore multiple identifiers. Cookie Syncing plays a key part in establishing these links, but frequently the challenge is to link data without such explicit ties, or to link online with offline customer data. This is the basis of what has been termed people-based marketing, which Facebook boasts as "We now can find a person instead of large groups such as "Adults 18-34" or "people who like coffee" and reach them on whatever device or platform they may be on".[105] In essence, this is marketing's use of big data analytics to target individuals.

[102] See https://duckduckgo.com/download/GDRP-CCPA-Hearing-Testimony_2019-03-12.pdf

[103] See https://support.google.com/google-ads/answer/2497941?hl=en-GB

[104] It could be argued Google has close to a complete view of some people's online interactions. However, their reach into offline activities is not as exhaustive, hence the need to purchase additional offline data See 100 .

[105] See https://en-gb.facebook.com/business/news/insights/the-future-of-marketing-people-based-planning-and-measurement

Big Data Analytics

Big Data Analytics can be considered to be an umbrella term that covers Big Data, Machine Learning, Predictive Analytics, and a number of other data mining techniques. One of the core concepts behind Big Data Analytics is the ability to analyse unstructured data to find latent patterns and relationships within that data. As such, it can be very effective at finding links between datasets that do not share an overt identifier. In such circumstances it is the data itself that is acting as the identifier. It is the capability that allows syncing data from different sources and building detailed profiles. It is also essential in syncing identifiers to ensure that any results can be appropriately targeted.

The relationship between Big Data, AI and Machine Learning was succinctly described in the UK ICO's report as "Big data can be seen as an asset that is difficult to exploit. AI can be seen as a key to unlocking the value of big data; and machine learning is one of the technical mechanisms that underpins and facilitates AI".[106] Each technology plays a part in collecting, processing, linking and analysing the data.

Once a suitable dataset has been constructed predictive analytics is applied to take the historic data and build a model that can take real-time data and make predictions about future actions, and how those actions can be influenced. Sometimes these models are relatively simple, for example, complementary product analysis, in which purchasing of one item leads to a high probability of purchasing another item. For example, if someone has purchased a flight there is a high probability they will also purchase a hotel. More complicated models can be built, for example, Netflix's recommendation model, which is believed to drive 80% of what is watched on the platform.[107] The common factor between the models is they are built on the basis of data that has been collected to show how individual behaviours are linked.

Once the predictive model has been created it can be used in real-time to create predictions. Those predictions are then targeted back at the user. In order for this to happen it is necessary to be able to recognise the person when they are next seen, for example, when they next view a website so an ad can be shown, or when they next use an app. This is where the online identifier again becomes critical, without it there would be no way to accurately target the individual. For example, there is not much value in knowing there is a high probability a user who books a flight will subsequently book a hotel, if there is no way of identifying that user to show them the relevant ad for hotels.

Behavioural Analytics

Such uses of online identifiers and big data analytics are common, but there are additional common uses for online identifiers in behavioural analytics. When performed at scale the analysis of behaviour can be used to fine-tune advertising campaigns, websites, and triggers, automatically. One of the most widespread and simplest forms of such data

[106] UK ICO, *Big Data, artificial intelligence, machine learning and data protection*, 2017, p.8; available at https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf
[107] See https://mobilesyrup.com/2017/08/22/80-percent-netflix-shows-discovered-recommendation/

collection and analysis is A/B testing. This is a form of analysis that shows two different variants of something to a target audience and records their reactions. For example, many news organisations run A/B testing on their headlines to determine the headline that gets the most clicks. In a study undertaken in 2018 the New York Times was found to be undertaking 91 A/B tests targeting 90 audience groups during widespread scan of the use of A/B testing.[108]

Fine-tuning headlines or layouts is relatively benign, although it does raise some ethical questions.[109] More troubling is the usage of behavioural analysis on digital platforms. A paper in 2014 conducted an experiment on Facebook users by altering their news feeds to either increase or decrease the appearance of emotional expressions in their news feeds. Future posts from the user were analysed to determine the impact, with the results indicating that emotional contagion can take place via news feeds.[110]

Such experimentation is not uncommon. In defence of the Facebook study the Co-founder of the dating site OKCupid wrote a blog post entitled “We Experiment On Human Beings!” in which he wrote “…guess what, everybody: if you use the Internet, you’re the subject of hundreds of experiments at any given time, on every site”.[111] The post proceeds to detail how they ran experiments on their users by creating artificial match recommendations, and manipulating the display of users’ profiles to prospective matches. Such experiments are only possible because of the ability to track users across interactions via online identifiers. Such identifiers could not be blocked, since they are essential to the functioning of those platforms. The concern is focussed more on the usage of the data collected as opposed to the identifier itself. In particular, whether the user is sufficiently capable and empowered to understand and control such insights and manipulation. If not there is not only a risk of privacy harm, but also broader financial and opportunity harm as well.

Metadata Collection and Usage

In addition to the content that is contained within a communication, any communication will generate further data about the communication itself, in the form of metadata. Such data is essential to the successful delivery of the communication, for example, to send a packet of data reliably across the internet, at the very least, a sender and a receiver must be included. The receiver is needed to correctly route the packet across the internet, and the sender is required to return an acknowledgement of safe delivery. On the internet those identifiers are known as IP addresses (Internet Protocol). IP addresses come in two forms, the original IPv4 addresses, and the new IPv6 addresses. IPv6 is required because the internet has

[108] Jiang, Shan, John Martin, and Christo Wilson. "Who's the Guinea Pig? Investigating Online A/B/n Tests in-the-Wild." Proceedings of the Conference on Fairness, Accountability, and Transparency. 2019. Available at https://shanjiang.me/publications/fat19_paper.pdf

[109] Jiang, Shan, John Martin, and Christo Wilson. "Who's the Guinea Pig? Investigating Online A/B/n Tests in-the-Wild." Proceedings of the Conference on Fairness, Accountability, and Transparency. 2019. Available at https://shanjiang.me/publications/fat19_paper.pdf

[110] Kramer, Adam DI, Jamie E. Guillory, and Jeffrey T. Hancock. "Experimental evidence of massive-scale emotional contagion through social networks." Proceedings of the National Academy of Sciences 111.24 (2014): 8788-8790. Available at https://www.pnas.org/content/111/24/8788.full

[111] See “We Experiment On Human Beings!” archived at: https://web.archive.org/web/20160317051625/http://blog.okcupid.com/index.php/we-experiment-on-human-beings/

effectively run out of IPv4 addresses. As such, reuse and sharing of Ipv4 addresses is commonplace. Additionally, consumers are not directly connecting to the internet, they are connecting through one or more private networks linked together. Rather than being a single large network, the internet is a network of networks. As such, there are levels of addressing and metadata beyond just what would be considered to be the public IP address.

Figure 5 shows a high-level abstraction of a typical home or Wi-Fi internet connection. As such, a single web request from a user's device will travel across 4 different networks to just get to the internet backbone, and potentially a further 4 networks to get to the destination. The addressing that is used on these different networks will vary, generally the same address will not be used across all the networks. Each network must maintain a suitable mapping to ensure responses can be correctly routed back to their source, and that information must be encoded in some way into the metadata attached to the request. As such, it is possible that networks several levels away will be able to identify traffic coming from the same source, even if it is unaware of its IP address on the other network.

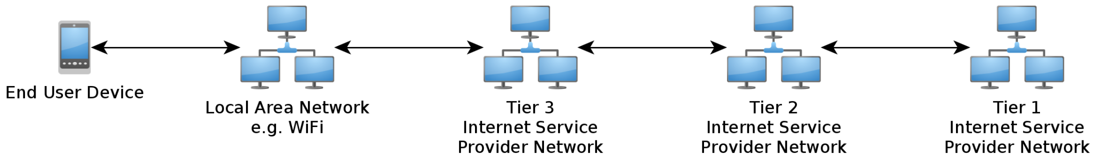
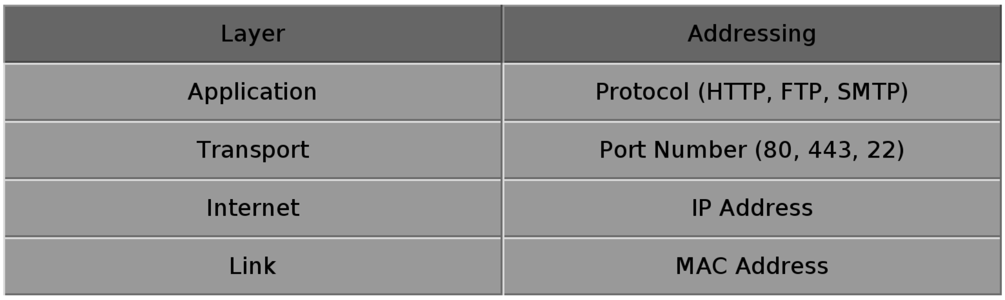

End User Device
Local Area Network e.g. WiFi
Tier 3 Internet Service Provider Network
Tier 2 Internet Service Provider Network
Tier 1 Internet Service Provider Network

Figure 5: Internet - Network of Networks

In addition to there being levels of networks, there are layers of networking and addressing operating on each of those networks. Figure 6 shows the different networking layers used in TCP/IP[112] – the protocol used on the internet. Of particular note are MAC addresses, which are used for the low-level connection between two machines on the same network. MAC Addresses do not propagate across the internet, they are only used within a single network, but those MAC addresses are intended to be globally unique. This allows devices to connect to a network without having to first be assigned an address, with a guarantee that no two devices will have the same address and therefore conflict.[113]



Layer	Addressing
Application	Protocol (HTTP, FTP, SMTP)
Transport	Port Number (80, 443, 22)
Internet	IP Address
Link	MAC Address

Figure 6: TCP/IP Networking Layers

[112] TCP/IP stands for Transmission Control Protocol and the Internet Protocol

[113] Due to the ability to configure some MAC addresses it is possible to spoof or have duplicate addresses, MAC addresses are intended to be universally unique: See http://www.ieee802.org/secmail/pdfocSP2xXA6d.pdf

MAC addresses in particular have raised privacy concerns, as they are broadcast whilst scanning for Wi-Fi and Bluetooth networks. As such, the combination of them being globally unique, and easily obtainable by listening for Wi-Fi or Bluetooth scans, made them a popular target for device tracking and localisation. When combined with a network of multiple scanners it is possible to derive location information about a device by just monitoring which scanners receive the MAC address broadcasts. More accurate location information can be obtained my also modelling the comparative signal strengths.

Awareness of the problem has led to the development of techniques to attempt to hide the true MAC address of a device when scanning for a network.[114] This is achieved by randomising the broadcast MAC address at regular intervals. Research has indicated that even with this additional step it is still possible to track devices via other network level metadata.[115] Even if such randomisation was effective, many MAC addresses are only randomised during scanning, with the real MAC address, or a derivative of it, used when connecting to the network. This potentially allows networks to share MAC address records to build a broader picture of device location and movement. It could be thought of as being somewhat analogous to a third-party cookie, particularly where a single provider is delivering the underlying Wi-Fi functionality across different sites and brands.[116]

To address this weakness, device operating systems are starting to randomise the MAC address on a per network basis. As such, the MAC address seen when connecting to the hotel Wi-Fi will be different to the one seen when connecting to the Airport Wi-Fi. This can be considered analogous to a first-party cookie. It offers some protection, but will still permit reoccurring tracking when accessing the same network. Much like in the case of cookies, further problems occur when such networks can link the MAC address to a platform identifier, for example, by requiring users to log in through social media accounts. Additionally, if the same network is deployed over a large area, for example, an entire city, then the effectiveness of the network isolation is reduced.

Much of the above has focussed on Wi-Fi MAC addresses, but Bluetooth MAC addresses are equally used for device tracking and localisation. It is for this reason that if an Android app wishes to access Bluetooth it must also seek permission from the user to know the location, since location is considered to be inferable from Bluetooth scanning data alone, generally via Bluetooth Beacons. However, that is not the only usage of Bluetooth scanning, increasingly networks of Bluetooth scanners are being deployed to track traffic movements.[117] Victoria has over 1,100 Bluetooth scanners deployed across the state.[118] They work by scanning for Bluetooth MAC addresses and monitoring how they move between scanning locations to determine how quickly the traffic is moving. Such scanning is not necessarily targeting just mobile phones, many cars will include Bluetooth to allow device connectivity. However, the operating systems on the car's in-car-entertainment

[114] See https://source.android.com/devices/tech/connect/wifi-mac-randomization
[115] Jeremy Martin, Travis Mayberry, Collin Donahue, Lucas Foppe, Lamont Brown, Chadwick Riggins, Erik C. Rye, Dane Brown, "A Study of MAC Address Randomization in Mobile Devices and When it Fails"
[116] See https://skyfii.io/
[117] See https://austraffic.com.au/projects/bluetooth-based-technology
[118] See https://vicroadsopendata-vicroadsmaps.opendata.arcgis.com/datasets/48fd4d7e1127453ea5f9bdc757ab00e7_0

system is less likely to be updated, and may not include MAC address randomisation, or even either ways to disable Bluetooth. If such technology is paired with other traffic monitoring techniques, for example, Automatic Number Plate Recognition cameras,[119] the MAC addresses can be tied to a longer-lived identifiers and additional data.

Similar identifiers are used in the cellular network (mobile phone network) and it can be viewed as being an alternative to the Local Area Network in Figure 5. In much the same way, the cellular network will connect through levels of providers to eventually connect to the internet backbone. Whilst conceptually similar to other forms of wireless networking, one important distinction is that the network operates as a series of "cells".[120] Each cell will normally have multiple transceivers (typically three) that are directed towards it. In much the same way as Wi-Fi localisation works, the cellular provide can determine in which cell the device is, and within that cell, broadly where the device is by evaluating signal strength. As new cellular standards have been developed, in particular LTE (4-generation) and the new 5th generation (5g) networks, they have prioritised device localisation as a service. A white paper on the proposed 5g network indicate an accuracy of between 10m and <1m on 80% of the occasions, and less than 1m indoors.[121] Such accuracy would be an improvement on consumer grade GPS signals,[122] particularly when indoors, where GPS is typically not an option due to the requirement for line-of-sight of the satellites.

Such information is available to network operators by virtue of being the metadata collected in order to run the network itself. Knowing the location of the device is necessary so that the signal can be directed most efficiently to the device, and handovers between cells can be seamlessly managed. However, the cellular providers are increasingly looking to commercialise this data through partnerships with location platforms and data analytics companies.[123,124] This demonstrates the dual role that metadata can have.

More broadly there are problems with defining exactly what is considered to be metadata. A good example of this is data associated with the Domain Name System (DNS). When a user types in a URL into their browser the first step is for the browser to query the Domain Name System for the IP address of the server associated with that URL. Abstractly the Domain Name System can be considered to be a giant address book that maps website URLs to server addresses. The issue is that the Domain Name System was designed when the internet was small and the notion of third-party tracking had not even materialised. As such, DNS queries (lookup requests) are not encrypted by default.

This lack of encryption allows the ISP or any other party that is able to monitor the network to record and even alter those queries. This functionality is used to implement some web filtering systems and parental controls. But it also has the potential to be used for analytics, profiling and targeted advertisements.[125] In effect, it allows an ISP to build a record of all

[119] See https://www.sensordynamics.com.au/
[120] See https://radio-waves.orange.com/en/how-does-a-mobile-network-work/
[121] Next Generation Mobile Networks Aliance, 5G White Paper, 2015, available at https://www.ngmn.org/wp-content/uploads/NGMN_5G_White_Paper_V1_0.pdf
[122] See https://www.gps.gov/systems/gps/performance/accuracy/
[123] See https://locationinsights.telstra.com/
[124] See https://www.vodafone.com.au/media/vodafone-and-nokia-develop-4g-incident-detection-prototype
[125] See https://www.theregister.com/2008/03/17/bt_phorm_lies/

the websites a particular device or subscriber has visited. Such information could be considered quite invasive of an individual's privacy, and may reveal sensitive attributes about themselves. It also allows them to redirect the user to target sites when no address is found.[126]

It is possible to assign a different DNS provider on a device, however, this would not stop an ISP from monitoring the DNS queries as they travel across their network. Recent proposals for encrypting DNS queries has gained traction, with support being included in leading browsers.[127] However, there has been significant push back from both ISPs and Governments at the possible loss of access to such data. As well as concern from some that enabling such security will give greater access to DNS query data to large US tech companies.[128] Ultimately, whichever organisation runs the DNS server a user connects to, has the potential to profile their web browsing habits.

In most cases the DNS will be provided by the ISP, and will come preconfigured on devices, for example, broadband routers or mobile phones. It is possible to change the DNS, although the difficulty of the process varies and requires some understanding of basic networking concepts. Furthermore, it is possible for the ISP to block third-party DNS queries if it wishes, although this has not been widely reported in Australia. However, given the increase usage of DNS filtering[129] to tackle malware and perform site blocking, it is not inconceivable blocking of third-party DNS could occur in the future.

Inspecting the contents of a DNS query is just part of a broader set of measures termed Deep Packet Inspection (DPI). DPI involves inspecting the full contents of the packet, including the data. It was designed as tool for better network security, allowing detecting of malicious or malformed packets, and blocking or redirecting them. However, the techniques have also been used for the purposes of profiling and targeted advertising.[130] Increased usage of encryption can limit the amount of data available to DPI, however, where traffic is channelled through a TLS proxy (see discussion above at *Intermediary Cookies.*) by the web server host or the network provider for the user, the capability to perform DPI is re-established. Such channelling is becoming common place for websites to protect against cyber security attacks and for some corporate networks, or even home networks where parental controls are in place, to inspect and filter traffic for safety and security reasons.

Depending on perspective, DNS queries could be viewed as metadata used to route traffic, alternatively, they could be considered as content, since the query itself contains a user specified or derived address. In either regard, the information contained within them reveals behaviour information about the user.

In addition to network level metadata, there is also latent metadata present in data itself. For example, a list of credit card transactions may not seem like it contains location data. However, by mapping the merchant IDs or names to the locations of the shop or businesses

[126] See https://www.crn.com.au/news/bigpond-redirects-typos-to-unethical-branded-search-page-160923
[127] See https://www.theregister.com/2019/11/19/microsoft_joins_doh/
[128] See https://www.theregister.com/2019/09/10/chrome_78_dnsoverhttps/
[129] See https://www.zdnet.com/article/telstra-steps-up-dns-filtering-to-fight-malware/
[130] See https://www.infosys.com/services/data-analytics/insights/documents/customer-behavior-analytics.pdf

where the transaction took place, it is possible to build a partial trajectory of the user, complete with timing information. This type of latent location information is very common, even fairly innocuous snippets of information can contain detailed location data. For example, a posting on social media about graffiti on a train on a particular line will allow inference of the train the user is on by cross-referencing train timetables and the time of the posting.[131]

Value of location data

In the context of network operations location data could be considered to be metadata, however, in other contexts it is arguably content data in its own right. In particular where apps deliver content based on current location; the location data could be considered as equivalent to search submissions, not just metadata. The collection of such data is common in mobile apps, with it often shared with multiple parties. A New York Times article in 2018 found one app to be sharing precise location with 40 companies.[132]

Location data is considered to be highly valuable for analytics and profiling. A significant advantage of location data is that it is provides insight into actual events.[133] There is more value in knowing someone has attended a car dealership, than knowing they searched for car dealerships online.[134] Likewise, it is more interesting to a life insurance company to know how often and for how long someone attends a gym, than just knowing they pay a monthly gym membership. In effect, location data paints a picture of someone's actual life, not just their aspirations or potential interests. But it is in this regard where privacy concerns occur, because location data alone represents the individual, irrespective of whether their name or another identifier is attached to the record. As a result it can form the link between other data sets.

Location data is often shared on the basis of it having been anonymised, which normally amounts to removing name and other known identifiers. However, the trajectory alone is often uniquely identifying. In 2013 a study of human mobility data for 1.5 million individuals over 15 months revealed how unique human mobility traces were:

> “in a dataset where the location of an individual is specified hourly and with a spatial resolution equal to that given by the carrier's antennas, four spatio-temporal points are enough to uniquely identify 95% of the individuals”.[135]

The reason such a degree of uniqueness is of concern is the power such data has to link between other supposedly anonymous or de-identified datasets; for example, combining location data from mobile phones with payment data from credit cards, a data source with

[131] Culnane, Chris, Benjamin I. P. Rubinstein, Vanessa Teague, “Stop the Open Data Bus, We Want to Get Off”, 2018, available at https://arxiv.org/abs/1908.05004

[132] See https://www.nytimes.com/interactive/2018/12/10/business/location-data-privacy-apps.html

[133] See https://blog.near.co/news/we-know-which-suburb-eats-more-pizza-by-analyzing-data-from-15-million-australians/

[134] See https://www.youtube.com/watch?v=BVZc86CZovU

[135] de Montjoye, Y., Hidalgo, C., Verleysen, M. et al. “Unique in the Crowd: The privacy bounds of human mobility”. Sci Rep 3, 1376 (2013). Available at https://doi.org/10.1038/srep01376

similar levels of uniqueness.[136] Cross-correlating the mobile phone locations with the locations of the merchants in the credit card data will potentially allow the two datasets to be linked, despite the fact both were supposed to be anonymous or de-identified.

Such linking is not even particularly secret, despite that fact it would seem to demonstrate the fallibility of the claimed anonymisation/de-identification used to originally justify the sharing of the data. For example, connecting to Westfield's Free Wi-Fi involves agreeing to a set of terms and conditions which include the following statement:

> "If you access or log-in to the Westfield Wi-Fi Service, and we hold other personally unidentifiable information that can be associated to you or the device on which you are accessing the Wi-Fi Service (including, but not limited to a device ID number (MAC address)), then that information may be linked with personal information we hold about you as set out in the Wi-Fi Privacy Terms or the Scentre Group Privacy Policy, and will be treated in the same manner as the personal information to which it has been linked".[137]

The fact that "personally unidentifiable information" could be associated to a person or device is evidence that it is not personally unidentifiable information in a global sense, only within the dataset it was shared in. But once combined with other datasets the data ceases to be unidentifiable.

The combining of different datasets into a single unified view provides an unprecedented insight into an individual's life. This linking is being enabled by advances in machine learning and Artificial Intelligence, using the data itself to link the records. One of the major players in the space is Near,[138] which boasts of having "the world's largest data set of people's behavior in the real-world" consisting of 1.6 billion users, across 44 countries, processing 5 billion events per day.[139] Near claims to have

> "...developed an in-house technology for ID unification as part of our ongoing efforts to build our Data Intelligence Platform. CrossMatrix embeds generic statistical matching techniques along with domain/channel specific heuristics to generate resolved identifiers".[140]

Near claims it processes the data in a privacy preserving way, which seems closely tied to the asserted absence of personally identifiable data. However, as in the case with cookies and other online identifiers, the fact the legal identity of the individual is not known does not prevent privacy harms from occurring, particularly when individuals can still be targeted based on their actions.

[136] de Montjoye, Yves-Alexandre, Laura Radaelli, and Vivek Kumar Singh. "Unique in the shopping mall: On the reidentifiability of credit card metadata." Science 347.6221 (2015): 536-539. Available at https://science.sciencemag.org/content/347/6221/536

[137] See https://www.westfield.com.au/terms-and-conditions#wi-fi-terms-of-use-and-privacy-terms

[138] See https://near.co/

[139] See https://near.co/data/

[140] See https://blog.near.co/technology/the-how-and-why-of-id-unification/

The process of ID unification based on the data itself erodes the distinction between data and identifier. If a new "super" identifier can be generated to link disparate datasets with non-unified IDs, the focus needs to shift towards the identifiability of the data, as opposed to the presence or absence of a known or even random identifier. Currently, the focus on identity is leading to contradictions in privacy, with assertions that data is anonymised whilst still permitting individual targeting.

A recent example from Near is the analysis conducted on movements during the COVID-19 isolation in Melbourne.[141] In the report it is stated:

> "Near has access to data from 17 million Australian devices, which it says provides a more accurate picture of people's movements.
>
> Near country manager Adam Boekeman stressed that the data was anonymized to protect privacy, which is the main reason for the slow take-up of the government's own app".[141]

However, in the same report the country manager is reported to have said: "We can support app adoption, saying to someone you've been to a postcode or a high-risk area and encourage them to download the app. That's quite easy to do".[141]

The two claims would appear to be contradictory. If the data is anonymised it should not be possible to subsequently notify an individual on the basis of their individual movements. If this is possible, it demonstrates why even without identity information the individual can remain identifiable and therefore can incur a privacy harm. If their device can be identified for the purpose of messaging it can be identified for the purpose of targeted adverts and profiling.

Unintended consequences of Bluetooth contact tracing

A timely reminder of the challenges of regulating and understanding the consequences of technology can be seen in the recent deployment of the Australian Government's COVIDSafe app. Evaluating and understanding the consequences of the app is a challenge, in part due to the lack of technical details and in part due to the rapid changes that are taking place. However, the underlying technology, Bluetooth Low Energy (BLE), itself presents a challenge to privacy. The widespread usage of beacons and Bluetooth scanners (see above at *Metadata Collection and Usage*) creates an environment that is intrinsically hostile to privacy. For example, on Android devices just enabling Bluetooth results in collection of scan data by Google. It was found that "[w]hen either Bluetooth or Bluetooth scanning is enabled, a report containing a list of nearby Bluetooth beacons is sent to Google any time an app refreshes Android location services".[142] This has a direct impact on the privacy impact of Bluetooth Contact Tracing, since the app requires Bluetooth to remain on, and will regularly perform scans.

[141] See https://blog.near.co/news/workers-tracked-20km-from-infected-abattoir/
[142] See https://qz.com/1169760/phone-data/

It is also not just Google that collects such information. Many apps will register for opportunistic Bluetooth scans,[143] in which apps receive the results of any Bluetooth scans undertaken by other apps. This is a way of reducing battery consumption associated with multiple apps conducting such scans themselves. However, increasing the number of scans, along with always-on Bluetooth will result in any such apps receiving an increased number of scan results, potentially allowing for finer grained tracking. If the user has not individually denied Bluetooth permissions to those apps, and has instead relied on disabling Bluetooth on the device as a whole when not in use, the privacy cost associated with always-on Bluetooth and regular scanning could be high.

Specific legislation to regulate the additional data being broadcast by COVIDSafe has been passed,[144] however, the legislation failed to address the secondary impacts of requiring everyone to enable Bluetooth all the time. Whilst it would have been desirable to have had such a regulation in place, the broader question is whether an essential communication technology, in this case Bluetooth, should have been permitted to be repurposed for widespread location tracking by third parties unrelated to the COVIDSafe purpose. Should consumers be denied the ability to use hands-free devices, and syncing with their car for fear of incurring a privacy harm?

The issues associated with the COVIDSafe app demonstrate the challenge in regulating technology as well as the current state of the technology environment, with tracking and profiling purposes consuming essential technologies that facilitate the connected world many people have come to enjoy and rely upon. Even if the COVIDSafe legislation had addressed the issue of Bluetooth tracking, it would have been a temporary respite, demonstrating the need to address the core issue of widespread tracking, which is often being undertaken in a way that is beyond the comprehension or capability to control of the average user. To further demonstrate this, even if Bluetooth is switched off on an Android device, Google continues to use it for location scanning142. In order to completely disable it requires the user to change location scanning settings. Those settings include an option, which defaults to on, with the label “Improve location by allowing the system services to scan for Bluetooth devices, even when Bluetooth is off”,[145] It seems unlikely the average user would understand that switching Bluetooth off didn’t actually switch if off *for Google*.

Advances in technology

Advances in AI and Machine Learning are outstripping advances in privacy protection. The ability to statistically analyse, link and model vast quantities of data was not a concern when many privacy protection laws were being conceived. The ability to construct longitudinal data across extended periods of time at a resolution often measured in minutes, is creating

[143] See https://developer.android.com/reference/android/bluetooth/le/ScanSettings#SCAN_MODE_OPPORTUNISTIC
[144] Privacy Amendment (Public Health Contact Information) Act 2020
[145] See https://www.androidpolice.com/2015/05/29/android-m-feature-spotlight-bluetooth-scanning-joins-wifi-to-improve-location-accuracy/

data that alone, without any recognisable identifiers, can be linked and built into a targeted individual profile.

Privacy Enhancing Technology

That is not to say there have not been advances in privacy enhancing technology. There have been advances in browser extensions that aim to prevent tracking.[146, 147] However, such tools require user awareness and knowledge to install, potentially limiting their uptake. There was also an attempt, which ultimately failed, to standardise user preferences regarding tracking, namely through the Do Not Track header.[148] The header was intended to be a browser configured option that allowed a user to express their tracking preferences on each request through one of three values, 1: not consenting to tracking, 0: consenting to tracking, or null when no preference has been set. Whilst support was built into browsers, the advertising industry largely ignored the header.[149] Google implemented support for Do Not Track into its Chrome browser, but openly acknowledged that it did not respect it on its own websites.[150]

Additionally, there have been advances in protecting data prior to collection or sharing, in the form of Differential Privacy, and longer term proposals for how to change the way data is handled on the web through the proposed SOLID platform.

Differential Privacy

Differential Privacy[151] is a privacy protection technique that is gaining mainstream attention and increased usage. The underlying mathematics of Differential Privacy is complicated. A high-level summary was provided in a paper titled “Differential Privacy: A Primer for a Non-technical Audience”, in which it is described as:

> “Differential privacy is a strong, mathematical definition of privacy in the context of statistical and machine learning analysis. It is used to enable the collection, analysis, and sharing of a broad range of statistical estimates, such as averages, contingency tables, and synthetic data, based on personal data while protecting the privacy of the individuals in the data”.[152]

The exact details of the approach are beyond the scope of this report, but fundamentally it provides a mathematically rigorous process for the sharing of data in a way which does not reveal whether or not any given individual is represented in the dataset. As such, it could be applied in scenarios where trends and aggregates statistics are being shared. However, the party applying the differential privacy still has access to the full dataset.

[146] See https://privacybadger.org/
[147] See https://github.com/gorhill/uBlock
[148] See https://w3c.github.io/dnt/drafts/tracking-dnt.html
[149] See https://gizmodo.com/do-not-track-the-privacy-tool-used-by-millions-of-peop-1828868324
[150] See https://support.google.com/chrome/answer/2790761
[151] Dwork, C. "Differential Privacy In: Proc. of the 33rd International Colloquium on Automata, Languages and Programming (ICALP), 1–12." (2006). Available at https://www.microsoft.com/en-us/research/publication/differential-privacy/
[152] Nissim, Kobbi, et al. "Differential privacy: A primer for a non-technical audience." Privacy Law Scholars Conf. 2017.

A stricter variant is Local Differential Privacy,[153] which aims to not only protect the outputs from the dataset as a whole, but also the individual responses that make up the dataset. In effect the privacy protection is applied by the client device before sharing any data. As such, the data from the user is protected at the point of collection.

This notion is not new; randomised responses were first developed in 1965.[154] A simple example might be someone wanting to know what percentage of their friends smoke without finding out about any individuals. The process can be setup as follows: If the question is "Do you smoke?", then first flip a coin in secret:

- If it comes up heads answer truthfully (yes or no)
- If it comes up tails, flip another coin
    - If it comes up heads answer yes
    - If it comes up tails answer no

It should be clear that 'noise' (in the sense of random data) has been added to the responses, and that any single response cannot be distinguished from being the true response or a random response. For example, in the above structure an actual smoker has a 0.75 (75%) chance of saying yes, (0.5 from the first coin flip, and 0.25 chance of getting a tail in the first and head in the second (0.5*0.5)). Provided enough people take part the noise will average out. For example, if someone asked 1,000 friends and got 700 answers Yes and 300 answers No. In that situation the person knows that on average 50% of the answers are random (tail on the first coin flip), they can remove 250 answers from each result, giving 450 Yes, and 50 No. The ratio between these will reflect the true answer, i.e. 90% Yes, 10% No.

The above is an extremely simplified example, and the actual percentages can be varied to adjust the trade-off between privacy and noise.

Such approaches are not purely theoretical, with a number of major companies developing approaches in recent years. Google proposed and built the RAPPOR[155] platform, which claims to be "... a novel privacy technology that allows inferring statistics about populations while preserving the privacy of individual users".[156] Google used the platform in a study of user set homepages in Chrome. Apple developed an approach to collect various system information values, including emoji use, learning new words to add to its global dictionary, and memory usage in the Safari Browser.[157] LinkedIn has also developed a differentially private analytics and reporting system.[158]

[153] Kasiviswanathan, Shiva Prasad; Lee, Homin K.; Nissim, Kobbi; Raskhodnikova, Sofya; Smith, Adam D. (2008). "What Can We Learn Privately?". 2008 49th Annual IEEE Symposium on Foundations of Computer Science. pp. 531–540

[154] Warner, Stanley L. "Randomized response: A survey technique for eliminating evasive answer bias." Journal of the American Statistical Association 60.309 (1965): 63-69.

[155] Erlingsson, Úlfar, Vasyl Pihur, and Aleksandra Korolova. "Rappor: Randomized aggregatable privacy-preserving ordinal response." Proceedings of the 2014 ACM SIGSAC conference on computer and communications security. 2014. Available at https://research.google/pubs/pub42852/

[156] See https://github.com/google/rappor

[157] See https://machinelearning.apple.com/2017/12/06/learning-with-privacy-at-scale.html

[158] See https://engineering.linkedin.com/blog/2019/04/privacy-preserving-analytics-and-reporting-at-linkedin

Differential Privacy can assist with the collection of aggregate trend data, but does not resolve the issues associated with sharing of data with multiple parties and platforms, which will remain a fundamental part of the web. For example, people will still want to share photos, news, and events. Likewise, they will still need to share contact information with e-commerce sites, as well as store and use personal, financial and health information online.

Decentralised data storage - SOLID

Tim Berners-Lee, the inventor of the World Wide Web, has proposed a radical rethink of how data is stored online to address the fundamental problem of wanting to share and protect data at the same time. The approach is known as Solid,[159] which he describes as:

> "Solid changes the current model where users have to hand over personal data to digital giants in exchange for perceived value. As we've all discovered, this hasn't been in our best interests. Solid is how we evolve the web in order to restore balance - by giving every one of us complete control over data, personal or not, in a revolutionary way".[160]

Solid proposes users store their data in PODS (Personal Online Data Stores) which the user can choose to host wherever they wish. A user might have different PODS for different information, for example, contact information, health information, and social information. When a user wants to share that information they provide permission for the other party to access the relevant information in the relevant PODS. The user retain control over what data is stored, who has access to it, and where it is located.

Such an approach would be a paradigm shift in how data is managed online and offers full control over data to the user. The ability to revoke access acts as an incentive to organisations to behave appropriately, through fear of losing access to all the data.

[159] See https://solid.mit.edu/

[160] See https://inrupt.com/blog/one-small-step-for-the-web

Chapter 3

Privacy risks posed by online identifiers

Privacy harms exist on a spectrum

Privacy laws exist to protect people from privacy harms. It is because of the scope to do harm to people that some practices are deserving of regulation.

Privacy harms exist across a spectrum, and include:

- tangible or 'material' harms at one end (such as physical harm or threats of violence, stalking and harassment, identity theft, financial loss and psychological damage),
- intangible or 'moral' harms in the middle (such as reputational damage, "creepy inferences", humiliation, embarrassment or anxiety, loss of autonomy, discrimination and social exclusion), and
- abstract or 'social' harms at the other end (such as the threats to democracy, chilling effect on free speech, loss of trust and social cohesion posed by a 'surveillance society', and by manipulation and amplification of political messaging on social media).[161]

Online micro-targeting

As outlined in Chapter 2, online identifiers allow many types of communications – not just advertisements – to be targeted at particular individuals. Personalisation means precise decisions are made at an individual level about who sees what, and equally what will be withheld from whom.

By facilitating exclusion, online identifiers also facilitate discrimination. Facebook has been caught allowing advertisers to target – and exclude – people on the basis of their 'racial affinity', amongst other social, demographic, racial and religious characteristics.[162] For example, a landlord with an advertisement for rental housing could prevent people profiled as 'single mothers' from ever seeing their ad; an employer could prevent people identifying as Jewish from seeing a job ad; or a bank could prevent people categorised as 'liking African American content' from seeing an ad for a home loan.[163]

[161] This spectrum of privacy harms is drawn from work by the former UK Information Commissioner, as well as the Future of Privacy Forum's paper, "Benefit-Risk Analysis for Big Data Projects", September 2014, available at www.futureofprivacy.org

[162] Julia Angwin, Ariana Tobin and Madeleine Varner, "Facebook (Still) Letting Housing Advertisers Exclude Users by Race", *ProPublica*, 17 November 2017; available at https://www.propublica.org/article/facebook-advertising-discrimination-housing-race-sex-national-origin

[163] Alex Hern, "Facebook lets advertisers target users based on sensitive interests", *The Guardian*, 16 May 2018; available at https://amp.theguardian.com/technology/2018/may/16/facebook-lets-advertisers-target-users-based-on-sensitive-interests

Existing patterns of social exclusion, economic inequality, prejudice and discrimination are further entrenched by micro-targeted advertising, which is hidden from public view and regulatory scrutiny. Preying on vulnerable individuals which could lead to physical, financial or social harm is also a risk of micro-targeting. For example a pharmaceutical company selling addictive opioid-based pain medication used Google's search terms data to target people with chronic pain, promoting ads of escalating intensity across multiple sites, despite laws prohibiting the advertising direct to consumers of prescription medication.[164] Even after the 'emotional contagion' scandal,[165] it was revealed in 2017 that Australian Facebook executives were promoting to advertisers their ability to target psychologically vulnerable teenagers.[166] Advertising mental health services is one thing; advertising pharmaceuticals is another; while advertising services such as high-stakes gambling to vulnerable individuals inferred by a digital platform to be in the midst of a manic episode is yet another.

'Personalisation' can also lead to price discrimination, like pricing based on an airline knowing this user has searched for a quote before; or market exclusion, like insurance products only being advertised to users already profiled as 'low risk', based on their online activities.[167]

Micro-targeting can also be used to manipulate behaviour, such as voting intentions.[168] Facebook in particular has been described as "a 'manipulation machine' that can be used to discourage black voters just as easily as to sell sneakers",[169] with researchers at the non-profit Data and Society Research Institute finding in 2018 that online behavioural advertising has been 'weaponised' to enable "political manipulation and other forms of anti-democratic strategic communication".[170]

The activities described above hold the potential to impact on individuals' autonomy, by narrowing or altering their market or life choices. Philosophy professor Michael Lynch has said that "taking you out of the decision-making equation" matters because "autonomy enables us to shape our own decisions and make ones that are in line with our deepest preferences and convictions. Autonomy lies at the heart of our humanity".[171]

[164] Alison Branley, "Google search data used by pharma giant to bombard users with ads for addictive opioids", ABC News Online, 13 July 2019; available at https://www.abc.net.au/news/2019-07-13/searches-data-mined-by-pharma-giant-to-promote-new-opioid/11300396

[165] Kramer, Adam DI, Jamie E. Guillory, and Jeffrey T. Hancock. "Experimental evidence of massive-scale emotional contagion through social networks." Proceedings of the National Academy of Sciences 111.24 (2014): 8788-8790. Available at https://www.pnas.org/content/111/24/8788.full

[166] Nitasha Tiku, "Get Ready for the Next Big Privacy Backlash Against Facebook", *Wired*, 21 May 2017; available at https://www.wired.com/2017/05/welcome-next-phase-facebook-backlash/

[167] Rafi Mohammed, "How Retailers Use Personalized Prices to Test What You're Willing to Pay", *Harvard Business Review*, 20 October 2017; available at https://hbr.org/2017/10/how-retailers-use-personalized-prices-to-test-what-youre-willing-to-pay

[168] Luke Dormehl, "Will Your Computer Tell You How to Vote?", *Politico Magazine*, 25 November 2014; available at https://www.politico.com/magazine/story/2014/11/computers-algorithms-tell-you-how-to-vote-113142

[169] Gilad Edelman, "Why Don't We Just Ban Targeted Advertising?", *Wired*, 22 March 2020; available at https://www.wired.com/story/why-dont-we-just-ban-targeted-advertising/

[170] Anthony Nadler, Matthew Crain, and Joan Donovan, Weaponizing the Digital Influence Machine: The Political Perils of Online Ad Tech, Data and Society Research Institute, 2018; available at https://datasociety.net/wp-content/uploads/2018/10/DS_Digital_Influence_Machine.pdf

[171] Michael Lynch, "Why does our privacy really matter?", *Christian Science Monitor*, 22 April 2016; available at https://www.csmonitor.com/World/Passcode/Security-culture/2016/0422/Why-does-our-privacy-really-matter

Rating the level of risk posed by online identifiers

As has been discussed above in Chapter 2, online identifiers are a tool for creating longitudinal records of individual behaviour. However, the privacy risk stems from the collated data, rather than the identifier itself. Whilst it is important to regulate the use of identifiers to stem data collection, it is also important to not focus solely on identifiers. In particular, if a data set is assembled through identifiers but shared without identifiers the privacy risks are not reduced, since the identifier would probably not have a played an important role in further linking of the data in any case.

It is also not possible to forbid the use of identifiers altogether, since their use is essential in authentication, session management, security management, and network routing. Whilst it is tempting to forbid long-life identifiers it is not clear what a 'safe' lifetime would be. Since the data itself could be the linkable identifier, the safe lifetime will be dependent on the quantity and resolution of the data being collected. For example, there is a high chance precise location data over a single day will be uniquely identifying, however, that probability is lower for course location data at postcode level. A further challenge is that neither probability is entirely predictable, for example, unusual movements will increase both probabilities, but a lack of movement would reduce them.

It is not possible to define any types of identifiers as overtly safe. Session identifiers are sometimes seen as low risk, but because the length of time they may operate for is unknown, particularly given the popularity of session restore in browsers,[172] and the nature of the data collecting could vary so no such guarantee can be made. Likewise, the purpose of the identifier is not fixed, so an identifier for authentication is essential, but the data collected through its use could be repurposed for tracking and profiling.

It could also be argued that some online identifiers are useful for storing preferences about targeted advertising and cookie consent. However, this approach seems flawed. Concerns about tracking protection being used as a tracking vector[173] has motivated the recent change to block all third-party cookies in the Safari browser.[174] Using identifiers to distinguish user level blocking risks doing the same thing. A better option would be to establish a safe baseline in the browser itself, with per site settings being modified on an opt-out basis, since someone opting in to tracking, and opting-out of the default protection, cannot expect their preference to not allow them to be tracked.

Clearly some identifiers are going to continue to be needed, however, a robust classification of identifiers as safe or unsafe seems infeasible, and may not be desirable, since it could be overly prescriptive allowing technological advances to bypass any restrictions. A report by Ireland's Data Protection Commission indicated problems with incorrect classification of

[172] See https://webkit.org/blog/10218/full-third-party-cookie-blocking-and-more/
[173] See https://webkit.org/blog/9661/preventing-tracking-prevention-tracking/
[174] See https://webkit.org/blog/10218/full-third-party-cookie-blocking-and-more/

cookies,[175] in particular around the definition of what is 'strictly necessary' in the context of the ePD.

A better approach would be to establish the principles in which an online identifier *should* operate. This Research Paper proposes that an online identifier should have the following properties:

- Observable
- Resettable
- Blockable

Observable
It is essential that a user is aware of online identifiers that are in use, so they can determine when they are being tracked. It would not be sufficient that the identifier is merely a part of the observable data, it must be readily observable as an online identifier. Of note, this would prohibit the use of caches and ETags, amongst others, for online identifiers. Likewise, it would prohibit the use of injected identifiers that the user never sees.

Resettable
It must be possible for a user to reset their online identifiers in a manner that will protect against their past, present, and future identifiers being able to be linked. If past, present, and future identifiers can be linked the reset power the user has diminishes to zero. Of particular note will be digital platform identifiers that currently are very effective at linking old and new identifiers after a reset. Likewise, biometrics, fingerprinting of devices or behaviours are inherently not resettable, it is not reasonable to expect someone to change their biometrics, buy a new device, or change their everyday behaviour to avoid tracking.

It is probably desirable to have a single location for the storage of identifiers to allow for easy resetting and deletion by the user. To some degree this is what cookies originally provided, in that the clearing of cookies used to be all that was required to remove online identifiers. However, with the development of new identifiers and technologies there are now multiple locations and services that need to be reset: advertising identifiers, cache identifiers, flash cookies, network identifiers, amongst others. If a user wishes to reset their identifiers they need to be able to reset them all at the same time, for example by using a single 'reset' button. Otherwise, the risk of zombie cookies remains. If a single location could be established in which identifiers could be stored, and therefore cleared or reset, it would assist the user in exercising control. Such a single defined store becomes even more important for smart devices and IoT devices, which may not have sophisticated user interfaces or provide sufficient access to clear identifiers from multiple locations. Furthermore, having a single unified approach would simplify user education and not require users to learn different procedures for different apps and devices.

[175] Data Protection Commission (Ireland), "Report by the Data Protection Commission on the use of cookies and other tracking technologies", 2020, Available at https://www.dataprotection.ie/sites/default/files/uploads/2020-04/Report%20by%20the%20DPC%20on%20the%20use%20of%20cookies%20and%20other%20tracking%20technologies.pdf

Blockable

It must be possible for the user to choose to block the collection of such identifiers. This may prevent access to a service, for example, if a user blocks an authentication token, however, the power to exercise a block should rest with the user. This would prohibit the current usage of AdvertisingIDs, which cannot be blocked, only reset. Likewise, biometrics cannot be easily blocked, nor can device or behaviour fingerprinting. Note, something should only be considered blockable if it can be blocked without undermining core functionality or incurring additional costs. For example, it could be argued that cache based identifiers are blockable by disabling the cache, however, to do so would incur additional bandwidth costs to the user, as well as slower performance and a worse user experience.

In essence what is required is transparency of identifier usage, combined with ultimate control resting with the end user to reset or block their usage.

Even if such a significant change in approach is not possible, the idealised properties set a baseline for evaluating the risk of current identifiers. As such, an identifier that is non-resettable, non-observable, and non-blockable would be considered to be high risk. By way of an example, biometric identifiers, and injected ISP identifiers would both fall into this category. As such, a classification of low to high risk could be established based on how many properties are delivered.

High Risk	Medium-High
• Biometrics • Injected ISP • DeviceID/MAC Address (where not-resettable) • Device Fingerprinting • User/Behaviour fingerprinting • ISP UIDH (super cookies)	• IP Address • Cache based Tracking (ETag) • AdvertisingID • Flash Cookies (considered not easily blockable or observable due to storage outside the browser)
Medium-Low	**Low**
• Local Storage (not considered observable) • Web Beacons/Pixel Tags (lack of observation of their use for cookie syncing)	• Session Cookies • Persistent cookies

However, even this classification could not be used universally. MAC addresses are clearly identifiers, are often non-resettable, non-blockable and not particularly observable. Nonetheless, they are essential to the function of the network. As such, blocking their usage outright is not an option.

A middle ground could be limiting tracking identifiers to those that meet the idealised attributes, combined with a shift to consider all data collected about the user or as a result of the user's actions (i.e. metadata) as being personal data.

It should also be noted that rating a particular identifier or technology as low risk does not imply it cannot be used for tracking, far from it. It merely indicates that the user can reasonably be assumed to be able to observe the usage, reset any identifiers stored, and outright block such identifiers if they choose. As such, effective user education, and the provision of tools and functionality on devices and in browsers that facilitate the exercising of such control is essential.

The risk associated with cookies is well established and increasingly well known. The public are adopting technologies to counter the effectiveness of cookies in tracking, including tools like Ad-Blockers – which have seen significant increase in usage over the past 10 years.[176] Recent announcements from Google to join other browsers in blocking third-party cookies may seem like an advantage for consumer privacy, however, it could be a reaction to more effective blocking of cookies and trackers, necessitating an alternative approach to be found. It may turn out that cookies were one of the easier identifiers to block and manage, and that whatever comes to replace them, for the purposes of tracking and profiling, leads to a worsening of individual privacy in the future.

It is therefore important to not focus solely on technical solutions, nor focus solely on identifiers. Attempting to regulate specific technologies will likely fail, as a result of the rapid development of technology presenting new or alternative approaches. Likewise, focusing solely on identifiers risks missing the fact that the data itself, once collected in sufficient quantity, becomes identifiable in and of itself.

The broader issue is the set of consequences arising from the collection and processing of data about users, combined with the ability to distinguish the target user from others in the future. The current capabilities of the major players in the advertising and profiling space vastly outstrip the capabilities of the average consumer. This creates an inherent power and information asymmetry, leading to engrained disadvantage for the end user. The lack of ability to control or prevent the collection of such data is one such example of this imbalance. As such, improved privacy protections are not just about privacy, they are as much about addressing the balance of power and information to better enable the end user to engage in the online environment on a level playing field.

[176] "Growth of the Blocked Web 2020 PageFair Adblock Report", Blockthrough; available at https://s3.amazonaws.com/media.mediapost.com/uploads/2020-PageFair_Blockthrough-Adblock-Report.pdf

Risks to children and young people

The issue of privacy for children and young people is closely tied with online identifiers due to the increased amount of time children and young people spend online. The 2017 Australian Child Health Poll revealed that “Almost all (94%) Australian teenagers and two-thirds (67%) of primary school-aged children and over a third (36%) of preschoolers have their own mobile screen-based device”[177] and that “Three in four teenagers and one in six primary school-aged children have their own social media accounts” with the average screen time, according to parents, being over 3 hours a day for the majority of children. With such extensive engagement, and usage of technology and online services, comes a greater risk of identifiers, tracking, and profiling.

Of particular concern is whether children are aware of the risks and sufficiently equipped to make the consent decisions being asked of them, when engaging with online services. Many platforms operate self-enforced restrictions, for example, asking the user for their age and not allowing children under the age of 13 access. This is in order to comply with the *Children*'*s Online Privacy Protection Act* in the USA,[178] However, the use of self-selected age fields motivate children to lie about their age in order to gain access. In some cases providing false information is facilitated by the app itself, for example, by preselecting a birth year above the age limit,[179] so just pressing next will allow access. Research on the US market indicates that a significant number of mobile apps are not compliant with US restrictions on tracking children.[180]

The significant amount of time children and young people are spending online, often unsupervised, brings into question whether parents are actually in a position to even be aware of what their children are ‘consenting’ to. This problem is not isolated to just their own devices. In 2015 researchers in Canada began a seven-year research project into “young people’s experiences of privacy and equality in networked spaces”,[181] which has raised particular concerns with the EdTech (Education Technology) market.[182] The rapid introduction of technology into EdTech, often by the same companies involved with tracking and profiling, is raising a number of concerns, including a 2018 FBI warning on the sensitivity and protection of the data being collected.[183] Frequently decisions about implementation are being undertaken at an institutional level, for example, by a school. As a result, parents may not be fully aware of the data being collected, and are often not even asked for their consent. That is not to say that decisions should be pushed onto parents either. Evaluating the privacy impact of new technologies is complicated, potentially beyond the capabilities of parents, teachers, and individual schools. Given the impact will be felt

[177] Rhodes, Anthea “Australian Child Health Poll 2017” See https://www.rchpoll.org.au/wp-content/uploads/2017/06/ACHP-Poll7_Detailed-Report-June21.pdf

[178] See https://www.ftc.gov/enforcement/rules/rulemaking-regulatory-reform-proceedings/childrens-online-privacy-protection-rule

[179] See https://www.nytimes.com/interactive/2018/09/12/technology/kids-apps-data-privacy-google-twitter.html

[180] Reyes, Irwin, et al. "“Won’t somebody think of the children?” examining COPPA compliance at scale." Proceedings on Privacy Enhancing Technologies 2018.3 (2018): 63-83.

[181] See http://www.equalityproject.ca/

[182] See https://theconversation.com/childrens-privacy-is-at-risk-with-rapid-shifts-to-online-schooling-under-coronavirus-135787

[183] See https://www.ic3.gov/media/2018/180913.aspx

across many children, it would seem some degree of central oversight and approval would be more appropriate.

More broadly, the long-term consequences of early use of online technologies, and the associated profiling, is not known. It poses many questions: Will children develop into adults who tolerate far greater invasion of their privacy because they have known no different? How will their life choices and opportunities be affected by having been profiled from a young age? Will they have the same opportunity to make mistakes and develop without fear that their actions are being recorded and profiled for future use by third parties?

It is not immediately clear how to resolve the challenge of protecting children and young people's online privacy. A purely technological solution seems unlikely, and the current approach of requiring or claiming parental consent seems deeply flawed. Technology may be able to assist in strengthening such consent procedures, for example, providing a single app or service for parents to provide consent for their children, moving away from the per app checkbox approach. However, any such implementation would need to be carefully overseen to ensure it did not become a source of privacy invasion itself. More broadly, this Research Paper argues the default protections offered to children and young people should be strengthened. It is not reasonable to expect children to be able to assert their rights themselves, necessitating a more constrained legislative environment to better protect their privacy.

Chapter 4

How online identifiers are currently regulated

Overview

This chapter demonstrates that a number of privacy statutes explicitly refer to online identifiers and/or location data within their definitions of personal information or personal data.

In addition many jurisdictions, whether through explicit mention in the statute, in case law or regulatory guidance, recognise that to the extent that a *device* can be identified (this unique mobile phone, that unique smart TV):

- a device identifier can be a proxy for an identifier for the individual who is the user of that device, and
- information about the use or physical movements of the device is information 'about' or 'relating to' the behaviour or physical movements of the individual who is the user of that device.

Nonetheless arguments are still being made by digital platforms, publishers, advertisers, ad brokers, data brokers and others that the data in which they trade is 'de-identified' or 'anonymised' or 'non-personal'.[184] These claims are made either to obfuscate the reality to consumers, or to claim to regulators that privacy or data protection laws do not apply to their practices.

In our previous Research Paper, we suggested that the Australian Privacy Act should no longer turn on such a narrow focus as 'identifiability', since identifiability of an individual is not the only vector for privacy harm in an online environment.

Australia

The Privacy Act 1988 defines 'personal information' as:

> "information or an opinion about an identified individual, or an individual who is reasonably identifiable:
>
> (a) whether the information or opinion is true or not; and

[184] Dr Katharine Kemp, "Submission in Response to the Australian Competition and Consumer Commission Ad Tech Inquiry Issues Paper", 26 April 2020; available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3587239

(b) whether the information or opinion is recorded in a material form or not".

While this definition of personal information does not call out cookies or online identifiers, where an online identifier meets the current test in the Privacy Act – i.e. it is found to be 'about' an individual who is at least 'reasonably identifiable' – then it is 'personal information'.

Other Australian jurisdictions would appear to agree with this assessment. In a Queensland case, an IP address was found not to be personal information, because even in the course of a police investigation, neither Telstra nor the police service could determine its originating source, and thus they could not reasonably identify the individual concerned.[185] However the implication was that so long as an individual's identity could be reasonably ascertained (the test in the Queensland statute), then IP addresses would be covered. Guidance from the Office of the Victorian Information Commissioner states that "unique machine addresses for computers connected to the internet (for example, IP addresses), 'cookies' and other monitoring software" will be included, to the extent that they enable an individual's identity to be reasonably ascertained.[186]

In relation to the use of cookies and online identifiers to collect personal information about individuals, there appears to be a significant disconnect between community expectations in Australia, and common business practices. Research conducted for the OAIC in 2017 found that Australians feel uncomfortable about online tracking, with 79% uncomfortable about targeted advertising based on their online activities, and 83% uncomfortable with social networking companies keeping databases of information about their online activities.[187] A survey conducted by Roy Morgan in 2018 found that almost 90% of Australians say it is unacceptable for social media and search companies to use their personal data in order to tailor ads and offers to consumers.[188]

European Union

The regulatory situation with regards to cookies and online identifiers within the EU is quite complex.

GDPR

Since May 2018, the GDPR has regulated the handling of 'personal data', which is defined to mean:

[185] *Lockyer Valley Regional Council and Queensland Police Service (311307)* at [26]; available at https://www.oic.qld.gov.au/decisions/lockyer-valley-regional-council-and-queensland-police-service.

[186] Office of the Victorian Information Commissioner, *Guidelines to the Information Privacy Principles*, 4th edition, November 2019, under 'Key concepts'; available at https://ovic.vic.gov.au/privacy/guidelines-to-the-information-privacy-principles/

[187] Office of the Australian Information Commissioner, *Community attitudes to privacy survey, Research report 2017*; see www.oaic.gov.au/resources/engage-with-us/community-attitudes/acaps-2017/acaps-2017-report.pdf

[188] See Roy Morgan, *Australians worried about online privacy but slow to act*, 6 July 2018; available at http://www.roymorgan.com/findings/7650-online-privacy-concerns-june-2018-201807060422

> "any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person".[189]

The GDPR's recitals further clarify that:

> "Natural persons may be associated with online identifiers provided by their devices, applications, tools and protocols, such as internet protocol addresses, cookie identifiers or other identifiers such as radio frequency identification tags. This may leave traces which, in particular when combined with unique identifiers and other information received by the servers, may be used to create profiles of the natural persons and identify them."[190]

The GDPR thus offers an explicit and expansive view of online identifiers as within the scope of its regulation. For example the UK's regulator, the Information Commissioner's Office (ICO), regards even so-called 'anonymous identifiers' such as advertising IDs as an example of an 'online identifier' which constitutes personal data under the GDPR.[191]

The GDPR sets out seven data protection principles, which regulate the processing (collection, use and disclosure) of personal data:

- Lawfulness, fairness and transparency in processing
- Purpose limitation
- Data minimisation
- Accuracy
- Storage limitation
- Integrity and confidentiality, and
- Accountability.

As such, the collection, use and disclosure of data in, from or via online identifiers must be lawful, fair and transparent, as well as necessary (or otherwise lawfully authorised) and proportionate. The purpose limitation principle means that data must be "collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes".[192] The data minimisation principle means that data should be "adequate, relevant and limited to what is necessary for the purposes for which

[189] Article 4 of the GDPR.

[190] Recital 30 of the GDPR.

[191] UK ICO, *Draft direct marketing code of practice*, v1.0 for public consultation, January 2020, p.96; available at https://ico.org.uk/media/about-the-ico/consultations/2616882/direct-marketing-code-draft-guidance.pdf

[192] Article 5 of the GDPR.

they are processed";[193] and should be processed "only if the purpose of the processing could not reasonably be fulfilled by other means".[194]

In addition to the seven data protection principles, the GDPR requires that any data processing (what we would call collection, use or disclosure) of personal data that is not from a special category (what we would call sensitive personal information) must only occur if it can rest on one of six grounds:[195]

- with the consent of the individual
- as necessary for the performance of a contract (where the data subject is or would be a party, including an employment contract)
- to fulfil legal obligations
- to protect vital interests (health and safety, etc)
- in the public interest or by an official authority acting under their local law, or
- for the legitimate interests of a private sector organisation, *except* where such interests are overridden by the interests or fundamental rights and freedoms of the data subject.

Direct marketing is mentioned as an example of a 'legitimate interest',[196] but nonetheless the GDPR requires that the organisation can justify (i) that it is pursuing a legitimate interest, (ii) that the use or disclosure is necessary for that purpose, and (iii) that the individual's privacy interests do not override the organisation's interests. Privacy advocates are focussing on this balancing act aspect of the GDPR to argue that online behavioural advertising using third party data is unlawful in the EU.[197]

ePD

In addition to the GDPR, there is the Directive on Privacy and Electronic Communications, which is commonly known as the ePrivacy Directive (the ePD). As a Directive, each EU Member State needed to implement its provisions into their local laws. An example is the UK's *Privacy and Electronic Communications Regulations* (PECR).

The ePD dates from 2002, and was amended in 2009. There is underway a process to replace the ePD with a new Regulation (the ePR), which as a Regulation like the GDPR would apply directly to all EU Member States. The drafting of the ePR was originally proceeding in parallel with the GDPR, but became mired in debates which considerably slowed its progress.

[193] Article 5 of the GDPR.
[194] Recital 39 of the GDPR.
[195] Article 6 of the GDPR.
[196] Recital 47 of the GDPR.
[197] Gilad Edelman, "Why Don't We Just Ban Targeted Advertising?", *Wired*, 22 March 2020; available at https://www.wired.com/story/why-dont-we-just-ban-targeted-advertising/

Unlike the GDPR, the ePD does not turn on a definition of ‘personal data’, and does not limit its focus to the rights of individuals. Its objectives include to protect the confidentiality and integrity of electronic communications, including between organisations, when using a public communications network and publicly available electronic communications services. It also covers broader telecommunication matters such as caller line identification and number portability.

Clause 5(3) of the ePD provides:

> “Member States shall ensure that the storing of information, or the gaining of access to information already stored, in the terminal equipment of a subscriber or user is only allowed on condition that the subscriber or user concerned has given his or her consent, having been provided with clear and comprehensive information, in accordance with Directive 95/46/EC, inter alia, about the purposes of the processing. This shall not prevent any technical storage or access for the sole purpose of carrying out the transmission of a communication over an electronic communications network, or as strictly necessary in order for the provider of an information society service explicitly requested by the subscriber or user to provide the service.”

The phrase “storing of information, or the gaining of access to information already stored, in the terminal equipment of a subscriber or user” encompasses the placing of cookies or other online identifiers on to an individual’s device.

Clause 6(3) of the ePD provides:

> “For the purpose of marketing electronic communications services or for the provision of value added services, the provider of a publicly available electronic communications service may process the data referred to in paragraph 1 to the extent and for the duration necessary for such services or marketing, if the subscriber or user to whom the data relate has given his or her prior consent. Users or subscribers shall be given the possibility to withdraw their consent for the processing of traffic data at any time.”

Thus both the placement of cookies (or other online identifiers) on to a user’s device such as their computer or smartphone, and the marketing of services to the individual via email or SMS, requires the individual’s prior consent, unless a cookie or similar was “strictly necessary” for the service to be delivered.

However up until recently, there was considerable variation in how the ePD’s provisions were incorporated into local legislation and enforced by European privacy regulators. There were different interpretations of what constituted a “strictly necessary” cookie; and there were different interpretations of what constituted a “consent” in relation to cookies. For example the UK, France and Germany suggested that consent to a cookie could be implied by the user’s continued use of a service after a cookie banner was displayed; others took the position that at least an opt-out option must be available to users; while Italy required proactive opt-in to cookies.

In October 2019 the Court of Justice of the European Union (CJEU) delivered a case which considered the question of what constitutes a valid consent in relation to cookies under the ePD.[198] The CJEU found that:

- opt out mechanisms (e.g. pre-ticked boxes) do not amount to valid consent, because they cannot demonstrate the voluntary intent of the individual: the CJEU stated that "consent given in the form of a preselected tick in a checkbox does not imply active behaviour on the part of a website user";[199]
- further, opt out mechanisms do not amount to valid consent, because they cannot demonstrate that the decision was informed: the CJEU stated that "it would appear impossible in practice to ascertain objectively whether a website user had actually given his or her consent to the processing of his or her personal data by not deselecting a pre-ticked checkbox nor, in any event, whether that consent had been informed. It is not inconceivable that a user would not have read the information accompanying the preselected checkbox, or even would not have noticed that checkbox, before continuing with his or her activity on the website visited";[200]
- bundled agreements do not amount to valid consent, because they do not meet the test of specificity: the CJEU stated that consent must "be 'specific' in the sense that it must relate specifically to the processing of the data in question and cannot be inferred from an indication of the data subject's wishes for other purposes";[201] and
- a lack of "clear and comprehensive" information about consequences means that a consent cannot be regarded as informed; a user must be "in a position to be able to determine easily the consequences of any consent he or she might give and ensure that the consent given is well informed".[202]

Although the Planet49 case was about an opt out mechanism, it would appear the same reasoning would apply to the use of cookie banners or notice mechanisms (e.g. claiming in a notice that consent is obtained by having users continuing to use a website or app), which do not even offer an opt out choice.

As a result of the commencement of the GDPR and/or the Planet49 case, a number of European privacy regulators including the UK, France, Germany and Spain have revised their guidance on cookies in the past year.[203] Most importantly, the European Data Protection Board (EDPB), which issues guidance on behalf of all EU privacy regulators as a group, issued guidance in May 2020 which effectively re-stated the impact of the Planet49

[198] See the CJEU's *Planet49* judgment at http://curia.europa.eu/juris/document/document.jsf?text=&docid=218462&pageIndex=0&doclang=EN&mode=req&dir=&occ=first&part=1&cid=1447493

[199] See para [52] of the CJEU's Planet49 judgment

[200] See para [55] of the CJEU's Planet49 judgment

[201] See para [58] of the CJEU's Planet49 judgment

[202] See para [74] of the CJEU's Planet49 judgment

[203] See for example the UK's guidance at https://ico.org.uk/for-organisations/guide-to-pecr/cookies-and-similar-technologies/ and France's guidance and consultation plans at https://www.cnil.fr/en/cnil-launches-public-consultation-its-draft-recommendation-cookies-and-other-trackers

case: the placement of cookies or other online identifiers on a user's device can only be done on the basis of a voluntary, informed, specific and proactive (opt-in) consent, in which the user is free to reject the cookies, and silence is an indication of rejection rather than acceptance.[204]

Significantly, the EDPB's May 2020 guidelines note that proactive consent is required not only in relation to cookies, but also "most online marketing messages … and online tracking methods including by the use of cookies or apps or other software".[205] The EDPB's 2019 position on the ePR drafting process also noted that "not only cookies, but every tracking technology is already subject to consent of the user or is subject to one of the exceptions specified in the ePrivacy Directive".[206] The UK ICO's 2019 guidance states that the scope of the PECR includes not only cookies but also "HTML5 local storage, Local Shared Objects and fingerprinting techniques … scripts, tracking pixels and plugins".[207]

Beyond online identifiers *per se*, in January 2020 the UK's ICO proposed a new code for the direct marketing industry, the effect of which would appear to be to make all online targeted advertising and messaging unlawful in the absence of a proactive consent from the user.[208]

It is therefore to be expected that even without (or prior to) the passage of the ePR, enforcement of the existing law within Europe and the UK is now expected to be significantly less tolerant of the use of online identifiers, in the absence of a valid consent, than was until recently the case. The progress of a number of representative complaints from privacy advocacy bodies also appears to be a factor.[209]

ePR

One of the reasons the ePD is in need of updating is that it currently only regulates what happens on public electronic communication networks. It therefore excludes from scope so-called 'over the top' services – i.e. IP-based communication services such as Apple's Facetime calls, online chat functionality within a digital platform, and Facebook's WhatsApp messaging – even if they are functionally equivalent to a service operated over traditional telephony networks.

[204] European Data Protection Board, *Guidelines 05/2020 on consent under Regulation 2016/679*, May 2020; available at https://edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-052020-consent-under-regulation-2016679_en

[205] See p.5 of the EDPB's May 2020 guidelines.

[206] EDPB, "Statement of the EDPB on the revision of the ePrivacy Regulation and its impact on the protection of individuals with regard to the privacy and confidentiality of their communications", March 2019, p.2; available at https://edpb.europa.eu/our-work-tools/our-documents/drugi/statement-edpb-revision-eprivacy-regulation-and-its-impact_en

[207] See the UK ICO's guidance at https://ico.org.uk/for-organisations/guide-to-pecr/cookies-and-similar-technologies/

[208] UK ICO, *Draft direct marketing code of practice*, v1.0 for public consultation, January 2020; available at https://ico.org.uk/media/about-the-ico/consultations/2616882/direct-marketing-code-draft-guidance.pdf

[209] See the rationale from the French regulator CNIL for its new regulatory stance at https://www.cnil.fr/en/cnil-launches-public-consultation-its-draft-recommendation-cookies-and-other-trackers

While negotiation over the drafting of the ePR is still on-going, the ePR is currently expected to:

- extend to 'over the top' services
- also regulate messages 'sent or presented to' individuals including via pop-ups, social media marketing, etc
- increase penalties in line with GDPR
- have extra-territorial reach in line with GDPR
- only allow secondary use of metadata / location data if anonymised
- re-define what is covered by the 'strictly necessary' exception; there is debate that this might allow audience measurement (e.g. first party website analytics) as well as security updates
- may introduce a new exception, to allow, as an alternative to user consent, 'legitimate interests'[210] (which is a ground under which data processing by the private sector is authorised under the GDPR), although this is opposed by the EDPB[211] and apparently some Member States[212]

Canada

Privacy Act

The Canadian Privacy Act 1985 regulates the federal public sector. Its definition of 'personal information' specifically includes "any identifying number, symbol or other particular assigned to the individual".[213]

PIPEDA

The Canadian Personal Information Protection and Electronic Documents Act 2000 (PIPEDA) regulates the handling of personal information by the private sector. Personal information is simply defined as "information about an identifiable individual".[214]

PIPEDA includes a 'fairness' gatekeeper provision. Section 5(3) of PIPEDA says: "An organization may collect, use or disclose personal information only for purposes that a reasonable person would consider are appropriate in the circumstances." The Office of the Privacy Commissioner of Canada (OPCC) publishes guidance on 'no-go zones', based on

[210] See latest May 2020 communique at https://data.consilium.europa.eu/doc/document/ST-9351-2019-INIT/en/pdf

[211] EDPB, "Statement of the EDPB on the revision of the ePrivacy Regulation and its impact on the protection of individuals with regard to the privacy and confidentiality of their communications", March 2019, p.3

[212] See https://www.huntonprivacyblog.com/2020/06/04/eu-council-presidency-releases-progress-report-on-draft-eprivacy-regulation/

[213] Section 3, Privacy Act 1985 (Canada)

[214] Section 2(1), Personal Information Protection and Electronic Documents Act 2000 (Canada)

court interpretations of s.5(3) as well as consultations with stakeholders and focus groups.[215] One such 'no-go zone' is "Profiling or categorization that leads to unfair, unethical or discriminatory treatment contrary to human rights law".

The OPCC issued a policy position in 2015 in relation to online behavioural advertising and the use of cookies, web beacons, device fingerprinting and other online identifiers.[216] Having consulted both industry and consumers, the OPCC noted "discrepancies in characterizing what is and is not personal information", as well as a "lack of transparency around tracking, profiling and targeting and what this means in terms of obtaining meaningful consent – a requirement under PIPEDA".

In reviewing the question as to whether online identifiers used for online behavioural advertising constitute personal information under PIPEDA, the OPCC concluded:

> "Much of the information at issue in OBA – third-party tracking cookies, IP addresses, browser settings – may not be personal information in and of itself, in that, alone, it may say nothing *about* an *identifiable* individual. However, when combined and used for the purpose of profiling a user in order to target advertisements to him or her based on inferred interests, the information can become information about an identifiable individual".

The OPCC's stated policy position is that:

> "Taking a broad, contextual view of the definition of personal information, the OPC will generally consider information collected for the purpose of OBA to be personal information, *given*: the fact that the purpose behind collecting information is to create profiles of individuals that in turn permit the serving of targeted ads; the powerful means available for gathering and analyzing disparate bits of data and the serious possibility of identifying affected individuals; and the potentially highly personalized nature of the resulting advertising."

The OPCC then considered whether or not the use of such personal information could be considered fair under s.5(3) of PIPEDA. While declining to call online behavioural advertising *per se* a 'no-go zone', the OPCC did warn that meaningful consent was required:

> "Given that some users may be uncomfortable with the notion of being "followed" around the web, yet think that advertisements geared to their interests are useful and, given that services are generally free and users ought to expect that some personal information may be needed to access services and information, OBA may be considered an appropriate purpose for the collection, use and/or disclosure of personal information from the perspective of the reasonable person. However, OBA should not be considered a term or condition for individuals to use the Internet

[215] See https://www.priv.gc.ca/en/privacy-topics/collecting-personal-information/consent/gd_53_201805/#s4
[216] See https://www.priv.gc.ca/en/privacy-topics/technology/online-privacy-tracking-cookies/tracking-and-ads/bg_ba_1206/

> generally. There are still other forms of advertising that web sites can rely on. There must also be meaningful consent, and there should be limitations on the types of information collected and used for profiling. Safeguarding the information is also vital, as is limiting the retention of the data to the least amount of time possible".

The OPCC then suggested that 'meaningful consent' could be opt-out, but only so long as:

- "Individuals are made aware of the purposes for the practice in a manner that is clear and understandable – the purposes must be made obvious and cannot be buried in a privacy policy. Organizations should be transparent about their practices and consider how to effectively inform individuals of their OBA practices, by using a variety of communication methods, such as online banners, layered approaches, and interactive tools;
- Individuals are informed of these purposes at or before the time of collection and provided with information about the various parties involved in OBA;
- Individuals are able to easily opt-out of the practice - ideally at or before the time the information is collected;
- The opt-out takes effect immediately and is persistent;
- The information collected and used is limited, to the extent practicable, to non-sensitive information (avoiding sensitive information such as medical or health information); and
- Information collected and used is destroyed as soon as possible or effectively de-identified".

However the OPCC also identified two practices which it considered non-compliant with s.5(3) of PIPEDA:

- online tracking of children and marketing targeted at children, and
- "Zombie cookies, supercookies, third-party cookies that appear to be first-party cookies, device fingerprinting, and other techniques that cannot be controlled by individuals".[217]

United States

CCPA

The 2018 California Consumer Privacy Act (CCPA) is, of the privacy statutes we have reviewed for this paper, the most explicit in its inclusion of online identifiers. The CCPA expressly includes, within its definition of personal information, data which is "capable of being associated with, or could reasonably be linked, directly or indirectly, with a particular consumer or household", without first needing to pass an identifiability test.[218]

[217] See https://www.priv.gc.ca/en/privacy-topics/technology/online-privacy-tracking-cookies/tracking-and-ads/bg_ba_1206/
[218] CCPA section 1798.140(o)(1)

The definition is fleshed out with a number of examples, including:

> (A) "Identifiers such as ... unique personal identifier, online identifier, Internet Protocol address..."

and

> (F) "Internet or other electronic network activity information, including, but not limited to, browsing history, search history, and information regarding a consumer's interaction with an Internet Web site, application or advertisement"

and

> (K) "Inferences drawn from any of the information identified in this subdivision to create a profile about a consumer reflecting the consumer's preferences, characteristics, psychological trends, predispositions, behaviour, attitudes, intelligence, abilities, and aptitudes".

This theme is further fleshed out within the definition of 'unique personal identifier', which means:

> "a persistent identifier that can be used to recognize a consumer, a family, or a device that is linked to a consumer or family, over time and across different services, including, but not limited to, a device identifier; an Internet Protocol address; cookies, beacons, pixel tags, mobile ad identifiers, or similar technology..."[219]

However the efficacy of the CCPA as a regulatory response to the privacy risks posed by online identifiers remains to be seen. Having only commenced in January 2020, enforcement outcomes are understandably not yet forthcoming. In our view, the CCPA's overall focus will limit its ability to effect real change in the online ecosystem. Rather than a set of privacy principles regulating the lifecycle of personal information processing, the CCPA creates a mechanism by which consumers can direct companies with a 'do not sell' communication.

CCPA has little to say about the lawfulness, proportionality or fairness of companies collecting, using or disclosing online identifiers in the first place; it merely requires privacy notices to be published, and creates an opportunity for consumers to 'opt out' of a particular type of practice. In doing so, the CCPA may end up legitimising much of the data broking behaviour it seeks to restrict, giving companies carte blanche to track and profile consumers, and monetise consumers' personal information, so long as an individual consumer has not found the company and exercised their right to opt out.

[219] CCPA section 1798.140(o)(1)(x)

COPPA

The Children's Online Privacy Protection Act 1998 applies to the online collection of personal information about children under 13 years of age.

Personal information is defined to mean:

> “individually identifiable information about an individual collected online, including:
>
> (1) …
>
> (7) A persistent identifier that can be used to recognize a user over time and across different Web sites or online services. Such persistent identifier includes, but is not limited to, a customer number held in a cookie, an Internet Protocol (IP) address, a processor or device serial number, or unique device identifier…”.[220]

The collection of personal information about children under 13 typically requires parental consent. While there are some exemptions, COPPA’s scope includes websites and online services operated for commercial purposes that are either directed towards children under 13, or where the operator has actual knowledge that children under 13 are providing information online.

The Federal Trade Commission is the regulator. To avoid the costs of compliance with COPPA – parental consent being difficult to obtain and verify online[221] – many online services such as social media sites simply state that users under 13 are not allowed to join or use their services. Companies which have been found in violation of COPPA include Music.ly (now TikTok), which knew - and promoted the fact - that many of its users were younger than 13;[222] an advertising network which tracked geolocation of users under 13 even if location tracking permission had been denied;[223] and YouTube which knowingly tracked and sold ads targeted at children, a case which was settled for a US$170M penalty and which has prompted recent changes in practices for users posting videos to YouTube.[224]

Nigeria

Within its definition of ‘personal data’, the 2019 Nigerian privacy regulation, issued by the National Information Technology Development Agency, defines “an identifiable natural person” as “one who can be identified, directly or indirectly, in particular by reference to an

[220] Section 312.2, Part 312 of Title 16: Commercial Practices in the Electronic Code of Federal Regulations; available at https://www.ecfr.gov/cgi-bin/text-idx?SID=4939e77c77a1a1a08c1cbf905fc4b409&node=16%3A1.0.1.3.36&rgn=div5#_top

[221] See the steps required at https://www.ftc.gov/tips-advice/business-center/guidance/childrens-online-privacy-protection-rule-six-step-compliance#step4

[222] See https://www.ftc.gov/news-events/blogs/business-blog/2019/02/largest-ftc-coppa-settlement-requires-musically-change-its

[223] See https://www.ftc.gov/news-events/press-releases/2016/06/mobile-advertising-network-inmobi-settles-ftc-charges-it-tracked

[224] See https://edition.cnn.com/2019/09/04/tech/google-youtube-ftc-settlement/index.html

identifier such as... an identification number (or) “online identifier”. A set of examples includes a “MAC address, IP address, IMEI number, IMSI number, SIM” as well as “information that can be used on its own or with other information to identify, contact, or locate a single person, or to identify an individual in a context”.[225]

The Regulation requires the use of technical methods to collect personal data, such as cookies and web tokens, to be outlined in the organisation’s privacy policy.[226] Individuals have the right to object to the use of their personal data for direct marketing.[227]

Guidelines issued under the Regulation in May 2020 focus on the processing of personal data by public sector institutions, and include lawful grounds for processing which are similar to the GDPR.[228] We could not find that similar guidelines have yet been issued for private sector organisations.

Singapore

The Personal Data Protection Act 2012 of Singapore defines ‘personal data’ to mean

> “data, whether true or not, about an individual who can be identified —
>
> (a) from that data; or
>
> (b) from that data and other information to which the organisation has or is likely to have access”.[229]

Section 18 of the Act is a purpose limitation principle, which requires an organisation to only collect, use, and disclose personal data about an individual for purposes that a reasonable person would consider appropriate in the circumstances, and which, if applicable, have been notified to the individual concerned.

The Personal Data Protection Commission Singapore has issued guidelines which state:

> “An IP address, or any other network identifier such as an IMEI number, may not be personal data when viewed in isolation, because it simply identifies a networked device. However, IP addresses have the potential of identifying unique individuals through their activities, especially when combined with traces of information that individuals leave on these networked devices as they interact with the Internet.
>
> …

[225] Clause xix in Part 1.3 of the *Nigeria Data Protection Regulation 2019*; available at https://nitda.gov.ng/wp-content/uploads/2019/01/NigeriaDataProtectionRegulation.pdf
[226] Part 2.5 of the *Nigeria Data Protection Regulation 2019*
[227] Part 2.8 of the *Nigeria Data Protection Regulation 2019*
[228] See https://nitda.gov.ng/wp-content/uploads/2020/05/GuidelinesForImplementationOfNDPRInPublicInstitutionsFinal.pdf
[229] Section 2 of the Personal Data Protection Act 2012; available at https://sso.agc.gov.sg/Act/PDPA2012#pr2-

> Along with other information such as cookies… IP addresses can identify individuals, and are likely to be personal data in such context”.[230]

The guidance goes on to discuss cookies, and distinguishes between cookies which perform routine functions to enable websites to work properly or to reflect a user’s wishes, and non-essential tracking cookies. In particular, the Personal Data Protection Commission states:

> “Where behavioural targeting involves the collection and use of personal data, the individual’s consent is required”.[231]

While consent may be deemed for some activities, such as where cookies are essential to functioning and “the individual voluntarily provides the personal data for that purpose of the activity” (such as remembering a shopper’s details, at their request, for future use),[232] the guidance also notes that “the mere failure of an individual to actively manage his browser settings does not imply that the individual has consented to the collection, use and disclosure of his personal data by all websites for their stated purpose”.[233] While opt-out is not prohibited, opt-in is strongly preferred.

Further, under s.14 of the Act, consent cannot be a condition for the provision of a product or service, beyond what is reasonable to provide the product or service to the individual.

The summary position in Singapore therefore appears to be that cookies which are used to profile and target messaging will require both notice to the individual,[234] and some form of demonstrable (i.e. opt-in) consent from the individual.

Japan

The Act on the Protection of Personal Information of 2003[235] defines ‘personal information’ as follows:

> (1) “Personal information” in this Act means that information relating to a living individual which falls under any of each following item:

[230] Part 6.1-6.2 of the *Advisory Guidelines on the Personal Data Protection Act for Selected Topics*, Personal Data Protection Commission Singapore, October 2019; available at https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Advisory-Guidelines/AG-on-Selected-Topics/Chapter-6-9-Oct-2019.pdf?la=en

[231] Part 6.11

[232] Part 6.8

[233] Part 6.9

[234] See also the Commission’s guidance on notification at https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Other-Guides/Guide-to-Notification-260919.pdf

[235] See https://www.ppc.go.jp/files/pdf/Act_on_the_Protection_of_Personal_Information.pdf for a tentative translation issued by the Personal Information Protection Commission Japan

> > (i) those containing a name, date of birth, or other descriptions etc… whereby a specific individual can be identified (including those which can be readily collated with other information and thereby identify a specific individual)
> >
> > (ii) those containing an individual identification code
>
> (2) An "individual identification code" in this Act means those prescribed by cabinet order which are any character, letter, number, symbol or other codes falling under any of each following item.
>
> > (i) those able to identify a specific individual that are a character, letter, number, symbol or other codes into which a bodily partial feature of the specific individual has been converted in order to be provided for use by computers
> >
> > (ii) those character, letter, number, symbol or other codes which are assigned in regard to the use of services provided to an individual or to the purchase of goods sold to an individual, or which are stated or electromagnetically recorded in a card or other document issued to an individual so as to be able to identify a specific user or purchaser, or recipient of issuance by having made the said codes differently assigned or, stated or recoded for the said user or purchaser, or recipient of issuance".

It appears that several types of information have been prescribed by Cabinet Order as an "individual identification code", including DNA information and various types of information about an individual's face, eyes, gait and fingerprints.[236] However to date there appears to be no specific regulation or guidance issued in relation to cookies or other online identifiers.

News reports from late 2019 suggested that the Personal Information Protection Commission and/or the Fair Trade Commission of Japan were looking at regulating cookies, including the prospect of making cookies 'opt-in' only.[237]

On 5 June 2020, a bill to amend the Act was passed by the Diet (Parliament of Japan) and is set to come into effect within two years. While an English translation is not yet available, from specialist privacy law news reports it would appear that the amendments will have the effect of more directly addressing cookies and online identifiers:

> "even if a discloser cannot identify a specific individual based on cookies, the provision of cookies to third parties would become subject to consent requirements so long as it is obvious to the discloser that the recipient may identify an individual".[238]

[236] See https://www.ppc.go.jp/files/pdf/Cabinet_Order.pdf

[237] See https://asia.nikkei.com/Politics/Japan-weighs-tighter-protections-on-internet-user-data and http://www.asahi.com/ajw/articles/AJ201910290065.html

[238] Advice from DataGuidance, June 2020; available to subscribers at https://www.dataguidance.com/notes/japan-cookies-similar-technologies

Summary position

Globally, other jurisdictions have more modern definitions than that found in the Australian Privacy Act. Many clearly anticipate device identifiers, online identifiers and location data being used to identify (in its broadest sense) individuals, and thus trigger the application of privacy laws. Some jurisdictions (Europe, California) specifically regulate cookies and other online identifiers, while others apply existing privacy frameworks. Europe and Singapore require consent (opt-in) for non-essential cookies and other online identifiers, with Japan possibly soon to follow. Canada and the USA prohibit online behavioural advertising targeted at children, but otherwise take an opt-out stance. Europe and Canada also use overarching fairness frameworks to judge practices on a case-by-case basis.

We caution against following the European approach of attempting to regulate particular technologies, as the ePD does. In our view, by focussing on particular technologies, the ePD quickly became out-of-date (which is why it already requires replacing), and also fails to offer a risk-based approach to cookies and online identifiers. All implementations face the same 'strictly necessary or you need consent' regulatory stance, regardless of the actual privacy risk posed to any individual.

Nonetheless without even needing to wait for the ePD's replacement ePR, the combination of the GDPR, the Planet49 case and other privacy class actions waiting in the wings[239] – along with Google's decision to block third party cookies - is hastening the demise of third party cookie-based online behavioural advertising.

However as Chapter 2 demonstrates, cookies are by no means the only technology at play in this space. Arguably, the demise of third party advertising cookies will yield the highest dividends not for consumers, but for the two major digital platforms, Facebook and Google. As platforms with identifiable users, hosting content on their own websites and apps, the two dominant players can continue to track and profile the behaviour of billions of users,[240] and thus offer brands a direct way of targeting their advertising spend.

Facebook in particular offers a powerful combination of data about Facebook users, which includes not only what users provide directly (e.g. their posts, shares and 'likes'), but also what is observed about them (e.g. what posts a user clicks on, spends time reading, as well as what other third party websites the user frequents, purchases from, etc) and what is inferred about them as a result.

In our view, the Canadian fairness framework approach is preferable, because rather than focusing on any particular type of technology, it focuses on intent and outcomes, as well as whether there are meaningful opportunities for individuals to understand and control what is happening to their data.

[239] Gilad Edelman, "Why Don't We Just Ban Targeted Advertising?", *Wired*, 22 March 2020; available at https://www.wired.com/story/why-dont-we-just-ban-targeted-advertising/

[240] This tracking occurs even when a user is using incognito browsing; see https://www.itnews.com.au/news/google-faces-us5-billion-lawsuit-for-tracking-private-internet-use-548892

Chapter 5

Conclusions

Online identifiers facilitate privacy harms

Online identifiers have three roles to play in tracking, profiling and delivering messages to individuals; first to facilitate longitudinal data collection, second to aid the linking of disparate datasets, and third to allow the targeting of the user with the results of the analysis of the collected data.

However, the privacy risks stem from the data collected and generated from the use of online identifiers, rather than from identifiers themselves. Whilst it is important to regulate the use of identifiers to stem intrusive forms of data collection, it is also important to not focus solely on identifiers. In particular, if a data set is assembled through the use of online identifiers but later shared without the identifiers, the privacy risks are not necessarily reduced.

It is also not possible to prohibit the use of online identifiers altogether, since their use is essential in authentication, session management, security management, and network routing. Nor is it possible to define any types of identifiers as overtly ‘safe’. Even ‘essential’ identifiers can be repurposed for tracking and profiling.

As Chapter 3 has demonstrated, the privacy harms facilitated by the unfair and/or intrusive use of online identifiers to track, profile and micro-target individuals online - and even offline - include social and market exclusion, price discrimination resulting in economic inequality, prejudice and discrimination, manipulation leading to negative physical and mental health impacts, and manipulation of voting intentions.

New regulatory responses are necessary

Data is the lifeblood of the digital economy, and will increasingly power decision-making in all sectors of the economy.

Robust data protection regulation is necessary to achieve both consumer protection outcomes, and consistency of the playing field for industry. It will therefore be critical to ensure that the Privacy Act remains fit for its purpose of enabling effective regulation of personal information handling, in line with community and business expectations.

Whether or not any particular piece of data meets the definition of 'personal information' is a threshold legal issue for the operation of privacy law in Australia: the definition of 'personal information' determines the boundaries of what is regulated, and what is protected. Yet through its Digital Platforms Inquiry, the ACCC found that the current definition of 'personal information' suffers from a lack of certainty around its coverage of technical data, including online identifiers.

In recommending that the Privacy Act should be amended to more explicitly include online identifiers, the ACCC is not alone. There is increasingly global recognition that online identifiers pose privacy risks (in that they facilitate privacy harms), and require more consistent and robust regulation.

However we caution against a regulatory response which relies primarily on the regulation of any particular type of technology. As Chapter 2 has demonstrated, cookies are neither inherently good, nor inherently bad. Similarly, personalisation of messaging to users is neither inherently good, nor inherently bad. We therefore also caution against a regulatory response which overly restricts all practices, rather than focussing on the level of harm posed by any particular practice. For this reason, we have not recommended following the model currently proposed by the UK's ICO, the effect of which would appear to be to make all online targeted advertising and messaging unlawful in the absence of a proactive consent from the user.

Further, as Chapter 4 has demonstrated, the European approach to regulating cookies and other tracking technology via the ePrivacy Directive has produced legislation which is already out-of-date. The current reform process is mired in debates over what is or is not a 'strictly necessary' cookie, rather than debating what is fair or intrusive. Meanwhile that debate is increasingly moot; Chapter 2 provides a number of examples, such as Login Management and reCAPTCHA, in which 'strictly necessary' identifiers have been re-purposed for tracking.

In our view, the Canadian fairness framework approach is preferable, because rather than focusing on any particular type of technology, it focuses on intent and outcomes, as well as whether there are meaningful opportunities for individuals to understand and control what is happening to their data.

Guiding principles in framing our recommendations

For the reasons outlined above, we suggest that the appropriate regulatory response is to focus on regulating privacy-intrusive behaviours or practices, rather than any particular technology. As such, we suggest explicitly bringing online identifiers within the Privacy Act's scope, but then allowing the Australian Privacy Principles (APPs) to do the heavy lifting, in terms of determining what use cases will be considered lawful and fair, and what will not.

We also suggest that any proposed reform must be mindful of the need for global consistency, which is beneficial for consumers, regulated entities and regulators alike; but must also ensure that the definition is 'fit for purpose' for Australian conditions now and into the future.

Our recommendations in the next chapter have therefore been drafted with the following objectives:

(a) Resolving the lack of clarity around coverage of technical data, in line with the ACCC's recommendations

(b) Enabling consistency with the GDPR where suitable

(c) Maintaining the technological neutrality of the Privacy Act

(d) Making minimal regulatory change for maximum effectiveness

(e) Updating the Privacy Act to ensure it remains fit for purpose in protecting against multiple forms of privacy harms in digital environments, while

(f) Avoiding regulatory over-reach into technologies, practices or behaviours which do *not* pose risks of privacy harms.

Chapter 6

Recommendations

In this chapter we make a number of recommendations, in relation to:

- reforming the Privacy Act
- developing Codes and guidelines under the Privacy Act, and
- other policy responses the OAIC could take.

Following this chapter, in Appendix A, we have mapped out examples of how implementation of our recommendations for reform would likely impact on some of the practices outlined in this report.

Reforms to the Privacy Act

1. Amend the definition of personal information as per Recommendations 1 and 2 in our previous Research Paper

Rationale

So as to enable clarity and consistency in the application of privacy law, and to protect against the potential privacy harms enabled by tracking, profiling and targeting individuals online, our previous Research Paper concluded that the Australian Privacy Act should be amended, to incorporate a definition for the word 'identifiable':

> "(i) able to be identified, *or* (ii) able to be discerned or recognised as an individual distinct from others, regardless of whether their identity can be ascertained or verified"

Our previous Research Paper suggested that the test for identifiability should be that an individual will be considered "able to be discerned or recognised as an individual distinct from others":

> "if the individual, or a device linked to the individual, could (whether online or offline) be surveilled, tracked or monitored; or located, contacted or targeted; or profiled in order to be subjected to any action, decision or intervention including the provision or

withholding of information, content, advertisements or offers; or linked to other data which is about or relates to the individual".

This recommendation would ensure that to the extent that the various practices described in Chapter 2 are aimed at individuals, those practices should not escape regulation simply on the basis that they claim to be using data that is not 'personal information' because the individuals are not 'identifiable'.

The effect would be to treat most online identifiers as 'personal information', and thus their collection, use and disclosure would be regulated under the APPs.

This should also ensure that newer technologies which avoid sharing identifiers altogether, such as those discussed in Chapter 2 under *Non-cookie based tracking and identifiers*, are still appropriately regulated to the extent that they enable individuals to be profiled and targeted for messaging.

However the additional layer to the test for identifiability mentioned above aims to ensure that the scope of the regulation does not over-reach into technologies which do *not* pose risks of privacy harms, such as the use of sessional or load-balancing cookies which are necessary to make a website work, but which do not then continue to track the user.

2. Amend the definition of 'de-identified' as per Recommendation 3 in our previous Research Paper

Rationale

Together with recommendation 1 above, this recommendation would ensure that to the extent that the various practices described in Chapter 2 are aimed at individuals, those practices should not escape regulation simply on the basis that they claim to be using 'de-identified' or 'non-personal' data. This would also ensure that the data collected or generated about individuals via online identifiers is also regulated, even if the identifier is later stripped out from the dataset, so long as it is unit record level data.

3. **Amend the definition of 'consent' to mean a clear affirmative act of agreement that is freely given, specific, unambiguous and informed, current and given by a person with capacity.**

Rationale

The OAIC has noted that "'consent' is only a meaningful and effective privacy self-management tool where the individual actually has a choice and can exercise control over their personal information".[241]

The current definition of consent in the Privacy Act is simply "express consent or implied consent".[242] Recommendation 16(c) of the ACCC's Digital Platforms Inquiry was that the definition of 'consent' in the Privacy Act should be amended:

> "Valid consent should require a clear affirmative act that is freely given, specific, unambiguous and informed (including about the consequences of providing or withholding consent). This means that any settings for data practices relying on consent must be pre-selected to 'off' and that different purposes of data collection, use or disclosure must not be bundled. Where the personal information of children is collected, consents to collect the personal information of children must be obtained from the child's guardian".[243]

This amendment would bring the Privacy Act into line with the GDPR.

The various practices described in Chapter 2 should not escape regulation simply on the basis that they claim to be operating with the 'consent' of the individual, unless that consent was proactively expressed, genuinely voluntary, informed, specific, current and given by a person with capacity.

As such, a valid consent cannot form part of standard terms and conditions, it cannot be bundled with consent for other purposes, it cannot be opt-out, nor can it be implied or inferred from an individual's use of a service.

[241] See https://www.accc.gov.au/system/files/Office%20of%20the%20Australian%20Information%20Commissioner%20%28May%202019%29.pdf

[242] Section 6 of the Privacy Act.

[243] See https://www.accc.gov.au/publications/digital-platforms-inquiry-final-report

4. Introduce an overarching ‘fairness’ requirement on the APPs

a. Develop an overarching ‘fairness’ requirement, to which consent is not an exception

Rationale

We need a more wholistic and protective approach to privacy regulation, in which an organisation can only collect, use or disclose personal information when it is fair to do so. There are some practices so privacy invasive or socially damaging that even ‘consent’ should not be allowed to authorise them. The late Giovanni Buttarelli, European Data Protection Supervisor, argued that “The right to human dignity demands limits to the degree to which an individual can be scanned, monitored and monetised — irrespective of any claims to putative ‘consent’”.[244]

The OAIC has said that “Overreliance on consent shifts the burden to individuals to critically analyse and decide whether they should disclose their personal information in return for a service or benefit”.[245] Consent should be the last resort, not the first or only choice from a menu of regulatory or design responses to privacy problems. The responsibility for protecting privacy should fall on privacy regulators, government legislators, and organisations themselves – not on individual consumers or citizens.

b. The overarching ‘fairness’ requirement should be that all handling of personal information must be fair in intent, lawful, transparent, reasonable, proportionate, not unnecessarily intrusive, and not lead to unfair or discriminatory outcomes

Rationale

Recommendation 17.3 of the ACCC’s Digital Platforms Inquiry was that the review of the Privacy Act should consider:

> “whether the Privacy Act should set a higher standard of privacy protection, such as by requiring all use and disclosure of personal information to be by fair and lawful means”.[246]

[244] See https://iapp.org/media/pdf/resource_center/giovanni_manifesto.pdf

[245] OAIC’s submission to the ACCC in response to its Consumer Loyalty review, October 2019; available at https://www.accc.gov.au/system/files/Office%20of%20the%20Australian%20Information%20Commissioner%20-%20October%202019.pdf

[246] See https://www.accc.gov.au/publications/digital-platforms-inquiry-final-report

An overarching fairness requirement offers a way of addressing the issue of power imbalances between entities and consumers, as well as protecting the privacy of vulnerable Australians including children.[247]

European and Canadian privacy laws each have an overarching fairness requirement.

The Australian Privacy Act should be reformed to introduce an overarching fairness test, in which even 'consent' would not be sufficient to authorise data practices which would otherwise be unfair in intent, have discriminatory or unfair impacts, or diminish human dignity.

We suggest that simply requiring 'lawful and fair' conduct is not sufficient. The requirement should be that all collection, use or disclosure of personal information must be fair in intent, lawful, transparent, reasonable, proportionate, not unnecessarily intrusive, and not lead to unfair or discriminatory outcomes.

Such a framework could enable more robust review of practices such as marketing to children or other vulnerable populations, price discrimination, the commercial application of facial recognition technology, racial profiling or algorithms which lead to discriminatory outcomes.

c. apply this overarching requirement to APPs 2-9

Rationale

While the ACCC mentioned only use and disclosure, to adequately regulate practices which commence with data collection via online identifiers, we suggest that all those privacy principles which touch on collection, use and disclosure should be subject to the overarching fairness requirement.

APP 3.5 currently requires collection to be "by lawful and fair means". In our view, this principle has not been expressly strongly enough to prevent intrusive and arguably unfair data collection practices such as profiling of consumers' online behaviour without their knowledge.

d. enable the OAIC to issue binding Codes and/or non-binding Guidelines to specify 'no-go zones' and provide other examples of what is / is not acceptable under the overarching fairness requirement

[247] OAIC's submission to the ACCC in response to its Preliminary Report of the Digital Platforms Inquiry, May 2019; available at https://www.accc.gov.au/system/files/Office%20of%20the%20Australian%20Information%20Commissioner%20%28May%202019%29.pdf

Rationale

Canadian privacy law includes a 'fairness' gatekeeper provision. Section 5(3) of PIPEDA says: "An organization may collect, use or disclose personal information only for purposes that a reasonable person would consider are appropriate in the circumstances." The Canadian Privacy Commissioner publishes guidance on 'no-go zones', based on court interpretations of s.5(3) as well as consultations with stakeholders and focus groups.[248] One such 'no-go zone' is "Profiling or categorization that leads to unfair, unethical or discriminatory treatment contrary to human rights law".

While what is 'reasonable', 'proportionate' or 'fair' are subjective assessments, the Canadian model at least creates a space for reflecting community expectations in the application of legal tests.

By enabling the OAIC to make Codes or Guidelines to illuminate 'no-go zones' created by the overarching fairness test, the law can more easily reflect emerging technologies and shifting community expectations over time.

5. Repeal APP 7 (the Direct Marketing principle)

Rationale

The scope of APP 7 is no longer fit for purpose: its scope is too narrow to encompass the privacy harms posed by the use of online identifiers to directly contact individuals, which go beyond the marketing of goods and services, to also include automated decisions about the goods or services offered to a customer in the first place, automated decisions about prices offered to a customer, personalised content and political messaging.

Instead, APP 6 should cover the field for determining whether or not profiling and/or contacting an individual is appropriate. Direct marketing and similar activities would thus typically fall under APP 6.2(a) – the 'related secondary purpose' test.

This recommendation does not affect the operation of the Spam Act or Do Not Call Register Act, which regulate the use of particular communication channels.

[248] See https://www.priv.gc.ca/en/privacy-topics/collecting-personal-information/consent/gd_53_201805/#s4

6. Strengthen APP 6.2(a) (the related secondary purpose test)

a. remove the distinction at APP 6.2(a) between sensitive information and other types of personal information, by collapsing parts (i) and (ii) together, to make the test that all related secondary purposes must be 'directly' related to the primary purpose

> Rationale
>
> The various practices described in Chapter 2 should not escape regulation simply on the basis that they claim to be related secondary purposes. APP 6.2(a) should be tightened to require any such secondary uses to be *directly* related.
>
> This amendment would still allow non-invasive business practices, such as session cookies remembering what is in an online shopping cart, or website analytics being conducted at an aggregated level of analysis. However more intrusive forms of tracking customers' online behaviour would not meet the test for 'directly' related secondary purposes.

b. prohibit *disclosure* for the purpose of direct marketing, messaging, profiling, targeting or tracking individuals (whether online or offline) under the 'directly related secondary purpose' test

> Rationale
>
> This provision would halt the trade in personal information between different entities seeking to engage in data sharing or data-matching for the purpose of online behavioural advertising or other forms of direct marketing, messaging, profiling, targeting or tracking, *unless* another exception applied, such as 'with the consent of the individual' (or law enforcement, public interest research, etc).
>
> By limiting this requirement to practices otherwise authorised under the 'directly related secondary purpose' test, this requirement would not be imposed unnecessarily on practices covered by other parts of APP 6 or other laws, such as a telecommunications provider disclosing location data to law enforcement for the purpose of tracking a suspect.
>
> This provision would not impact the provision of customer data to a third party contractor such as a mailing house, because the provision of data to a contracted service provider would not constitute a 'disclosure' but a 'use'.

c. prohibit any *use* of ‘sensitive personal information’ for direct marketing, messaging, profiling, targeting or tracking of individuals (whether online or offline) under the ‘directly related secondary purpose’ test

Rationale

This provision would replicate APP 7’s requirements that the use of *sensitive* personal information for direct marketing requires the consent of the individual. This is also consistent with the Canadian and European approaches.

By limiting this requirement to practices otherwise authorised under the ‘directly related secondary purpose’ test, this requirement still allows for use under another exception, such as ‘with consent’.

d. require all *use* of personal information for direct marketing, messaging, profiling, targeting or tracking of individuals (whether online or offline) under the ‘directly related secondary purpose’ test to include a clear, accessible and functional opt-out mechanism

Rationale

This provision would replicate the position in relation to direct marketing (whether considering APP 7 or the Spam Act) now, but extend it to other forms of direct messaging, profiling, targeting or tracking of individuals.

Similar to the ‘unsubscribe’ link on an email newsletter, online behavioural advertising or political targeted messaging (as opposed to contextual advertising) should be clearly identified to the user as an ad or message shown to *this user*, along with a mechanism providing that user with the opportunity to easily opt out of all such future direct marketing, messaging, profiling, targeting or tracking.

This position is consistent with the Canadian approach of allowing most forms of online behavioural advertising, so long as individuals can *easily* opt out. The effect would be to prohibit practices which do not support user control, such as zombie cookies.

By limiting this requirement to practices authorised under the ‘directly related secondary purpose’ test, this requirement would not be imposed unnecessarily on practices covered by other parts of APP 6 or other laws, such as law enforcement tracking of a suspect; or a telecommunications provider recording geolocation data as part of their *primary* purpose of delivering a mobile phone service.

e. prohibit direct marketing, messaging, profiling, targeting or tracking individuals (whether online or offline) if the individual has opted-out

Rationale

This provision would replicate the position in relation to direct marketing in APP 7 (and the Spam Act) now, and closes the loop on the provision above.

Codes and guidelines made under the Privacy Act

7. OAIC to create a binding Code to specify 'no-go zones' under the fairness requirement

a. any personalised or individualised messaging, marketing, profiling, targeting or tracking (online or offline) of children under 16, where the entity knew or ought to have known the subject was a child under 16

Rationale

A Code could set out appropriate rules - if any – to allow for profiling of or marketing to children.

b. any personalised or individualised messaging, marketing, profiling, targeting or tracking (online or offline) using biometrics

Rationale

A Code could allow for law enforcement and national security purposes, but not for commercial or employment purposes.

c. collection of personal information via a third party for the purpose of personalised or individualised messaging, marketing, profiling, targeting or tracking (online or offline)

Rationale

APP 3.5 currently requires collection to be "by lawful and fair means", while APP 3.6(b) currently prohibits the collection of personal information via a third party unless "it is unreasonable or impracticable" to ask the individual directly.

Yet Chapter 2 of this report highlights numerous unfair practices in which online identifiers are used to collect personal information about individuals in circumstances they do not know about and cannot control.

Creating a Code to prohibit indirect collection for certain purposes (individualised marketing etc not supported by broader public interest grounds) would have the effect of banning third party list broking, and other intrusive and unfair practices such as Facebook's online tracking of users across third party sites via the mere existence of a 'social sharing' plug-in on a third party website.

This would effectively require individualised marketing and similar practices to only be based on first party-collected data, in circumstances which the customer would expect and was notified about, such as where there is an existing customer relationship with the individual.

8. OAIC to develop Guidelines in relation to online identifiers

The OAIC should publish non-binding Guidelines on how it views practices or behaviours when collecting, using or disclosing personal information via online identifiers. Alternatively, guidance could be incorporated into the Binding Online Privacy Code we understand is being developed as part of the Privacy Act reforms.[249]

Rationale

Rather than prohibiting all or certain types of identifiers (e.g. the ePD approach of categorising all cookies as either 'strictly necessary' or not), a better approach would be to establish the principles in which an online identifier *should* operate.

For example, the OAIC could state that an online identifier *should* typically have the following properties:

- Observable
- Resettable
- Blockable

Without being unnecessarily prescriptive or proscriptive such as to impact on non-intrusive uses of online identifiers such as MAC addresses, the OAIC could promote these three features, by noting that in the absence of

[249] Australian Government, *Regulating in the digital age: Government Response and Implementation Roadmap for the Digital Platforms Inquiry*, December 2019, p.3; available at https://treasury.gov.au/sites/default/files/2019-12/Government-Response-p2019-41708.pdf

all three features, the OAIC would err on the side of calling an identifier unfair and intrusive, and thus less likely to comply with APPs 3 and/or 6.

By effectively promoting these three features as industry 'best practice', a favourable industry response would then create an incentive for browsers to enable a simple 'reset all' button, which would be a significant win for consumers.

Other policy responses

9. OAIC to consider additional non-legislative responses

Without wishing to be prescriptive about non-legislative steps that the OAIC could take, we have identified three possible areas of focus:

a. Promote privacy-enhancing technologies, and other alternatives to intrusive or unfair practices involving online identifiers.

b. Work with the Australian Government to adopt a 'model privacy protective' stance throughout the federal public sector, for example through the development of new public sector online design standards to prevent users of Australian Government websites from being tracked by third parties.

c. Engage in dialogue with other regulators about how to best regulate micro-targeting of political messaging.

In relation to this final topic, we note that the Privacy Act does not currently regulate the handling of personal information by political parties, and that regardless, any regulation of political messaging via micro-targeting would like fall under more specific electoral laws. Nonetheless, to the extent that the OAIC is in a position to influence debate on this topic, we draw the OAIC's attention to a number of concrete proposals to regulate political micro-targeting, made by a coalition of researchers led by the privacy non-profit Panoptykon Foundation.[250]

These include:

- full transparency of political ad targeting processes
- all political ads should be archived in libraries with legislated technical standards and APIs to make them accessible
- transparency over the criteria used in ad optimisation models

[250] Panoptykon Foundation, Who (really) targets you? Facebook in Polish election campaigns, undated; available at https://panoptykon.org/political-ads-report

- users to have access to, and control over, their marketing profiles
- require a separate opt-in to behavioural data for political micro-targeting (i.e. separate to any opt-in for online behavioural advertising)
- a prohibition on using certain characteristics / types of data in political micro-targeting
- define political micro-targeting as a high risk application of AI, triggering mandatory Privacy Impact Assessment, and
- constraints on the financing of online political campaigns (e.g. invoices from advertisers / digital platforms to be submitted to a regulator).

Appendix A

Examples of recommendations applied in practice

Practice	Example	Allowed (if fair)	OK if not sensitive + directly related + expected + opt-out + fair	Needs consent + fair	Prohibited
Marketing to own customer mailing list	A clothing store emails its own customer mailing list A department store mails a catalogue to its own customer mailing list		X		
Providing own customer list to contractor to act on our behalf	A department store contracts a mailing house to mail out catalogues to its own customer mailing list	X			
Buying a mailing list from a third party	A start-up food delivery service buys a list of customer names and email addresses from a wine delivery service			X	
Providing own customer list to third party for List-based marketing	A clothing store gives its own customer mailing list to a social media			X	

	company to find and show ads to those customers (or: to past customers who have unsubscribed from their mailing list)				
Providing own customer list to third party for Lookalike audience marketing	A chain of gyms provides its own customer mailing list and profiles to a social media company, for the social media company to find and show ads to people *not* on the list but who have similar profiles			X	
Providing data about own website users to a third party	A cosmetics retailer uses 'social sharing' plug-ins on its website, which enable social media companies to collect data about viewers of that website, even if the viewer is not logged in to the social media platform and does not click on any 'Like' / 'Tweet' / 'Share' buttons			X	
Contextual ads	A free 'catch up' TV streaming service places the same ad in the ad break for all viewers of a particular TV show An airline pays a newspaper to show its ad for flights from Sydney to Queenstown to all online readers of a travel story about skiing in New Zealand	X			

Profile-based marketing based on first party-provided data	An airline asks a social media company to show its ad for flights from Sydney to Queenstown to people whose user-generated profile includes “enjoys skiing” and “lives in Sydney”		X		
Profile-based marketing based on other data sources (e.g. observed/inferred or collected from third parties)	An airline asks a social media company to show its ad for flights from Sydney to Queenstown to people who the social media company has inferred from the user’s online activities that they are in the market for a ski holiday			X	
Affinity audience marketing	A recruitment company asks a social media company to only show its job ad to white men				X
Personalised recommendations based on first party-provided data	A subscription TV streaming service recommends ‘other shows for you’, based on that user’s viewing habits on that service alone		X		
Personalised recommendations based on other data (inferred, or provided by a third party)	A subscription TV streaming service recommends ‘other shows for you’, based on *other* data collected about that user			X	
Personalised ads using first party-provided data only	A free ‘catch up’ TV streaming service determines what ad is shown in ad		X		

	breaks for *this user*, based on that user’s viewing habits on this service, and/or other user-provided data (e.g. user self-described as male aged 50-59 living in WA) A newspaper determines what ads to show *this reader*, based on that subscriber’s reading habits on the newspaper’s own website, and/or user-provided data (e.g. subscriber self-described as female aged 20-29 living in Melbourne)				
Personalised ads based on other data (inferred, or provided by a third party)	A free ‘catch up’ TV streaming service determines what ad is shown in ad breaks for *this user*, based on *other* data collected or inferred about that user (e.g. an ad broker tells the TV channel that based on their search history, *this user* is in the market for a new car) A newspaper determines what ad is shown to *this user*, based on *other* data collected or inferred about that user (e.g. an ad broker tells the newspaper publisher that *this user* owns a dog and enjoys camping trips)			X	

Contextual political content	A political party places an ad on a newspaper website to be shown to all readers of a story about climate change	X			
Personalised political content	A political party places an ad on a newspaper website to be shown to readers known to have posted comments online in support of a climate change rally			X	
Personalised content based on expressed wishes	Social media news feed, determined on basis of who this user ‘follows’	X			
Personalised content based on first party data	Social media news feed, determined on basis of user’s behaviour on that social media platform alone		X		
Personalised content based on other data	Social media news feed, determined on basis of user’s behaviour elsewhere online or offline			X	
Personalised offers based on first party data	A supermarket offers discounts on nappies directly to some of its customers based on loyalty card data about their purchasing habits		X		
Personalised offers based on other data (inferred, or provided by a third party)	An insurance company offers a car insurance product only to customers it			X	

	has determined are low risk, based on inferred or third party-provided data				
Offline tracking using online identifiers	A shopping centre uses a Bluetooth beacon from a customer's phone to track their movements through the shopping centre			X	
Offline tracking using biometrics	A shopping centre uses facial or gait detection to track customers' movements through the shopping centre				X
Contextual advertising to children	An ad for a toy is placed on a free 'catch up' TV streaming service during a kid's TV program	X			
Online behavioural advertising to children based on first party data	A free puzzle app designed for pre-schoolers shows different ads to users, according to profiling users built on how that user uses this app				X
Online behavioural advertising to children based on other data (inferred, or provided by a third party)	A free puzzle app designed for pre-schoolers shows different ads to users, according to profiling of each user, built on data collected via third parties using a cookie				X

Session cookies	An online retailer uses cookies to remember what is in their shopping cart during this session	X			
Cookies for recording preferences	An online retailer uses cookies to remember a shopper’s language preference over repeat visits	X			
Cookies for personalising marketing	An online retailer uses cookies to build a profile of what a customer looks at on their own website, in order to recommend products in an email newsletter or on the next online visit		X		
Site analytics	An online retailer uses website analytics to review number of hits on different pages, or what countries users are in (based on their IP address), in aggregated reports	X			
Sharing de-identified but unit-record level data for generating audience insights	A telco partners with a bank to share customer profiles at an individual record level to generate new insights about individual customers			X	
Sharing de-identified but unit-record level data for machine learning dataset compilation and/or building algorithms	A health insurer partners with a supermarket chain to build algorithms predicting health risk based on shopping habits; identifiers are removed from the dataset first			X	

Qualifications & confidentiality

The analysis in this report does not constitute legal advice, and should not be construed or relied upon as legal advice by any party. Legal professional privilege does not apply to this report.

This report is made on a confidential basis to our client. It is for the sole discretion of our client to determine whether it will waive confidentiality and provide this report to any other party. The contents of this report will not be divulged to any third party by Salinger Privacy without the express and written permission of our client.

About the authors

This report has been prepared by Anna Johnston, Principal, Salinger Privacy, and Dr Chris Culnane, Principal of Castellate Consulting.

Ms Johnston was previously the Deputy Privacy Commissioner of NSW. She holds a first class honours degree in Law, a Masters of Public Policy with honours, a Graduate Certificate in Management, a Graduate Diploma of Legal Practice, and a Bachelor of Arts. Ms Johnston was admitted as a Solicitor of the Supreme Court of NSW in 1996. She established Salinger Privacy in 2004.

Dr Culnane is an expert in cyber security and privacy. He has previously held academic posts at the University of Surrey and the University of Melbourne, where his research focus included the integrity of electronic voting, and the privacy and cyber security of open data releases. Chris holds a first class honours Bachelor of Science in Computing, a Master of Science in Internet Computing, and a PhD in Computer and Information Systems Security. For more information see www.castellate.com

Salinger Privacy offers a range of privacy guidance publications, eLearning and face-to-face compliance training options, and Privacy Tools such as templates and checklists. See our website for more information.

Salinger Privacy also offers specialist consulting services on privacy and data governance matters, including Privacy Impact Assessments and privacy audits, and the development of privacy-related policies and procedures.

SalingerPrivacy

We know privacy inside and out.

Salinger Consulting Pty Ltd
ABN 84 110 386 537
PO Box 1250, Manly NSW 1655
www.salingerprivacy.com.au