TV Ads, Fingerprinting and Differential Privacy
Review of recent articles

Florin Dobrian Oleg White

Data Analytics Group

February 25, 2020
rough draft version

Contents

1 Browser Fingerprinting, Analysis

(1) How Unique Is Your Web Browser?, Peter Eckersley, Electronic Frontier Foundation, 2010

This paper is widely considered as a foundational investigation in the field of browser fingerprinting. It is frequently cited and is a delight to read.

From the abstract:

We investigate the degree to which modern web browsers are subject to "device fingerprinting" via the version and configuration information that they will transmit to websites upon request. We implemented one possible fingerprinting algorithm, and collected these fingerprints from a large sample of browsers that visited our test side, `panopticlick.eff.org`. We observe that the distribution of our fingerprint contains at least 18.1 bits of entropy, meaning that if we pick a browser at random, at best we expect that only one in 286,777 other browsers will share its fingerprint. Among browsers that support Flash or Java, the situation is worse, with the average browser carrying at least 18.8 bits of identifying information. 94.2% of browsers with Flash or Java were unique in our sample.

By observing returning visitors, we estimate how rapidly browser fingerprints might change over time. In our sample, fingerprints changed quite rapidly, but even a simple heuristic was usually able to guess when a fingerprint was an "upgraded" version of a previously observed browser's fingerprint, with 99.1% of guesses correct and a false positive rate of only 0.86%.

We discuss what privacy threat browser fingerprinting poses in practice, and what countermeasures may be appropriate to prevent it. There is a tradeoff between protection against fingerprintability and certain kinds of debuggability, which in current browsers is weighted heavily against privacy. Paradoxically, anti-fingerprinting privacy technologies can be self-defeating if they are not used by a sufficient number of people; we show that some privacy measures currently fall victim to this paradox, but others do not.

Additional notes from the paper:

Fingerprints as Global Identifiers: Global identifier fingerprints are a worst case for privacy. But even users who are not globally identified by a particular fingerprint may be vulnerable to more context-specific kinds of tracking by the same fingerprint algorithm, if the print is used in combination with other data.

Fingerprint + IP address as Cookie Regenerators: Fingerprints may pose a 'cookie regeneration' threat, even if those fingerprints are not globally identifying. In particular, a fingerprint that carries no more than 15-20 bits of identifying information will in almost all cases be sufficient to uniquely identify a particular browser, given its IP address, its subnet, or even just its Autonomous System Number. If the user deletes their cookies while continuing to use an IP address, subnet or ASN that they have used previously, the cookie-setter could, with high probability, link their new cookie to the old one.

Fingerprint + IP address in the Absence of Cookies: A final use for fingerprints is as a means of distinguishing machines behind a single IP address, even if those machines block cookies entirely. It is very likely that fingerprinting will work for this purpose in all but a tiny number of cases.

(2) PriVaricator: Deceiving Fingerprinters with Little White Lies, February 2014

An important observation is that making fingerprints non-deterministic also makes them hard to link across subsequent web site visits. Our key insight is that when it comes to web tracking, the real problem with fingerprinting is not uniqueness of a fingerprint, it is linkability, i.e. the ability to connect the same fingerprint across multiple visits.

Third-party tracking: Probably the most common use of fingerprinting involves tracking the user across multiple, possibly unrelated, web sites to construct an interest profile for the user; this profile can then be employed to deliver targeted ads. As has been argued before, fingerprinting is an effective and stealthy alternative to stateful cookie-based user tracking. In a sense, the better fingerprinting works, the more information is learned about the user with a higher degree of reliability, leading to better ad targeting and thus to higher conversion rates for the advertisers. This creates a direct incentive for ad delivery networks to invest in better fingerprinting strategies, especially given that fingerprinters might not necessarily obey browser-provided Do-Not-Track (DNT) headers.

Fraud prevention: Third-party tracking, an activity that has provoked much outrage on the part of both privacy advocates and some users, is not the only raison d'ˆetre behind fingerprinting. It is sometimes argued that fingerprints can be used for fraud prevention. Advocates of fingerprinting claim that a device fingerprint is a powerful tool for finding related transactions either as an identifier in itself or as a means of finding transactions with related characteristics. Fingerprints also can be used to find out when account information is being shared illegally. The gathered fingerprints can be augmented with device reputation information and be used to blacklist fraudulent users and their activities.

Opt-out: While some of the aforementioned fingerprinting companies offer opt-out pages for the user, it is highly non-obvious what a successful opt-out really means. Ironically, to know that a user has opted-out of tracking, the fingerprinters still first need to compute the fingerprint (assuming cookies are disabled) and then, if they are honest, proceed to disregard information from that session.

`navigator.plugins` and installed fonts, timezone, screen dimensions, math constants

(3) (Cross-)Browser Fingerprinting via OS and Hardware Level Features, February 2017

(4) Analysis of the Effectiveness of Browser Fingerprinting at Large Scale, April 2018

(5) Browser Fingerprinting: A survey, November 2019 and this HN discussion

2 Browser Fingerprinting, Solutions

(1) Modern & flexible browser fingerprinting library fingerprintjs2 on GitHub

(2) UAParser.js - JavaScript library to detect browser, engine, OS, CPU, and device type/model from userAgent string.

3 Browser Fingerprinting, Defence

(1) Read more about Panopticlick's methodology, statistical results, and some defenses against fingerprinting here.

(2) The design and implementation document of the Tor Browser's anti-fingerprinting project

Strategies for Defense: Randomization versus Uniformity

When applying a form of defense to a specific fingerprinting vector or source, there are two general strategies available: either the implementation for all users of a single browser version can be made to behave as uniformly as possible, or the user agent can attempt to randomize its behavior so that each interaction between a user and a site provides a different fingerprint.

Although some research suggests that randomization can be effective, so far striving for uniformity has generally proved to be a better strategy for Tor Browser (see the list of reasons inside the link).

4 Browser Fingerprinting, General

(1) What you need to know about your browser's digital fingerprints, June 2018

Your browser has a fingerprint. It's not as obvious as the real ones on your fingertips, but it exists nonetheless. And advertising networks can use it to track your browsing. That's when a tracker tries to identify your machine based on the details it can glean about it, using seemingly innocuous details like the system fonts you have installed.

If you're interested in seeing what elements of your browser are trackable, head over to a Electronic Frontier Foundation's page called Panopticlick; click "test me," and then check out the results. Click "Show full results for fingerprinting" to see what kind of details the test noticed about your browser. In testing it on Chrome, Safari, and Firefox, I noticed that none of them performed well when it comes to fingerprinting protection.

(2) Everything You Need To Know About Fingerprinting After The Chrome Crackdown, May 2019

JavaScript usually detects device attributes to properly render a webpage or an application so the content looks right, the language is correct or, for example, to ensure that a mobile version of a site loads when it's accessed on a phone.

When a browser loads something, it acts as an agent on behalf of a user – hence the term "user agent" – to retrieve requested content. The browser, as the user agent, sees information about a device and the network it's on, which the developer can use to customize the on-site experience to the visitor's browser.

Every browser has a unique user agent string so that a server can know which browser it's negotiating with to load content.

Browser fingerprinting links device attributes and compresses that information into a hashed ID, usually in the form of a short numerical string.

Advertisers use two types of fingerprinting: a basic version that only uses two fields (IP address and user agent string to create a very rough identifier) and more sophisticated techniques that use JavaScript to read many different settings and configurations.

The latter could include the collection of hundreds of seemingly generic data signals that wouldn't mean much on their own, but together can be used to probabilistically determine identity and create persistent statistical identifiers in the absence of cookies.

In addition to the user agent string, sophisticated browser fingerprinting relies on collating everything from language settings, screen resolution, color depth, time zone, underlying operating system, the OS version and device type to the plug-ins, the type of graphics hardware being used and even whether someone has Do Not Track enabled (not that anyone actually respects it).

Another approach involves a practice called canvas fingerprinting that involves exploiting an element within HTML5 that helps graphics appear on a webpage, and can be used to create a unique fingerprint of visitors. Unlike cookies, this data is not stored locally on a device, so users can't opt out or delete the information.

That lack of transparency about where and how the data is sourced, how it's commingled and how the models are developed is what has made fingerprinting such a black box.

Even though fingerprinting and tracking cookies are in the same boat (persona non grata on the privacy front), some ad tech companies are toiling away on persistent ID solutions that use fingerprinting, a move directly related to the limitations of cookie tracking.

Compared with deterministically matched attribution, fingerprinting is 98% accurate when the fingerprint match is made within the first 10 minutes, which is also when the majority (56%) of attribution occurs. If the attribution window is between 10 minutes and three hours, accuracy drops to 80%. Between three and 24 hours, using fingerprinting logic is a coin flip – only 50% accurate.

When the attribution window is longer than a day, forget about it. Once you get outside of 24 hours, it's more wrong than right. The point being that fingerprinting can be accurate, but only within a narrow timeframe.

Impression tracking, frequency capping, sequential targeting, multitouch attribution – anything that relies on persistent identity – will become more difficult or even impossible, and the open web will probably have to go back to earlier forms of targeting, like contextual, which is happening in any case thanks to stringent privacy regs like GDPR.

(3) 'Fingerprinting' to Track Us Online Is on the Rise. Here's What to Do., July 2019

(4) Think you're anonymous online? A third of popular websites are 'fingerprinting' you., October 2019

5 TV Ads

(1) Hulu Says 70% of Its 82 Million Viewers Are on Ad-Supported Plan, May 2019

Hulu has previously disclosed subscriber numbers – announcing 28 million customer accounts earlier this month – but hasn't broken those out by plan type.

Now Hulu, which in the past month became fully ensconced under Disney's wing, has provided some context around the size of its audience base. Overall, it has 82 million viewers (meaning there's an average of 2.9 viewers per Hulu account). And of those, about 7%, or 58 million, are on the ad-supported plan.

Hulu's ad business is a significant source of revenue, generating almost $1.5 billion in ad revenue in 2018. To that end, Hulu strives to make the way it presents advertising is viewer-friendly – otherwise it risks pushing those subscribers to the zero-advertising tier or losing them altogether.

(2) Can Comcast's Blockgraph Bring Data Matching And Crypto Tech To TV?, February 2020

Last year, Blockgraph ran the first campaign where a marketer's first-party CRM data was directly matched to a TV distributor's targetable data without a third-party onboarder.

Blockgraph's task now is to make that work when multiple potential broadcasters pool their audience data to enable targeting for a single advertiser. Blockgraph is positioned as a data onboarding and clean room technology, more akin to how platforms such as Google and Amazon merge their audience data with marketer segments.

It's a peer-to-peer version of data onboarding. So with a company like LiveRamp, both parties would onboard data and audience files to the third-party company, which sends back synthetic IDs that work amongst themselves.

Blockgraph replicates that, but instead of the data going to a trusted third party, they'd run our software in their systems to anonymize audience records. From there, the data can be matched and shared across our common identity layer.

Another important difference is that we provide household-based identity. Most digital identifiers are focused on being people-based, whereas in TV and streaming video media is still transacted at the household level.

(3) Comcast Doubles Down On Ad-Supported Streaming With Deal For Xumo, February 2020

Lightshed Partners, a media and tech research firm, estimates ad-supported streaming platforms earned about $3 billion in advertising in 2019. Most of that went to Hulu, but it's still a pittance next to linear TV's $70 billion market in the United States.

The category of ad-supported video on demand (AVOD) is growing fast, though. Lightshed said AVOD will add more than $1 billion in new revenues this year and growth will accelerate even more in 2021.

Most smart-TV viewing happens through a cable subscription or an OTT platform like Roku, Apple TV or Amazon's Fire TV. Xumo's pitch to smart TV manufacturers, which are trying to build their own media and advertising businesses, is that people can turn on the TV and the experience isn't so different from linear TV. Viewers can click through Xumo channels using their normal remote (instead of the much-maligned remotes that many Apple or Roku stick users must use), without plugging in OTT hardware or setting up a TV bundle.

(4) Fox Looks to Buy Streaming Service Tubi, February 2020

Entertainment giants have been scooping up ad-supported video platforms to complement subscription offerings. Tubi is an advertiser-supported streaming service that carries reruns of television shows and movies.

Free, ad-supported video platforms have a viable future even as a flurry of new subscription-based streaming services enter the market.

Tubi, Pluto TV, Xumo LLC, Crackle of Chicken Soup for the Soul Entertainment Inc.

Almost all major entertainment companies have launched or are preparing subscription streaming service to compete with Netflix Inc. But several companies also see the need to offer free, ad-supported tiers to reach other consumers.

(5) ViacomCBS's Pluto TV Launches $30 Million Ad Campaign, Touts Enhanced Features, March 2020

Pluto TV, a central plank in ViacomCBS's overall streaming strategy, last year grew active monthly users 75% to 22.4 million in the U.S. The service provides over 250 live streaming channels, organized into a TV-like channel guide.

Other players in the space include Tubi – which Fox Corp. is reportedly circling as an acquisition target – the Roku Channel, Amazon's IMDb TV, and Walmart's Vudu, which NBCUniversal is kicking the tires on. Comcast last week bought free streaming platform Xumo for a reported $100 million-plus, and this summer Comcast and NBCU are planning to launch a free, ad-supported version of the Peacock streaming service nationwide.

6 Differential Privacy

(1) Why Every Ad Tech Company Must Understand Differential Privacy, February 2020

Big tech is all in on differential privacy.

It's a foundational concept within Google's Privacy Sandbox; Apple applies it to the study of diagnostic device, health and web browsing data; and, just last week, Facebook used differential privacy to protect a trove of data it made available to researchers analyzing the effect of sharing misinformation on elections.

Uber employs differential privacy to detect statistical trends in its user base without exposing personal information. Amazon's AI systems tap it to prevent data leakage. Snapchat has used differential privacy to train machine learning models. And Salesforce uses DP filters in its reporting logs.

Differential privacy is a set of cryptographic properties that can be applied to machine learning algorithms in order to set a limit on how much information can be extracted from data before it's possible to draw inferences about individuals. In practice, that means the data owner purposely adds noise or randomness into a data set so that it's simultaneously possible to learn something about a population from the data without identifying any of the individuals included in the group.

Consider a pollster gathering statistical information about embarrassing behavior, like drug use or cheating. To protect their privacy, respondents flip a coin before answering without revealing the result to the pollster. If the coin lands on tails, they are asked to respond truthfully. If it's heads, they flip a second coin and answer "yes" for heads and "no" for tails. This introduces randomness, or plausible deniability, into the eventual outcomes of the study. But because the researcher knows how the errors were introduced, he or she can later work backward to systematically remove them in the aggregate and still glean something useful from the data.

There is no way to know whether an answer is random or not. But because we know the process by which noise is added to the response, it's possible to subtract the noise and learn the average. The caveat is that researchers require larger data sets to study in order to make up for the deliberate randomness.

In the blog post announcing Chrome's intention to deprecate third-party cookies by 2022, Justin Schuh, Chrome's director of engineering, explicitly called out differential privacy as a building block for a future in which ads can be delivered to "large groups of similar people without letting individually identifying data ever leave your browser."

7 Data Privacy and Ethics

(1) Why LiveRamp Quietly Sold Its Location Data Business Last Year, January 2020

LiveRamp is distancing itself from location data, but is keeping the rest of the business, which includes technology to connect first-party web and app data with unique hashed identifiers.

Geolocation data is covered under the California Consumer Privacy Act and under the General Data Protection Regulation in Europe; smartphone-derived location data is the target of scathing news reports; and new iOS features are coming out that aim to cut down on location tracking.

"We sold off our data manufacturing business and focused our strategy on operating a neutral and open marketplace to connect high-quality, ethically-sourced data sources with data buyers," LiveRamp CEO Scott Howe said.

LiveRamp positions itself as a neutral and agnostic party that facilitates identity resolution but doesn't actually own any data assets.

(2) The Untold History of Facebook's Most Controversial Growth Tool, February 2020

"People You May Know" helped the social media giant grow exponentially.

The concept of the Monthly Active User started in 2007. Other internet businesses counted how many people were on the site each day, or how many had signed up in total. But monthly was a better indicator, because someone consistently on the service for a full month was likely there to stay. Thus the number took into account the "churn" – how many people were leaving Facebook. It was proposed to be utterly obsessive about MAUs – to look at every part of Facebook's business in light of this metric, to learn what can drive MAUs, to fix things that don't increase it, and to build new parts of the company to boost MAUs even higher.

The masterpiece of MAU Growth is a feature called People You May Know, referred to internally by the acronym PYMK. Officially launched in 2008, People You May Know is a feature that identifies personally selected prospects for one's friend list. It wasn't a Facebook invention – LinkedIn did it first – but PYMK proved to be one of Growth Circle's most effective tools, and also one of its most controversial ones, a symbol of how the dark art of growth hacking can lead to unexpected consequences.

For many people, PYMK is a welcome feature: a helpful prompt to get in touch with connections who would help them get value from their Facebook experience. But sometimes PYMK can be unsettling, raising questions of what caused those cameo appearances on your News Feed by people whose connection to you was obscure, and sometimes downright unwelcome. A sex worker found Facebook recommending her clients, who did not know her true identity. A sperm donor got a suggestion for the biological child he never met. A psychiatrist learned that Facebook was recommending that some of her patients friend each other on the service – even though the psychiatrist did not friend her patients on Facebook. And millions of people went Ew! as Facebook suggested they develop relationships with friends of their children, spouses of their casual acquaintances, or disastrous blind dates of a decade ago. The story of the woman who got a Facebook suggestion that she friend the mistress of her long-absent father. Another woman was stunned to find that someone on her own PYMK suggestions turned out to be a great-aunt she'd never met.

Dark profiles (data on people not signed up on Facebook) did exist, and the Growth team took advantage of them. Facebook would take out search ads on Google using the names of Facebook holdouts as keywords. The ads would link to those dark profiles of nonusers that supposedly do not exist. "You would search for your own name on the internet and you'd land on a dark profile on Facebook," Chamath Palihapitiya says. "And then you'd be like well, fuck it, you'd fill it in and then PYMK would kick in and we would show you a bunch of your friends."