This paper appears at the 25th USENIX Security Symposium (2016).

Internet Jones and the Raiders of the Lost Trackers: An Archaeological Study of Web Tracking from 1996 to 2016

Ada Lerner*, Anna Kornfeld Simpson*, Tadayoshi Kohno, Franziska Roesner
University of Washington
{lerner,aksimpso,yoshi,franzi}@cs.washington.edu

Abstract

Though web tracking and its privacy implications have received much attention in recent years, that attention has come relatively recently in the history of the web and lacks full historical context. In this paper, we present longitudinal measurements of third-party web tracking behaviors from 1996 to present (2016). Our tool, TrackingExcavator, leverages a key insight: that the Internet Archive's Wayback Machine opens the possibility for a retrospective analysis of tracking over time. We contribute an evaluation of the Wayback Machine's view of past third-party requests, which we find is imperfect — we evaluate its limitations and unearth lessons and strategies for overcoming them. Applying these strategies in our measurements, we discover (among other findings) that third-party tracking on the web has increased in prevalence and complexity since the first third-party tracker that we observe in 1996, and we see the spread of the most popular trackers to an increasing percentage of the most popular sites on the web. We argue that an understanding of the ecosystem's historical trends — which we provide for the first time at this scale in our work — is important to any technical and policy discussions surrounding tracking.

1 Introduction

Third-party web tracking is the practice by which third parties like advertisers, social media widgets, and website analytics engines — embedded in the first party sites that users visit directly — re-identify users across domains as they browse the web. Web tracking, and the associated privacy concerns from tracking companies building a list of sites users have browsed to, has inspired a significant and growing body of academic work in the computer security and privacy community, attempting to understand, measure, and defend against such tracking (e.g., [3, 4, 6, 8, 14, 15, 18–20, 22, 24, 25, 27–30, 32–34, 37, 39–43, 45, 46, 51, 57, 60, 61, 64–66, 70, 71]).

*Co-first authors listed in alphabetical order.

However, the research community's interest in web tracking comes relatively recently in the history of web. To our knowledge, the earliest measurement studies began in 2005 [42], with most coming after 2009 — while display advertising and the HTTP cookie standard date to the mid-1990s [44, 48]. Though numerous studies have now been done, they typically consist of short-term measurements of specific tracking techniques. We argue that public and private discussions surrounding web tracking — happening in technical, legal, and policy arenas (e.g., [49, 72]) — ought to be informed not just by a single snapshot of the web tracking ecosystem but by a comprehensive knowledge of its trajectory over time. We provide such a comprehensive view in this paper, conducting a measurement study of third-party web tracking across 20 years since 1996.

Measurement studies of web tracking are critical to provide transparency for users, technologists, policymakers, and even those sites that include trackers, to help them understand how user data is collected and used, to enable informed decisions about privacy, and to incentivize companies to consider privacy. However, the web tracking ecosystem is continuously evolving, and others have shown that web privacy studies at a single point in time may only temporarily reduce the use of specific controversial tracking techniques [63]. While one can study tracking longitudinally starting in the present, as we and others have (e.g., [42, 63]), ideally any future developments in the web tracking ecosystem can be contextualized in a comprehensive view of that ecosystem over time — i.e., since the very earliest instance of tracking on the web. We provide that longitudinal, historical context in this paper, asking: how has the third-party web tracking ecosystem evolved since its beginnings?

To answer this question, we apply a key insight: the Internet Archive's *Wayback Machine* [31] enables a retrospective analysis of third-party tracking on the web

over time. The Wayback Machine[1] contains archives of full webpages, including JavaScript, stylesheets, and embedded resources, dating back to 1996. To leverage this archive, we design and implement a retrospective tracking detection and analysis platform called *TrackingExcavator* (Section 3), which allows us to conduct a longitudinal study of third-party tracking from 1996 to present (2016). TrackingExcavator logs in-browser behaviors related to web tracking, including: third-party requests, cookies attached to requests, cookies programmatically set by JavaScript, and the use of other relevant JavaScript APIs (e.g., HTML5 LocalStorage and APIs used in browser fingerprinting [15, 57], such as enumerating installed plugins). TrackingExcavator can run on both live as well as archived versions of websites.

Harnessing the power of the Wayback Machine for our analysis turns out to be surprisingly challenging (Section 4). Indeed, a key contribution of this paper is our evaluation of the historical data provided by the Wayback Machine, and a set of lessons and techniques for extracting information about trends in third-party content over time. Through a comparison with ground truth datasets collected in 2011 (provided to us by the authors of [60]), 2013, 2015, and 2016, we find that the Wayback Machine's view of the past, as it relates to included third-party content, is imperfect for many reasons, including sites that were not archived due to `robots.txt` restrictions (which are respected by the Wayback Machine's crawlers), the Wayback Machine's occasional failure to archive embedded content, as well as site resources that were archived at different times than the top-level site. Though popular sites are typically archived at regular intervals, their embedded content (including third-party trackers) may thus be only partially represented. Whereas others have observed similar limitations with the Wayback Machine, especially as it relates to content visible on the top-level page [10, 38, 53], our analysis is focused on the technical impact of missing third-party elements, particularly with respect to tracking. Through our evaluation, we characterize what the Wayback Machine lets us measure about the embedded third parties, and showcase some techniques for best using the data it provides and working around some of its weaknesses (Section 4).

After evaluating the Wayback Machine's view into the past and developing best practices for using its data, we use TrackingExcavator to conduct a longitudinal study of the third-party web tracking ecosystem from 1996-2016 (Sections 5). We explore how this ecosystem has changed over time, including the prevalence of different web tracking behaviors, the identities and scope of popular trackers, and the complexity of relationships within the ecosystem. Among our findings, we identify the earliest tracker in our dataset in 1996 and observe the rise and fall of important players in the ecosystem (e.g., the rise of Google Analytics to appear on over a third of all popular websites). We find that websites contact an increasing number of third parties over time (about 5% of the 500 most popular sites contacted at least 5 separate third parties in early 2000s, whereas nearly 40% do so in 2016) and that the top trackers can track users across an increasing percentage of the web's most popular sites. We also find that tracking behaviors changed over time, e.g., that third-party popups peaked in the mid-2000s and that the fraction of trackers that rely on referrals from other trackers has recently risen.

Taken together, our findings show that third-party web tracking is a rapidly growing practice in an increasingly complex ecosystem — suggesting that users' and policymakers' concerns about privacy require sustained, and perhaps increasing, attention. Our results provide hitherto unavailable historical context for today's technical and policy discussions.

In summary, our contributions are:

1. TrackingExcavator, a **measurement infrastructure** for detecting and analyzing third-party web tracking behaviors in the present and — leveraging the Wayback Machine — in the past (Section 3).
2. An **in-depth analysis** of the scope and accuracy of the Wayback Machine's view of historical web tracking behaviors and trends, and techniques for working around its weaknesses (Section 4).
3. A **longitudinal measurement study** of third-party cookie-based web tracking from 1996 to present (2016) — to the best of our knowledge, the longest longitudinal study of tracking to date (Section 5).

This paper and any updates, including any data or code we publish, will be made available at `http://trackingexcavator.cs.washington.edu/`.

2 Background and Motivation

Third-party web tracking is the practice by which entities ("trackers") embedded in webpages re-identify users as they browse the web, collecting information about the websites that they visit [50, 60]. Tracking is typically done for the purposes of website analytics, targeted advertising, and other forms of personalization (e.g., social media content). For example, when a user visits `www.cnn.com`, the browser may make additional requests to `doubleclick.net` to load targeted ads and to `facebook.com` to load the "Like" button; as a result, Doubleclick and Facebook learn about that user's visit to CNN. Cookie-based trackers re-identify users by setting unique identifiers in browser cookies, which are then automatically included with requests to the tracker's domain. Figure 1 shows a basic example; we discuss

[1] `https://archive.org`

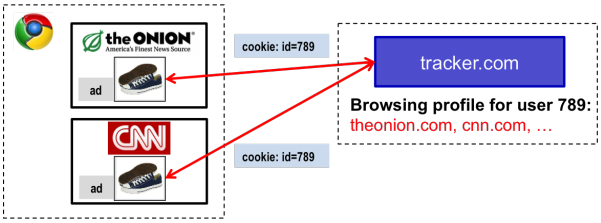


Figure 1: Overview of basic cookie-based web tracking. The third-party domain `tracker.com` uses a browser cookie to re-identify users on sites that embed content from `tracker.com`. This example shows *vanilla* tracking according to the taxonomy from [60]; other behaviors are described in Section 3.

more complex cookie-based tracking behaviors in Section 3. Though cookie-based tracking is extremely common [60], other types of tracking behaviors have also emerged, including the use of other client-side storage mechanisms, such as HTML5 LocalStorage, or the use of browser and/or machine fingerprinting to re-identify users without the need to store local state [15, 57].

Because these embedded trackers are often invisible to users and not visited intentionally, there has been growing concern about the privacy implications of third-party tracking. In recent years, it has been the subject of repeated policy discussions (Mayer and Mitchell provide an overview as of 2012 [50]); simultaneously, the computer science research community has studied tracking mechanisms (e.g., [50, 57, 60, 71]), measured their prevalence (e.g., [3, 20, 42, 60]), and developed new defenses or privacy-preserving alternatives (e.g., [6, 22, 25, 61, 64]). We discuss related works further in Section 6.

However, the research community's interest in web tracking is relatively recent, with the earliest measurements (to our knowledge) beginning in 2005 [42], and each study using a different methodology and measuring a different subset of known tracking techniques (see Englehardt et al. [18] for a comprehensive list of such studies). The practices of embedding third-party content and targeted advertising on websites predate these first studies [48], and longitudinal studies have been limited. However, longitudinal studies are critical to ensure the sustained effects of transparency [63] and to contextualize future measurements. Thus, to help ground technical and policy discussions surrounding web tracking in historical trends, we ask: how has the third-party tracking ecosystem evolved over the lifetime of the web?

We investigate questions such as:

- How have the **numbers, identities, and behaviors** of dominant trackers changed over time?
- How has the **scope** of the most popular trackers (i.e., the number of websites on which they are embedded) changed over time?
- How has the **prevalence** of tracking changed over time? For example, do websites include many more third-party trackers now than they did in the past?
- How have the **behaviors** of web trackers (e.g., JavaScript APIs used) changed over time?

By answering these questions, we are to able provide a systematic and longitudinal view of third-party web tracking over the last 20 years, retroactively filling this gap in the research literature, shedding a light on the evolution of third-party tracking practices on the web, and informing future technical and policy discussions.

The Wayback Machine. To conduct our archeological study, we rely on data from the Internet Archive's Wayback Machine (`https://archive.org`). Since 1996, the Wayback Machine has archived full webpages, including JavaScript, stylesheets, and any resources (including third-party JavaScript) that it can identify statically from the site contents. It mirrors past snapshots of these webpages on its own servers; visitors to the archive see the pages as they appeared in the past, make requests for all resources from the Wayback Machine's archived copy, and execute all JavaScript that was archived. We evaluate the completeness of the archive, particularly with respect to third-party requests, in Section 4.

3 Measurement Infrastructure: TrackingExcavator

To conduct a longitudinal study of web tracking using historical data from the Wayback Machine, we built a tool, TrackingExcavator, with the capability to (1) detect and analyze third-party tracking-related behaviors on a given web page, and (2) run that analysis over historical web pages archived and accessed by the Wayback Machine. In this section, we introduce TrackingExcavator. Figure 2 provides and overview of TrackingExcavator, which is organized into four pipeline stages:

(1) Input Generation (Section 3.1): TrackingExcavator takes as input a list of top-level sites on which to measure tracking behaviors (such as the Alexa top 500 sites), and, in "Wayback mode," a timestamp for the desired archival time to create `archive.org` URLs.

(2) Data Collection (Section 3.2): TrackingExcavator includes a Chrome browser extension that automatically visits the pages from the input set and collects tracking-relevant data, such as third-party requests, cookies, and the use of certain JavaScript APIs.

(3) Data Analysis (Section 3.3): TrackingExcavator processes collected measurement events to detect and categorize third-party web tracking behaviors.

(4) Data Visualization: Finally, we process our results into visual representations (included in Section 5).

3.1 Input Generation

In the input generation phase, we provide TrackingExcavator with a list of top-level sites to use for measurement.

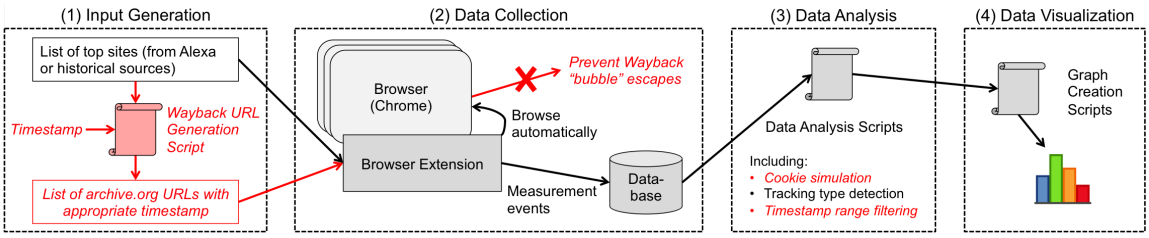


Figure 2: Overview of our infrastructure, TrackingExcavator, organized into four pipeline stages. Red/italic elements apply only to "Wayback mode" for historical measurements, while black/non-italics elements apply also to present-day measurements.

For historical measurements, TrackingExcavator must take a list of top-level URLs along with historical timestamps and transform them into appropriate URLs on `archive.org`. For example, the URL for the Wayback Machine's February 10, 2016 snapshot of `https://www.usenix.org/conference/usenixsecurity16` is `https://web.archive.org/web/20160210050636/https://www.usenix.org/conference/usenixsecurity16`.

We use the Memento API to find the nearest archived snapshot of a website occurring before the specified measurement date [36]. Though this process ensures a reasonable timestamp for the top-level page, embedded resources may have been archived at different times [5]. During analysis, we thus filter out archived resources whose timestamps are more than six months from our measurement timestamp, to ensure minimal overlap and sufficient spacing between measurements of different years.

3.2 Data Collection

To collect data, TrackingExcavator uses a Chrome extension to automatically visit the set of input sites. Note that we cannot log into sites, since the Wayback Machine cannot act as the original server. Our browser is configured to allow third-party cookies as well as pop-ups, and we visit the set of sites twice: once to prime the cache and the cookie store (to avoid artifacts of first-time browser use), and once for data collection. During these visits, we collect the following information relevant to third-party web tracking and store it in a local database:

- All request and response headers (including `set-cookie`).
- All cookies programmatically set by JavaScript (using `document.cookie`).
- All accesses to fingerprint-related JavaScript APIs, as described below.
- For each request: the requested URL, (if available) the referrer, and (if available) information about the originating tab, frame, and window.

We later process this data in the analysis phase of TrackingExcavator's pipeline (Section 3.3 below).

Fingerprint-Related APIs. Since cookie-based web tracking is extremely common (i.e., it is "classic" web tracking), we focus largely on it — and third-party requests in general — to capture the broadest view of the web tracking ecosystem over time. However, we also collect information about the uses of other, more recently emerged tracking-related behaviors, such as JavaScript APIs that may be used to create browser or machine fingerprints [15, 57]. To capture any accesses a webpage makes to a fingerprint-related JavaScript API (such as `navigator.userAgent`), TrackingExcavator's Chrome extension Content Script overwrites these APIs on each webpage to (1) log the use of that API and (2) call the original, overwritten function. The set of APIs that we hook was collected from prior work on fingerprint-based tracking [3, 4, 15, 56, 57] and is provided in Appendix A.

Preventing Wayback "Escapes". In archiving a page, the Wayback Machine transforms all embedded URLs to archived versions of those URLs (similar to our own process above). However, sometimes the Wayback Machine fails to properly identify and rewrite embedded URLs. As a result, when that archived page is loaded on `archive.org`, some requests may "escape" the archive and reference resources on the live web [9, 38]. In our data collection phase, we block such requests to the live web to avoid anachronistic side effects. However, we record the domain to which such a request was attempted, since the archived site did originally make that request, and thus we include it in our analysis.

3.3 Data Analysis

In designing TrackingExcavator, we chose to separate data collection from data analysis, rather than detecting and measuring tracking behaviors on the fly. This modular architecture simplifies data collection and isolates it from possible bugs or changes in the analysis pipeline — allowing us to rerun different analyses on previously collected data (e.g., to retroactively omit certain domains).

"Replaying" Events. Our analysis metaphorically "replays" collected events to simulate loading each page in the measurement. For historical measurements, we modify request headers to replace "live web" `Set-Cookie` headers with `X-Archive-Orig-Set-Cookie` headers

added by `archive.org`, stripping the Wayback Machine prefixes from request and referrer URLs, and filling our simulated cookie jar (described further below). During the replay, TrackingExcavator analyzes each event for tracking behaviors.

Classifying Tracking Behaviors. For *cookie-based trackers*, we base our analysis on a previously published taxonomy [60].[2] We summarize — and augment — that taxonomy here. Note that a tracker may fall into multiple categories, and that a single tracker may exhibit different behaviors across different sites or page loads:

1. *Analytics Tracking*: The tracker provides a script that implements website analytics functionality. Analytics trackers are characterized by a script, sourced from a third party but run in the first-party context, that sets first-party cookies and later leaks those cookies to the third-party domain.
2. *Vanilla Tracking*: The tracker is included as a third party (e.g., an iframe) in the top-level page and uses third-party cookies to track users across sites.
3. *Forced Tracking*: The tracker forces users to visit its domain directly — for example, by opening a popup or redirecting the user to a full-page ad — allowing it to set cookies from a first-party position.
4. *Referred Tracking*: The tracker relies on another tracker to leak unique identifiers to it, rather than on its own cookies. In a hypothetical example, `adnetwork.com` might set its own cookie, and then explicitly leak that cookie in requests to referred tracker `ads.com`. In this case, `ads.com` need not set its own cookies to perform tracking.
5. *Personal Tracking*: The tracker behaves like a Vanilla tracker but is visited by the user directly in other contexts. Personal trackers commonly appear as social widgets (e.g., "Like" or "tweet" buttons).

In addition to these categories previously introduced [60], we discovered an additional type of tracker related to but subtly different from Analytics tracking:

6. *Referred Analytics Tracking:* Similar to an Analytics tracker, but the domain which sets a first-party cookie is *different* from the domain to which the first-party cookie is later leaked.

Beyond cookie-based tracking behaviors, we also consider the use of fingerprint-related JavaScript APIs, as described above. Though the use of these APIs does not necessarily imply that the caller is fingerprinting the user — we know of no published heuristic for determining fingerprinting automatically — but the use of many such APIs may suggest *fingerprint-based tracking*.

Finally, in our measurements we also consider *third-party requests* that are not otherwise classified as trackers. If contacted by multiple domains, these third-parties have the *ability* to track users across sites, but may or may not actually do so. In other words, the set of all domains to which we observe a third-party request provides an upper bound on the set of third-party trackers.

We tested TrackingExcavator's detection and classification algorithms using a set of test websites that we constructed and archived using the Wayback Machine, triggering each of these tracking behaviors.

Reconstructing Archived Cookies. For many tracking types, the presence or absence of cookies is a key factor in determining whether the request represents a tracking behavior. In our live measurements, we have the actual `Cookie` headers attached by Chrome during the crawl. On archived pages, the Wayback Machine includes past `Set-Cookie` headers as `X-Archive-Orig-Set-Cookie` headers on archived responses. To capture the cookies that would have actually been set during a live visit to that archived page, TrackingExcavator must simulate a browser cookie store based on these archival cookie headers and JavaScript cookie set events recorded during data collection.

Unfortunately, cookie engines are complicated and standards non-compliant in major browsers, including Chrome [11]. Python's cookie storage implementation is compliant with RFC 2965, obsoleted by RFC 6265, but these standards proposals do not accurately represent modern browser practices [7, 13, 21]. For efficiency, we nevertheless use Python's cookie jar rather than attempting to re-implement Chrome's cookie engine ourselves.

We found that Python's cookie jar computed cookies exactly matching Chrome's for only 71% of requests seen in a live run of the top 100. However, for most types of tracking, we only need to know whether *any* cookies would have been set for the request, which we correctly determine 96% of the time. Thus our tool captures most tracking despite using Python's cookie jar.

Classifying Personal Trackers in Measurements. For most tracker types, classification is independent of user behaviors. Personal trackers, however, are distinguished from Vanilla trackers based on whether the user visits that domain as a top-level page (e.g., Facebook or Google). To identify likely Personal trackers in automated measurement, we thus develop a heuristic for user browsing behaviors: we use popular sites from each year, as these are (by definition) sites that many users visited.

Alexa's top sites include several that users would not typically visit directly, e.g., `googleadservices.com`. Thus, we manually examined lists of popular sites for each year to distinguish between domains that users *typically* visit intentionally (e.g., Facebook, Amazon) from those which ordinary users never or rarely visit intentionally (e.g., ad networks or CDNs). Two researchers

[2] We are not aware of other taxonomies of this granularity for cookie-based tracking.

independently classified the domains on the Alexa top 100 sites for each year where we have Alexa data, gathering information about sites for which they were unsure. The researchers examined 435 total domains: for the top 100 domains in 2016, they agreed on 100% and identified 94 sites as potential Personal trackers; for the 335 additional domains in the previous years' lists, they agreed on 95.4% and identified 296 Personal tracker domains.

4 Evaluating the Wayback Machine as an Archaeological Data Source for Tracking

The Wayback Machine provides a unique and comprehensive source of historical web data. However, it was not created for the purpose of studying third-party web tracking and is thus imperfect for that use. Nevertheless, the only way to study web tracking *prior* to explicit measurements targeting it is to leverage materials previously archived for other purposes. Therefore, before using the Wayback Machine's archived data, it is essential to systematically characterize and analyze its capabilities and flaws in the context of third-party tracking.

In this section we thus study the extent to which data from the Wayback Machine allows us to study historical web tracking behaviors. Beyond providing confidence in the trends of web tracking over time that we present in Section 5, we view this evaluation of the Wayback Machine as a contribution of this paper. While others have studied the quality of the Wayback Machine's archive, particularly with respect to the quality of the archived content displayed on the top-level page (e.g., [10, 38, 53]), we are the first to systematically study the quality of the Wayback Machine's data about *third-party requests*, the key component of web tracking.

To conduct our evaluation, we leverage four ground truth data sets collected from the live web in 2011, 2013, 2015, and 2016. The 2011 data was originally used in [60] and provided to us by those authors. All datasets contain classifications of third-party cookie-based trackers (according to the above taxonomy) appearing on the Alexa top 500 sites (from the time of each measurement). The 2015 and 2016 data was collected by TrackingExcavator and further contains all HTTP requests, including those not classified as tracking.[3] We plan to release our ground truth datasets from 2013, 2015, and 2016.

We organize this section around a set of lessons that we draw from this evaluation. We apply these lessons in our measurements in Section 5. We believe our findings can assist future researchers seeking to use the Wayback Machine as a resource for studying tracking (or other web properties relying on third-party requests) over time.

[3]For comparison, the published results based on the 2011 dataset [60] measured tracking on the homepages of the top 500 websites as well as four additional pages on that domain; for the purposes of our work, we re-analyzed the 2011 data using only homepages.

	August 1	August 25	September 1
All Third-Parties	324	304	301
Analytics	7	13	11
Vanilla	127	115	108
Forced	0	0	0
Referred	3	3	3
Personal	23	21	21
Referred Analytics	21	17	18

Table 1: Natural variability in the trackers observed on different visits to the Alexa top 100 in 2015. This variability can result from non-static webpage content, e.g., ad auctions that result in different winners.

4.1 Lesson (Challenge): The Wayback Machine provides a partial view of third-party requests

A key question for using the Wayback Machine for historical measurements is: how complete is the archive's view of the past, both for the top-level pages and for the embedded content? In this lesson, we explore why its view is incomplete, surfacing challenges that we will overcome in subsequent lessons. We identify several reasons for the differences between the live and Wayback measurements, and quantify the effects of each.

Variation Between Visits. Different trackers and other third parties may appear on a site when it is loaded a second time, even if these views are close together; an example of this variation would be disparity in tracking behaviors between ads in an ad network.

To estimate the degree of variation between page views, we compare three live runs from August-September 2015 of the Alexa top 100 sites (Table 1). We find that variation between runs even a week apart is notable (though not enough to account for all of the differences between Wayback and live datasets). For the number of Vanilla trackers found, the August 25th and September 1st runs vary by 7 trackers, or 6%.

Non-Archived and Blocked Requests. There are several reasons that the Wayback Machine may fail to archive a response to a request, or provide a response that TrackingExcavator must ignore (e.g., from a far different time than the one we requested or from the live web). We describe these conditions here, and evaluate them in the context of a Wayback Machine crawl of the top 500 pages archived in 2016, according to the 2016 Alexa top 500 rankings; we elaborate on this dataset in Section 5. Table 2 summarizes how often the various conditions occur in this dataset, for requests, unique URLs, and unique domains. In the case of domains, we count only those domains for which *all* requests are affected, since those are the cases where we will *never* see a cookie or any other subsequent tracking indicators for that domain.

Robots.txt Exclusions (403 errors). If a domain's `robots.txt` asks that it not be crawled, the Wayback Machine will respect that restriction and thus not archive

Type of Blocking		Fraction Missed
Robots Exclusions	Requests	1115 / 56,173 (2.0%)
	URLs	609 / 27,532 (2.2%)
	Domains	18 / 1150 (1.6%)
Not Archived	Requests	809 / 56,173 (1.4%)
	URLs	579 / 27,532 (2.1%)
	Domains	8 / 1150 (0.7%)
Wayback Escapes	Requests	9025 / 56,173 (16.1%)
	URLs	4730 / 27,532 (17.2%)
	Domains	132 / 1150 (11.5%)
Inconsistent Timestamps	Requests	404 / 56,173 (0.7%)
	URLs	156 / 27,532 (0.6%)
	Domains	55 / 1150 (4.8%)

Table 2: For the archived versions of the Alexa top 500 sites from 2016, the fraction of requests, unique URLs, and unique domains affected by robots exclusion (403 errors), not archived (404), Wayback escapes (blocked by TrackingExcavator), or inconsistent timestamps (filtered by TrackingExcavator).

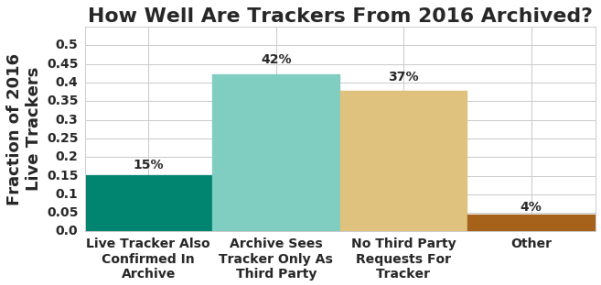


Figure 3: The fraction of domains categorized as Vanilla trackers in the live 2016 crawl which, in the archival 2016 crawl, (1) set and leaked cookies and thus were confirmed as trackers, (2) were only third-party requests (had at least one third-party request but no cookies), (3) did not appear at all, or (4) other (e.g., had cookies but not at the time of a third-party request, or cookies were not attached due to a cookie simulation bug).

the response. As a result, we will not receive any information about that site (including cookies, or use of Javascript) nor will we see any subsequent resources that would have resulted from that request.

We find that only a small fraction of all requests, unique URLs, and (complete) domains are affected by robots exclusion (Table 2). We note that robots exclusions are particularly common for popular trackers. Of the 20 most popular trackers on the 2016 live dataset, 12 (60%) are blocked at least once by robots.txt in the 2016 Wayback measurement. By contrast, this is true for only 74/456, or 16.23%, of all Vanilla trackers seen in live.

Other Failures to Archive (404 errors). The Wayback Machine may fail to archive resources for any number of reasons. For example, the domain serving a certain resource may have been unavailable at the time of the archive, or changes in the Wayback Machine's crawler may result in different archiving behaviors over time. As shown in Table 2, missing archives are rare.

URL Rewriting Failures (Wayback "Escapes"). Though the Wayback Machine's archived pages execute the corresponding archived JavaScript within the browser when TrackingExcavator visits them, the Wayback Machine does not execute JavaScript during its archival crawls of the web. Instead, it attempts to statically extract URLs from HTML and JavaScript to find additional sites to archive. It then modifies the archived JavaScript, rewriting the URLs in the included script to point to the archived copy of the resource. This process may fail, particularly for dynamically generated URLs. As a result, when TrackingExcavator visits archived pages, dynamically generated URLs not properly redirected to their archived versions will cause the page to attempt to make a request to the live web, i.e., "escape" the archive. TrackingExcavator blocks such escapes (see Section 3). As a result, the script never runs on the archived site, never sets a cookie or leaks it, and thus TrackingExcavator does not witness the associated tracking behavior.

We find that Wayback "escapes" are more common than robots exclusion or missing archives (Table 2): 16.1% of all requests attempted to "escape" (i.e., were not properly rewritten by the Wayback Machine) and were blocked by TrackingExcavator.

Inconsistent Timestamps. As others have documented [10], embedded resources in a webpage archived by the Wayback Machine may occasionally have a timestamp far from the timestamp of the top-level page. As described in Section 3, we ignore responses to requests for resources with timestamps more than six months away.

Cascading Failures. Any of the above failures can lead to cascading failures, in that non-archived responses or blocked requests will result in the omission of any subsequent requests or cookie setting events that would have resulted from the success of the original request. The "wake" of a single failure cannot be measured within an archival dataset, because events following that failure are simply missing. To study the effect of these cascading failures, we must compare an archival run to a live run from the same time; we do so in the next subsection.

4.2 Lesson (Opportunity): Consider all third-party requests, in addition to confirmed trackers

In the previous section, we evaluated the Wayback Machine's view of third-party requests *within* an archival measurement. For requests affected by the issues in Table 2, TrackingExcavator observes the existence of these requests — i.e., counts them as third parties — but without the corresponding response may miss additional information (e.g., set cookies) that would allow it to confirm these domains as trackers according to the taxonomy presented earlier. However, this analysis cannot give us a sense of how many third-party requests are entirely absent from Wayback data due to cascading failures, nor a sense of any other data missing from the archive, such as

	2011	2013	2015	2016
Wayback (All Third Parties)	553	621	749	723
Wayback (Vanilla+Personal)	47	49	92	90
Live (Vanilla+Personal)	370	419	493	459
Wayback-to-Live Ratio (Vanilla+Personal)	0.13	0.12	0.19	0.20

Table 3: We compare the prevalence of the most common tracking types (Vanilla and Personal) over the four years for which we have data from the live web. Though the Wayback Machine provides only partial data on trackers, it nevertheless illuminates a general upward trend reflected in our ground truth data.

missing cookie headers on otherwise archived responses. For that, we must compare directly with live results.

We focus our attention on unique trackers: we attempt to identify which live trackers are missing in the 2016 Wayback dataset, and why. For each tracker we observe in our 2016 live measurement, Figure 3 identifies whether we (1) also observe that tracker in "Wayback mode," (2) observe only a third-party request (but no confirmed cookie-based tracking behavior, i.e., we classify it only as a third-party domain), or (3) do not observe any requests to that tracker at all.

We conclude two things from this analysis. First, because the Wayback Machine may fail to provide sufficient data about responses or miss cookies even in archived responses, many trackers confirmed in the live dataset appear as simple third-party requests in the Wayback data (the second column in Figure 3). For example, `doubleclick.net`, one of the most popular trackers, appears as only a third party in Wayback data because of its `robots.txt` file. Thus, we learn that to study third-party web tracking in the past, due to missing data in the archive, *we must consider all third-party requests*, not only those confirmed as trackers according to the taxonomy. Though considering only third-party requests will overcount tracking in general (i.e., not all third parties on the web are trackers), we find that it broadens our view of tracking behaviors in the archive.

Second, we find that a non-trivial fraction of trackers are missing entirely from the archive (the third column in Figure 3). In the next subsection, we show that we can nevertheless draw conclusions about trends over time, despite the fact that the Wayback Machine underrepresents the raw number of third parties contacted.

4.3 Lesson (Opportunity): The Wayback Machine's data allows us to study trends over time

As revealed above, the Wayback Machine's view of the past may miss the presence of some third parties entirely. Thus, one unfortunately cannot rely on the archive to shed light on the exact raw numbers of trackers and other third parties over time. Instead, we ask: does the Wayback Machine's data reveal genuine historical *trends*?

To investigate trends, we compare all of our live datasets (2011, 2013, 2015, and 2016) to their Wayback counterparts. Table 3 compares the number of Vanilla and Personal trackers (the most prevalent types) detected in each dataset. For the purposes of this comparison, we sum the two types, since their distinction depends only on the user's browsing behaviors. We also include the number of all third parties in the Wayback datasets, based on the previous lesson. Though not all of these third parties represent trackers in live data, they help illuminate trends in third party prevalence over time.

We draw two conclusions from this comparison. First, we find that we *can* rely on the archive to illuminate general trends over time. Although confirmed trackers in "Wayback mode" (as expected from our earlier lessons) underrepresent the number of confirmed trackers found on the live web — and third parties in the archive overestimate confirmed trackers in the live data — we find that the trends we see over time are comparable in both sets of measurements. Critically, we see that *the upward trend in our archival view is not merely the result of improvements in archive quality over time or other factors — we indeed observe this trend reflected in ground truth data.* We gain further confidence in these trends in Section 5, where we see a rise in tracking behaviors since 1996 that corresponds with our intuition. The absence of any large vertical steps in the figures in Section 5 further suggests that the trends we identify are artifacts of the web evolving as opposed to any significant changes in the Wayback Machine archival process.

Second, however, we find that — although long-term trends appear to be meaningfully represented by the Wayback Machine — one should not place too much confidence into *small* variations in trends. For example, the Wayback Machine's data in 2013 appears to be worse than in other years, under-representing the number of confirmed trackers more than average. Thus, in Section 5, we do not report on results that rely on small variations in trends unless we have other reasons to believe that these variations are meaningful.

4.4 Lesson (Opportunity): Popular trackers are represented in the Wayback Machine's data

Because popular trackers, by definition, appear on many sites that users likely browse to, they have a strong effect on user privacy and are particularly important to examine. We find that although the Wayback Machine misses some trackers (for reasons discussed above), *it does capture a large fraction of the most popular trackers* — likely because the Wayback Machine is more likely to have correctly archived at least one of each popular tracker's many appearances.

Specifically, when we examine the 2016 archival and live datasets, we find that 100% of the top 20 trackers from the live dataset are represented as either confirmed

trackers or other third parties in the Wayback data. In general, more popular trackers are better represented in Wayback data: 75% of the top 100 live trackers, compared to 53% of all live trackers. Tracker popularity drops quickly — the first live tracker missing in Wayback data is #22, which appears on only 22 of the top 500 websites; the 100th most popular tracker appears on only 4 sites. By contrast, the top tracker appears on 208 sites. In other words, those trackers that have the greatest impact on user privacy do appear in the archive.

Based on this lesson, we focus part of Section 5's analysis in on popular trackers, and we *manually label* those that the Wayback Machine only sees as third parties but that we know are confirmed trackers in live data.

4.5 Lesson (Opportunity): The Wayback Machine provides additional data beyond requests

Thus far, we have considered third-party requests and confirmed cookie-based trackers. However, the Wayback Machine provides, and TrackingExcavator collects, additional data related to web tracking behaviors, particularly the use of various JavaScript APIs that allow third parties to collect additional information about users and their machines (e.g., to re-identify users based on fingerprints). For JavaScript correctly archived by the Wayback Machine, TrackingExcavator observes accesses to the supported APIs (Appendix A). For example, we observe uses of `navigator.userAgent` as early as 1997.

4.6 Summary

In summary, we find that the Wayback Machine's view of the past is incomplete, and that its weaknesses particularly affect the third-party requests critical for evaluating web tracking over time. We identified and quantified those weaknesses in Section 4.1, and then introduced findings and strategies for mitigating these weaknesses in Sections 4.2-4.5, including considering third-party requests as well as confirmed trackers, manually labeling known popular trackers, and studying general trends over time instead of raw numbers. We leverage these strategies in our own measurements. By surfacing and evaluating these lessons, we also intend to help guide future researchers relying on data from the Wayback Machine.

We focus on the Wayback Machine since it is to our knowledge the most comprehensive web archive. Applying our approach to other, more specialized archives [58], if relevant for other research goals, would necessitate a new evaluation of the form we presented here.

5 Historical Web Tracking Measurements

We now turn to our longitudinal study of third-party cookie-based web tracking from 1996-2016.

Datasets. We focus our investigation on the most popular websites each year, for two reasons: first, trackers on these sites are (or were) able to collect information about the greatest number of users; second, popular sites are crawled more frequently by the Wayback Machine (if permitted by `robots.txt`). We thus need historical lists of the top sites globally on the web.

2003-2016: Alexa. For 2010-2016, we use Wayback Machine archives of Alexa's top million sites list (a `csv` file). For 2003-2009, we approximate the top 500 by scraping Alexa's own historical API (when available) and archives of individual Alexa top 100 pages. Because of inconsistencies in those sources, our final lists contain 459-500 top sites for those years.

1996-2002: Popular Links from Homepages. In 2002, only the Alexa top 100 are available; before 2002, we only have ComScore's list of 20 top sites [69]. Thus, to build a list of 500 popular sites for the years 1996-2002, we took advantage of the standard practice at the time of publishing links to popular domains on personal websites. Specifically, we located archives of the People pages of the Computer Science or similar department at the top 10 U.S. CS research universities as of 1999, as reported in that year by U.S. News Online [2]. We identified the top 500 domains linked to from the homepages accessible from those People pages, and added any ComScore domains that were not found by this process. We ran this process using People pages archived in 1996 and 1999; these personal pages were not updated or archived frequently enough to get finer granularity. We used the 1996 list as input to our 1996, 1997 and 1998 measurements, and the 1999 list as input for 1999-2002.

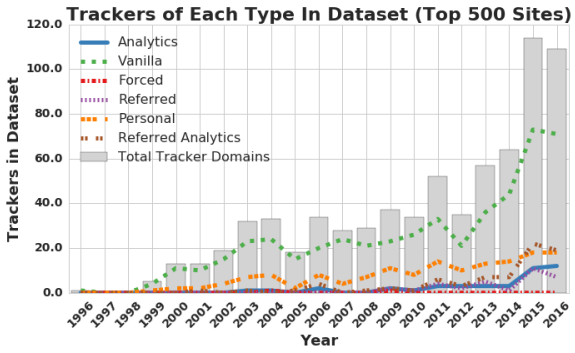


Figure 4: Evolution of tracker types over time. The grey bars show the total number of tracking domains present in each dataset, and the colored lines show the numbers of trackers with each type of tracking behavior. A single tracker may have more than one behavior in the dataset (e.g., both Vanilla and Analytics), so the sum of the lines might be greater than the bar.

5.1 Prevalence of Tracking Behaviors over Time

We begin by studying the prevalence of tracking behaviors over time: how many unique trackers do we observe, what types of tracking behaviors do those trackers exhibit, and how many trackers appear on sites over time?

Prevalence and Behaviors of Unique Trackers. Figure 4 shows the total number of unique trackers observed over time (the grey bars) and the prevalence of different tracking behavior types (the lines) for the top 500 sites from 1996-2016. Note that trackers may exhibit more than one behavior across sites or on a single site, so the sum of the lines may be greater than the height of the bar. We note that the particularly large bars in 2015 and 2016 may reflect not only a change in tracking prevalence but also changes in the way the Wayback Machine archived the web. See Table 3 for validation against live data which suggest that actual growth may have been smaller and more linear, similar to past years.

We make several observations. First, we see the emergence of different tracking behaviors: the first cookie-based tracker in our data is from 1996: `microsoft.com` as a Vanilla tracker on `digital.net`. The first Personal tracker to appear in our dataset is in 1999: `go.com` shows up on 5 different sites that year, all also owned by Disney: `disney.com`, `espn.com`, `sportszone.com`, `wbs.net`, and `infoseek.com` (acquired by Disney mid-1999 [1], before the date of our measurement). The existence of a Personal tracker that only appeared on sites owned by the same company differs from today's Personal tracking ecosystem, in which social media widgets like the Facebook "Like" button appear on many popular sites unaffiliated with that tracker (Facebook, in this case) [60].

More generally, we see a marked increase in quantities of trackers over time, with rises in all types of tracking behavior. One exception is Forced trackers — those relying on popups — which are rare and peaked in the early 2000s before popup blockers became default (e.g., in 2004 for Internet Explorer [54]). Indeed, we see third-party popups peak significantly in 2003 and 2004 (17 and 30 popups, respectively, compared to an annual mean of about 4), though we could not confirm all as trackers for Figure 4. Additionally, we see an increasing variety of tracking behavior over time, with early trackers nearly all simply Vanilla, but more recent rises in Personal, Analytics, and Referred tracking.

We can also consider the complexity of individual trackers, i.e., how many distinct tracking behaviors they exhibit over each year's dataset. (Note that some behaviors are exclusive, e.g., a tracker cannot be both Personal and Vanilla, but others are nonexclusive.) Table 4 suggests that there has been some increase in complexity in recent years, with more trackers exhibiting two or even three behaviors. Much of this increase is due to the rise in Referred or Referred Analytics trackers, which receive cookie values shared explicitly by other trackers in addition to using their own cookies in Vanilla behavior.

Fingerprint-Related APIs. We measured the use of Javascript APIs which can be used to fingerprint browsers and persist identifiers even across cookie deletion. Though the use of these APIs does not necessarily imply that they are used for tracking (and we know of no published heuristic for correlating API use with genuine fingerprinting behaviors), the use of these APIs nevertheless allows third parties to gather potentially rich information about users and their machines. The full list of 37 fingerprint-related APIs we measure (based on prior work [3, 4, 15, 56, 57]) is in Appendix A.

We now consider third parties that are prolific users of fingerprint-related APIs, calling many APIs on each site. Table 5 shows the tracker in each year that calls the most APIs on a single site. Ties are broken by the choosing the third party that appears on the largest number of sites. Maximum usage of APIs has increased over time, but we observe that the most prolific API users are not the most popular cookie-based trackers. Although we only identify API uses within JavaScript, and not how their results are used, we note that increasing use

Year	1Type	2Type	3Type	4Type
1996	100.00% (1)	0	0	0
1998	0	0	0	0
2000	100.00% (13)	0	0	0
2002	100.00% (19)	0	0	0
2004	96.97% (32)	3.03% (1)	0	0
2006	100.00% (34)	0	0	0
2008	100.00% (29)	0	0	0
2010	94.12% (32)	2.94% (1)	2.94% (1)	0
2012	88.57% (31)	11.43% (4)	0	0
2014	93.75% (60)	4.69% (3)	1.56% (1)	0
2016	86.24% (94)	11.01% (12)	2.75% (3)	0

Table 4: Complexity of trackers, in terms of the percentage (and number) of trackers displaying one or more types of tracking behaviors across the top 500 sites.

Year	Most Prolific API-user	Num APIs Used	Coverage
1998	realhollywood.com	2	1
1999	go2net.com	2	1
2000	go.com	6	2
2001	akamai.net	8	15
2002	go.com	10	2
2003	bcentral.com	5	1
2004	163.com	9	3
2005	163.com	8	1
2006	sina.com.cn	11	2
2007	googlesyndication.com	8	24
2008	go.com	12	1
2009	clicksor.com	10	2
2010	tribalfusion.com	17	1
2011	tribalfusion.com	17	2
2012	imedia.cz	12	1
2013	imedia.cz	13	1
2014	imedia.cz	13	1
2015	aolcdn.com	25	5
2016	aolcdn.com	25	3

Table 5: Most prolific API-users, with ties broken by coverage (number of sites on which they appear) for each year. The maximum number of APIs used increases over time, but the max API users are not necessarily the most popular trackers.

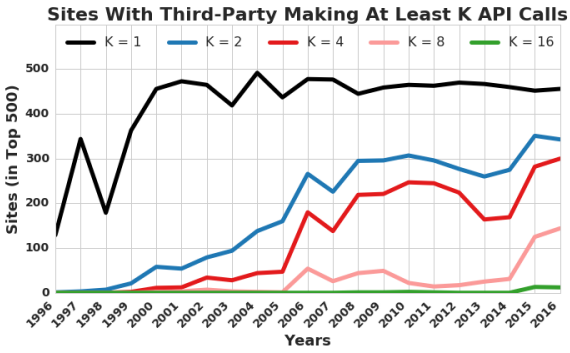


Figure 5: Number of sites in each year with a tracker that calls (on that site) at least K (of our 37) fingerprint-related APIs.

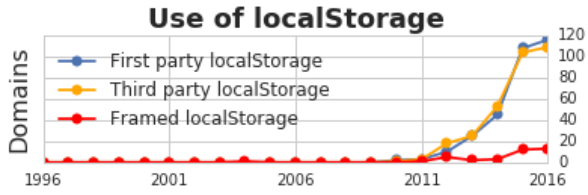


Figure 6: Domains using `window.localStorage`. First party usages are uses in the top frame of a web page by a script loaded from the web page's own domain. Third party usages are those also in the top frame of a page but by a script loaded from a third party. Framed uses are those inside of an iframe.

of these APIs implies increased power to fingerprint, especially when combined with non-Javascript signals such as HTTP headers and plugin behavior. For example, Panopticlick derived 18 bits of entropy about remote browsers from a subset of these APIs plus HTTP headers and information from plugins [15].

Beyond the power of the most prolific fingerprint-related API users growing, we also find that more sites include more trackers using these APIs over time. Figure 5 shows the number of sites in each year containing a tracker that calls, on that site, at least K of the 37 fingerprinting APIs. Although many sites contain and have contained trackers that use at least 1 API (typically `navigator.userAgent`, common in browser compatibility checks), the number of sites containing trackers that call 2 or more APIs has risen significantly over time.

In addition to fingerprint-related APIs, we also examine the use of HTML5 LocalStorage, a per-site persistent storage mechanism standardized in 2009 in addition to cookies. Figure 6 shows that the use of the `localStorage` API rises rapidly since its introduction in 2009, indicating that tracking defenses should increasingly consider on storage mechanisms beyond cookies.

Third Parties Contacted. We now turn our attention to the number of third parties that users encounter as they browse the web. Even third parties not confirmed as trackers have the *potential* to track users across the web, and as we discovered in Section 4, many third parties in archived data may in fact be confirmed trackers for which the Wayback Machine simply archived insufficient information. Figure 7 thus shows the distributions of how many third parties the top 500 sites contacted in each year. We see a rise in the median number of third parties contacted — in other words, more sites are giving more third parties the opportunity to track users.

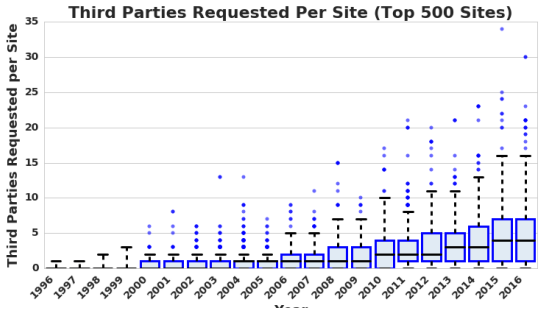


Figure 7: Distributions of third-party requests for the top 500 sites 1996-2016. Center box lines are medians, whiskers end at 1.5*IQR. The increase in both medians and distributions of the data show that more third-parties are being contacted by popular sites in both the common and extreme cases.

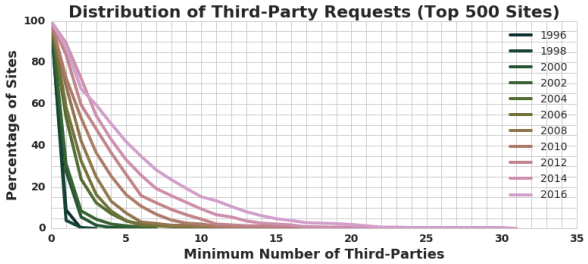


Figure 8: Distribution of top sites for each year by number of unique third-parties (tracking-capable domains) they contact. In later years, more sites appear to contact more third parties.

Figure 8 provides a different view of similar data, showing the distribution of the top sites for each year by number of distinct third parties contacted. In the early 2000s, only about 5% of sites contacted at least 5 third parties, while in 2016 nearly 40% of sites did so. We see a maximum in 2015, when one site contacted 34 separate third-parties (a raw number that is likely underestimated by the Wayback Machine's data)!

5.2 Top Trackers over Time

We now turn to an investigation of the top trackers each year: who are the top players in the ecosystem, and how wide is their view of users' browsing behaviors?

Coverage of Top Trackers. We define the *coverage* of a set of trackers as the percentage of total sites from the dataset for which at least one of those trackers appears. For a single tracker, its coverage is the percentage of sites on which it appears. Intuitively, coverage suggests the concentration of tracking ability — greater coverage al-

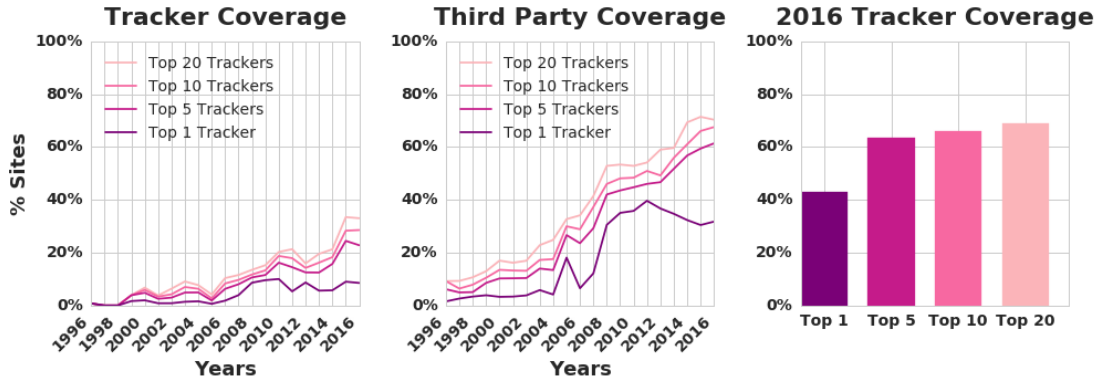


Figure 9: The growth in the coverage (percentage of top 500 sites tracked) of the top 1/5/10/20 trackers for each year is shown in the first and second panels, for all confirmed trackers and for all third parties respectively. The right hand panel shows the values on the live web for confirmed trackers, with the top 5 trackers covering about 70% of all sites in the dataset. Note that top third party coverage in the archive is an excellent proxy for modern confirmed tracker coverage today.

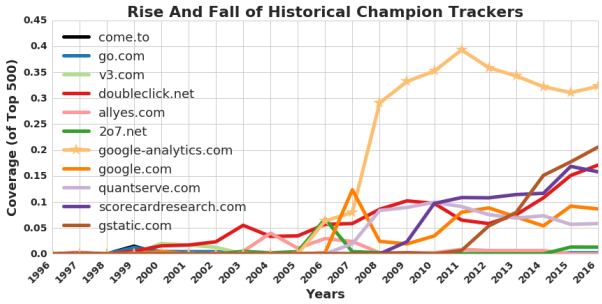


Figure 10: This figure depicts variations in site coverage for a number of the most popular confirmed trackers from years across the studied period. We call the two trackers embedded on the most sites in a given year the "champions" of that year, filtered by manual classification as described in the text.

lows trackers to build larger browsing profiles. This metric reaches toward the core privacy concern of tracking, that certain entities may know nearly everything a person does on the web. We consider trackers by domain name, even though some trackers are in fact owned by the same company (e.g., Google owns `google-analytics.com`, `doubleclick.net`, and the "+1" button served from `google.com`), because a business relationship does not imply that the entities share data, though some trackers may indeed share information out of public view.

Figure 9 illustrates the growth of tracker coverage over time. It considers both the single domain with the highest coverage for each year (Top 1 Tracker) as well as the combined coverage of the union of the top 5, 10 and 20 trackers. Confirming the lesson from Section 4.2, the coverage rates we see for third party domains in the archive are similar to live coverage of *confirmed* Vanilla cookie-based trackers.

Clearly, the coverage of top trackers has risen over time, suggesting that a small number of third parties can observe an increasing portion of user browsing histories.

Popular Trackers over Time. Who are these top trackers? Figure 10 shows the rise and fall of the top two trackers ("champions") for each year. To create this figure, we make use of the lesson in Section 4.4 to manually label known popular confirmed trackers. We identified the two domains with the highest third-party request coverage for each year, omitting cases where the most popular tracker in a year appeared on only one site. We manually verified that 12/19 of these domains were in fact trackers by researching the domain, owning company, archived behavior and context, and modern behaviors (if applicable). Based on this analysis, we are able to assess the change in tracking behaviors even of domains for whom cookies are lost in the archive (e.g., `doubleclick.net`). In particular, this analysis reveals trends in the trackers with the most power to capture profiles of user behavior across many sites.

We find that in the early 2000s, no single tracker was present on more than 10% of top sites, but in recent years, `google-analytics.com` has been present on nearly a third of top sites and 2-4 others have been present on more than 10% and growing. Some, such as `doubleclick.net` (acquired by Google in 2008) have been popular throughout the entire time period of the graph, while others, such as `scorecardresearch.com`, have seen a much more recent rise.

We note that `google-analytics.com` is a remarkable outlier with nearly 35% coverage in 2011. Google Analytics is also an outlier in that it is one of only two non-cross-site trackers among the champions (`gstatic.com`, a Referred Analytics tracker, is the other). As an Analytics type tracker, Google Analytics trackers users only within a single site, meaning that its "coverage" is arguably less meaningful than that of a cross-site tracker. However, we observe that Google Analytics *could* track users across sites via fingerprinting or by changing its behavior to store tracking cookies. This observation highlights the need for repeated

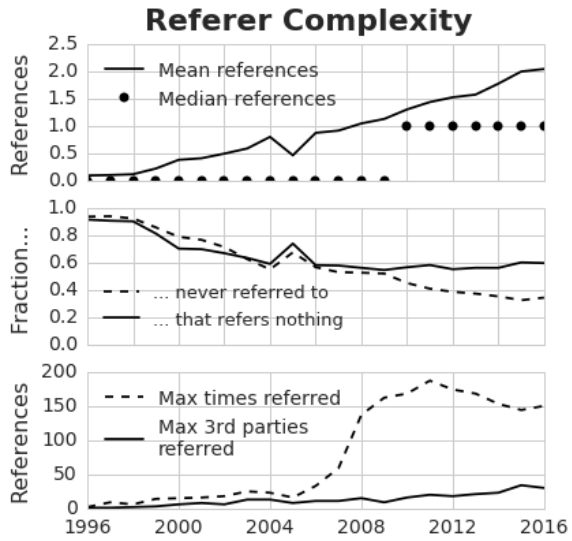


Figure 11: Changes in the frequency with which domains are referred to or refer to other domains (based on HTTP Referer).

measurements studies that provide transparency on the web: with a simple change to its tracking infrastructure, Google Analytics could begin to track users across 40% of the most popular sites on the web overnight. Thus, Google's decision not to structure Google Analytics in this way has a tremendous impact on user privacy.

5.3 Evolution of the Tracking Ecosystem

Finally, we consider the tracking ecosystem as a whole, focusing on relationships between different trackers. We find a remarkable increase in the complexity of these relationship over time. Again we consider only relationships observable using TrackingExcavator, not external information about business relationships.

To study these relationships, we construct the graph of referring relationships between elements on pages. For example, if we observe a third-party request from `example.com` to `tracker.com`, or from `tracker.com` referring to `tracker2.com`, the nodes for those domains in the graph will be connected by edges.

We find a significant increase in complexity over time by examining several properties of this graph (Figure 11). Over time, the mean number of referrals outward from domains increases (top of Figure 11), while the number of domains that are never referred to by other domains or never refer outward steadily decreases (middle of Figure 11). Meanwhile, the maximum number of domains that refer to a single domain increases dramatically, suggesting that individual third parties in the web ecosystem have gradually gained increasing prominence and coverage. This reflects and confirms trends shown by other aspects of our data (Figures 10 and 9). These trends illuminate an ecosystem of generally increasingly connected relationships and players growing in size and influence. Appendix B shows this evolution in graph form; the increase in complexity over time is quite striking.

5.4 Summary and Discussion

We have uncovered trends suggesting that tracking has become more prevalent and complex in the 20 years since 1996: there are now more unique trackers exhibiting more types of behaviors; websites contact increasing numbers of third parties, giving them the opportunity to track users; the scope of top trackers has increased, providing them with a broader view of user browsing behaviors; and the complexity and interconnectedness of the tracking ecosystem has increased markedly.

From a privacy perspective, our findings show that over time, more third parties are in a position to gather and utilize increasing amounts of information about users and their browsing behaviors. This increase comes despite recent academic, policy, and media attention on these privacy concerns and suggests that these discussions are far from resolved. As researchers continue to conduct longitudinal measurements of web tracking going forward, our work provides the necessary historical context in which to situate future developments.

6 Additional Related Work

Tracking and Defenses. Third-party tracking has been studied extensively in recent years, particularly through analysis and measurements from 2005 to present [18, 19, 24, 30, 32–34, 40–43, 60]. A few studies have considered mobile, rather than desktop, browser tracking [20, 27]. Beyond explicit stateful (e.g., cookie-based) tracking, recent work has studied the use of browser and machine fingerprinting techniques to re-identify and track users [3, 4, 15, 37, 57, 71]. Others have studied the possible results of tracking, including targeted ads [45, 70], personalized search [29], and price discrimination [66].

User-facing defenses against tracking range from browser extensions like Ghostery [23] and Privacy Badger [16] to research proposals (e.g. [8, 28]). Researchers have also designed privacy-preserving alternatives including privacy-preserving ads [22, 25, 59, 64], social media widgets [14, 39, 61], and analytics [6]. Others have studied user attitudes towards tracking and targeted advertising (e.g., [46, 51, 65]). Our study shows the increased prevalence of tracking over time, suggesting that designing and supporting these defenses for privacy-sensitive users is as important as ever.

Wayback Machine and other Longitudinal Measurements. Others have used the Wayback Machine for historical measurements to predict whether websites will become malicious [62] and to study JavaScript inclusion [55] and website accessibility [26]; to recover medical references [67]; to analyze social trends [35]; and as evidence in legal cases [17]. Others [53] found that

websites are accurately reflected in the archive. These studies noted similar limitations as we did, as well as ways it has changed over time [38]. Finally, researchers have studied other aspects of the web and Internet longitudinally without the use of archives, including IPv6 adoption [12], search-engine poisoning [47], privacy notices [52], and botnets [68].

7 Conclusion

Though third-party web tracking and its associated privacy concerns have received attention in recent years, the practice long predates the first academic measurements studies of tracking (begun in 2005). Indeed, in our measurements we find tracking behaviors as early as 1996. We introduce TrackingExcavator, a measurement infrastructure for third-party web tracking behaviors that leverages `archive.org`'s Wayback Machine to conduct historical studies. We rigorously evaluate the Wayback Machine's view of past third-party requests and develop strategies for overcoming its limitations.

We then use TrackingExcavator to conduct the most extensive longitudinal study of the third-party web tracking ecosystem to date, retrospectively from 1996 to present (2016). We find that the web tracking ecosystem has expanded in scope and complexity over time: today's users browsing the web's popular sites encounter more trackers, with more complex behaviors, with wider coverage, and with more connections to other trackers, than at any point in the past 20 years. We argue that understanding the trends in the web tracking ecosystem over time — provided for the first time at this scale by our work — is important to future discussions surrounding web tracking, both technical and political.

Beyond web tracking, there are many questions about the history and evolution of the web. We believe our evaluation of the Wayback Machine's view of the past, as well as TrackingExcavator, which we plan to release with this paper, will aid future study of these questions.

Acknowledgements

We thank individuals who generously offered their time and resources, and organizations and grants that support us and this work. Jason Howe of UW CSE offered invaluable technical help. Camille Cobb, Peter Ney, Will Scott, Lucy Simko, and Paul Vines read our drafts thoroughly and gave insightful feedback. We thank our colleagues from the UW Tech Policy Lab, particularly Ryan Calo and Emily McReynolds, for their thoughts and advice. This work was supported in part by NSF Grants CNS-0846065 and IIS-1302709, an NSF Graduate Research Fellowship under Grant No. DGE-1256082, and the Short-Dooley Professorship.

References

[1] Disney absorbs Infoseek, July 1999. `http://money.cnn.com/1999/07/12/deals/disney/`.

[2] Grad School Rankings, Engineering Specialties: Computer, 1999. `https://web.archive.org/web/19990427094034/http://www4.usnews.com/usnews/edu/beyond/gradrank/gbengsp5.htm`.

[3] ACAR, G., EUBANK, C., ENGLEHARDT, S., JUAREZ, M., NARAYANAN, A., AND DIAZ, C. The Web Never Forgets: Persistent Tracking Mechanisms in the Wild. In *Proceedings of the ACM Conference on Computer and Communications Security* (2014).

[4] ACAR, G., JUAREZ, M., NIKIFORAKIS, N., DIAZ, C., GÜRSES, S., PIESSENS, F., AND PRENEEL, B. FPDetective: Dusting the web for fingerprinters. In *20th ACM Conference on Computer and Communications Security* (2013), ACM.

[5] AINSWORTH, S. G., NELSON, M. L., AND VAN DE SOMPEL, H. Only One Out of Five Archived Web Pages Existed as Presented. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media* (2015), ACM, pp. 257–266.

[6] AKKUS, I. E., CHEN, R., HARDT, M., FRANCIS, P., AND GEHRKE, J. Non-tracking web analytics. In *Proceedings of the ACM Conference on Computer and Communications Security* (2012).

[7] BARTH, A. HTTP State Management Mechanism, Apr. 2011. `https://tools.ietf.org/html/rfc6265`.

[8] BAU, J., MAYER, J., PASKOV, H., AND MITCHELL, J. C. A Promising Direction for Web Tracking Countermeasures. In *Web 2.0 Security and Privacy* (2013).

[9] BRUNELLE, J. F. 2012-10-10: Zombies in the Archives. `http://ws-dl.blogspot.com/2012/10/2012-10-10-zombies-in-archives.html`.

[10] BRUNELLE, J. F., KELLY, M., SALAHELDEEN, H., WEIGLE, M. C., AND NELSON, M. L. Not All Mementos Are Created Equal : Measuring The Impact Of Missing Resources Categories and Subject Descriptors. *International Journal on Digital Libraries* (2015).

[11] CHROMIUM. CookieMonster. `https://www.chromium.org/developers/design-documents/network-stack/cookiemonster`.

[12] CZYZ, J., ALLMAN, M., ZHANG, J., IEKEL-JOHNSON, S., OSTERWEIL, E., AND BAILEY, M. Measuring IPv6 Adoption. *ACM SIGCOMM Computer Communication Review 44*, 4 (2015), 87–98.

[13] D. KRISTOL, L. M. HTTP State Management Mechanism, Oct. 2000. `https://tools.ietf.org/html/rfc2965.html`.

[14] DHAWAN, M., KREIBICH, C., AND WEAVER, N. The Priv3 Firefox Extension. `http://priv3.icsi.berkeley.edu/`.

[15] ECKERSLEY, P. How unique is your web browser? In *Proceedings of the International Conference on Privacy Enhancing Technologies* (2010).

[16] ELECTRONIC FRONTIER FOUNDATION. Privacy Badger. `https://www.eff.org/privacybadger`.

[17] ELTGROTH, D. R. Best Evidence and the Wayback Machine: a Workable Authentication Standard for Archived Internet Evidence. *78 Fordham L. Rev. 181*. (2009), 181–215.

[18] ENGLEHARDT, S., EUBANK, C., ZIMMERMAN, P., REISMAN, D., AND NARAYANAN, A. OpenWPM: An automated platform for web privacy measurement. Tech. rep., Princeton University, Mar. 2015.

[19] ENGLEHARDT, S., REISMAN, D., EUBANK, C., ZIMMERMAN, P., MAYER, J., NARAYANAN, A., AND FELTEN, E. W. Cookies That Give You Away: The Surveillance Implications of Web Tracking. In *Proceedings of the 24th International World Wide*

Web Conference (2015).

[20] EUBANK, C., MELARA, M., PEREZ-BOTERO, D., AND NARAYANAN, A. Shining the Floodlights on Mobile Web Tracking — A Privacy Survey. In *Proceedings of the IEEE Workshop on Web 2.0 Security and Privacy* (2013).

[21] FOUNDATION, P. S. 21.24. http.cookiejar Cookie handling for HTTP clients, Feb. 2015. `https://docs.python.org/3.4/library/http.cookiejar.html`.

[22] FREDRIKSON, M., AND LIVSHITS, B. RePriv: Re-Envisioning In-Browser Privacy. In *Proceedings of the IEEE Symposium on Security and Privacy* (2011).

[23] GHOSTERY. Ghostery. `https://www.ghostery.com`.

[24] GUHA, S., CHENG, B., AND FRANCIS, P. Challenges in measuring online advertising systems. In *Proceedings of the ACM Internet Measurement Conference* (2010).

[25] GUHA, S., CHENG, B., AND FRANCIS, P. Privad: Practical Privacy in Online Advertising. In *Proceedings of the USENIX Symposium on Networked Systems Design and Implementation* (2011).

[26] HACKETT, S., PARMANTO, B., AND ZENG, X. Accessibility of Internet Websites Through Time. In *Proceedings of the 6th International ACM SIGACCESS Conference on Computers and Accessibility* (New York, NY, USA, 2004), Assets '04, ACM, pp. 32–39.

[27] HAN, S., JUNG, J., AND WETHERALL, D. A Study of Third-Party Tracking by Mobile Apps in the Wild. Tech. Rep. UW-CSE-12-03-01, University of Washington, Mar. 2012.

[28] HAN, S., LIU, V., PU, Q., PETER, S., ANDERSON, T. E., KRISHNAMURTHY, A., AND WETHERALL, D. Expressive Privacy Control with Pseudonyms. In *SIGCOMM* (2013).

[29] HANNAK, A., SAPIEŻYŃSKI, P., KAKHKI, A. M., KRISHNAMURTHY, B., LAZER, D., MISLOVE, A., AND WILSON, C. Measuring Personalization of Web Search. In *Proceedings of the International World Wide Web Conference* (2013).

[30] IHM, S., AND PAI, V. Towards Understanding Modern Web Traffic. In *Proceedings of the ACM Internet Measurement Conference* (2011).

[31] INTERNET ARCHIVE. Wayback Machine. `https://archive.org/`.

[32] JACKSON, C., BORTZ, A., BONEH, D., AND MITCHELL, J. C. Protecting Browser State From Web Privacy Attacks. In *Proceedings of the International World Wide Web Conference* (2006).

[33] JANG, D., JHALA, R., LERNER, S., AND SHACHAM, H. An empirical study of privacy-violating information flows in JavaScript web applications. In *Proceedings of the ACM Conference on Computer and Communications Security* (2010).

[34] JENSEN, C., SARKAR, C., JENSEN, C., AND POTTS, C. Tracking website data-collection and privacy practices with the iWatch web crawler. In *Proceedings of the Symposium on Usable Privacy and Security* (2007).

[35] JOHN, N. A. Sharing and Web 2.0: The emergence of a keyword. *New Media & Society* (2012).

[36] JONES, S. M., NELSON, M. L., SHANKAR, H., AND DE SOMPEL, H. V. Bringing Web Time Travel to MediaWiki: An Assessment of the Memento MediaWiki Extension. *CoRR abs/1406.3876* (2014).

[37] KAMKAR, S. Evercookie — virtually irrevocable persistent cookies. `http://samy.pl/evercookie/`.

[38] KELLY, M., BRUNELLE, J. F., WEIGLE, M. C., AND NELSON, M. L. On the Change in Archivability of Websites Over Time. *CoRR abs/1307.8067* (2013).

[39] KONTAXIS, G., POLYCHRONAKIS, M., KEROMYTIS, A. D., AND MARKATOS, E. P. Privacy-preserving social plugins. In *USENIX Security Symposium* (2012).

[40] KRISHNAMURTHY, B., NARYSHKIN, K., AND WILLS, C. Privacy Leakage vs. Protection Measures: The Growing Disconnect. In *Proceedings of the IEEE Workshop on Web 2.0 Security and Privacy* (2011).

[41] KRISHNAMURTHY, B., AND WILLS, C. On the leakage of personally identifiable information via online social networks. In *Proceedings of the ACM Workshop on Online Social Networks* (2009).

[42] KRISHNAMURTHY, B., AND WILLS, C. Privacy Diffusion on the Web: a Longitudinal Perspective. In *Proceedings of the International World Wide Web Conference* (2009).

[43] KRISHNAMURTHY, B., AND WILLS, C. E. Generating a Privacy Footprint on the Internet. In *Proceedings of the ACM Internet Measurement Conference* (2006).

[44] KRISTOL, D., AND MONTULLI, L. RFC 2109 - HTTP State Management Mechanism, 1997. `https://tools.ietf.org/html/rfc2109`.

[45] LÉCUYER, M., DUCOFFE, G., LAN, F., PAPANCEA, A., PETSIOS, T., SPAHN, R., CHAINTREAU, A., AND GEAMBASU, R. XRay: Enhancing the Web's Transparency with Differential Correlation. In *23rd USENIX Security Symposium* (2014).

[46] LEON, P. G., UR, B., WANG, Y., SLEEPER, M., BALEBAKO, R., SHAY, R., BAUER, L., CHRISTODORESCU, M., AND CRANOR, L. F. What Matters to Users? Factors that Affect Users' Willingness to Share Information with Online Advertisers. In *Symposium on Usable Privacy and Security* (2013).

[47] LEONTIADIS, N., MOORE, T., AND CHRISTIN, N. A nearly four-year longitudinal study of search-engine poisoning. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security* (2014), ACM, pp. 930–941.

[48] LUND, A. The History of Online Ad Targeting, 2014. `http://www.sojern.com/blog/history-online-ad-targeting/`.

[49] MAYER, J., AND NARAYANAN, A. Do Not Track. `http://donottrack.us/`.

[50] MAYER, J. R., AND MITCHELL, J. C. Third-Party Web Tracking: Policy and Technology. In *Proceedings of the IEEE Symposium on Security and Privacy* (2012).

[51] MCDONALD, A. M., AND CRANOR, L. F. Americans' Attitudes about Internet Behavioral Advertising Practices. In *Proceedings of the Workshop on Privacy in the Electronic Society* (2010).

[52] MILNE, G. R., AND CULNAN, M. J. Using the content of online privacy notices to inform public policy: A longitudinal analysis of the 1998-2001 US Web surveys. *The Information Society 18*, 5 (2002), 345–359.

[53] MURPHY, J., HASHIM, N. H., AND OCONNOR, P. Take Me Back: Validating the Wayback Machine. *Journal of Computer-Mediated Communication 13*, 1 (2007), 60–75.

[54] NARAINE, R. Windows XP SP2 Turns 'On' Pop-up Blocking, 2004. `http://www.internetnews.com/dev-news/article.php/3327991`.

[55] NIKIFORAKIS, N., INVERNIZZI, L., KAPRAVELOS, A., VAN ACKER, S., JOOSEN, W., KRUEGEL, C., PIESSENS, F., AND VIGNA, G. You Are What You Include: Large-scale Evaluation of Remote Javascript Inclusions. In *Proceedings of the ACM Conference on Computer and Communications Security* (2012).

[56] NIKIFORAKIS, N., JOOSEN, W., AND LIVSHITS, B. Privaricator: Deceiving fingerprinters with little white lies. In *Proceedings of the 24th International Conference on World Wide Web* (2015), International World Wide Web Conferences Steering Committee, pp. 820–830.

[57] NIKIFORAKIS, N., KAPRAVELOS, A., JOOSEN, W., KRUEGEL, C., PIESSENS, F., AND VIGNA, G. Cookieless Monster: Exploring the Ecosystem of Web-based Device Fingerprinting. In *Proceedings of the IEEE Symposium on Security and Privacy* (2013).

[58] RESEARCH LIBRARY OF LOS ALAMOS NATIONAL LABORATORY. Time Travel. http://timetravel.mementoweb.org/about/.

[59] REZNICHENKO, A., AND FRANCIS, P. Private-by-Design Advertising Meets the Real World. In *Proceedings of the ACM Conference on Computer and Communications Security* (2014).

[60] ROESNER, F., KOHNO, T., AND WETHERALL, D. Detecting and Defending Against Third-Party Tracking on the Web. In *Proceedings of the USENIX Symposium on Networked Systems Design and Implementation* (2012).

[61] ROESNER, F., ROVILLOS, C., KOHNO, T., AND WETHERALL, D. ShareMeNot: Balancing Privacy and Functionality of Third-Party Social Widgets. *USENIX ;login: 37* (2012).

[62] SOSKA, K., AND CHRISTIN, N. Automatically detecting vulnerable websites before they turn malicious. In *23rd USENIX Security Symposium (USENIX Security 14)* (2014), pp. 625–640.

[63] STEVEN ENGLEHARDT. Do privacy studies help? A Retrospective look at Canvas Fingerprinting. https://freedom-to-tinker.com/blog/englehardt/retrospective-look-at-canvas-fingerprinting/.

[64] TOUBIANA, V., NARAYANAN, A., BONEH, D., NISSENBAUM, H., AND BAROCAS, S. Adnostic: Privacy Preserving Targeted Advertising. In *Proceedings of the Network and Distributed System Security Symposium* (2010).

[65] UR, B., LEON, P. G., CRANOR, L. F., SHAY, R., AND WANG, Y. Smart, useful, scary, creepy: perceptions of online behavioral advertising. In *8th Symposium on Usable Privacy and Security* (2012).

[66] VISSERS, T., NIKIFORAKIS, N., BIELOVA, N., AND JOOSEN, W. Crying wolf? on the price discrimination of online airline tickets. In *HotPETS* (2014).

[67] WAGNER, C., GEBREMICHAEL, M. D., TAYLOR, M. K., AND SOLTYS, M. J. Disappearing act: decay of uniform resource locators in health care management journals. *Journal of the Medical Library Association : JMLA 97*, 2 (2009), 122–130.

[68] WANG, D. Y., SAVAGE, S., AND VOELKER, G. M. Juice: A Longitudinal Study of an SEO Botnet. In *NDSS* (2013).

[69] WASHINGTON POST. From Lycos to Ask Jeeves to Facebook: Tracking the 20 most popular web sites every year since 1996. https://www.washingtonpost.com/news/the-intersect/wp/2014/12/15/from-lycos-to-ask-jeeves-to-facebook-tracking-the-20-most-popular-web-sites-every-year-since-1996/.

[70] WILLS, C. E., AND TATAR, C. Understanding what they do with what they know. In *Proceedings of the ACM Workshop on Privacy in the Electronic Society* (2012).

[71] YEN, T.-F., XIE, Y., YU, F., YU, R. P., AND ABADI, M. Host Fingerprinting and Tracking on the Web: Privacy and Security Implications. In *Proceedings of the Network and Distributed System Security Symposium* (2012).

[72] ZACK WHITTAKER. PGP co-founder: Ad companies are the biggest privacy problem today, not governments, 2016. www.zdnet.com/article/pgp-co-founder-the-biggest-privacy-issue-today-are-online-ads/.

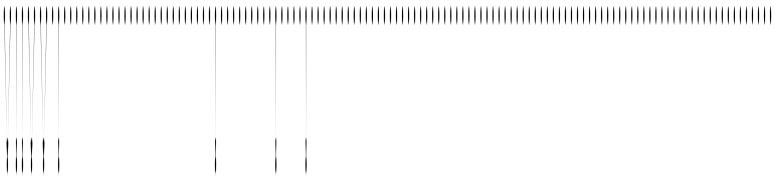
A Fingerprint-Related JavaScript APIs

As described in Section 3, TrackingExcavator hooks a number of JavaScript APIs that may be used in fingerprint-based tracking and drawn from prior work [3, 4, 15, 56, 57]. The complete list:

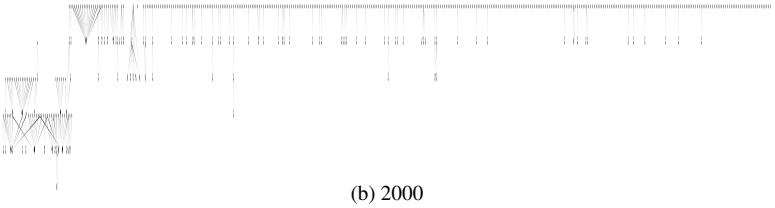
- `navigator.appCodeName`
- `navigator.appName`
- `navigator.appVersion`
- `navigator.cookieEnabled`
- `navigator.doNotTrack`
- `navigator.language`
- `navigator.languages`
- `navigator.maxTouchPoints`
- `navigator.mediaDevices`
- `navigator.mimeTypes`
- `navigator.platform`
- `navigator.plugins`
- `navigator.product`
- `navigator.productSub`
- `navigator.userAgent`
- `navigator.vendor`
- `navigator.vendorSub`
- `screen.availHeight`
- `screen.availLeft`
- `screen.availTop`
- `screen.availWidth`
- `screen.colorDepth`
- `screen.height`
- `screen.orientation`
- `screen.pixelDepth`
- `screen.width`
- `CanvasRenderingContext2D.getImageData`
- `CanvasRenderingContext2D.fillText`
- `CanvasRenderingContext2D.strokeText`
- `WebGLRenderingContext.getImageData`
- `WebGLRenderingContext.fillText`
- `WebGLRenderingContext.strokeText`
- `HTMLCanvasElement.toDataURL`
- `window.TouchEvent`
- `HTMLElement.offsetHeight`
- `HTMLElement.offsetWidth`
- `HTMLElement.getBoundingClientRect`

B Ecosystem Complexity

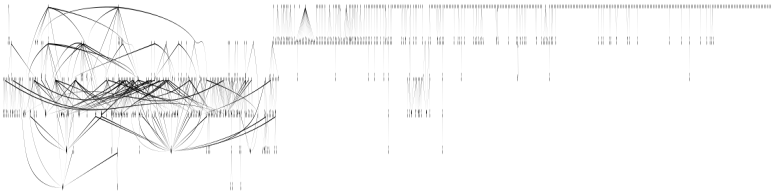
Figure 12 (on the next page) visually depicts the connections between entities in the tracking ecosystem that we observe in our datasets for 1996, 2000, 2004, 2008, 2012, and 2016: domains as nodes, and referral relationships as edges. Note that the visual organization of these graphs (with nodes in multiple tiers) is not meaningful and simply an artifact of the graph visualization software. Over time, the complexity and interconnectedness of relationships between third-party domains on the top 450 websites has increased dramatically.



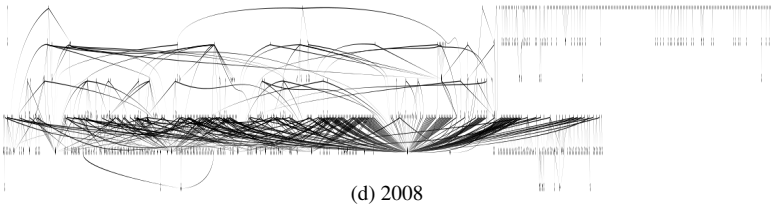
(a) 1996



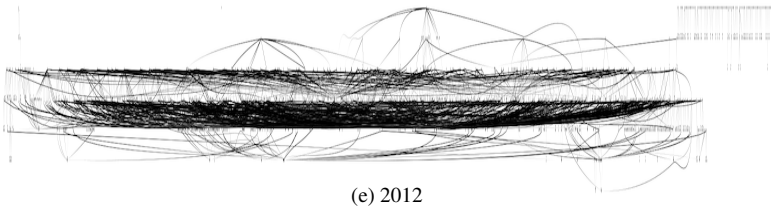
(b) 2000



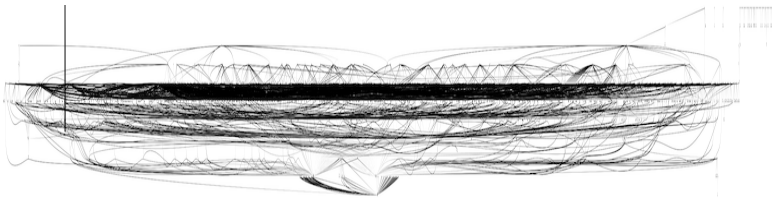
(c) 2004



(d) 2008



(e) 2012



(f) 2016

Figure 12: Referrer graphs for the top 450 sites in 1996, 2000, 2004, 2008, 2012 and 2016 as seen in the Wayback Machine's archive. An edge from a domain `referrer.com` to another domain `referred.com` is included if any URL from `referrer.com` is seen to be the referrer for any request to `referred.com`. Note the increasing complexity of the graph over time.