Excavating Web Trackers Using Web Archaeology

ADA LERNER, ANNA KORNFELD SIMPSON, TADAYOSHI KOHNO, AND FRANZISKA ROESNER

Ada Lerner is a PhD candidate in the Department of Computer Science & Engineering at the University of Washington. They received a BA from Amherst College. Their research is broadly within computer security and privacy, including the intersection of technical, social, and legal concerns. lerner@cs.washington.edu

Anna Kornfeld Simpson is a third-year PhD student at the University of Washington. She is an NSF Graduate Fellow and earned a BSE in computer science from Princeton University in 2014. She is broadly interested in security, privacy, and building secure systems informed by tech policy. aksimpso@cs.washington.edu

Tadayoshi Kohno is the Short-Dooley Professor in the Department of Computer Science & Engineering at the University of Washington. He received his PhD from the University of California San Diego and his BS from the University of Colorado. His research focuses on computer security, broadly defined. yoshi@cs.washington.edu

Franziska Roesner is an Assistant Professor in the Department of Computer Science & Engineering at the University of Washington. She received her PhD from the University of Washington and her BS from The University of Texas at Austin. Her research focuses broadly on topics in computer security and privacy, with a particular interest in understanding and improving security and privacy for end users of existing and emerging technologies. franzi@cs.washington.edu

Third-party Web tracking has recently become a frequent source of privacy concerns among researchers, policymakers, and the public. But how long has tracking been a part of the Web, and how has it changed over the years? These questions led us to build a tool, Tracking Excavator, which time travels using the Wayback Machine's archive of the Web. We were able to collect data on Web tracking over nearly the whole history of the Web, back to 1996, showing that archive-based measurements of the history of the Web are not only possible but are a powerful technique for examining Web tracking and other trends in the Web over its history.

A common problem we face as security and privacy researchers is our recurring need for time travel. Since security is hard to retrofit, we want to know what technologies and threats will be important in the future, so we can start studying and securing things now. This is particularly true for the Web, which changes very rapidly. Time travel into the future would be valuable, but at this time, the only technique we have is to wait.

We also sometimes wish to time travel backwards. Researchers like to measure the trends of how a technology or attack rises to prominence, but by the time one is ubiquitous, rapid changes in the Web have swept away the evidence, making it too late to measure how it became so popular. In this article we'll tell you about a paper we wrote when we found out how to time travel into the past of the Web, and the measurements we made using this technique of third-party Web tracking over the past 20 years [1].

Why Web Tracking?

Third parties are domains that appear on multiple different Web sites in order to provide valuable services, such as analytics, content delivery, social media integrations, and advertising. Many third parties track Web site visitors across the sites they visit, building up a profile of the Web sites they visit. This third-party Web tracking has become a major feature of the Web's economy: tracking underlies targeted advertising, making it a linchpin of Web sites funded by advertising revenue and enabling many of the services we all use so often. However, its importance and ubiquity has made some ask what privacy we're giving up in return, and news media, scientific researchers, policymakers, and the public as a whole have taken interest. The earliest measurements of third party, cookie-based tracking we are aware of came from the FTC in 2000 [2], while academics seem to have first begun to publish on the topic in 2009 [3].

Diverse people use the Web in many sectors of their lives, personal and professional. Recreation and commerce are common, but the Web also influences and enables sensitive activities, both intimate and public. On the intimate side, people explore their gender and sexuality, study religious and spiritual beliefs, and research their physical and mental health. Meanwhile, on the public side, the Web is important to our society, our politics, and our democracy, to the way people consume news, debate politics, and influence policymakers by public discourse and public comment. Engaging with these topics on the Web can provide great

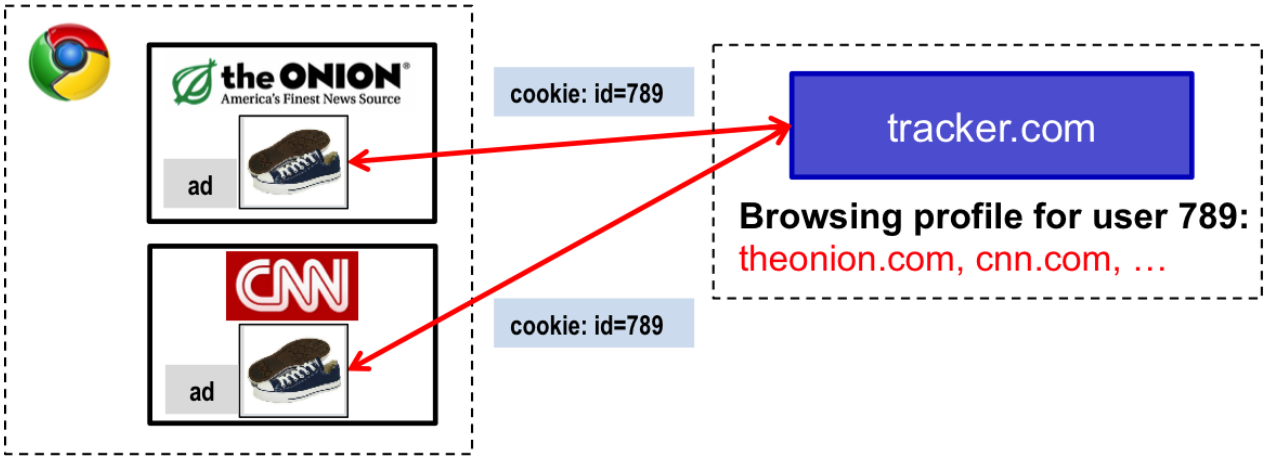

Figure 1 depicts a hypothetical tracker, tracker.com, which appears as a third party on both theonion.com and cnn.com. A cookie set by tracker.com when you visit The Onion is later sent to tracker.com when you visit CNN, allowing tracker.com to know which sites you visit.

benefits to individuals and society, but tracking people while they do so creates the possibility for privacy violations, chilling effects, and harm to those among us who may be most vulnerable. To balance these concerns, we need to understand how tracking happens so we can make informed choices.

Thinking of these issues, privacy researchers have been studying Web tracking for some time. Unfortunately, these measurements often use different methodologies, making them hard to compare or aggregate longitudinally. We wanted a longitudinal study of tracking in order to understand its trends and how it has responded to changes in the Web, such as in the way browsers are designed and in the technologies commonly used. To build this longitudinal picture, we needed a way to go back in time. Fortunately, the Wayback Machine (https://archive.org/web/) exists. It is a publicly available Web archive, an extensive collection of Web-page snapshots from the past 20 years, reaching back to 1996. The Wayback Machine serves the archived pages just like live Web sites; as an example, check out the *;login:* home page from 1997 at https://web.archive.org/web/19970606050039/http://www.usenix.org/publications/login/index.html.

The Wayback Machine contains both the contents and formatting of Web pages, as well as the JavaScript code and HTTP headers that were sent along with the page. This means that we can study not only the content of pages—the words and images—but also how pages were constructed technically and how they behaved dynamically, including the ways they may have used cookies and tracked visitors. Seeing this, we built a tool, Tracking Excavator, that uses the Wayback Machine to travel back in time and show us how Web tracking has evolved since its early days.

How Tracking Works

We measured third-party Web tracking over the past 20 years, focusing on cookie-based tracking. Cookie-based tracking occurs when a third party on the Web, such as an advertiser or social network provider, labels your browser with a unique identifier by asking your browser to store that identifier in a cookie. The third party then uses the identifier to recognize you in future, allowing it to build up a list of the places you go on the Web. We refer to this list of places you go as the browsing profile that a tracker learns about the people it tracks.

How does cookie-based tracking work technically? The following process is depicted in Figure 1. When you go to a Web site by typing its name (e.g., example.com) or clicking a link, we call that Web site a first party. First parties are the Web sites you visit intentionally, and you often have some direct relationship with them: you may purchase products there or sign up for an account. However, most first-party sites today also include one or more third parties. These third parties host some of the Web site's parts, such as scripts, images, style information (.css files), social media integrations, and advertisements. Your browser automatically contacts these third parties as part of its normal process of loading the first-party Web site. For example, when you visit example.com, your browser may load an advertisement from advertisements.com. Third parties can appear on many different first-party Web sites, so advertisements.com may also provide advertisements on other domains, like example2.com and example3.com. By appearing in these separate contexts, a third party may be able to track you across those sites, building up a browsing profile that describes where you've been on the Web.

When your browser makes requests to a particular third-party tracker for the first time, the tracker sends your browser a cookie that contains a unique identifier, to label your browser. Your browser stores cookies in a file and recalls these cookies,

sending them whenever it makes future requests to that same third party, allowing the third party to recognize you when it serves your browser again. Since these requests to Web servers typically include information about the first-party site you're visiting, the third party learns about the sites you visit, and can associate that browsing profile with the unique identifier in your cookie. It's important to remember that they associate the browsing profile only with a unique identifier—often trackers don't know your name or who you are in real life, and are limited only to a (sometimes quite detailed!) picture of your interests and activities.

Tracking in All Its Glory

Some trackers are more complex or sophisticated than the simple description above. That said, when someone mentions "Web tracking," you should probably think of what we've described here. Our study shows that "vanilla" tracking—our name for the simple form described above—has been and still remains the most common type of cookie-based tracking. We studied more complicated types of tracking as well. For example, *referred trackers* are those trackers that share and exchange identifiers with one another, expanding their power to follow you to sites with different third parties. And we call trackers that you also sometimes visit as a first party, such as social network sites, *social trackers*, since your first-party relationship with them might allow them to link your browsing profile to your real life identity.

We classified the trackers we studied using a taxonomy we developed in earlier work [4]. That taxonomy classifies trackers by the way they behave and how those behaviors allow them to track people. The taxonomy we use also separates trackers according to the amount of extra information they have about the users they track (e.g., social trackers with whom users may have an account, such as facebook.com or google.com) and according to whether they share information with other trackers (referred trackers).

How to Time Travel (for Science!)

In this work, we developed archive-based, time-travel-capable measurement tools and used them to study the prevalence of Web tracking over the past 20 years. Our tool—Tracking Excavator—automatically browses the Wayback Machine's archive of the Web, collecting data about how the Web and its trackers used to behave. We analyzed this time-travel data to draw a picture of the history of tracking. In the future, our tool and our analysis techniques will allow us and other researchers to ask many other questions about topics, from security and privacy to software engineering and performance across the whole history of the Web. For example, recent work in 2014 and 2016 has provided two datapoints showing a downward trend of browser fingerprinting—how does that trend look over the entire lifespan of browser fingerprinting [5, 6]?

Figure 2 shows an example of missing resources in the archive. Here images are missing from a 2003 snapshot of the University of Washington's CSE homepage (https://web.archive.org/web/20031001160742/http://www.cs.washington.edu/).

Getting here wasn't easy. We analyzed and quantified the limitations of the Wayback Machine and developed techniques for performing retrospective measurements in the presence of these limitations so that our measurements accurately reflected the Web of the past. The Wayback Machine is extensive, containing over 10 petabytes of Web site snapshots, but the world isn't perfect, and archives are no exception. The Wayback Machine is sometimes incomplete (as in Figure 2), with missing Web site data and cookies. Other times, the Wayback Machine makes mistakes that cause Web site snapshots to be *inconsistent*. Rather than reflecting the way a page looked at a single point in the past, it sometimes mashes up anachronistic resources from different points in time.

Other studies have relied on the Wayback Machine (for example, to predict whether Web sites will become malicious [7] and to study JavaScript inclusion [8]) and noted similar limitations. One of our goals was to systematically evaluate and develop mitigations for these limitations to enable future studies.

Given that our archival data is sometimes incomplete and sometimes anachronistic, we put a lot of work into ensuring that our results accurately reflected tracking that really happened in the past.

We crawled many archival Web pages to learn how and when these types of errors occur to ensure that our results accurately reflected the way tracking happened in the past. First, we measured the ways the Wayback Machine can have missing or anachronistic data, quantified them, and reasoned through the effects those errors would have on our analysis. Then we incorporated that understanding into Tracking Excavator and into our data analysis to winnow our observations to only what really existed. Finally, we think about the limitations of our data whenever we share or interpret our results. For example, since we must ignore many anachronistic requests, our measurements

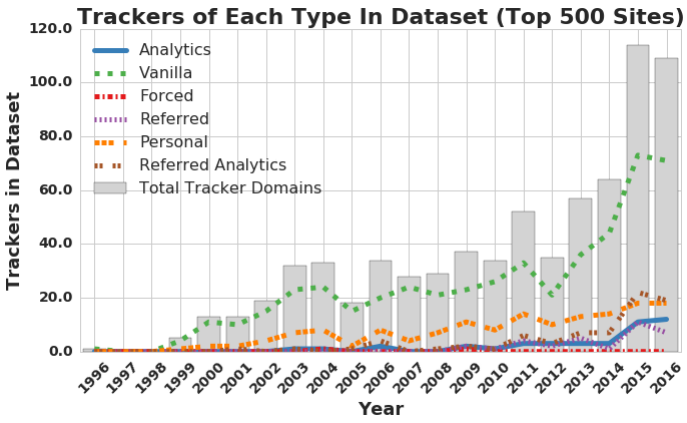

Figure 3: Number of tracking domains (gray bars) present on the 500 most popular sites of each year. The various line-styles represent different cookie-based tracking behaviors in our taxonomy; a single tracking domain may have multiple behaviors.

generally underestimate the amount of tracking that really happened.

We knew that our results undercounted the number and power of trackers, but by how much? Fortunately, we have been studying Web tracking for a few years, and we have data from the past five years. We used the same methods and taxonomy on live Web pages to measure Web tracking in 2011, 2013, 2015, and 2016. By comparing these past, ground-truth measurements to our archival results for the same periods, we were able to understand how our archival trends reflected real ones. The relationship between archival trackers and real trackers turns out to be quite consistent—see our paper for details of this validation.

Digging through the limitations of the Wayback Machine didn't just allow us to understand Web tracking. We hope that understanding and mitigating these limitations will allow other researchers to use these techniques in the future. We think that archive-based retrospective measurements are a powerful tool, and we hope our work will aid and inspire others to try measuring all sorts of aspects of security and privacy and other technical topics on the Web.

Our Results

Many people would believe intuitively that "Web tracking has increased over time" based on their experience of browsing the Web and reading the news. We shared this intuition too, but we found it valuable and informative to confirm this intuition scientifically. Quantifying this increase is a contribution to our understanding and debate over tracking. Using Tracking Excavator, we unveiled trends in the number of trackers, the extent of their coverage, and the prevalence and evolution of different tracking techniques. We discovered that the number of tracker domains present on the 500 most popular sites of each year has been increasing, that these domains are demonstrating increasingly complex tracking behaviors, and that the most prevalent domains are achieving greater coverage of popular Web sites and so can build a greater profile of users' browsing activity. We used a separate set of 500 popular sites for each year we studied, and those input datasets and our analysis code are available at our Web site, trackingexcavator.cs.washington.edu.

Thinking back to our challenges, we recall that the numbers which follow are generally undercounts, due to missing and inconsistent data in the archive. Additionally, we point out that our method measures only client-side, cookie-based behaviors that would enable tracking, but we cannot tell whether domains actually track users based on their cookies. We also don't look for alternative methods tracking, such as browser fingerprinting [9].

Quantifying Overall Tracking Behavior

The gray bars in Figure 3 show the overall number of tracker domains we saw for each year, while the different lines break those trackers down into the types of our taxonomy. Recall that these numbers were measured on the 500 Web sites that were most popular in that year. A single tracking domain can display multiple behaviors. The steadily increasing trend demonstrates that new players have continued to enter the game of tracking, developing profiles of users' browsing. Additionally, tracking domains are starting to use more complex techniques, such as sharing their data with additional third parties (as "referred" or "referred analytics" trackers) that allow further collection and dissemination of users' browsing histories. Refer to our paper [1] for more details about these types of trackers.

Figure 3 also shows an increase in personal trackers. These are tracking domains that users browse in their own right, as first parties, and that may have collected additional profiling information about a user. Some of the most popular sites on the Web, such as Facebook and Twitter, are included in this category. Personal trackers often appear today as social media buttons, such as the Facebook "Like" button or the Twitter "Tweet" button. Social trackers are particularly powerful because in addition to the browsing profile they build of you, they may also know a great deal about you through the profile, posts, and friendships you maintain on their social media site. By connecting their pseudonymous profile of browsing history to a social media profile, potentially across multiple machines, social trackers may be able to build deeper, longer lived, and more personal profiles of people.

The increase in the number of tracking domains corresponds with a general increase in the inclusion of third-party content over time. Figure 4 shows the distribution of the number of third parties on Web sites. Each line represents the data from one year, and the farther out from the axes the line lies, the larger the fraction of sites (y-axis) that had a larger number of third parties (x-axis). Although this distribution includes third parties such as content delivery networks (CDNs) that do not necessarily

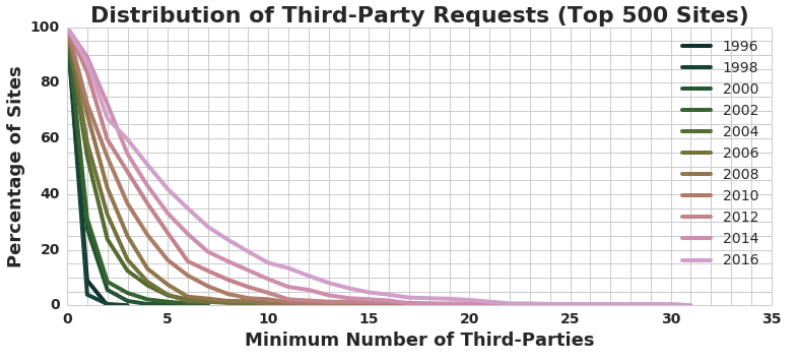

Figure 4: Distribution of included third parties for the top 500 most popular sites of each year. In 1996, less than 5% of sites included content from one or more third parties, while in 2016, 90% of sites included content from at least one third party, and 50% of sites included content from at least four third parties.

include any tracking behavior, it does capture sites whose cookies were not properly archived by the Wayback Machine that are missing from Figure 3. Of the 500 most popular sites in 1996, less than 5% of them included content from one or more third parties. By 2006, more than 50% included content from at least one third party, but few included content from more than five third parties. In archival data from 2016, 90% of the most popular sites include content from at least one third party, 50% of sites include content from at least four third parties, and 15% of sites include content from 10 or more third-party domains! Much as we found it valuable to confirm and quantify our intuitions about tracking, we also find it useful to quantify the increasing complexity of Web sites and the number of third parties that have the opportunity to observe people's browsing.

Quantifying the Most Powerful Trackers

While the number of trackers has grown over time, the above results don't tell us how large of a browsing profile each of those trackers can build. A third-party tracker can only track people who visit the sites where it appears, so we may be less concerned about the privacy implications of a large number of trackers, each of which appear on a small number of sites, and more concerned with a single tracker that appears on many sites. Therefore, we also examined the most prevalent trackers: the ones that show up on the greatest number of sites in the top 500. Figure 5 shows the most and second-most prevalent third parties each year. In the first decade of tracking that we measure, no tracker could directly track a person across more than 10% of the most popular sites. However, their reach increases in the second decade of our measurements: individual companies now have the ability to build significantly larger profiles of our browsing history.

One domain stands out in Figure 5: google-analytics.com (represented by the line at the top with stars through it) appears on nearly 200 of the top 500 most popular sites in 2011, and approximately a third of the sites in the years following.

We note that Google Analytics is designed as an on-site analytics script rather than a cross-site tracker, which means that its primary purpose is to provide analytics for a single domain rather than build cross-site browsing profiles. However, we observe that its high prevalence on popular sites gives it a large amount of power: its choices—and its transparency about its tracking policies and data sharing—can have a large effect on user privacy.

Changing Tracking Behavior

One encouraging anecdote comes from the mid-2000s. The archives of the early 2000s featured a number of pop-up advertisements, which we captured since our browser running Tracking Excavator allowed pop-ups. Pop-ups have a place in the Web tracking discussion because the browser treats the popped-up site as a first party (a site the user chose to go to) rather than a third party for the purposes of setting cookies. Therefore, tracking defenses that block third-party cookies are not effective against third-party pop-ups. From 2000–2004, we saw a significant increase in third-party pop-ups, but in 2005 the number of third-party pop-ups dropped dramatically. Some digging revealed that this change coincided with the decision by the developers of the Internet Explorer browser in 2004 (soon followed by other browsers) to block pop-ups by default. While it is likely that the browser manufacturers made this decision to improve user experience on the Web rather than for reasons related to tracking, it nevertheless also had positive effects for defending against tracking, because sites could not so easily evade third-party cookie blocking. As a result, the trackers were forced to implement other, more complex, techniques that we see later in the 2000s and 2010s.

Conclusion

In this study, we showed that it's possible to gather longitudinal data from the Wayback Machine in order to measure third-party Web tracking over nearly the whole history of the Web. In order to do so, we quantified and evaluated the challenges of using archival data in measurements, and developed techniques for mitigating those challenges, incorporating those techniques into our tool for retrospective measurements, Tracking Excavator. These techniques and this tool are not specific to Web tracking, and we hope that we've made a new type of measurement possible for others, enabling and inspiring them to go out and measure all kinds of properties of the Web retrospectively. If we're lucky, our minor form of Web-based time travel can support us in gathering the data we need as technologists, policymakers, and the public to make decisions about the way we can best make our technologies work for us while preserving our security and privacy.

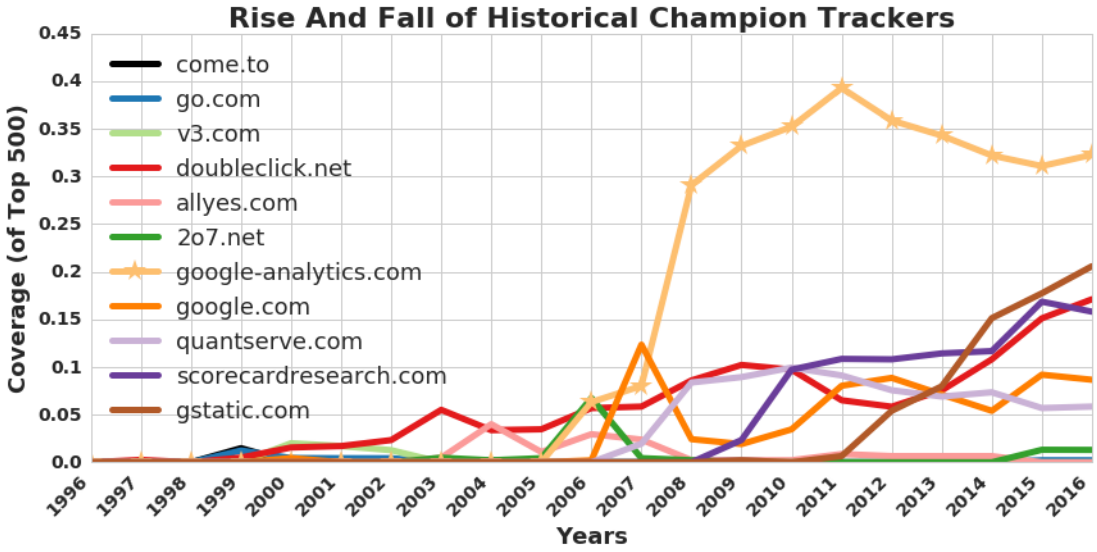

Figure 5: The third parties with the greatest or second-greatest presence on the top 500 most popular sites of each year. We refer to the number of sites a tracker appears on as its "coverage." Before 2007, we measured no domain on more than 10% of popular sites; now, several third parties appear on nearly 20% of sites.

References

[1] A. Lerner, A. Kornfeld Simpson, T. Kohno, and F. Roesner, "Internet Jones and the Raiders of the Lost Trackers: An Archaeological Study of Web Tracking from 1996 to 2016," in *Proceedings of the 25th USENIX Security Symposium (USENIX Security '16)*: https://trackingexcavator.cs.washington.edu/InternetJonesAndTheRaidersOfTheLostTrackers.pdf.

[2] Federal Trade Commission, "Privacy Online: Fair Information Practices in the Electronic Marketplace: A Report to Congress," May 2000: http://www.ftc.gov/reports/privacy2000/privacy2000.pdf.

[3] B. Krishnamurthy and C. Wills, "Privacy Diffusion on the Web: A Longitudinal Perspective," in *Proceedings of the 18th International Conference on World Wide Web (WWW '09)*: http://web.cs.wpi.edu/~cew/papers/www09.pdf.

[4] F. Roesner, T. Kohno, and D. Wetheral, "Detecting and Defending Against Third-Party Tracking on the Web," in *Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation (NSDI '12)*: http://www.franziroesner.com/pdf/webtracking-NSDI2012.pdf.

[5] S. Englehardt and A. Narayanan, "Online Tracking: A 1-Million-Site Measurement and Analysis," in *Proceedings of the 23rd ACM Conference on Computer and Communications Security (ACM CCS 2016)*: http://randomwalker.info/publications/OpenWPM_1_million_site_tracking_measurement.pdf.

[6] G. Acar, C. Eubank, S. Englehardt, M. Juarez, A. Narayanan, and C. Diaz, "The Web Never Forgets: Persistent Tracking Mechanisms in the Wild," in *Proceedings of the 21st ACM Conference on Computer and Communications Security (ACM CCS 2014)*: https://securehomes.esat.kuleuven.be/~gacar/persistent/the_web_never_forgets.pdf.

[7] K. Soska and N. Christin, "Automatically Detecting Vulnerable Websites Before They Turn Malicious," in *Proceedings of the 23rd USENIX Security Symposium (USENIX Security '14)*: https://www.usenix.org/system/files/conference/usenixsecurity14/sec14-paper-soska.pdf.

[8] N. Nikiforakis, L. Invernizzi, A. Kapravelos, S. Van Acker, W. Joosen, C. Kruegel, F. Piessens, and G. Vigna. "You Are What You Include: Large-Scale Evaluation of Remote JavaScript Inclusions," in *Proceedings of the 19th ACM Conference on Computer and Communications Security (ACM CCS 2012)*: https://seclab.cs.ucsb.edu/media/uploads/papers/jsinclusions.pdf.

[9] N. Nikiforakis, A. Kapravelos, W. Joosen, C. Kruegel, F. Piessens, and G. Vigna, "Cookieless Monster: Exploring the Ecosystem of Web-Based Device Fingerprinting," in *Proceedings of the IEEE Symposium on Security and Privacy (2013)*: https://seclab.cs.ucsb.edu/media/uploads/papers/sp2013_cookieless.pdf.