Behind the One-Way Mirror

A DEEP DIVE INTO THE TECHNOLOGY OF CORPORATE SURVEILLANCE

ELECTRONIC FRONTIER FOUNDATION

Author: Bennett Cyphers and Gennie Gebhart

A publication of the Electronic Frontier Foundation, 2019.

View this report online: https://www.eff.org/wp/behind-the-one-way-mirror

Behind the One-Way Mirror

A Deep Dive Into the Technology of Corporate Surveillance

BENNETT CYPHERS AND GENNIE GEBHART

December 2, 2019

Introduction

Trackers are hiding in nearly every corner of today's Internet, which is to say nearly every corner of modern life. The average web page shares data with dozens of third-parties. The average mobile app does the same, and many apps collect highly sensitive information like location and call records even when they're not in use. Tracking also reaches into the physical world. Shopping centers use automatic license-plate readers to track traffic through their parking lots, then share that data with law enforcement. Businesses, concert organizers, and political campaigns use Bluetooth and WiFi beacons to perform passive monitoring of people in their area. Retail stores use face recognition to identify customers, screen for theft, and deliver targeted ads.

The tech companies, data brokers, and advertisers behind this surveillance, and the technology that drives it, are largely invisible to the average user. Corporations have built a hall of one-way mirrors: from the inside, you can see only apps, web pages, ads, and yourself reflected by social media. But in the shadows behind the glass, trackers quietly take notes on nearly everything you do. These trackers are not omniscient, but they are widespread and indiscriminate. The data they collect and derive is not perfect, but it is nevertheless extremely sensitive.

This paper will focus on corporate "third-party" tracking: the collection of personal information by companies that users don't intend to interact with. It will shed light on the technical methods and business practices behind third-party tracking. For journalists, policy makers, and concerned consumers, we hope this paper will demystify the fundamentals of third-party tracking, explain the scope of the problem, and suggest ways for users and legislation to fight back against the status quo.

Part 1 breaks down "identifiers," or the pieces of information that trackers use to keep track of who is who on the web, on mobile devices, and in the physical world. Identifiers let trackers link behavioral data to real people.

Part 2 describes the techniques that companies use to collect those identifiers and other information. It also explores how the biggest trackers convince other businesses to help them build surveillance networks.

Part 3 goes into more detail about how and why disparate actors share information with each other. Not every tracker engages in every kind of tracking. Instead, a fragmented web of companies collect data in different contexts, then share or sell it in order to achieve specific goals.

Finally, Part 4 lays out actions consumers and policy makers can take to fight back. To start, consumers can change their tools and behaviors to block tracking on their devices. Policy makers must adopt comprehensive privacy laws to rein in third-party tracking.

First-party vs. third-party tracking

The biggest companies on the Internet collect vast amounts of data when people use their services. Facebook knows who your friends are, what you "Like," and what kinds of content you

read on your newsfeed. Google knows what you search for and where you go when you're navigating with Google Maps. Amazon knows what you shop for and what you buy.

The data that these companies collect through their own products and services is called "first-party data." This information can be extremely sensitive, and companies have a long track record of mishandling it. First-party data is sometimes collected as part of an implicit or explicit contract: choose to use our service, and you agree to let us use the data we collect while you do. More users are coming to understand that for many free services, they are the product, even if they don't like it.

However, companies collect just as much personal information, if not more, about people who *aren't* using their services. For example, Facebook collects information about users of other websites and apps with its invisible "conversion pixels." Likewise, Google uses location data to track user visits to brick and mortar stores. And thousand of other data brokers, advertisers, and other trackers lurk in the background of our day-to-day web browsing and device use. This is known as "third-party tracking." Third-party tracking is much harder to identify without a trained eye, and it's nearly impossible to avoid completely.

What do they know?

Many consumers are familiar with the most blatant privacy-invasive potential of their devices. Every smartphone is a pocket-sized GPS tracker, constantly broadcasting its location to parties unknown via the Internet. Internet-connected devices with cameras and microphones carry the inherent risk of conversion into silent wiretaps. And the risks are real: location data has been badly abused in the past. Amazon and Google have both allowed employees to listen to audio recorded by their in-home listening devices, Alexa and Home. And front-facing laptop cameras have been used by schools to spy on students in their homes.

But these better known surveillance channels are not the most common, or even necessarily the most threatening to our privacy. Even though we spend many of our waking hours in view of our devices' Internet-connected cameras, it's exceedingly rare for them to record anything without a user's express intent. And to avoid violating federal and state wiretapping laws, tech companies typically refrain from secretly listening in on users' conversations. As the rest of this paper will show, trackers learn more than enough from thousands of less dramatic sources of data. The unsettling truth is that although Facebook doesn't listen to you through your phone, that's just because it doesn't need to.

The most prevalent threat to our privacy is the slow, steady, relentless accumulation of relatively mundane data points about how we live our lives. This includes things like browsing history, app usage, purchases, and geolocation data. These humble parts can be combined into an exceptionally revealing whole. Trackers assemble data about our clicks, impressions, taps, and movement into sprawling *behavioral profiles*, which can reveal political affiliation, religious belief, sexual identity and activity, race and ethnicity, education level, income bracket, purchasing habits, and physical and mental health.

Despite the abundance of personal information they collect, tracking companies frequently use this data to derive conclusions that are inaccurate or wrong. Behavioral advertising is the practice of using data about a user's behavior to predict what they like, how they think, and what they are likely to buy, and it drives much of the third-party tracking industry. While

behavioral advertisers sometimes have access to precise information, they often deal in sweeping generalizations and "better than nothing" statistical guesses. Users see the results when both uncannily accurate and laughably off-target advertisements follow them around the web. Across the marketing industry, trackers use petabytes of personal data to power digital tea reading. Whether trackers' inferences are correct or not, the data they collect represents a disproportionate invasion of privacy, and the decisions they make based on that data can cause concrete harm.

Part 1: Whose Data is it Anyway: How Do Trackers Tie Data to People?

Most third-party tracking is designed to build profiles of real people. That means every time a tracker collects a piece of information, it needs an *identifier*—something it can use to tie that information to a particular person. Sometimes a tracker does so indirectly: by correlating collected data with a particular device or browser, which might in turn later be correlated to one person or perhaps a small group of people like a household.

To keep track of who is who, trackers need identifiers that are unique, persistent, and available. In other words, a tracker is looking for information (1) that points *only* to you or your device, (2) that won't change, and (3) that it has easy access to. Some potential identifiers fit all three of these requirements, but trackers can still make use of an identifier that checks only two of these three boxes. And trackers can combine multiple weak identifiers to create a single, strong one.

An identifier that checks all three boxes might be a name, an email, or a phone number. It might also be a "name" that the tracker itself gives you, like "af64a09c2" or "921972136.1561665654". What matters most to the tracker is that the identifier points to you and only you. Over time, it can build a rich enough profile about the person known as "af64a09c2"—where they live, what they read, what they buy—that a conventional name is not necessary. Trackers can use artificial identifiers, like cookies and mobile ad IDs, to reach users with targeted messaging. And data that isn't tied to a real name is no less sensitive: "anonymous" profiles of personal information can nearly always be linked back to real people.

Some types of identifiers, like cookies, are features built into the tech that we use. Others, like browser fingerprints, emerge from the way those technologies work. This section will break down how trackers on the web and in mobile apps are able to identify and attribute data points.

This section will describe a representative sample of identifiers that third-party trackers can use. It is not meant to be exhaustive; there are more ways for trackers to identify users than we can hope to cover, and new identifiers will emerge as technology evolves. The tables below give a brief overview of how unique, persistent, and available each type of identifier is.

Web Identifiers	Unique	Persistent	Available
Cookies	Yes	Until user deletes	In some browsers without tracking protection
IP address	Yes	On the same network, may persist for weeks or months	Always
TLS state	Yes	For up to one week	In most browsers
Local storage super cookie	Yes	Until user deletes	Only in third-party IFrames; can be blocked by tracker blockers
Browser fingerprint	Only on certain browsers	Yes	Almost always; usually requires JavaScript access, sometimes blocked by tracker blockers

Phone Identifiers	Unique	Persistent	Available
Phone number	Yes	Until user changes	Readily available from data brokers; only visible to apps with special permissions
IMSI and IMEI number	Yes	Yes	Only visible to apps with special permissions
Advertising ID	Yes	Until user resets	Yes, to all apps
MAC address	Yes	Yes	To apps: only with special permissions To passive trackers: visible unless OS performs randomization or device is in airplane mode

Other Identifiers	Unique	Persistent	Available
License plate	Yes	Yes	Yes
Face print	Yes	Yes	Yes
Credit card number	Yes	Yes, for months or years	To any companies involved in payment processing

Identifiers on the Web

Browsers are the primary way most people interact with the Web. Each time you visit a website, code on that site may cause your browser to make dozens or even hundreds of *requests* to hidden third parties. Each request contains several pieces of information that can be used to track you.

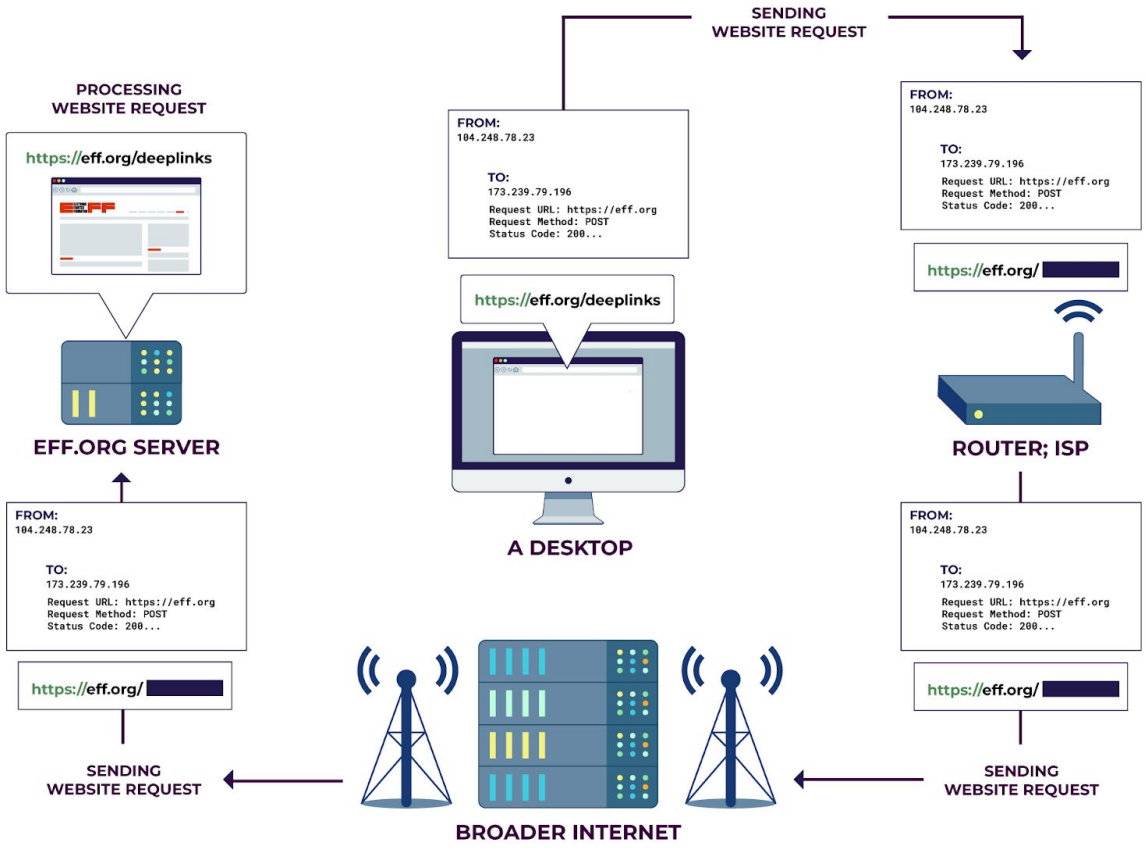
Anatomy of a Request

Almost every piece of data transmitted between your browser and the servers of the websites you interact with occurs in the form of an *HTTP request*. Basically, your browser asks a *web server* for content by sending it a particular URL. The web server can respond with content, like text or an image, or with a simple acknowledgement that it received your request. It can also respond with a *cookie*, which can contain a unique identifier for tracking purposes.

Each website you visit kicks off dozens or hundreds of different requests. The URL you see in the address bar of your browser is the address for the first request, but hundreds of other requests are made in the background. These requests can be used for loading images, code, and styles, or simply for sharing data.

Parts of a URL. The domain tells your computer where to send the request, while the path and parameters carry information that may be interpreted by the receiving server however it wants.

The URL itself contains a few different pieces of information. First is the *domain*, like "nytimes.com". This tells your browser which server to connect to. Next is the *path*, a string at the end of the domain like "/section/world.html". The server at nytimes.com chooses how to interpret the path, but it usually specifies a piece of content to serve—in this case, the world news section. Finally, some URLs have *parameters* at the end in the form of "?key1=value1&key2=value2". The parameters usually carry extra information about the request, including queries made by the user, context about the page, and tracking identifiers.



The path of a request. After it leaves your machine, the request is redirected by your router to your ISP, which sends it through a series of intermediary routing stations in "the Internet." Finally, it arrives at the server specified by the domain, which can decide how (or if) to respond.

The URL isn't all that gets sent to the server. There are also *HTTP headers*, which contain extra information about the request like your device's language and security settings, the "referring" URL, and *cookies*. For example, the *User-Agent* header identifies your browser type, version, and operating system. There's also lower-level information about the connection, including *IP address* and shared encryption state. Some requests contain even more configurable information in the form of POST data. POST requests are a way for websites to share chunks of data that are too large or unwieldy to fit in a URL. They can contain just about anything.

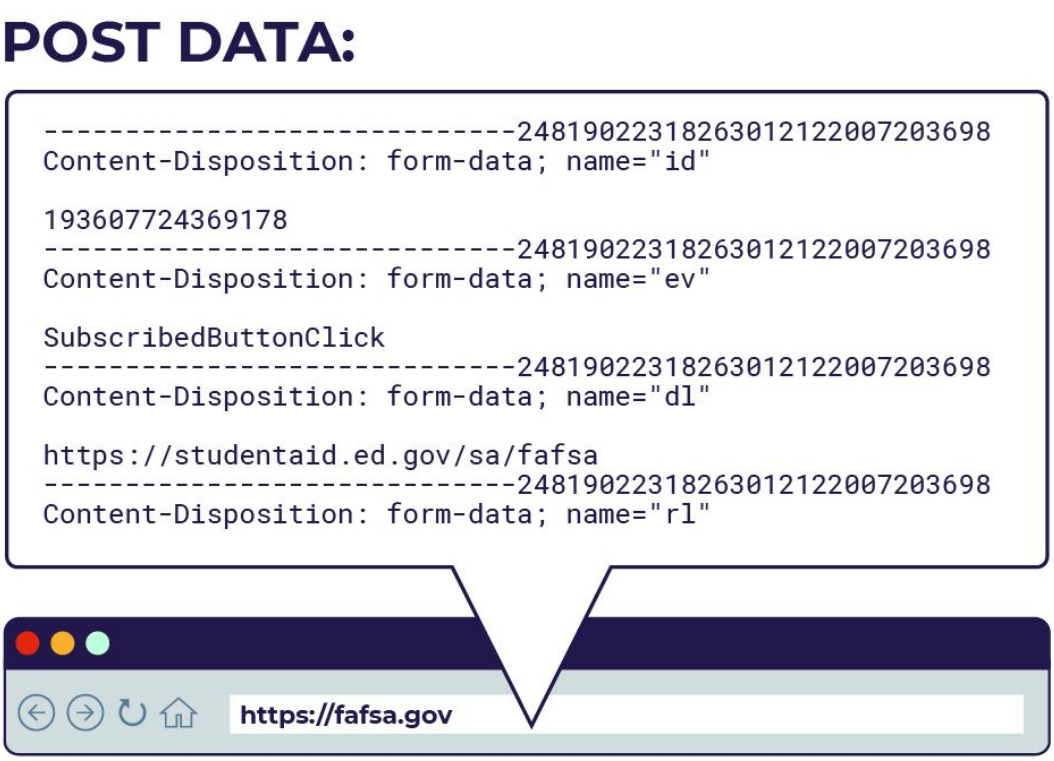
Some of this information, like the URL and POST data, is specifically tailored for each individual request; other parts, like your IP address and any cookies, are sent automatically by your machine. Almost all of it can be used for tracking.

GENERAL:

Request URL: https://www.facebook.com/tr/
Request Method: POST
Status Code: 200
Remote Address: 31.13.70.36:443

REQUEST HEADERS:

Host: www.facebook.com
User-Agent: Mozilla/5.0 (X11; Ubuntu; Linux x86_64; rv:70.0) Gecko/20100101 Firefox/70.0
Accept-Language: en-US,en;q=0.5
Origin: https://studentaid.ed.gov
Referer: https://studentaid.ed.gov/sa/fafsa
Cookie: fr=0epKOG2iQOjkdaV1v.AWXmwPoddUQIrlWxPuEr6Z6Pk-Ws.BcuRD0.U9.Fzl.0.0.BdzbjY.; sb=QquIXAabRhVGyXwmKnA51zbn; datr=XauIXP-tHmI5gBqNyN7IcHRx; c_user=100026095248544; xs=42%3AFQTekubGMuw6yQ%3A2%3A1554857792%3A-1%3A-1

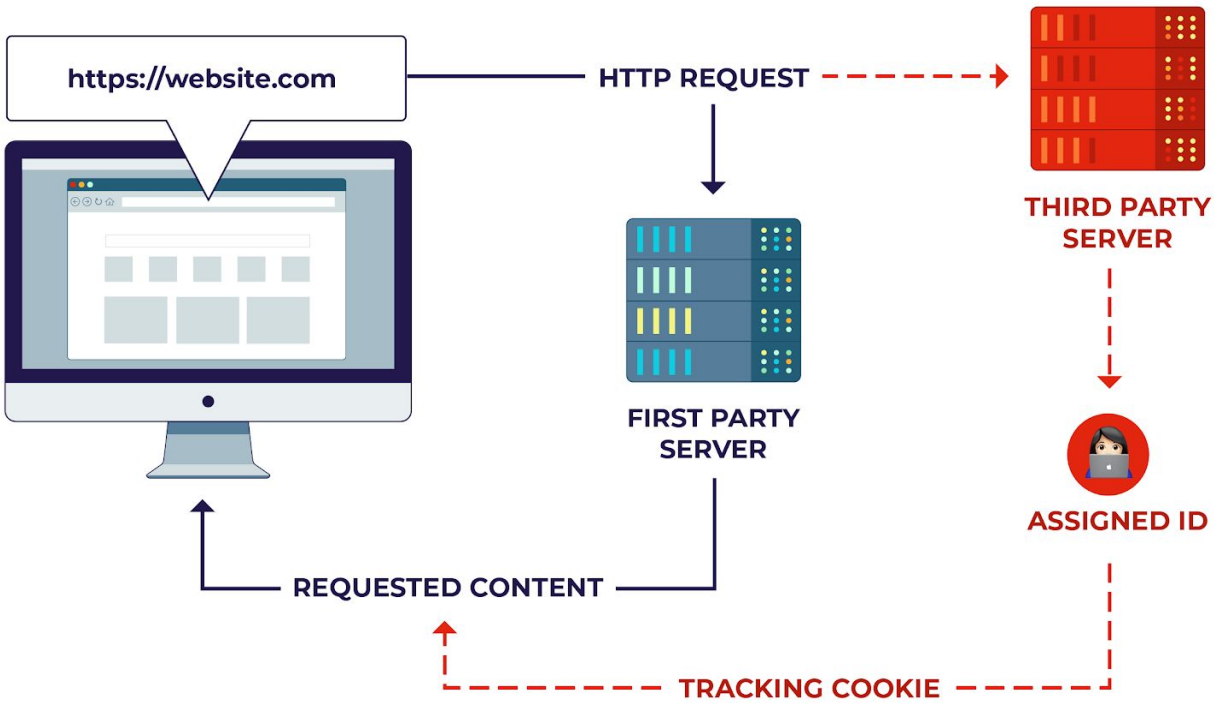


Data included with a background request. In the image, although the user has navigated to fafsa.gov, the page triggers a third-party request to facebook.com in the background. The URL isn't the only information that gets sent to the receiving server; HTTP Headers contain information like your User Agent string and cookies, and POST data can contain anything that the server wants.

The three images above contain data we collected directly from a normal version of Firefox. If you want to check it out for yourself, you can. All major browsers have an "inspector" or "developer" mode which allows users to see what's going on behind the scenes, including all requests coming from a particular tab. In Chrome and Firefox, you can access this interface with Crtl+Shift+I (or ⌘+Shift+I on Mac). The "Network" tab has a log of all the requests made by a particular page, and you can click on each one to see where it's going and what information it contains.

Identifiers shared automatically

Some identifiable information is shared automatically along with each request. This is either by necessity—as with IP addresses, which are required by the underlying protocols that power the Internet—or by design—as with cookies. Trackers don't need to do anything more than trigger a request, *any* request, in order to collect the information described here.

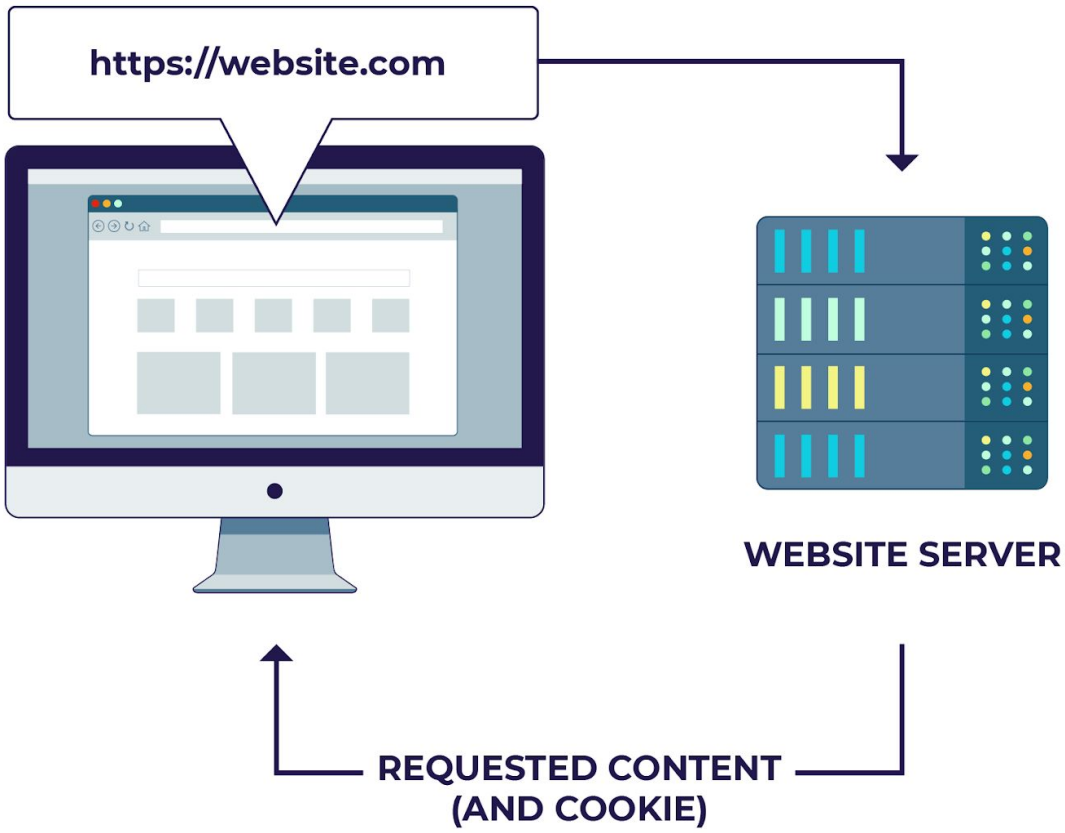


Each time you visit a website by typing in a URL or clicking on a link, your computer makes a request to that website's server (the "first party"). It may also make dozens or hundreds of requests to other servers, many of which may be able to track you.

Cookies

The most common tool for third-party tracking is the *HTTP cookie*. A cookie is a small piece of text that is stored in your browser, associated with a particular domain. Cookies were invented to help website owners determine whether a user had visited their site before, which makes them ideal for behavioral tracking. Here's how they work.

The first time your browser makes a request to a domain (like www.facebook.com), the server can attach a *Set-Cookie* header to its reply. This will tell your browser to store whatever value the website wants—for example, `c_user:"100026095248544"` (an actual Facebook cookie taken from the author's browser). Then, every time your browser makes a request to www.facebook.com in the future, it sends along the cookie that was set earlier. That way, every time Facebook gets a request, it knows which individual user or device it's coming from.



The first time a browser makes a request to a new server, the server can reply with a "Set-Cookie" header that stores a tracking cookie in the browser.

Not every cookie is a tracker. Cookies are also the reason that you don't have to log in every single time you visit a website, as well as the reason your cart doesn't empty if you leave a website in the middle of shopping. Cookies are just a means of sharing information from your browser to the website you are visiting. However, they are designed to be able to carry tracking information, and third-party tracking is their most notorious use.

Luckily, users can exercise a good deal of control over how their browsers handle cookies. Every major browser has an optional setting to disable third-party cookies (though it is usually turned off by default.) In addition, Safari and Firefox have recently started restricting access to third-party cookies for domains they deem to be trackers. As a result of this "cat and mouse game" between trackers and methods to block them, third-party trackers are beginning to shift away from relying solely on cookies to identify users, and are evolving to rely on other identifiers.

Cookies are always unique, and they normally persist until a user manually clears them. Cookies are always available to trackers in unmodified versions of Chrome, but third-party cookies are no longer available to many trackers in Safari and Firefox. Users can always block cookies themselves with browser extensions.

IP Address

Each request you make over the Internet contains your *IP address*, a temporary identifier that's unique to your device. Although it is unique, it is not necessarily persistent: your IP address changes every time you move to a new network (e.g., from home to work to a coffee shop). Thanks to the way IP addresses work, it may change even if you stay connected to the same network.

There are two types of IP addresses in widespread use, known as IPv4 and IPv6. IPv4 is a technology that predates the Web by a decade. It was designed for an Internet used by just a few hundred institutions, and there are only around 4 billion IPV4 addresses in the world to serve over 22 billion connected devices today. Even so, over 70% of Internet traffic still uses IPv4.

As a result, IPv4 addresses used by consumer devices are constantly being reassigned. When a device connects to the Internet, its internet service provider (ISP) gives it a "lease" on an IPv4 address. This lets the device use a single address for a few hours or a few days. When the lease is up, the ISP can decide to extend the lease or grant it a new IP. If a device remains on the same network for extended periods of time, its IP may change every few hours -- or it may not change for months.

IPv6 addresses don't have the same scarcity problem. They do not *need* to change, but thanks to a privacy-preserving extension to the technical standard, most devices generate a new, random IPv6 address every few hours or days. This means that IPv6 addresses may be used for short-term tracking or to link other identifiers, but cannot be used as standalone long-term identifiers.

IP addresses are not perfect identifiers on their own, but with enough data, trackers can use them to create long-term profiles of users, including mapping relationships between devices. You can hide your IP address from third-party trackers by using a trusted VPN or the Tor browser.

IP addresses are always unique, and always available to trackers unless a user connects through a VPN or Tor. Neither IPv4 nor IPv6 addresses are guaranteed to persist for longer than a few days, although IPv4 addresses may persist for several months.

TLS State

Today, most traffic on the web is encrypted using Transport Layer Security, or TLS. Any time you connect to a URL that starts with "https://" you're connecting using TLS. This is a very good thing. The encrypted connection that TLS and HTTPS provide prevents ISPs, hackers, and governments from spying on web traffic, and it ensures that data isn't being intercepted or modified on the way to its destination.

However, it also opens up new ways for trackers to identify users. TLS *session IDs* and *session tickets* are cryptographic identifiers that help speed up encrypted connections. When you connect to a server over HTTPS, your browser starts a new *TLS session* with the server.

The session setup involves some expensive cryptographic legwork, so servers don't like to do it more often than they have to. Instead of performing a full cryptographic "handshake" between the server and your browser every time you reconnect, the server can send your browser a

session ticket that encodes some of the shared encryption state. The next time you connect to the same server, your browser sends the session ticket, allowing both parties to skip the handshake. The only problem with this is that the session ticket can be exploited by trackers as a unique identifier.

TLS session tracking was only brought to the public's attention recently in an academic paper, and it's not clear how widespread its use is in the wild.

Like IP addresses, session tickets are always unique. They are available unless the user's browser is configured to reject them, as Tor is. Server operators can usually configure session tickets to persist for up to a week, but browsers do reset them after a while.

Identifiers created by trackers

Sometimes, web-based trackers want to use identifiers beyond just IP addresses (which are unreliable and not persistent), cookies (which a user can clear or block), or TLS state (which expires within hours or days). To do so, trackers need to put in a little more effort. They can use *JavaScript* to save and load data in *local storage* or perform *browser fingerprinting.*

Local storage "cookies" and IFrames

Local storage is a way for websites to store data in a browser for long periods of time. Local storage can help a web-based text editor save your settings, or allow an online game to save your progress. Like cookies, local storage allows third-party trackers to create and save unique identifiers in your browser.

Also like cookies, data in local storage is associated with a specific domain. This means if example.com sets a value in your browser, only example.com web pages and example.com's *IFrames* can access it. An IFrame is like a small web page within a web page. Inside an IFrame, a third-party domain can do almost everything a first-party domain can do. For example, embedded YouTube videos are built using IFrames; every time you see a YouTube video on a site other than YouTube, it's running inside a small page-within-a-page. For the most part, your browser treats the YouTube IFrame like a full-fledged web page, giving it permission to read and write to YouTube's local storage. Sure enough, YouTube uses that storage to save a unique "device identifier" and track users on any page with an embedded video.

Local storage "cookies" are unique, and they persist until a user manually clears their browser storage. They are only available to trackers which are able to run JavaScript code inside a third-party IFrame. Not all cookie-blocking measures take local storage cookies into account, so local storage cookies may sometimes be available to trackers for which normal cookie access is blocked.

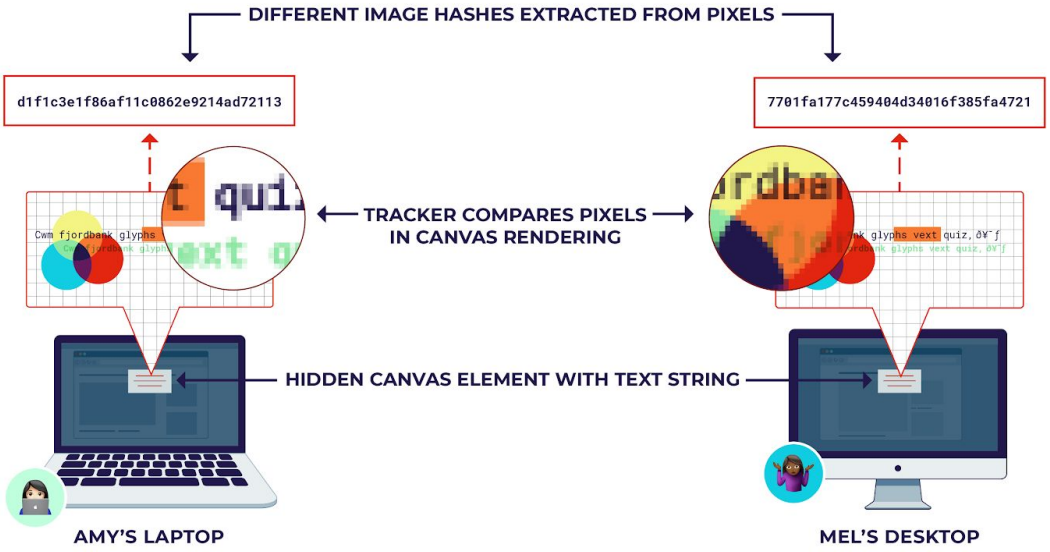
Fingerprinting

Browser fingerprinting is one of the most complex and insidious forms of web-based tracking. A browser fingerprint consists of one or more attributes that, on their own or when combined, uniquely identify an individual browser on an individual device. Usually, the data that go into a fingerprint are things that the browser can't help exposing, because they're just part of the way it interacts with the web. These include information sent along with the request made every

time the browser visits a site, along with attributes that can be discovered by running JavaScript on the page. Examples include the resolution of your screen, the specific version of software you have installed, and your time zone. Any information that your browser exposes to the websites you visit can be used to help assemble a browser fingerprint. You can get a sense of your own browser's fingerprint with EFF's Panopticlick project.

The reliability of fingerprinting is a topic of active research, and must be measured against the backdrop of ever-evolving web technologies. However, it is clear that new techniques increase the likelihood of unique identification, and the number of sites that use fingerprinting is increasing as well. A recent report found that at least a third of the top 500 sites visited by Americans employ some form of browser fingerprinting. The prevalence of fingerprinting on sites also varies considerably with the category of website.

Researchers have found *canvas fingerprinting* techniques to be particularly effective for browser identification. The *HTML Canvas* is a feature of HTML5 that allows websites to render complex graphics inside of a web page. It's used for games, art projects, and some of the most beautiful sites on the Web. Because it's so complex and performance-intensive, it works a little bit differently on each different device. Canvas fingerprinting takes advantage of this.



Canvas fingerprinting. A tracker renders shapes, graphics, and text in different fonts, then computes a "hash" of the pixels that get drawn. The hash will be different on devices with even slight differences in hardware, firmware, or software.

A tracker can create a "canvas" element that's invisible to the user, render a complicated shape or string of text using JavaScript, then extract data about exactly how each pixel on the canvas is rendered. The operating system, browser version, graphics card, firmware version, graphics driver version, and fonts installed on your computer all affect the final result.

For the purposes of fingerprinting, individual characteristics are hardly ever measured in isolation. Trackers are most effective in identifying a browser when they combine multiple characteristics together, stitching the bits of information left behind into a cohesive whole.

Even if one characteristic, like a canvas fingerprint, is itself not enough to uniquely identify your browser, it can usually be combined with others -- your language, time zone, or browser settings -- in order to identify you. And using a combination of simple bits of information is much more effective than you might guess.

Fingerprints are often, but not always, unique. Some browsers, like Tor and Safari, are specifically designed so that their users are more likely to look the same, which removes or limits the effectiveness of browser fingerprinting. Browser fingerprints tend to persist as long as a user has the same hardware and software: there's no setting you can fiddle with to "reset" your fingerprint. And fingerprints are usually available to any third parties who can run JavaScript in your browser.

Identifiers on mobile devices

Smartphones, tablets, and ebook readers usually have web browsers that work the same way desktop browsers do. That means that these types of connected devices are susceptible to all of the kinds of tracking described in the section above.

However, mobile devices are different in two big ways. First, users typically need to sign in with an Apple, Google, or Amazon account to take full advantage of the devices' features. This links device identifiers to an account identity, and makes it easier for those powerful corporate actors to profile user behavior. For example, in order to save your home and work address in Google Maps, you need to turn on Google's "Web and App Activity," which allows it to use your location, search history, and app activity to target ads.

Second, and just as importantly, most people spend most of their time on their mobile device in apps outside of the browser. Trackers in apps can't access cookies the same way web-based trackers can. But by taking advantage of the way mobile operating systems work, app trackers can still access unique identifiers that let them tie activity back to your device. In addition, mobile phones—particularly those running the Android and iOS operating systems—have access to a unique set of identifiers that can be used for tracking.

In the mobile ecosystem, most tracking happens by way of third-party software development kits, or SDKs. An SDK is a library of code that app developers can choose to include in their apps. For the most part, SDKs work just like the Web resources that third parties exploit, as discussed above: they allow a third party to learn about your behavior, device, and other characteristics. An app developer who wants to use a third-party analytics service or serve third-party ads downloads a piece of code from, for example, Google or Facebook. The developer then includes that code in the published version of their app. The third-party code thus has access to all the data that the app does, including data protected behind any permissions that the app has been granted, such as location or camera access.

On the web, browsers enforce a distinction between "first party" and "third party" resources. That allows them to put extra restrictions on third-party content, like blocking their access to browser storage. In mobile apps, this distinction doesn't exist. You can't grant a privilege to an app without granting the same privilege to all the third party code running inside it.

Phone numbers

The phone number is one of the oldest unique numeric identifiers, and one of the easiest to understand. Each number is unique to a particular device, and numbers don't change often. Users are encouraged to share their phone numbers for a wide variety of reasons (e.g., account verification, electronic receipts, and loyalty programs in brick-and-mortar stores). Thus, data brokers frequently collect and sell phone numbers. But phone numbers aren't easy to access from inside an app. On Android, phone numbers are only available to third-party trackers in apps that have been granted certain permissions. iOS prevents apps from accessing a user's phone number at all.

Phone numbers are unique and persistent, but usually not available to third-party trackers in most apps.

Hardware identifiers: IMSI and IMEI

Every device that can connect to a mobile network is assigned a unique identifier called an International Mobile Subscriber Identity (IMSI) number. IMSI numbers are assigned to users by their mobile carriers and stored on SIM cards, and normal users can't change their IMSI without changing their SIM. This makes them ideal identifiers for tracking purposes.

Similarly, every mobile device has an International Mobile Equipment Identity (IMEI) number "baked" into the hardware. You can change your SIM card and your phone number, but you can't change your IMEI without buying a new device.

IMSI numbers are shared with your cell provider every time you connect to a cell tower—which is all the time. As you move around the world, your phone sends out pings to nearby towers to request information about the state of the network. Your phone carrier can use this information to track your location (to varying degrees of accuracy). This is not quite third-party tracking, since it is perpetrated by a phone company that you have a relationship with, but regardless many users may not realize that it's happening.

Software and apps running on a mobile phone can also access IMSI and IMEI numbers, though not as easily. Mobile operating systems lock access to hardware identifiers behind permissions that users must approve and can later revoke. For example, starting with Android Q, apps need to request the "READ_PRIVILEGED_PHONE_STATE" permission in order to read non-resettable IDs. On iOS, it's not possible for apps to access these identifiers at all. This makes other identifiers more attractive options for most app-based third-party trackers. Like phone numbers, IMSI and IMEI numbers are unique and persistent, but not readily available, as most trackers have a hard time accessing them.

Advertising IDs

An advertising ID is a long, random string of letters and numbers that uniquely identifies a mobile device. Advertising IDs aren't part of any technical protocols, but are built in to the iOS and Android operating systems.

Ad IDs on mobile phones are analogous to cookies on the Web. Instead of being stored by your browser and shared with trackers on different websites like cookies, ad IDs are stored by your

phone and shared with trackers in different apps. Ad IDs exist for the sole purpose of helping behavioral advertisers link user activity across apps on a device.

Unlike IMSI or IMEI numbers, ad IDs can be changed and, on iOS, turned off completely. Ad IDs are enabled by default on both iOS and Android, and are available to all apps without any special permissions. On both platforms, the ad ID does not reset unless the user does so manually.

Both Google and Apple encourage developers to use ad IDs for behavioral profiling in lieu of other identifiers like IMEI or phone number. Ostensibly, this gives users more control over how they are tracked, since users can reset their identifiers by hand if they choose. However, in practice, even if a user goes to the trouble to reset their ad ID, it's very easy for trackers to identify them across resets by using other identifiers, like IP address or in-app storage. Android's developer policy instructs trackers not to engage in such behavior, but the platform has no technical safeguards to stop it. In February 2019, a study found that over 18,000 apps on the Play store were violating Google's policy.

Ad IDs are unique, and available to all apps by default. They persist until users manually reset them. That makes them very attractive identifiers for surreptitious trackers.

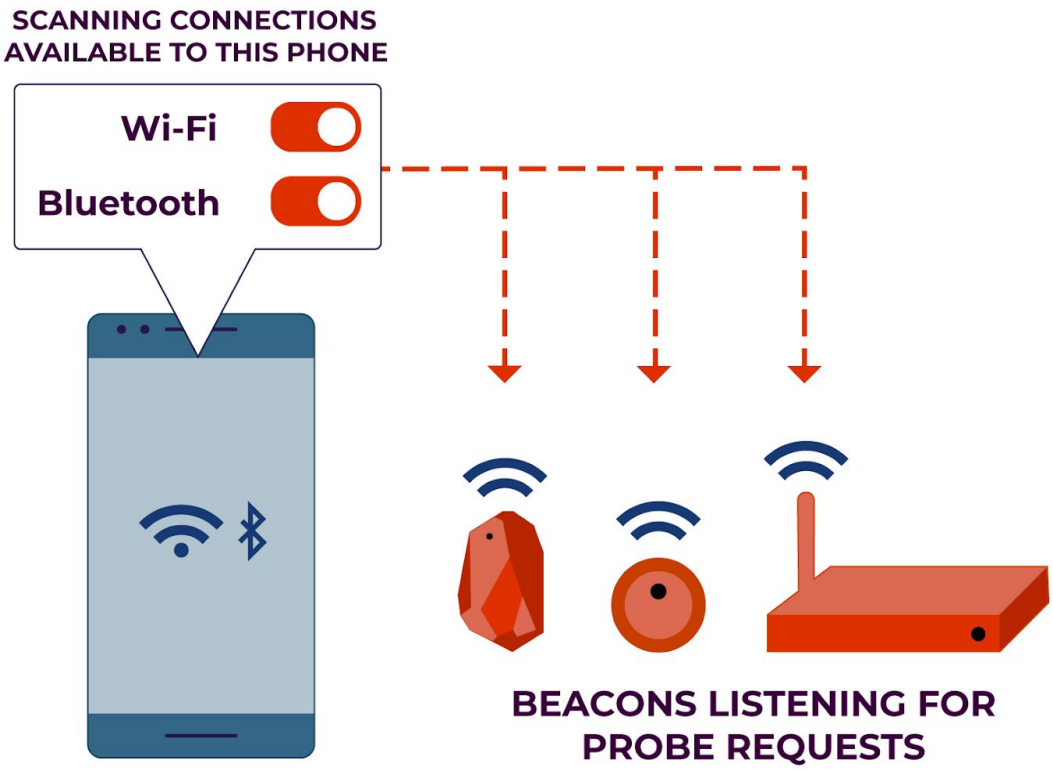
MAC addresses

Every device that can connect to the Internet has a hardware identifier called a Media Access Control (MAC) address. MAC addresses are used to set up the initial connection between two wireless-capable devices over WiFi or Bluetooth.

MAC addresses are used by all kinds of devices, but the privacy risks associated with them are heightened on mobile devices. Websites and other servers you interact with over the Internet can't actually see your MAC address, but any networking devices in your area can. In fact, you don't even have to connect to a network for it to see your MAC address; being nearby is enough.

Here's how it works. In order to find nearby Bluetooth devices and WiFi networks, your device is constantly sending out short radio signals called *probe requests*. Each probe request contains your device's unique MAC address. If there is a WiFi hotspot in the area, it will hear the probe and send back its own "probe response," addressed with your device's MAC, with information about how you can connect to it.

But other devices in the area can see and intercept the probe requests, too. This means that companies can set up wireless "beacons" that silently listen for MAC addresses in their vicinity, then use that data to track the movement of specific devices over time. Beacons are often set up in businesses, at public events, and even in political campaign yard signs. With enough beacons in enough places, companies can track users' movement around stores or around a city. They can also identify when two people are in the same location and use that information to build a social graph.



In order to find nearby Bluetooth devices and WiFi networks, your device is constantly sending out short radio signals called probe requests. Each probe request contains your device's unique MAC address. Companies can set up wireless "beacons" that silently listen for MAC addresses in their vicinity, then use that data to track the movement of specific devices over time.

This style of tracking can be thwarted with *MAC address randomization*. Instead of sharing its true, globally unique MAC address in probe requests, your device can make up a new, random, "spoofed" MAC address to broadcast each time. This makes it impossible for passive trackers to link one probe request to another, or to link them to a particular device. Luckily, the latest versions of iOS and Android both include MAC address randomization by default.

MAC address tracking remains a risk for laptops, older phones, and other devices, but the industry is trending towards more privacy-protective norms.

Hardware MAC addresses are globally unique. They are also persistent, not changing for the lifetime of a device. They are not readily available to trackers in apps, but are available to passive trackers using wireless beacons. However, since many devices now obfuscate MAC addresses by default, they are becoming a less reliable identifier for passive tracking.

Real-world identifiers

Many electronic device identifiers can be reset, obfuscated, or turned off by the user. But real-world identifiers are a different story: it's illegal to cover your car's license plate while

driving (and often while parked), and just about impossible to change biometric identifiers like your face and fingerprints.

License plates

Every car in the United States is legally required to have a license plate that is tied to their real-world identity. As far as tracking identifiers go, license plate numbers are about as good as it gets. They are easy to spot and illegal to obfuscate. They can't be changed easily, and they follow most people wherever they go.

Automatic license plate readers, or ALPRs, are special-purpose cameras that can automatically identify and record license plate numbers on passing cars. ALPRs can be installed at fixed points, like busy intersections or mall parking lots, or on other vehicles like police cars. Private companies operate ALPRs, use them to amass vast quantities of traveler location data, and sell this data to other businesses (as well as to police).

Unfortunately, tracking by ALPRs is essentially unavoidable for people who drive. It's not legal to hide or change your license plate, and since most ALPRs operate in public spaces, it's extremely difficult to avoid the devices themselves.

License plates are unique, available to anyone who can see the vehicle, and extremely persistent. They are ideal identifiers for gathering data about vehicles and their drivers, both for law enforcement and for third-party trackers.

Face biometrics

Faces are another class of unique identifier that are extremely attractive to third-party trackers. Faces are unique and highly inconvenient to change. Luckily, it's not illegal to hide your face from the general public, but it is impractical for most people to do so.

Everyone's face is unique, available, and persistent. However, current face recognition software will sometimes confuse one face for another. Furthermore, research has shown that algorithms are much more prone to making these kinds of errors when identifying people of color, women, and older individuals.

Facial recognition has already seen widespread deployment, but we are likely just beginning to feel the extent of its impact. In the future, facial recognition cameras may be in stores, on street corners, and docked on computer-aided glasses. Without strong privacy regulations, average people will have virtually no way to fight back against pervasive tracking and profiling via facial recognition.

Credit/debit cards

Credit card numbers are another excellent long-term identifier. While they can be cycled out, most people don't change their credit card numbers nearly as often as they clear their cookies. Additionally, credit card numbers are tied directly to real names, and anyone who receives your credit card number as part of a transaction also receives your legal name.

What most people may not understand is the amount of hidden third parties involved with each credit card transaction. If you buy a widget at a local store, the store probably contracts with a payment processor who provides card-handling services. The transaction also must be verified by your bank as well as the bank of the card provider. The payment processor in turn may employ other companies to validate its transactions, and all of these companies may receive information about the purchase. Banks and other financial institutions are regulated by the Gramm-Leach-Bliley Act, which mandates data security standards, requires them to disclose how user data is shared, and gives users the right to opt out of sharing. However, other financial technology companies, like payment processors and data aggregators, are significantly less regulated.

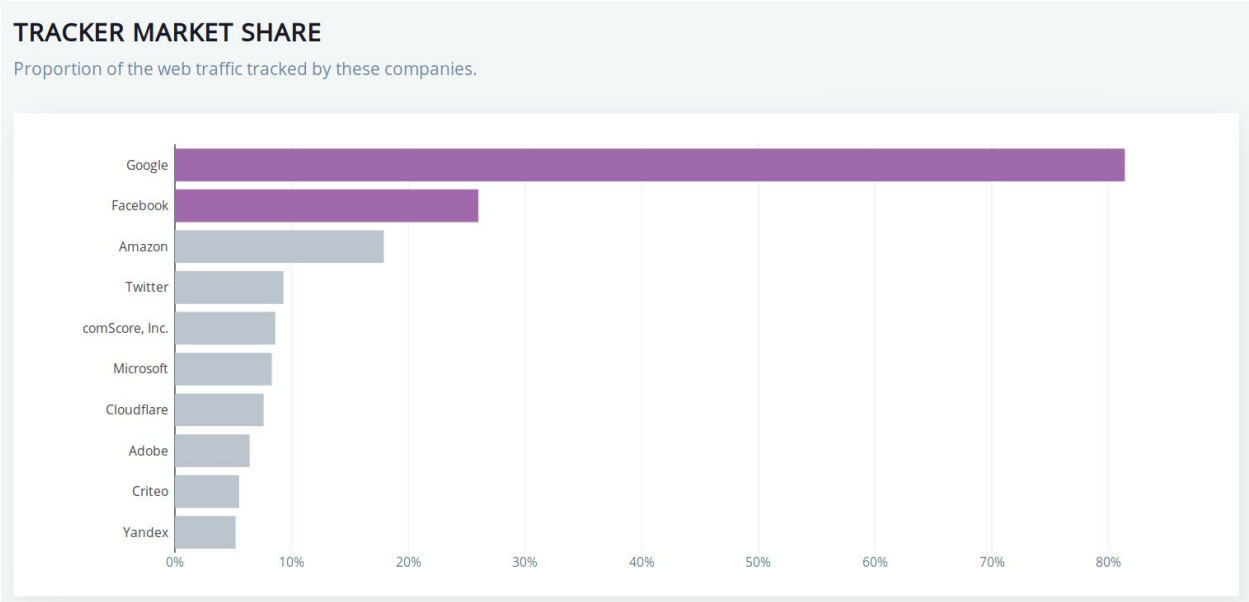
Linking identifiers over time

Often, a tracker can't rely on a single identifier to act as a stable link to a user. IP addresses change, people clear cookies, ad IDs can be reset, and more savvy users might have "burner" phone numbers and email addresses that they use to try to separate parts of their identity. When this happens, trackers don't give up and start a new user profile from scratch. Instead, they typically combine several identifiers to create a unified profile. This way, they are less likely to lose track of the user when one identifier or another changes, and they can link old identifiers to new ones over time.

Trackers have an advantage here because there are *so many* different ways to identify a user. If a user clears their cookies but their IP address doesn't change, linking the old cookie to the new one is trivial. If they move from one network to another but use the same browser, a browser fingerprint can link their old session to their new one. If they block third-party cookies and use a hard-to-fingerprint browser like Safari, trackers can use first-party cookie sharing in combination with TLS session data to build a long-term profile of user behavior. In this cat-and-mouse game, trackers have technological advantages over individual users.

Part 2: From bits to Big Data: What do tracking networks look like?

In order to track you, most tracking companies need to convince website or app developers to include custom tracking code in their products. That's no small thing: tracking code can have a number of undesirable effects for publishers. It can slow down software, annoy users, and trigger regulation under laws like GDPR. Yet the largest tracking networks cover vast swaths of the Web and the app stores, collecting data from millions of different sources all the time. In the physical world, trackers can be found in billboards, retail stores, and mall parking lots. So how and why are trackers so widespread? In this section, we'll talk about what tracking networks look like in the wild.

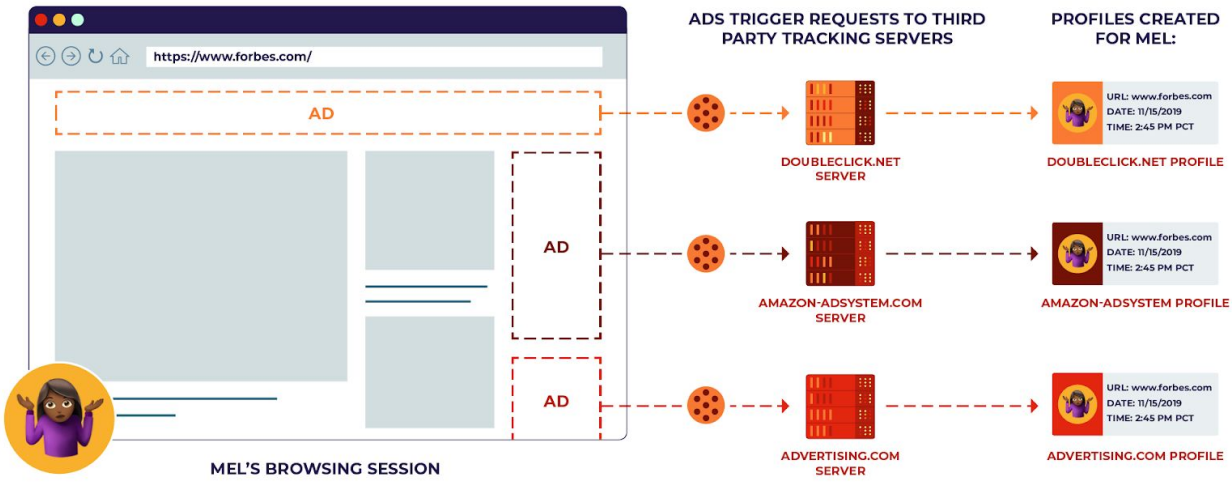


Top trackers on the Web, ranked by the proportion of web traffic that they collect data from. Google collects data about over 80% of measured web traffic. Source: WhoTracks.me, by Cliqz GBMH.

Tracking in software: Websites and Apps

Ad networks

The dominant market force behind third-party tracking is the advertising industry, as discussed below in Part 3. So it's no surprise that online ads are one of the primary vectors for data collection. In the simplest model, a single third-party ad network serves ads on a number of websites. Each publisher that works with the ad network must include a small snippet of code on their website that will load an ad from the ad server. This triggers a request to the ad server each time a user visits one of the cooperating sites, which lets the ad server set third-party cookies into users' browsers and track their activity across the network. Similarly, an ad server might provide an ad-hosting software development kit (SDK) for mobile app developers to use. Whenever a user opens an app that uses the SDK, the app makes a request to the ad server. This request can contain the advertising ID for the user's device, thus allowing the ad server to profile the user's activity across apps.



Each ad your browser loads may come from a different advertising server, and each server can build its own profile of you based on your activity. Each time you connect to that server, it can use a cookie to link that activity to your profile.

In reality, the online ad ecosystem is even more complicated. Ad *exchanges* host "real time auctions" for individual ad impressions on web pages. In the process, they may load code from several other third-party advertising providers, and may share data about each impression with many potential advertisers participating in the auction. Each ad you see might be responsible for sharing data with dozens of trackers. We'll go into more depth about Real Time Bidding and other data-sharing activities in Part 3.

Analytics and tracking pixels

Tracking code often isn't associated with anything visible to users, like a third-party ad. On the web, a significant portion of tracking happens via invisible, 1-pixel-by-1-pixel "images" that exist only to trigger requests to the trackers. These "tracking pixels" are used by many of the most prolific data collectors on the web, including Google Analytics, Facebook, Amazon, and DoubleVerify.

When website owners install a third party's tracking pixels, they usually do so in exchange for access to some of the data the third party collects. For example, Google Analytics and Chartbeat use pixels to collect information, and offer website owners and publishers insights about what kinds of people are visiting their sites. Going another level deeper, advertising platforms like Facebook also offer "conversion pixels," which allow publishers to keep track of how many click-throughs their own third-party ads are getting.

The biggest players in web-based analytics offer similar services to mobile apps. Google Analytics and Facebook are two of the most popular SDKs on both Android and iOS. Like their counterparts on the Web, these services silently collect information about users of mobile apps and then share some of that information with the app developers themselves.

Mobile third-party trackers convince app developers to install their SDKs by providing useful features like analytics or single sign-on. SDKs are just big blobs of code that app developers add to their projects. When they compile and distribute an app, the third-party code ships with it.

Unlike Web-based tools, analytics services in mobile apps don't need to use "pixels" or other tricks to trigger third-party requests.

Another class of trackers work on behalf of advertisers rather than first-party sites or apps. These companies work with advertisers to monitor where, how, and to whom their ads are being served. They often don't work with first-party publishers at all; in fact, their goal is to gather data about publishers as well as users.

DoubleVerify is one of the largest such services. Third-party advertisers inject DoubleVerify code alongside their ads, and DoubleVerify estimates whether each impression is coming from a real human (as opposed to a bot), whether the human is who the advertiser meant to target, and whether the page around the ad is "brand safe." According to its privacy policy, the company measures "how long the advertisement was displayed in the consumer's browser" and "the display characteristics of the ad on the consumer's browser." In order to do all that, DoubleVerify gathers detailed data about users' browsers; it is by far the largest source of third-party browser fingerprinting on the web. It collects location data, including data from other third-party sources, to try to determine whether a user is viewing an ad in the geographic area that the advertiser targeted.

Other companies in the space include Adobe, Oracle, and Comscore.

Embedded media players

Sometimes, third-party trackers serve content that users actually want to see. On the web, embedding third-party content is extremely common for blogs and other media sites. Some examples include video players for services like YouTube, Vimeo, Streamable, and Twitter, and audio widgets for Soundcloud, Spotify, and podcast-streaming services. These media players nearly always run inside IFrames, and therefore have access to local storage and the ability to run arbitrary JavaScript. This makes them well-suited to tracking users as well.

Social media widgets

Social media companies provide a variety of services to websites, such as Facebook Like buttons and Twitter Share buttons. These are often pitched as ways for publishers to improve traffic numbers on their own platforms as well as their presence on social media. Like and Share buttons can be used for tracking in the same way that pixels can: the "button" is really an embedded image which triggers a request to the social media company's server.

More sophisticated widgets, like comment sections, work more like embedded media players. They usually come inside of IFrames and enjoy more access to users' browsers than simple pixels or images. Like media players, these widgets are able to access local storage and run JavaScript in order to compute browser fingerprints.

Finally, the biggest companies (Facebook and Google in particular) offer account management services to smaller companies, like "Log in with Google." These services, known as "single sign-on," are attractive to publishers for several reasons. Independent websites and apps can offload the work of managing user accounts to the big companies. Users have fewer username/password pairs to remember, and less frequently go through annoying sign up/log-in flows. But for users, there is a price: account management services allow log-in

providers to act as a third party and track their users' activity on *all* of the services they log into. Log-in services are more reliable trackers than pixels or other simple widgets because they force users to confirm their identity.

CAPTCHAs

CAPTCHAs are a technology that attempts to separate users from robots. Publishers install CAPTCHAs on pages where they want to be particularly careful about blocking automated traffic, like sign-up forms and pages that serve particularly large files.

Google's ReCAPTCHA is the most popular CAPTCHA technology on the web. Every time you connect to a site that uses recaptcha, your browser connects to a *.google.com domain in order to load the CAPTCHA resources and shares all associated cookies with Google. This means that its CAPTCHA network is another source of data that Google can use to profile users.

While older CAPTCHAs asked users to read garbled text or click on pictures of bikes, the new ReCAPTCHA v3 records "interactions with the website" and silently guesses whether a user is human. ReCAPTCHA scripts don't send raw interaction data back to Google. Rather, they generate something akin to a behavioral fingerprint, which summarizes the way a user has interacted with a page. Google feeds this into a machine-learning model to estimate how likely the user is to be human, then returns that score to the first-party website. In addition to making things more convenient for users, this newer system benefits Google in two ways. First, it makes CAPTCHAS invisible to most users, which may make them less aware that Google (or anyone) is collecting data about them. Second, it leverages Google's huge set of behavioral data to cement its dominance in the CAPTCHA market, and ensures that any future competitors will need their own tranches of interaction data in order to build tools that work in a similar way.

Session replay services

Session replay services are tools that website or app owners can install in order to actually record how users interact with their services. These services operate both on websites and in apps. They log keystrokes, mouse movements, taps, swipes, and changes to the page, then allow first-party sites to "re-play" individual users' experiences after the fact. Often, users are given no indication that their actions are being recorded and shared with third parties.

These creepy tools create a massive risk that sensitive data, like medical information, credit card numbers, or passwords, will be recorded and leaked. The providers of session replay services usually leave it up to their clients to designate certain data as off-limits. But for clients, the process of filtering out sensitive information is subtle, painstaking, and time-consuming, and it clashes with replay services' promises to get set up "in a matter of seconds." As a result, independent auditing has found that sensitive data ends up in the recordings, and that session replay service providers often fail to secure that data appropriately.

Passive, real-world tracking

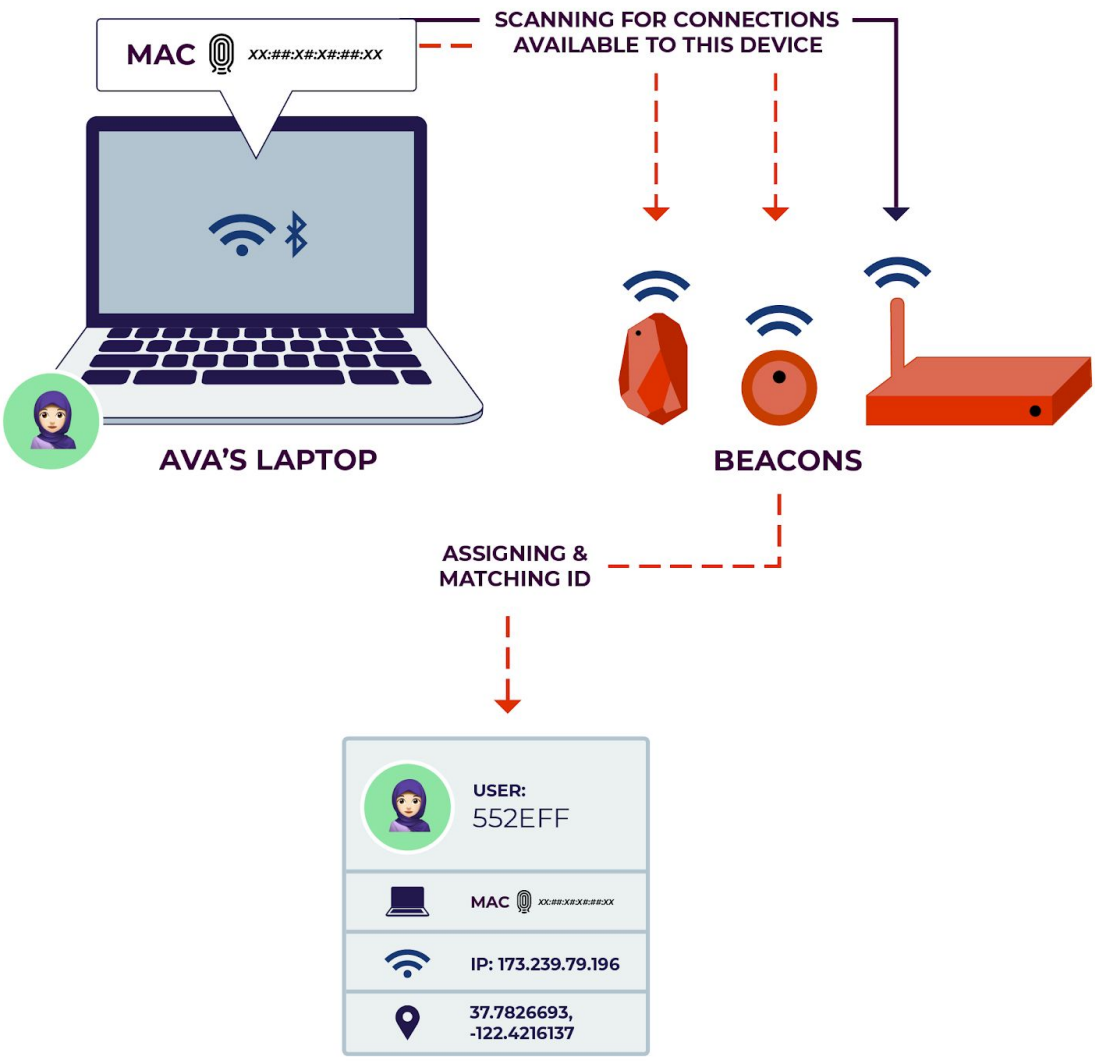
WiFi hotspots and wireless beacons

Many consumer devices emit wireless "probe" signals, and many companies install commercial beacons that intercept these probes all over the physical world. Some devices randomize the unique MAC address device identifiers they share in probes, protecting themselves from passive tracking, but not all do. And connecting to an open WiFi network or giving an app Bluetooth permissions always opens a device up to tracking.

As we discussed above, WiFi hotspots, wireless beacons, and other radio devices can be used to "listen" for nearby devices. Companies like Comcast (which provides XFinity WiFi) and Google (which provides free WiFi in Starbucks and many other businesses) have WiFi hotspots installed all over the world; Comcast alone boasts over 18 million XFinity WiFi installations. Dozens of other companies that you likely haven't heard of provide free WiFi to coffee shops, restaurants, events, and hotels.

Companies also pay to install wireless beacons in real-world businesses and public spaces. Bluetooth-enabled beacons have been installed around retail stores, at political rallies, in campaign lawn signs, and on streetlight poles.

Wireless beacons are capable of tracking on two levels. First, and most concerning, wireless beacons can passively monitor the "probes" that devices send out all the time. If a device is broadcasting its hardware MAC address, companies can use the probes they collect to track its user's movement over time.



WiFi hotspots and bluetooth beacons can listen for probes that wireless devices send out automatically. Trackers can use each device's MAC address to create a profile of it based on where they've seen that device.

Second, when a user connects to a WiFi hotspot or to a Bluetooth beacon, the controller of the hotspot or beacon can connect the device's MAC address to additional identifiers like IP address, cookies, and ad ID. Many WiFi hotspot operators also use a sign-in page to collect information about users' real names or email addresses. Then, when users browse the web from that hotspot, the operator can collect data on all the traffic coming from the user's device, much like an ISP. Bluetooth beacons are used slightly differently. Mobile phones allow apps to access the Bluetooth interface with certain permissions. Third-party trackers in apps with Bluetooth permissions can automatically connect to Bluetooth beacons in the real world, and they can use those connections to gather fine-grained location data.

Thankfully, both iOS and Android devices now send obfuscated MAC addresses with probes by default. This prevents the first, passive style of tracking described above.

But phones aren't the only devices with wireless capability. Laptops, e-readers, wireless headphones, and even cars are often outfitted with Bluetooth capability. Some of these devices don't have the MAC randomization features that recent models of smartphones do, making them vulnerable to passive location tracking.

Furthermore, even devices with MAC randomization usually share static MAC addresses when they actually connect to a wireless hotspot or Bluetooth device. This heightens the risks of the second style of tracking described above, which occurs when the devices connect to public WiFi networks or local Bluetooth beacons.

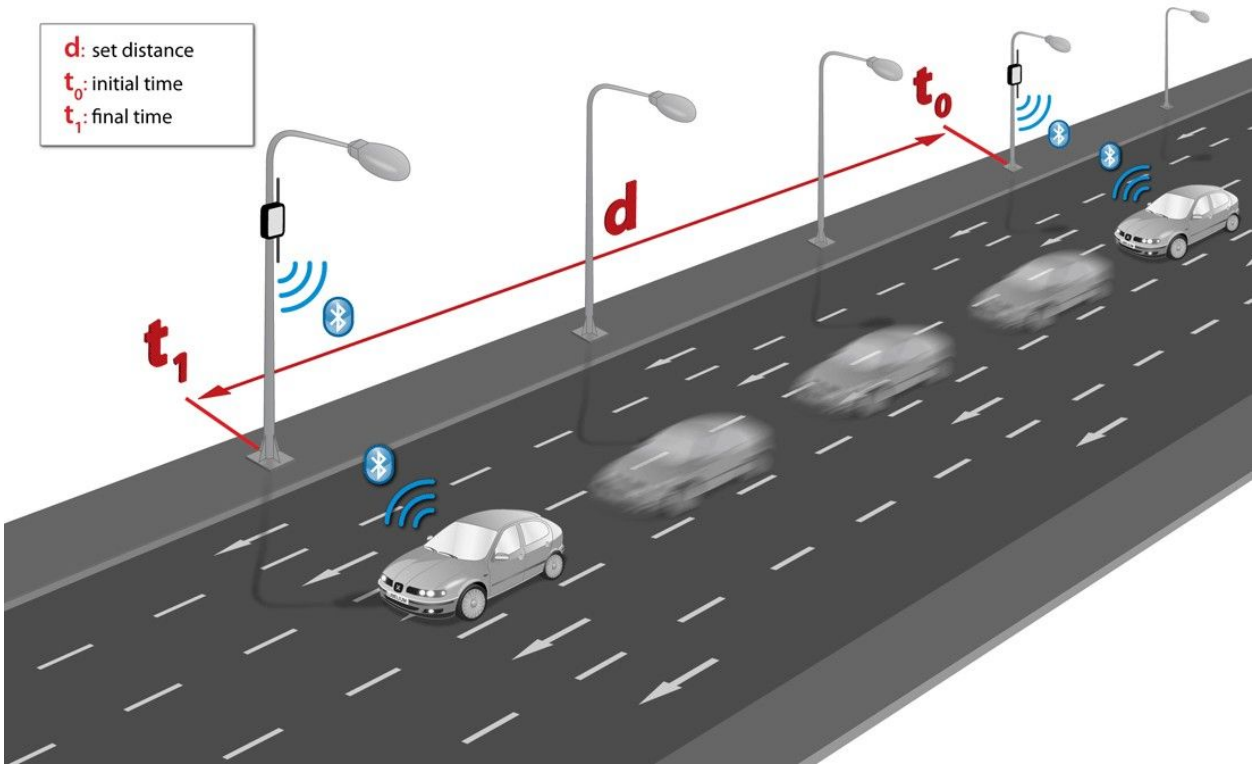
Vehicle tracking and ALPRs

Automated license plate readers, or ALPRs, are cameras outfitted with the ability to detect and read license plates. They can also use other characteristics of cars, like make, model, color, and wear, in order to help identify them. ALPRs are often used by law enforcement, but many ALPR devices are owned by private companies. These companies collect vehicle data indiscriminately, and once they have it, they can re-sell it to whomever they want: local police, federal immigration enforcement agencies, private data aggregators, insurance companies, lenders, or bounty hunters.

Different companies gather license plate data from different sources, and sell it to different audiences. Digital Recognition Network, or DRN, sources its data from thousands of repossession agencies around the country, and sells data to insurance agencies, private investigators, and "asset recovery" companies. According to an investigation by Motherboard, the vast majority of individuals about whom DRN collects data are not suspected of a crime or behind on car payments. The start-up Flock Safety offers ALPR-powered "neighborhood watch" services. Concerned homeowners can install ALPRs on their property in order to record and share information about cars that drive through their neighborhood.

DRN is owned by VaaS International Holdings, a Fort Worth-based company that brands itself as "the preeminent provider of license plate recognition ('LPR') and facial recognition products and data solutions." It also owns Vigilant Solutions, another private purveyor of ALPR technology. Vigilant's clients include law enforcement agencies and private shopping centers. Vigilant pools data from thousands of sources around the country into a single database, which it calls "PlateSearch." Scores of law enforcement agencies pay for access to PlateSearch. According to EFF's research, approximately 99.5% of the license plates recorded by Vigilant are not connected to a public safety interest at the time they are scanned.

Cameras and machine vision aren't the only technologies enabling vehicle tracking. Passive MAC address tracking can also be used to track vehicle movement. Phones inside of vehicles, and sometimes the vehicles themselves, broadcast probe requests including their MAC addresses. Wireless beacons placed strategically around roads can listen for those signals. One company, Libelium, sells a wireless beacon that is meant to be installed on streetlights in order to track nearby traffic.



Marketing material from the tracking hardware company Libelium, illustrating how its bluetooth beacons can be used to track the movement of cars over time.

Face recognition cameras

Face recognition has been deployed widely by law enforcement in some countries, including China and the UK. This has frightening implications: it allows mass logging of innocent people's activities. In China, it has been used to monitor and control members of the Uighur minority community.

We've covered the civil liberties harms associated with law enforcement use of face recognition extensively in the past. But face recognition also has been deployed in a number of private industries. Airlines use face recognition to authenticate passengers before boarding. Concert venues and ticket sellers have used it to screen concert-goers. Retailers use face recognition to identify people who supposedly are greater risks for shoplifting, which is especially concerning considering that the underlying mugshot databases are riddled with unfair racial disparities, and the technology is more likely to misidentify people of color. Private security companies sell robots equipped with face recognition to monitor public spaces and help employers keep tabs on employees. And schools and even summer camps use it to keep tabs on kids.

Big tech companies have begun investing in facial recognition for payment processing, which would give them another way to link real-world activity to users' online personas. Facebook has filed a patent on a system that would link faces to social media profiles in order to process payments. Also, Amazon's brick-and-mortar "Go" stores rely on biometrics to track who enters and what they take in order to charge them accordingly.

In addition, many see facial recognition as a logical way to bring targeted advertising to the physical world. Face recognition cameras can be installed in stores, on billboards, and in malls to profile people's behavior, build dossiers on their habits, and target messages at them. In January 2019, Walgreens began a pilot program using face recognition cameras installed on LED-screen fridge doors. The idea is that, instead of looking through a plate of glass to see the contents of a fridge, consumers can look at a screen which will display graphics indicating what's inside. The camera can perform facial recognition on whoever is standing in front of the fridge, and the graphics can be dynamically changed to serve ads targeted to that person. Whether or not Walgreens ends up deploying this technology at a larger scale, this appears to be one direction retailers are heading.

Payment processors and financial technology

Financial technology, or "fintech," is a blanket term for the burgeoning industry of finance-adjacent technology companies. Thousands of relatively new tech companies act as the technological glue between old-guard financial institutions and newer technologies, including tracking and surveillance. When they are regulated, fintech companies are often subject to less government oversight than traditional institutions like banks.

Payment processors are companies that accept payments on behalf of other businesses. As a result, they are privy to huge amounts of information about what businesses sell and what people buy. Since most financial transactions involve credit card numbers and names, it is easy for payment processors to tie the data they collect to real identities. Some of these companies are pure service providers, and don't use data for any purposes other than moving money from one place to another. Others build profiles of consumers or businesses and then monetize that data. For example, Square is a company that makes credit card readers for small businesses. It also uses the information it collects to serve targeted ads from third parties and to underwrite loans through its Square Capital program.

Some fintech companies offer financial services directly to users, like Intuit, the company behind TurboTax and Mint. Others provide services to banks or businesses. In the fintech world, "data aggregators" act as intermediaries between banks and other services, like money management apps. In the process, data aggregators gain access to all the data that passes through their pipes, including account balances, outstanding debts, and credit card transactions for millions of people. In addition, aggregators often collect consumers' usernames and passwords in order to extract data from their banks. Yodlee, one of the largest companies in the space, sells transaction data to hedge funds, which mine the information to inform stock market moves. Many users are unaware that their data is used for anything other than operating the apps they have signed up for.

Tracking and corporate power

Many of the companies that benefit most from data tracking have compelling ways to entice web developers, app creators, and store managers to install their tracking technology. Companies with monopolies or near-monopolies can use their market power to build tracking networks, monitor and inhibit smaller competitors, and exploit consumer privacy for their own

economic advantage. Corporate power and corporate surveillance reinforce one another in several ways.

First, dominant companies like Google and Facebook can pressure publishers into installing their tracking code. Publishers rely on the world's biggest social network and the world's biggest search engine to drive traffic to their own sites. As a result, most publishers need to advertise on those platforms. And in order to track how effective their ads are, they have no choice but to install Google and Facebook's conversion measurement code on their sites and apps. Google, Facebook, and Amazon also act as third-party ad networks, together controlling over two-thirds of the market. That means publishers who want to monetize their content have a hard time avoiding the big platforms' ad tracking code.

Second, vertically integrated tech companies can gain control of both sides of the tracking market. Google administers the largest behavioral advertising system in the world, which it powers by collecting data from its Android phones and Chrome browser—the most popular mobile operating system and most popular web browser in the world. Compared to its peer operating systems and browsers, Google's user software makes it easier for its trackers to collect data.

When the designers of the Web first described browsers, they called them "user agents:" pieces of software that would act on their users' behalf on the Internet. But when a browser maker is also a company whose main source of revenue is behavioral advertising, the company's interest in user privacy and control is pitted against the company's interest in tracking. The company's bottom line usually comes out on top.

Third, data can be used to profile not just people, but also competitor companies. The biggest data collectors don't just know how we act, they also know more about the market—and their competitors—than anyone else. Google's tracking tools monitor over 80% of traffic on the Web, which means it often knows as much about it's competitors' traffic as its competitors do (or more). Facebook (via third-party ads, analytics, conversion pixels, social widgets, and formerly its VPN app Onavo) also monitors the use and growth of websites, apps, and publishers large and small. Amazon already hosts a massive portion of the Internet in its Amazon Web Services computing cloud, and it is starting to build its own formidable third-party ad network. These giants use this information to identify nascent competitors, and then buy them out or clone their products before they become significant threats. According to confidential internal documents, Facebook used data about users' app habits from Onavo, its VPN, to inform its acquisition of WhatsApp.

Fourth, as tech giants concentrate tracking power into their own hands, they can use access to data as an anticompetitive cudgel. Facebook was well aware that access to its APIs (and the detailed private data that entailed) were invaluable to other social companies. It has a documented history of granting or withholding access to user data in order to undermine its competition.

Furthermore, Google and Facebook have both begun adopting policies that restrict competitors' access to their data without limiting what they collect themselves. For example, most of the large platforms now limit the third-party trackers on their own sites. In its own version of RTB, Google has recently begun restricting access to ad identifiers and other information that would allow competing ad networks to build user profiles. And following the Cambridge Analytica incident, Facebook started locking down access to third-party APIs, without meaningfully

changing anything about the data that Facebook itself collects on users. On the one hand, restricting third-party access can have privacy benefits. On the other, kicking third-party developers and outside actors off Facebook's and Google's platform services can make competition problems worse, give incumbent giants sole power over the user data they have collected, and cement their privacy-harmful business practices. Instead of seeing competition and privacy as isolated concerns, empowering users requires addressing *both* to reduce large companies' control over users' data and attention.

Finally, big companies can acquire troves of data from other companies in mergers and acquisitions. Google Analytics began its life as the independent company Urchin, which Google purchased in 2005. In 2007, Google supercharged its third-party advertising networks by purchasing Doubleclick, then as now a leader in the behaviorally targeted ad market. In late 2019, it purchased the health data company Fitbit, merging years of step counts and exercise logs into its own vast database of users' physical activity.

In its brief existence, Facebook has acquired 67 other companies. Amazon has acquired 91, and Google, 214—an average of over 10 per year. Many of the smaller firms that Facebook, Amazon, or Google have acquired had access to tremendous amounts of data and millions of active users. With each acquisition, those data sources are folded into the already-massive silos controlled by the tech giants. And thanks to network effects, the data becomes more valuable when it's all under one roof. On its own, Doubleclick could assemble pseudonymous profiles of users' browsing history. But as a part of Google, it can merge that data with real names, locations, cross-device activity, search histories, and social graphs.

Multi-billion dollar tech giants are not the only companies tracking us, nor are they the most irresponsible actors in the space. But the bigger they are, the more they know. And the more kinds of data a company has access to, the more powerful its profiles of users and competitors will be. In the new economy of personal information, the rich are only getting richer.

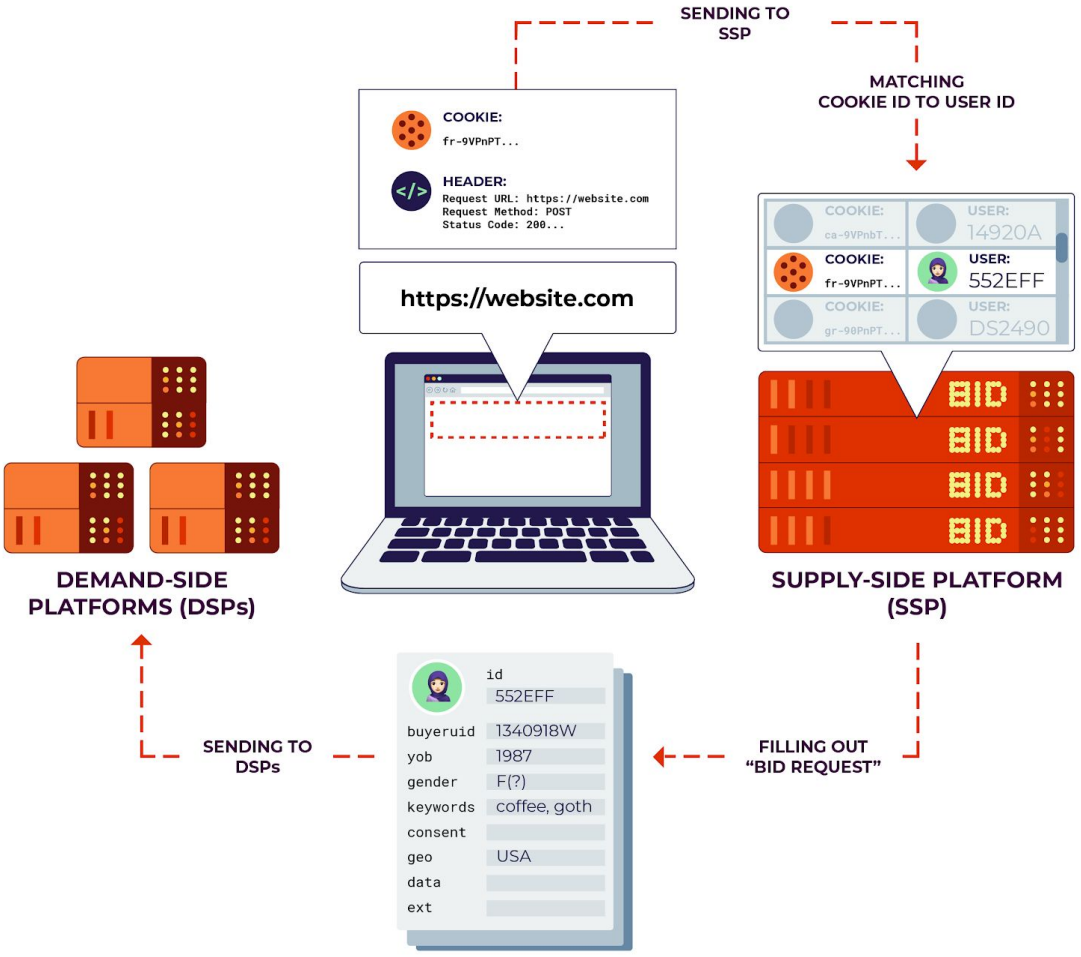
Part 3: Data sharing: Targeting, brokers, and real-time bidding

Where does the data go when it's collected? Most trackers don't collect every piece of information by themselves. Instead, companies work together, collecting data for themselves and sharing it with each other. Sometimes, companies with information about the same individual will combine it only briefly to determine which advertiser will serve which ad to that person. In other cases, companies base their entire business model on collecting and selling data about individuals they never interact with. In all cases, the type of data they collect and share can impact their target's experience, whether by affecting the ads they're exposed to or by determining which government databases they end up cataloged in. Moreover, the more a user's data is spread around, the greater the risk that they will be affected by a harmful data breach. This section will explore how personal information gets shared and where it goes.

Real-time bidding

Real-time bidding is the system that publishers and advertisers use to serve targeted ads. The unit of sale in the Internet advertising world is the "impression." Every time a person visits a web page with an ad, that person views an ad impression. Behind the scenes, an advertiser pays an ad network for the right to show you an ad, and the ad network pays the publisher of the web page where you saw the ad. But before that can happen, the publisher and the ad network have to decide *which* ad to show. To do so, they conduct a milliseconds-long auction, in which the auctioneer offers up a user's personal information, and then software on dozens of corporate servers bid on the rights to that user's attention. Data flows in one direction, and money flows in the other.

Such "real-time bidding" is quite complex, and the topic could use a whitepaper on its own. Luckily, there are tremendous, in-depth resources on the topic already. Dr. Johnny Ryan and Brave have written a series on the privacy impact of RTB. There is also a doctoral thesis on the privacy implications of the protocol. This section will give a brief overview of what the process looks like, much of which is based on Ryan's work.



Supply-side platforms use cookies to identify a user, then distribute "bid requests" with information about the user to potential advertisers.

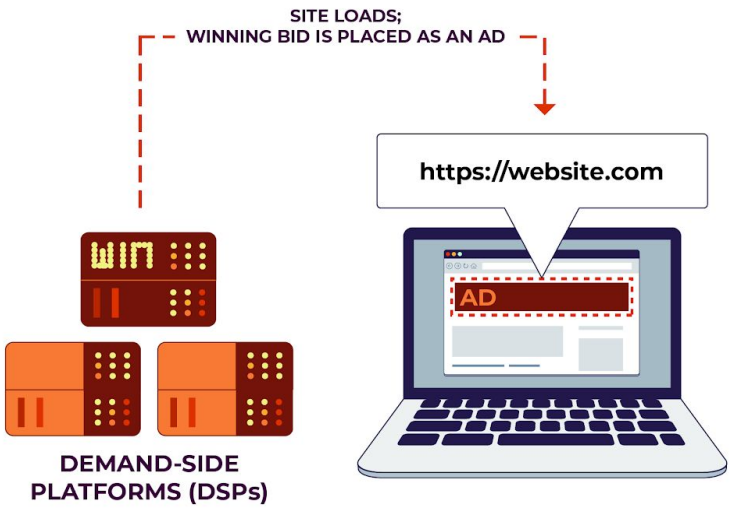
First, data flows from your browser to the ad networks, also known as "supply-side platforms" (SSPs). In this economy, your data and your attention are the "supply" that ad networks and SSPs are selling. Each SSP receives your identifying information, usually in the form of a cookie, and generates a "bid request" based on what it knows about your past behavior. Next, the SSP sends this bid request to each of the dozens of advertisers who have expressed interest in showing ads.

Attribute	Type	Definition
`id`	string; recommended	Vendor-specific ID for the user. At least one of `id` or `buyeruid` is strongly recommended.
`buyeruid`	string; recommended	Buyer-specific ID for the user as mapped by an exchange for the buyer. At least one of `id` or `buyeruid` is strongly recommended.
`yob`	integer	Year of birth as a 4-digit integer.
`gender`	string	Gender, where "M" = male, "F" = female, "O" = known to be other (i.e., omitted is unknown).
`keywords`	string	Comma separated list of keywords, interests, or intent.
`consent`	string	GDPR consent string if applicable, complying with the comply with the IAB standard Consent String Format in the Transparency and Consent Framework technical specifications.
`geo`	object	Location of the user's home base (i.e., not necessarily their current location). Refer to Object: Geo.
`data`	object array	Additional user data. Each `Data` object represents a different data source. Refer to Object: Data.
`ext`	object	Optional vendor-specific extensions.

The `user` object in an OpenRTB bid request contains the information a particular supply-side platform knows about the subject of an impression, including one or more unique IDs, age, gender, location, and interests.

Source: ***https://github.com/InteractiveAdvertisingBureau/AdCOM/blob/master/AdCOM%20v1.0%20FINAL.md#object--user-***

The bid request contains information about your location, your interests, and your device, and includes your unique ID. Figure X (above) shows the information included in an OpenRTB bid request.



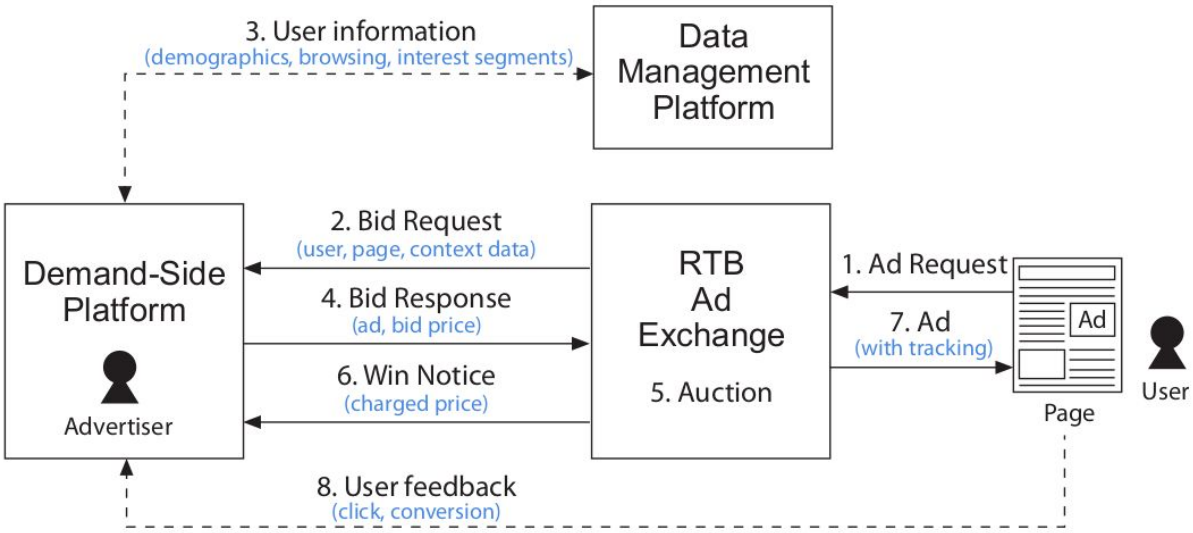
After the auction is complete, winning bidders pay supply-side platforms, SSPs pay the publisher, and the publisher shows the user an ad. At this point, the winning advertiser can collect even more information from the user's browser.

Finally, it's the bidders' turn. Using automated systems, the advertisers look at your info, decide whether they'd like to advertise to you and which ad they want to show, then respond to the SSP with a bid. The SSP determines who won the auction and displays the winner's ad on the publisher's web page.

All the information in the bid request is shared *before* any money changes hands. Advertisers who don't win the auction still receive the user's personal information. This enables "shadow bidding." Certain companies may pretend to be interested in buying impressions, but intentionally bid to lose in each auction with the goal of collecting as much data as possible as cheaply as possible.

Furthermore, there are several layers of companies that participate in RTB between the SSP and the advertisers, and each layer of companies also vacuums up user information. SSPs interface with "ad exchanges," which share data with "demand side platforms" (DSPs), which also share and purchase data from data brokers. Publishers work with SSPs to sell their ad space, advertisers work with DSPs to buy it, and ad exchanges connect buyers and sellers. You can read a breakdown of the difference between SSPs and DSPs, written for advertisers, here. Everyone involved in the process gets to collect behavioral data about the person who triggered the request.

During the bidding process, advertisers and the DSPs they work with can use third-party data brokers to augment their profiles of individual users. These data brokers, which refer to themselves innocuously as "data management platforms" (DMPs), sell data about individuals based on the identifiers and demographics included in a bid request. In other words, an advertiser can share a user ID with a data broker and receive that user's behavioral profile in return.



Source: Zhang, W., Yuan, S., Wang, J., and Shen, X. (2014b). Real-time bidding benchmarking with ipinyou dataset. arXiv preprint arXiv:1407.7073.

The diagram above gives another look at the flow of information and money in a single RTB auction.

In summary: (1) a user's visit to a page triggers an ad request from the page's publisher to an ad exchange. This is our real-time bidding "auctioneer." The ad exchange (2) requests bids from advertisers and the DSPs they work with, sending them information about the user in the process. The DSP then (3) augments the bid request data with more information from data brokers, or DMPs. Advertisers (4) respond with a bid for the ad space. After (5) a millisecond-long auction, the ad exchange (6) picks and notifiers the winning advertiser. The ad exchange (7) serves that ad to the user, complete with the tracking technology described above. The advertiser will (8) receive information about how the user interacted with the ad, e.g. how long they looked at it, what they clicked, if they purchased anything, etc. That data will feed back into the DSP's information about that user and other users who share their characteristics, informing future RTB bids.

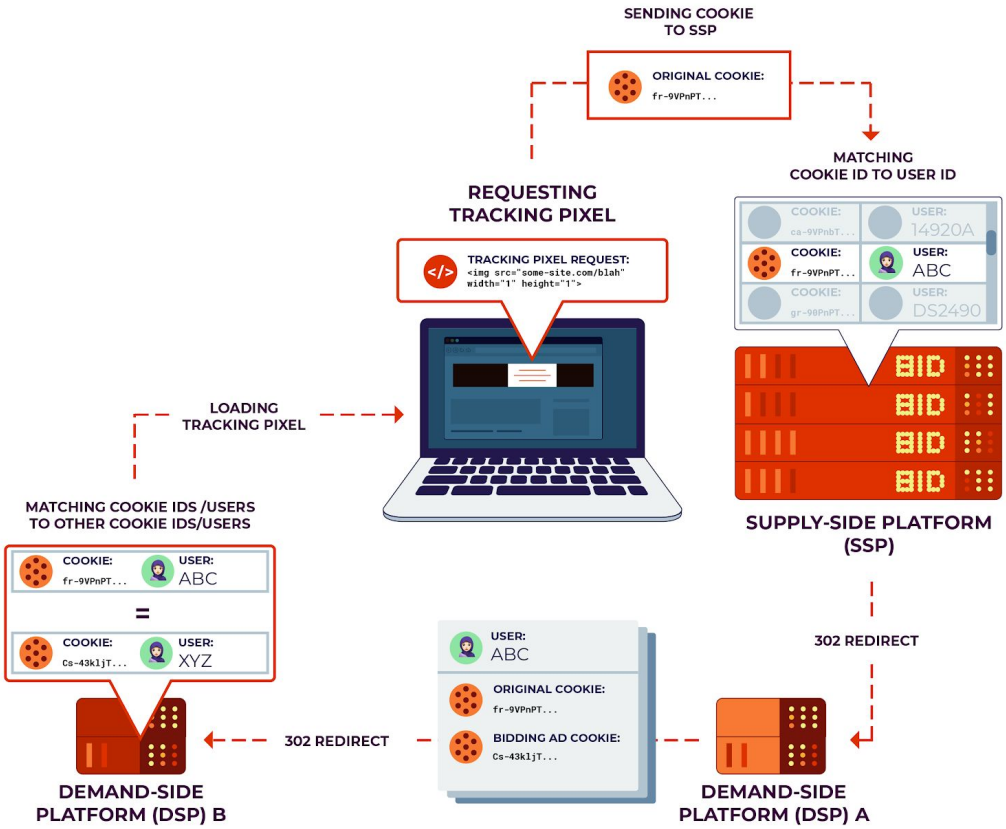
From the perspective of the user who visited the page, RTB causes two discrete sets of privacy invasions. First, before they visited the page, an array of companies tracked their personal information, both online and offline, and merged it all into a sophisticated profile about them. Then, during the RTB process, a different set of companies used that profile to decide how much to bid for the ad impression. Second, as a result of the user's visit to the page, the RTB participants harvest additional information from the visiting user. That information is injected into the user's old profile, to be used during subsequent RTBs triggered by their next page visits. Thus, RTB is both a cause of tracking and a means of tracking.

RTB on the web: cookie syncing

Cookie syncing is a method that web trackers use to link cookies with one another and combine the data one company has about a user with data that other companies might have.

Mechanically, it's very simple. One tracking domain triggers a request to another tracker. In the request, the first tracker sends a copy of its own tracking cookie. The second tracker gets both its own cookie and the cookie from the first tracker. This allows it to "compare notes" with the other tracker while building up its profile of the user.

Cookie sharing is commonly used as a part of RTB. In a bid request, the SSP shares its *own* cookie ID with all of the potential bidders. Without syncing, the demand side platforms might have their own profiles about users linked to their own cookie IDs. A DSP might not know that the user "abc" from Doubleclick (Google's ad network) is the same as its own user "xyz". Cookie syncing lets them be sure. As part of the bidding process, SSPs commonly trigger cookie-sync requests to many DSPs at a time. That way, the next time that SSP sends out a bid request, the DSPs who will be bidding can use their own behavioral profiles about the user to decide how to bid.



Cookie syncing. An invisible 'pixel' element on the page triggers a request to an ad exchange or SSP, which redirects the user to a DSP. The redirect URL contains information about the SSP's cookie that lets the DSP link it to its own identifier. A single SSP may trigger cookie syncs to many different DSPs at a time.

RTB in mobile apps

RTB was created for the Web, but it works just as well for ads in mobile apps. Instead of cookies, trackers use ad IDs. The ad IDs baked into iOS and Android make trackers' jobs easier. On the web, each advertiser has its own cookie ID, and demand-side platforms need to sync data with DMPs and with each other in order to tie their data to a specific user.

But on mobile devices, each user has a single, universal ad ID that is accessible from every app. That means that the syncing procedures described above on the web are not necessary on mobile; advertisers can use ad IDs to confirm identity, share data, and build more detailed profiles upon which to base bids.

Group targeting and look-alike audiences

Sometimes, large platforms do not disclose their data; rather, they lease out temporary access to their data-powered tools. Facebook, Google, and Twitter all allow advertisers to target categories of people with ads. For example, Facebook lets advertisers target users with certain "interests" or "affinities."

The companies do not show advertisers the actual identities of individuals their campaigns target. If you start a Facebook campaign targeting "people interested in Roller Derby in San Diego," you can't see a list of names right away. However, this kind of targeting does allow advertisers to reach out directly to roller derby-going San Diegans and direct them to an outside website or app. When targeted users click on an ad, they are directed off of Facebook and to the advertiser's domain. At this point, the advertiser knows they came from Facebook and that they are part of the targeted demographic. Once users have landed on the third-party site, the advertiser can use data exchange services to match them with behavioral profiles or even real-world identities.

In addition, Facebook allows advertisers to build "look-alike audiences" based on other groups of people. For example, suppose you're a payday loan company with a website. You can install an invisible Facebook pixel on a page that your debtors visit, make a list of people who visit that page, and then ask Facebook to create a "look-alike" audience of people who Facebook thinks are "similar" to the ones on your list. You can then target those people with ads on Facebook, directing them back to your website, where you can use cookies and data exchanges to identify who they are.

These "look-alike" features are black boxes. Without the ability to audit or study them, it's impossible to know what kinds of data they use and what kinds of information about users they might expose. We urge advertisers to disclose more information about them and to allow independent testing.

Data brokers

Data brokers are companies that collect, aggregate, process, and sell data. They operate out of sight from regular users, but in the center of the data-sharing economy. Often, data brokers have no direct relationships with users at all, and the people about whom they sell data may not

be aware they exist. Data brokers purchase information from a variety of smaller companies, including retailers, financial technology companies, medical research companies, online advertisers, cellular providers, Internet of Things device manufacturers, and local governments. They then sell data or data-powered services to advertisers, real estate agents, market research companies, colleges, governments, private bounty hunters, and other data brokers.

This is another topic that is far too broad to cover here, and others have written in depth about the data-selling ecosystem. Cracked Labs' report on corporate surveillance is both accessible and in-depth. Pam Dixon of the World Privacy Forum has also done excellent research into data brokers, including a report from 2014 and testimony before the Senate in 2015 and 2019.

The term "data broker" is broad. It includes "mom and pop" marketing firms that assemble and sell curated lists of phone numbers or emails, and behemoths like Oracle that ingest data from thousands of different streams and offer data-based services to other businesses.

Some brokers sell raw streams of information. This includes data about retail purchase behavior, data from Internet of Things devices, and data from connected cars. Others act as clearinghouses between buyers and sellers of all kinds of data. For example, Narrative promises to help sellers "unlock the value of [their] data" and help buyers "access the data [they] need." Dawex describes itself as "a global data marketplace where you can meet, sell and buy data directly."

Another class of companies act as middlemen or "aggregators," licensing raw data from several different sources, processing it, and repackaging it as a specific service for other businesses. For example, major phone carriers sold access to location data to aggregators called Zumigo and Microbilt, which in turn sold access to a broad array of other companies, with the resulting market ultimately reaching down to bail bondsmen and bounty hunters (and an undercover reporter). EFF is now suing AT&T for selling this data without users' consent and for misleading the public about its privacy practices.

Many of the largest data brokers don't sell the raw data they collect. Instead, they collect and consume data from thousands of different sources, then use it to assemble their own profiles and draw inferences about individuals. Oracle, one of the world's largest data brokers, owns Bluekai, one of the largest third-party trackers on the web. Credit reporting agencies, including Equifax and Experian, are also particularly active here. While the U.S. Fair Credit Reporting Act governs how credit raters can share specific types of data, it doesn't prevent credit agencies from selling most of the information that trackers collect today, including transaction information and browsing history. Many of these companies advertise their ability to derive *psychographics*, which are "innate" characteristics that describe user behavior. For example, Experian classifies people into financial categories like "Credit Hungry Card Switcher," "Disciplined, Passive Borrower," and "Insecure Debt Dependent," and claims to cover 95% of the U.S. population. Cambridge Analytica infamously used data about Facebook likes to derive "OCEAN scores"—ratings for openness, conscientiousness, extraversion, agreeableness, and neuroticism—about millions of voters, then sold that data to political campaigns.

Finally, many brokers use their internal profiles to offer "identity resolution" or "enrichment" services to others. If a business has one identifier, like a cookie or email address, it can pay a data broker to "enrich" that data and learn other information about the person. It can also link data tied to one identifier (like a cookie) to data from another (like a mobile ad ID). In the real-time bidding world, these services are known as "data management platforms." Real-time

bidders can use these kinds of services to learn who a particular user is and what their interests are, based only on the ID included with the bid request.

For years, data brokers have operated out of sight and out of mind of the general public. But we may be approaching a turning point. In 2018, Vermont passed the nation's first law requiring companies that buy and sell third-party data to register with the secretary of state. As a result, we now have access to a list of over 120 data brokers and information about their business models. Furthermore, when the California Consumer Privacy Act goes into effect in 2020, consumers will have the right to access the personal information that brokers have about them for free, and to opt out of having their data sold.

Data consumers

So far, this paper has discussed how data is collected, shared, and sold. But where does it end up? Who are the consumers of personal data, and what do they do with it?

Targeted advertising

By far the biggest, most visible, and most ubiquitous data consumers are targeted advertisers. Targeted advertising allows advertisers to reach users based on demographics, psychographics, and other traits. Behavioral advertising is a subset of targeted advertising that leverages data about users' past behavior in order to personalized ads.

The biggest data collectors are also the biggest targeted advertisers. Together, Google and Facebook control almost 60% of the digital ad market in the U.S., and they use their respective troves of data in order to do so. Google, Facebook, Amazon, and Twitter offer end-to-end targeting services where advertisers can target high-level categories of users, and the advertisers don't need to have access to any data themselves. Facebook lets advertisers target users based on location; demographics like age, gender, education, and income; and interests like hobbies, music genres, celebrities, and political leaning. Some of the "interests" Facebook uses are based on what users have "liked" or commented on, and others are derived based on Facebook's third-party tracking. While Facebook uses its data to match advertisers to target audiences, Facebook does not share its data with those advertisers.

Real-time bidding (RTB) involves more data sharing, and there are a vast array of smaller companies involved in different levels of the process. The big tech companies offer services in this space as well: Google's Doubleclick Bid Manager and Amazon DSP are both RTB demand-side platforms. In RTB, identifiers are shared so that the advertisers themselves (or their agents) can decide whether they want to reach each individual and what ad they want to show. In the RTB ecosystem, advertisers collect their own data about how users behave, and they may use in-house machine learning models in order to predict which users are most likely to engage with their ads or buy their products.

Some advertisers want to reach users on Facebook or Google, but don't want to use the big companies' proprietary targeting techniques. Instead, they can buy lists of contact information from data brokers, then upload those lists directly to Facebook or Google, who will reach those users across all of their platforms. This system undermines big companies' efforts to rein in discriminatory or otherwise malicious targeting. Targeting platforms like Google and Facebook do not allow advertisers to target users of particular ethnicities with ads for jobs, housing, or

credit. However, advertisers can buy demographic information about individuals from data brokers, upload a list of names who happen to be from the same racial group, and have the platform target those people directly. Both Google and Facebook forbid the use of "sensitive information" to target people with contact lists, but it's unclear how they enforce these policies.

Political campaigns and interest groups

Companies aren't the only entities that try to benefit from data collection and targeted advertising. Cambridge Analytica used ill-gotten personal data to estimate "psychographics" for millions of potential voters, then used that data to help political campaigns. In 2018, the group CatholicVote used cell-phone location data to determine who had been inside a Catholic church, then targeted them with "get out the vote" ads. Anti-abortion groups used similar geo-fencing technology to target ads to women while they were at abortion clinics..

And those incidents are not isolated. Some non-profits that rely on donations buy data to help narrow in on potential donors. Many politicians around the country have used open voter registration data to target voters. The Democratic National Committee is reportedly investing heavily in its "data warehouse" ahead of the 2020 election. And Deep Root Analytics, a consulting firm for the Republican party, was the source of the largest breach of US voter data in history; it had been collecting names, registration details, and "modeled" ethnicity and religion data about nearly 200 million Americans.

Debt collectors, bounty hunters, and fraud investigators

Debt collectors, bounty hunters, and repossession agencies all purchase and use location data from a number of sources. EFF is suing AT&T for its role in selling location data to aggregators, which enabled a secondary market that allowed access by bounty hunters. However, phone carriers aren't the only source of that data. The bail bond company Captira sold location data gathered from cell phones and ALPRs to bounty hunters for as little as $7.50. And thousands of apps collect "consensual" location data using GPS permissions, then sell that data to downstream aggregators. This data can be used to locate fugitives, debtors, and those who have not kept up with car payments. And as investigations have shown, it can also be purchased—and abused—by nearly anyone.

Cities, law enforcement, intelligence agencies

The public sector also purchases data from the private sector for all manner of applications. For example, U.S. Immigration and Customs Enforcement bought ALPR data from Vigilant to help locate people the agency intends to deport. Government agencies contract with data brokers for myriad tasks, from determining eligibility for human services to tax collection, according to the League of California Cities, in a letter seeking an exception from that state's consumer data privacy law for contracts between government agencies and data brokers. Advocates have long decried these arrangements between government agencies and private data brokers as a threat to consumer data privacy, as well as an end-run around legal limits on governments' own databases. And of course, national security surveillance often rests on the data mining of private companies' reservoirs of consumer data. For example, as part of the PRISM program revealed by

Edward Snowden, the NSA collected personal data directly from Google, YouTube, Facebook, and Yahoo.

Part 4: Fighting back

You might want to resist tracking to avoid being targeted by invasive or manipulative ads. You might be unhappy that your private information is being bartered and sold behind your back. You might be concerned that someone who wishes you harm can access your location through a third-party data broker. Perhaps you fear that data collected by corporations will end up in the hands of police and intelligence agencies. Or third-party tracking might just be a persistent nuisance that gives you a vague sense of unease.

But the unfortunate reality is that tracking is hard to avoid. With thousands of independent actors using hundreds of different techniques, corporate surveillance is widespread and well-funded. While there's no switch to flip that can prevent every method of tracking, there's still a lot that you can do to take back your privacy. This section will go over some of the ways that privacy-conscious users can avoid and disrupt third-party tracking.

Each person should decide for themselves how much effort they're willing to put into protecting their privacy. Small changes can seriously cut back on the amount of data that trackers can collect and share, like installing EFF's tracker-blocker extension Privacy Badger in your browser and changing settings on a phone. Bigger changes, like uninstalling third-party apps and using Tor, can offer stronger privacy guarantees at the cost of time, convenience, and sometimes money. Stronger measures may be worth it for users who have serious concerns.

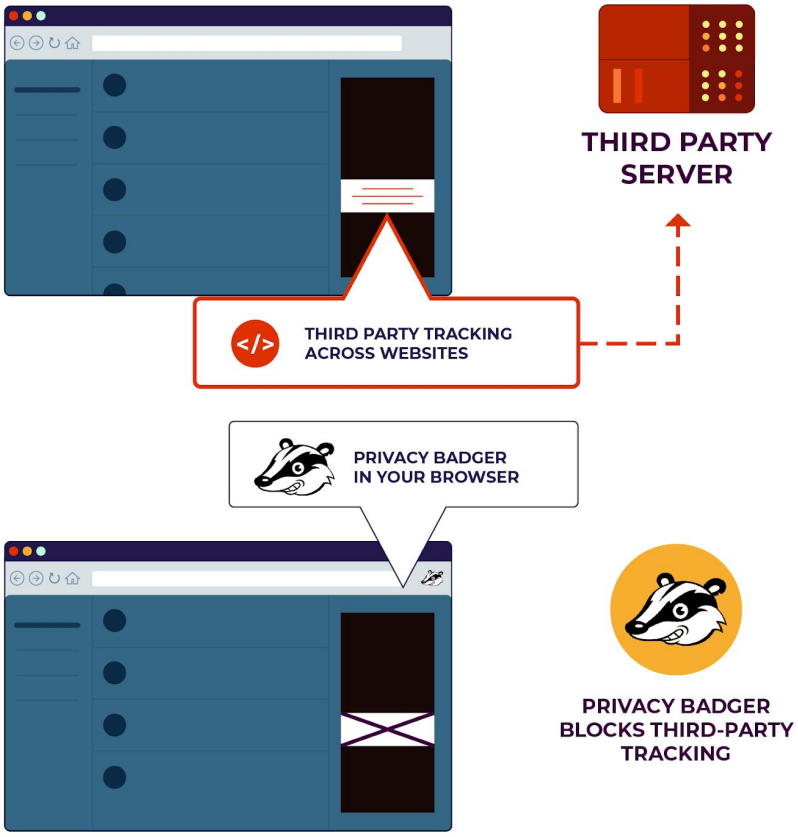
Finally, keep in mind that *none of this is your fault*. Privacy shouldn't be a matter of personal responsibility. It's not your job to obsess over the latest technologies that can secretly monitor you, and you shouldn't have to read through a quarter million words of privacy-policy legalese to understand how your phone shares data. Privacy should be a right, not a privilege for the well-educated and those flush with spare time. Everyone deserves to live in a world—online and offline—that respects their privacy.

In a better world, the companies that we choose to share our data with would earn our trust, and everyone else would mind their own business. That's why EFF files lawsuits to compel companies to respect consumers' data privacy, and why we support legislation that would make privacy the law of the land. With the help of our members and supporters, we are making progress, but changing corporate surveillance policies is a long and winding path. So for now, let's talk about how you can fight back.

On the web

There are several ways to limit your exposure to tracking on the Web. First, your choice of browser matters. Certain browser developers take more seriously their software's role as a "user agent" acting on your behalf. Apple's Safari takes active measures against the most common forms of tracking, including third-party cookies, first-to-third party cookie sharing, and fingerprinting. Mozilla's Firefox blocks third-party cookies from known trackers by default, and Firefox's Private Browsing mode will block requests to trackers altogether.

Browser extensions like EFF's Privacy Badger and uBlock Origin offer another layer of protection. In particular, Privacy Badger learns to block trackers using heuristics, which means it might catch new or uncommon trackers that static, list-based blockers miss. This makes Privacy Badger a good supplement to the built-in protections offered by Firefox, which rely on the Disconnect list. And while Google Chrome does not block any tracking behavior by default, installing Privacy Badger or another tracker-blocking extension in Chrome will allow you to use it with relatively little exposure to tracking. (However, planned changes in Chrome will likely affect the security and privacy tools that many use to block tracking.)



Browser extensions like EFF's Privacy Badger offer a layer of protection against third-party tracking on the web. Privacy Badger learns to block trackers using heuristics, which means it might catch new or uncommon trackers that static, list-based blockers miss.

No tracker blocker is perfect. All tracker blockers must make exceptions for companies that serve legitimate content. Privacy Badger, for example, maintains a list of domains which are known to perform tracking behaviors *as well as* serving content that is necessary for many sites to function, such as content delivery networks and video hosts. Privacy Badger restricts those domains' ability to track by blocking cookies and access to local storage, but dedicated trackers can still access IP addresses, TLS state, and some kinds of fingerprintable data.

If you'd like to go the extra mile and are comfortable with tinkering, you can install a network-level filter in your home. Pi-hole filters all traffic on a local network at the DNS level. It acts as a personal DNS server, rejecting requests to domains which are known to host

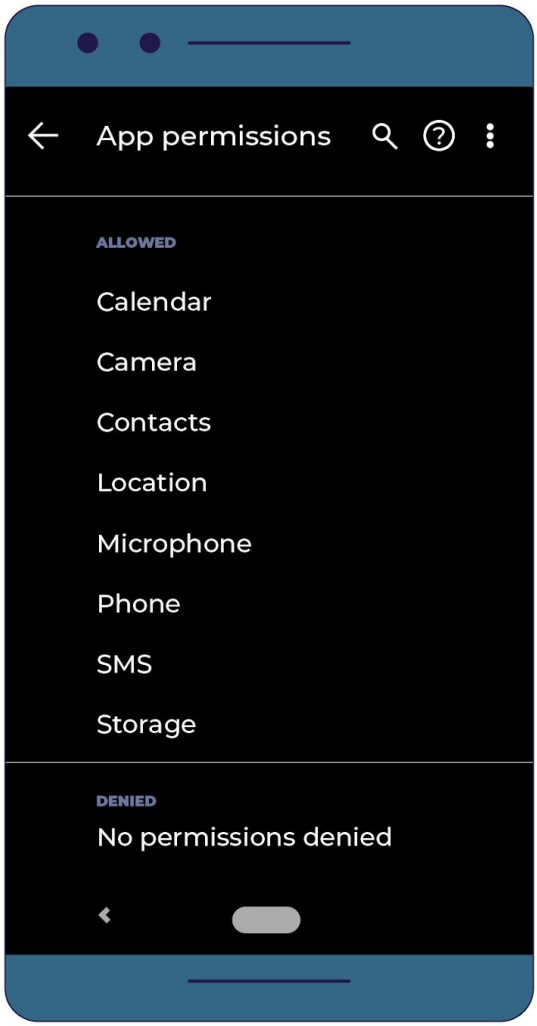
trackers. Pi-hole blocks tracking requests coming from devices which are otherwise difficult to configure, like smart TVs, game consoles, and Internet of Things products.

For people who want to reduce their exposure as much as possible, Tor Browser is the gold standard for privacy. Tor uses an *onion routing* service to totally mask its users' IP addresses. It takes aggressive steps to reduce fingerprinting, like blocking access to the HTML canvas by default. It completely rejects TLS session tickets and clears cookies at the end of each session.

Unfortunately, browsing the web with Tor in 2019 is not for everyone. It significantly slows down traffic, so pages take much longer to load, and streaming video or other real-time content is very difficult. Worse, much of the modern web relies on invisible CAPTCHAs that block or throttle traffic from sources deemed "suspicious." Traffic from Tor is frequently classified as high-risk, so doing something as simple as a Google search with Tor can trigger CAPTCHA tests. And since Tor is a public network which attackers also use, some websites will block Tor visitors altogether.

On mobile phones

Blocking trackers on mobile devices is more complicated. There isn't one solution, like a browser or an extension, that can cover many bases. And unfortunately, it's simply not possible to control certain kinds of tracking on certain devices.



The first line of defense against tracking is your device's settings.

Both iOS and Android let users view and control the permissions that each app has access to. You should check the permissions that your apps have, and remove the permissions that aren't needed. While you are at it, you might simply remove the apps you are not using. In addition to per-app settings, you can change global settings that affect how your device collects and shares particularly sensitive information, like location. You can also control how apps are allowed to access the Internet when they are not in use, which can prevent passive tracking.

Both operating systems also have options to reset your device's ad ID in different ways. On iOS, you can remove the ad ID entirely by setting it to a string of zeros. (Here are some other ways to block ad tracking on iOS.) On Android, you can manually reset it. This is equivalent to clearing your cookies,

but not blocking new ones: it won't disable tracking entirely, but will make it more difficult for trackers to build a unified profile about you.

Android also has a setting to "opt out of interest-based ads." This sends a signal to apps that the user does not want to have their data used for targeted ads, but it doesn't actually stop the apps from doing so by means of the ad ID. Indeed, recent research found that tens of thousands of apps simply ignore the signal.

On iOS, there are a handful of apps that can filter tracking activity from other apps. On Android, it's not so easy. Google bans ad- and tracker-blockers from its app store, the Play Store, so it has no officially vetted apps of this kind. It's possible to "side-load" blockers from outside of the Play Store, but this can be very risky. Make sure you only install apps from publishers you trust, preferably with open source code.

You should also think about the networks your devices are communicating with. It is best to avoid connecting to unfamiliar public WiFi networks. If you do, the "free" WiFi probably comes at the cost of your data.

Wireless beacons are also trying to collect information from your device. They can only collect identifying information if your devices are broadcasting their hardware MAC addresses. Both iOS and Android now randomize these MAC addresses by default, but other kinds of devices may not. Your e-reader, smart watch, or car may be broadcasting probe requests that trackers can use to derive location data. To prevent this, you can usually turn off WiFi and Bluetooth or set your device to "airplane mode." (This is also a good way to save battery!)

Finally, if you really need to be anonymous, using a "burner phone" can help you control tracking associated with inherent hardware identifiers.

IRL

In the real world, opting out isn't so simple.

As we've described, there are many ways to modify the way your devices work to prevent them from working against you. But it's almost impossible to avoid tracking by face recognition cameras and automatic license plate readers. Sure, you can paint your face to disrupt face recognition algorithms, you can choose not to own a car to stay out of ALPR companies' databases, and you can use cash or virtual credit cards to stop payment processors from profiling you. But these options aren't realistic for most people most of the time, and it's not feasible for *anyone* to avoid all the tracking that they're exposed to.

Knowledge is, however, half the battle. For now, face recognition cameras are most likely to identify you in specific locations, like airports, during international travel. ALPR cameras are much more pervasive and harder to avoid, but if absolutely necessary, it is possible to use public transit or other transportation methods to limit how often your vehicle is tracked.

In the legislature

Some jurisdictions have laws to protect users from tracking. The General Data Protection Regulation (GDPR) in the European Union gives those it covers the right to access and delete

information that's been collected about them. It also requires companies to have a legitimate reason to use data, which could come from a "legitimate interest" or opt-in consent. The GDPR is far from perfect, and its effectiveness will depend on how regulators and courts implement it in the years to come. But it gives meaningful rights to users and prescribes real consequences for companies who violate them.

In the U.S., a smattering of state and federal laws offer specific protections to some. Vermont's data privacy law brings transparency to data brokers. The Illinois Biometric Information Protection Act (BIPA) requires companies to get consent from users before collecting or sharing biometric identifiers. In 2020, the California Consumer Privacy Act (CCPA) will take effect, giving users there the right to access their personal information, delete it, and opt out of its sale. Some communities have passed legislation to limit government use of face recognition, and more plan to pass it soon.

At the federal level, some information in some circumstances is protected by laws like HIPAA, FERPA, COPPA, the Video Privacy Protection Act, and a handful of financial data privacy laws. However, these sector-specific federal statutes apply only to specific types information about specific types of people when held by specific businesses. They have many gaps, which are exploited by trackers, advertisers, and data brokers.

To make a long story very short, most third-party data collection in the U.S. is unregulated. That's why EFF advocates for new laws to protect user privacy. People should have the right to know what personal information is collected about them and what is done with it. We should be free from corporate processing of our data unless we give our informed opt-in consent. Companies shouldn't be able to charge extra or degrade service when users choose to exercise their privacy rights. They should be held accountable when they misuse or mishandle our data. And people should have the right to take companies to court when their privacy is violated.

The first step is to break the one-way mirror. We need to shed light on the tangled network of trackers that lurk in the shadows behind the glass. In the sunlight, these systems of commercial surveillance are exposed for what they are: Orwellian, but not omniscient; entrenched, but not inevitable. Once we, the users, understand what we're up against, we can fight back.