First and last name, email address, and phone number of researcher(s) making the Request.

Ibrahim Altaweel
abealtaweel@eecs.berkeley.edu
925-413-0097

Nathaniel Good
Nathan@goodresearch.com
(510) 527-1537

An abstract, draft, or completed research paper describing your privacy or data security research:

Abstract: Web Privacy Census v 3.0, by Nathaniel Good and Ibrahim Altaweel

Please note: we will provide the final paper on request, we are finalizing it this week.

Public policymakers are proposing measures to give consumers more privacy rights online. These measures are based upon the assumption that the web privacy landscape has become worse for consumers; that their online activities are tracked more pervasively now than they were in the past. As policymakers consider different approaches for addressing internet privacy, it is critical to understand how interventions such as negative press attention, self-regulation, Federal Trade Commission enforcement actions, and direct regulation affect tracking. As early as 1995, Beth Givens of the Privacy Rights Clearinghouse suggested that federal agencies create benchmarks for online privacy. The first attempts of web measurement, discussed in our literature review, found relatively little tracking online in 1997–only 23 of the most popular websites were using cookies on their homepages. But within a few years, tracking for network advertising was present on many websites, and by 2011, all of the most popular websites employed cookies.

The Web Privacy Census is intended to formalize the benchmarking process and measure internet tracking consistently over time. We seek to explore:

- How many entities are tracking users online?
- What vectors (technologies) are most popular for tracking users?
- Is there displacement (i.e. a shift from one tracking technology to another) in tracking practices?
- Is there greater concentration of tracking companies online?
- What entities have the greatest potential for online tracking and why?

Our Findings

In the October 2015 Web Privacy Census, we found that users who merely visit the homepages of the top 100 most popular sites would collect over 6,000 HTTP cookies in the process—twice as many as we detected in 2012. If the user browsed to just two more links, the number of HTTP cookies would double. Eighty-three percent of cookies were set by third party hosts, and just in visiting the homepage of popular sites, users would have cookies placed by 275 third-party hosts.

Some popular websites use a lot of cookies. In just visiting the homepage of popular sites, we found that we found 24 websites that placed over 100 cookies, 6 websites that placed over 200 cookies, and 3 websites placed over 300.

We also found that more sites are using HTML5 storage, which enables websites to store more information about consumers.

Google, through its various affiliated companies and services, has tracking infrastructure on 85 of the top 100 most popular websites, and 844 of the top 1,000 websites. This means that Google's ability to track on popular websites is unparalleled and it approaches the level of surveillance that only an ISP can achieve. While it is claimed that this tracking is not identifiable, any user who ever authenticates (signs in) to a Google service provides directly personally-identifiable information than can be linked across services, and backwards in time through Google's retained records.

Our Methods

Data were collected on the United States top 100, 1,000 and 25,000 websites as ranked on Quantcast's top 1 million websites in the United States. These data were collected using two processes: 1) A shallow automated crawl of the top 100, 1000, and 25,000 sites, which consisted of visiting only the homepage of the domain obtained from Quantcast's rankings, and 2) A deep automated crawl of the top 100 and 1000 sites which consisted of visiting the homepage and 2 randomly selected links from the homepage. After visiting the first link, the crawler returned to the homepage before selecting the second link. Both links were on the same domain as the homepage.

The Crawler

The crawler used was OpenWPM, a flexible and scalable platform written in Python. This crawler offers features such as collecting HTTP cookies, Flash cookies, HTML5 local storage objects, and the ability to perform deep crawls by visiting links. OpenWPM allows the crawl to be run in either Firefox or Chrome, and can be run with or without add-ons.

All crawls were run using a Firefox version 39 browser with no add-ons, with Flash turned on, and in headless mode. The following information was collected from each crawled domain visit: HTTP cookies, HTML5 local storage objects, Flash cookies, and HTTP requests and responses (including headers). Each crawl was run four times and the average was taken for each tracking method.

Shallow Automated Crawl

The shallow crawls were run with a clean browser instance that was cleared of all tracking data. The crawler visited each URL homepage, waited for the page to load, and then dumped all tracking data obtained from that URL into a database. The crawler would then close that browser tab, open a new tab, then continue this process with the next URL on the Quantcast list.

Deep Automated Crawl

The deep crawls were run with a clean browser instance that was cleared of all tracking data. The crawler visited each URL homepage and waited for the page to load. It would then randomly select a link on the homepage and visit that site. After the linked page finished loading, the crawler would go back to the previous page and visit a second randomly selected link. After the second link finished loading, the crawler would dump all tracking data obtained from those three URLs into a database. The crawler would then close that browser tab, open a new tab, then continue this process with the next URL on the Quantcast list.

LIMITATIONS

Limitations of crawler methods

For the October 2015 report, the crawler did not "log in" to any sites, nor bypass any modal dialogs, and therefore our data does not record how cookies changed based on additional information provided by users logging into third party services or requesting further access to the main site. Additionally, as the crawler automated selection of URLs for deep crawls, any retargeting that was based on a human action (e.g. adding items to a shopping cart) was not necessarily captured in this crawl. Deep crawls were also limited to HTML anchor tags found and did not follow links set by JavaScript. Additionally, links obtained by the Deep Crawler were selected at random from links on the page, and consequently did not take into account page layout and visual layout in the selection process. The crawl was run using Firefox with no add-ons.

Limitations of data collection methods

The identification and classification of third and first party cookies is a complex task. Many tracking and advertising companies are owned by other sites that have different domain names. For example, DoubleClick is owned by Google. For consistency in categorizing third party cookies, the public suffix list was leveraged to combine suffixes consistent with previous work. Cookies from the top level domain were classified as first party, while cookies from a domain outside of the top level domain were classified as third party. Analysis of third party domains is therefore limited to domains that are syntactically considered to be third parties, and not reflective of any underlying agreements or connections that may exist between multiple domains, through "DNS aliasing," for instance, where a primary domain assigns a subdomain to a tracking company. Under such an arrangement, ordinary third party cookies would be instantiated in a first-party fashion. The ranking list used was Quantcast's top 1 million sites in the United States. This ranking may be different in other countries.

How our research differs from prior research in the area.

Our research represents an academic effort to quantify how websites are tracking users and the amount of tracking that occurs online. It has no commercial sponsors or influence.

How our work satisfies the selection criteria:

- This is our own research that we designed and conducted.
- We receive no corporate funding for this report. We are supported by NSF-TRUST.
- This research presents objective facts about the presence of HTTP cookies, HTML5, and Flash objects in the browser, rather than our opinions about the internet.
- There is no promotional or commercial aspect to our research or our intended presentation.
- Our research does not present any security vulnerabilities.
- Our research is the product of a multi-year, academic effort. It has been reported upon by the technical press. No one, to our knowledge, has criticized its methods or findings.