The Threat to Our Online Anonymity:
How 3rd Parties Are Building Profiles of Consumers

COMP116: Introduction to Computer Security

By: Evgeni Dobranov | December 13th, 2017

ABSTRACT

On the internet, no one knows you're a dog. Or at least this was something we had all taken for granted a long time in the past, resting assured that our online identities remain hidden if we so choose to make them that way. The details of our personal lives that are shared should be wholly at our voluntary discretion and personal judgement, but we're finding that this is not the reality we face today or moving forward as 3rd parties collect our information and build profiles of us from our living habits.

This presents risks for many reasons, namely on the account of inevitable security breaches and leaking of profiles that are built from sensitive information. Moreover, these 3rd parties are discovering new, far more pervasive ways in which to track its users as new technologies come to fruition and become more ubiquitous. Understanding these methods and the forces that drive them are a critical first steps towards recognizing the significance of safeguarding your online identity.

INTRODUCTION

Connectivity with the outside world is becoming more of a right in our lives day by day. It is nearly impossible to function offline in a nation where >95% of those under 50 years of age use the internet.[1] Seeing as it permeates nearly all facets of our lives, we should have control over and be aware of the details surrounding our usage. Yet every time we turn on our devices and use any internet service, the information is unknowingly extrapolated from us is far greater than we realize. As Edward Snowden

[1] Anderson, Monica, and Andrew Perrin. "13% Of Americans Don't Use the Internet. Who Are They?" *Pew Research Center*, 7 Sept. 2016, www.pewresearch.org/fact-tank/2016/09/07/some-americans-dont-use-the-internet-who-are-they/

revealed in an interview post-exile[2], simple Google searches or just having your phone turned on can be used to track and reveal over time qualities that should remain confidential like age, gender, marital status, location and living style. This can be used to build a demographic of exactly who you are, something that Snowden calls "pattern of life."

Barring government entities, many of the other 3rd party entities that provide us their services for free upfront have an obligation to maintain their source of revenue through adverts, but oftentimes gather more information than anyone would be comfortable with along the way. Global digital advertising is a $200 billion industry in 2017[3] and expected to only grow moving forward. With such a lucrative market at hand and web browsers being the primary way we all browse the internet, it should come as no surprise that companies target the intricacies that lie in these pieces of software to benefit their revenue. The umbrella term for this *browser fingerprinting* and is the focus of this paper.

TO THE COMMUNITY

For many, the web browser is the main point of access to the internet. Many of us hop on it within an hour of waking up and use it throughout the day until we go to bed at night. Yet few are aware of the underlying mechanisms that allow device information to be extracted from it, and the subsequent extrapolation of browsing patterns. Factors like screen resolution, canvas fingerprinting, HTTP headers, system locale/time zone, platform and language settings only scrape the surface of what is available to a server whenever you click something on the web. More importantly, front end technologies are practically evolving on a weekly basis as new frameworks and APIs constantly arise. There isn't a feasible limit on the techniques that will develop in the future, which is why remaining conscientious about online presence is far more crucial now than every before.

---

[2] "Inside the Mind of Edward Snowden, part 3" *NBC News*, 29 May 2014, https://www.nbcnews.com/video/inside-the-mind-of-edward-snowden-part-3-269115971954 0:35

[3] "U.S. Digital Advertising Industry - Statistics & Facts." *Statista*, www.statista.com/statistics/456679/digital-advertising-revenue-format-digital-market-outlook-worldwide/

While gathering browsing patterns and analyzing them can assist in thwarting identity theft and falsified click bots, any and all information extracted from a web browser is a reflection of whoever is doing the browsing – click bot or not. When it is you, the people doing the collecting have ownership of your information and can do with it as they see fit, including storing it. You might think "why should I care if I have nothing to hide?" This is generally a deeply complex question with many facets to philosophize about, but the short and defensive answer is that you won't always get to choose where your information goes and how it's interpreted. Aside from having obvious sensitive information fall into the wrong hands (e.g. SSN, credit card numbers), online habits may be used to develop an image of you and permanently logged. Once the information is out there, it's unlikely to every permanently disappear. To paraphrase Supreme Court Justice Stephen Breyer, "the complexity of modern federal criminal law … makes it difficult for anyone to know, in advance, just when a particular set of statements might appear to be relevant to [an] investigation."[4] The information that's collected from you might superficially appear harmless now, but there is potential harm in some undetermined future.

STATIC BROWSER FINGERPRINTING OVERVIEW

As mentioned above, browser fingerprinting largely pertains to the information that a client browser leaks to a server in exchanging resources. This includes the User-Agent request header (which directly gives off the application type, OS and software versions of the agent making the request), browser locale, language and time zone, installed fonts, screen resolution and browser plugins. These are largely harmless when treated independently and are often used to enhance a user's experience on a site, such as when Google defaults to the country version depending on where your IP address is located.

But when treated in conjunction with one another, all of these elements can be combined to create an identification vector of the person behind they keyboard. Of course, it is entirely possible to

[4] Marlinspike, Moxie. "Why 'I Have Nothing to Hide' Is the Wrong Way to Think About Surveillance." *Wired*, Conde Nast, 6 June 2013, www.wired.com/2013/06/why-i-have-nothing-to-hide-is-the-wrong-way-to-think-about-surveillance/

have multiple web browsers with the same values for all of those categories, but the probability of such an event is overwhelmingly small. Based on Peter Eckersley's 2013 DEF CON presentation[5] on browser uniqueness, 94.2% of "typical desktop browsers" were found to be unique in one such experiment. Many other experiments lead to >90% uniqueness results too. On a site that receives a medium amount of traffic, the chances of seeing two browsers with perfectly matching identification hashes is near zero.

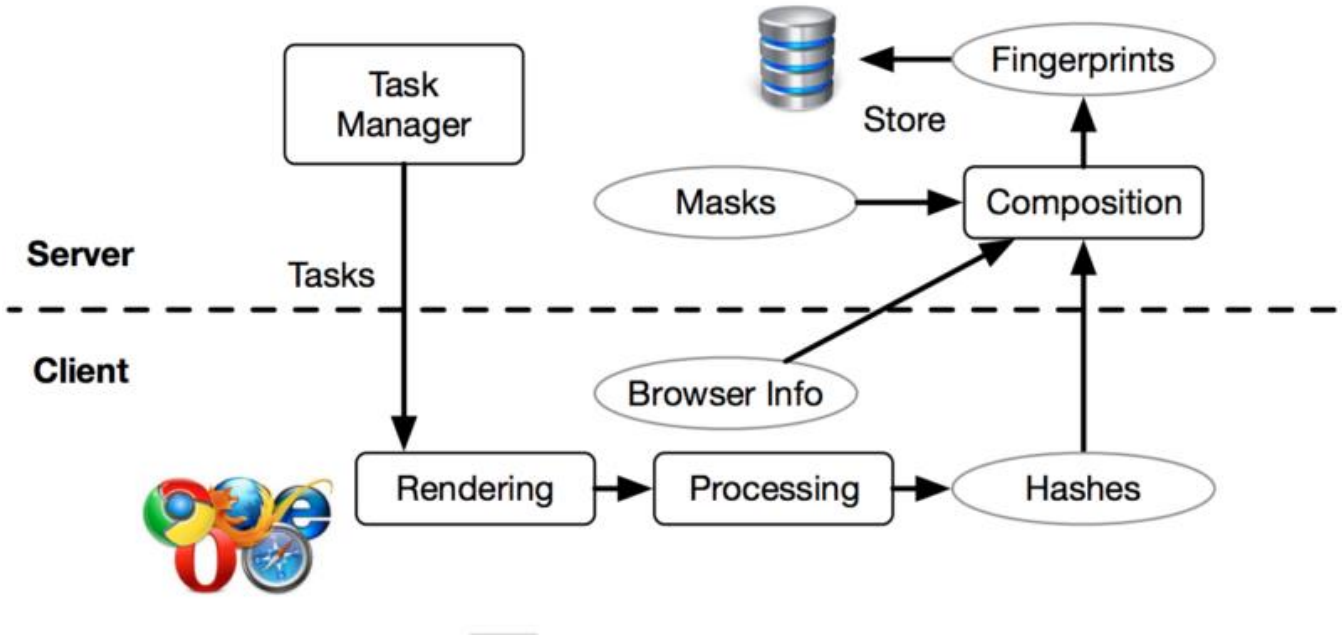
THE DYNAMIC SIDE OF FINGERPRINTING

While the methods may seem to end there, they actually only depict the static side of browser fingerprinting – all of the information gathered there is simply stored on the client end and directly extracted by the server side. There is a dynamic facet to fingerprinting that uses computational heuristics in extrapolating even more data on who the client is, and further lowering the probability that two browsers overlap in any categories. This comes in the form of canvas fingerprinting and benchmark tests, which run server-specified code on the client's platform and targets a computer component (e.g. CPU cores, GPU), measures performance and time, and incorporates these numbers into the overall fingerprint. This can be anything from a simple JavaScript code snippet to a more in-depth 3D WebGL operation that spans more resources.

A key concept to notice here is that this allows a server to discern a user regardless of the web browser they are using, resulting in something called cross-browser fingerprinting. The heuristics gathered transcend the web browser's configuration settings and now dig into the hardware level of a machine, effectively exploring the depths of the OSI model. Thus, switching from Chrome to Firefox to fool a website might not work if it employs these techniques in its client-browsing experience. Like with the static methods and different browsers, it's unlikely that two machines will have the exact same

[5] Eckersley, Peter. "How Unique is Your Browser?" *DEF CON*, August 2013, https://www.defcon.org/images/defcon-18/dc-18-presentations/Eckersley/DEFCON-18-Eckersley-Panopticlick.pdf | https://www.youtube.com/watch?v=6gM9r8jNMFU

hardware specifications and perform identically on a computational task. A simplified diagram from a Lehigh University research paper[6] by lead researcher Yinzhi Cao is shown below. It demonstrates the dispatching flow of these types of tasks with a browser and incorporating that into a fingerprint along with the browser info would work:



To get the most accuracy out of a web browser, the server would typically append the browser's static information to the heuristics of the rendering task, yielding an extremely precise and practically unique identification hash of the user.

AN INTERESTING AND UNFORESEEN EXAMPLE

The Battery Status API was a web standard integrated into most web browsers. It came out approximately five years ago and was supported until recently when it was marked as obsolete in 2015. On some web browsers, the API could still retrieve a good amount of information on a device's power

[6] Cao, Yinzhi. Li, Song. Erik, Wijmans. "(Cross-) Browser Fingerprinting via OS and Hardware Level Features" February 2017, http://yinzhicao.org/TrackingFree/crossbrowsertracking_NDSS17.pdf

status, such as charging status, battery percentage level, and time left until the battery died. It was designed with the intention of giving developers a tool to intelligently choose whether or not to provide a more minimalistic version of their website to users whose phones and laptops were low on battery. Lo and behold, developers interested in tracking began figuring out how to use this as an invariant when users hopped across web pages, making it part of the browser fingerprinting ensemble.[7]

There are many factors might give each device a unique ratio between these numbers, such as the device's power usage, battery size and battery health. According to the article cited above, the number of unique combinations is on the order of 14 million, easily adding another massive factor into the already large number of combinations from the categories discussed above. There is a small HTML site accompanying this paper to show how simple it is to use this API in just a few dozen lines of code.

It takes no stretch of the imagination to know that mobile applications have similar capabilities. They can use this information to influence users' actions and benefit themselves as well. If the next update of the mobile Uber application took into account your battery percentage and raised the rates by 50% when it saw your battery was below 5%, you wouldn't have much of a choice but request a ride if you wanted to get home. Web sites can twist this information just as easily in very similar ways.

THE PROTECTIVE MEASURES

With all of this said, there are some measures you can take to protect yourself from leaking away information. According to Panopticlick[8], there are a few predominant ways to do this, but each comes with some tradeoffs. By far the most effective method is using the Tor browser. While its architecture is worth a whole separate paper in itself, the gist is that it decentralizes the browsing experience into nodes between users where each is granted a degree of anonymity. It has an inherent anti-browser

[7] Hern, Alex. "Your Battery Status Is Being Used to Track You Online." *The Guardian*, Guardian News and Media, 2 Aug. 2016, www.theguardian.com/technology/2016/aug/02/battery-status-indicators-tracking-online
[8] https://panopticlick.eff.org/about#browser-fingerprinting

fingerprinting nature built into its design and will refrain from sharing many of the headers or fonts that a user might have installed. The downside is that the decentralization makes it slower than a regular web browser, but that is the price to pay for privacy.

Of course, disabling JavaScript and re-enabling it manually on a site-by-site basis is an option, but that can be tedious even with extensions like NoScript[9] and might break the functionality of many sites. Using multiple computers is also technically an option, but also largely infeasible for the average consumer. Rather, browsing through a virtual machine may offer a good temporary choice if there is a site that should be browsed with a fresh environment.

CONCLUSION

In this paper, we've covered the most pervasive methods out there being used to track people through their browsing habits. We've seen just how much information browsers can communicate over the web to servers on behalf of their users, ranging from browser and OS versions to hardware specifications to battery percentages. While these are largely intended to be used to enhance the users' experiences, there is little stopping developers from aggregating all of this information into a profile for each user. Historically we've seen instances of those with too much information being present for one's good, and how leaking it to the wrong people could have potentially disastrous results.

These 3rd party entities need to start becoming more responsible, transparent and honest with exactly what information they are collecting from users and how it is being used. Employing sneaky JavaScript tricks behind users' backs is often perceived as a violation of trust and demonstrates poor ethical code. Until these 3rd parties become more upfront with the browsing experience they are offering, users must remain aware and vigilant in their efforts to protect their online identities for both their present and future selves.

[9] https://noscript.net/

REFERENCES

[1] Anderson, Monica, and Andrew Perrin. "13% Of Americans Don't Use the Internet. Who Are They?" *Pew Research Center*, 7 Sept. 2016, www.pewresearch.org/fact-tank/2016/09/07/some-americans-dont-use-the-internet-who-are-they/

[2] "Inside the Mind of Edward Snowden, part 3" *NBC News*, 29 May 2014, https://www.nbcnews.com/video/inside-the-mind-of-edward-snowden-part-3-269115971954 0:35

[3] "U.S. Digital Advertising Industry - Statistics & Facts." *Statista*, www.statista.com/statistics/456679/digital-advertising-revenue-format-digital-market-outlook-worldwide/

[4] Marlinspike, Moxie. "Why 'I Have Nothing to Hide' Is the Wrong Way to Think About Surveillance." *Wired*, Conde Nast, 6 June 2013, www.wired.com/2013/06/why-i-have-nothing-to-hide-is-the-wrong-way-to-think-about-surveillance/

[5] Eckersley, Peter. "How Unique is Your Browser?" *DEF CON*, August 2013, https://www.defcon.org/images/defcon-18/dc-18-presentations/Eckersley/DEFCON-18-Eckersley-Panopticlick.pdf | https://www.youtube.com/watch?v=6gM9r8jNMFU

[6] Cao, Yinzhi. Li, Song. Erik, Wijmans. "(Cross-) Browser Fingerprinting via OS and Hardware Level Features" February 2017, http://yinzhicao.org/TrackingFree/crossbrowsertracking_NDSS17.pdf

[7] Hern, Alex. "Your Battery Status Is Being Used to Track You Online." *The Guardian*, Guardian News and Media, 2 Aug. 2016, www.theguardian.com/technology/2016/aug/02/battery-status-indicators-tracking-online

[8] https://panopticlick.eff.org/about#browser-fingerprinting

[9] https://noscript.net/