Web Privacy Census: HTML5 Storage Takes the Spotlight As Flash Returns

Ibrahim Altaweel[1], Jaime Cabrera[2], Hen Su Choi[3], Katie Ho[4], Nathan Good[5], Chris Hoofnagle[5]

Diablo Valley College[1], California State University Fullerton[2], University of California San Diego[3], Mount Holyoke College[4], University of California Berkeley[5]

OVERVIEW

The purpose of the Web Privacy Census is to measure internet tracking consistently over time and raise awareness of various new techniques and tools available to users to control tracking online [1].

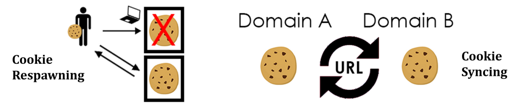
Since the 2012 report, we have noted a growth in alternative methods, besides HTTP cookies, that trackers have used to gather information from unsuspecting users on the internet. Our aim in this report is to compare data with the 2012 Web Privacy Census and discuss the patterns and trends we see surrounding the current state of web privacy.

Through this study, we sought to explore:
- How many entities are tracking users online?
- What vectors are most popular for tracking users?
- Are there displacements in tracking practices?
- What diffuse or concentrated is tracking?

BACKGROUND

Characteristics of HTTP Cookies, Flash Cookies, HTML5 Storage [2]

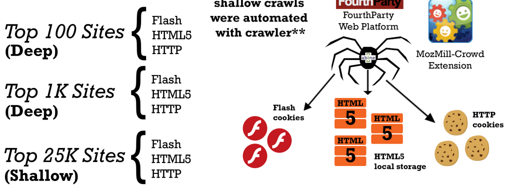
Type of Tracking Object	Storage size	Characteristics
HTTP Cookie	4 KB	• Small • Easy to delete • Expires when the session ends by default
Flash Cookie	100 KB	• Uses a common directory • Permanent by default • Can be accessed for multiple browsers
HTML5 Storage	5 MB	• Data are stored in name/value pairs • Permanent by default • More secure because data is not included with every server request, only sent when asked for



1997 -- First attempts of web measurement (only 23 of the most popular websites were using cookies on their homepages)
2009 -- Flash cookies present on popular sites to track users, evidence of cookie respawning and cookie syncing [2]
2011 -- HTTP cookies present on all popular websites
2012 -- Decrease in Flash cookies present on popular websites

METHODS

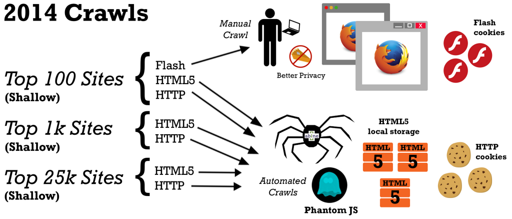
2012 Crawls



Top 100 Sites (Deep): Flash, HTML5, HTTP
Top 1K Sites (Deep): Flash, HTML5, HTTP
Top 25K Sites (Shallow): Flash, HTML5, HTTP

Both deep and shallow crawls were automated with crawler

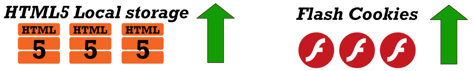
2014 Crawls



Top 100 Sites (Shallow): Flash, HTML5, HTTP
Top 1k Sites (Shallow): HTML5, HTTP
Top 25k Sites (Shallow): HTML5, HTTP

The crawler runs using PhantomJS and Webkit, an open source browser engine. PhantomJS is a Java Script tool that deploys headless web browsers to simulate user activity. However, because headless web browsing has some limitations, including not being able to collect Flash cookies, we conducted multiple crawls and manual samples to help yield more accurate results.

Compared to the deep crawls of various sets of sites we ran in 2012, for the 2014 report, we only administered shallow crawls with larger sets of sites. Our shallow crawl of the top 25,000 sites revealed that 88% had HTTP cookies, and 35% had HTML5 storage objects. There was also a marked increase in Flash cookies.



This work was supported in part by TRUST, Team for Research in Ubiquitous Secure Technology, which receives support from the National Science Foundation (NSF award number CCF-0424422). Special thanks to Chris Hoofnagle, Nathan Good, Abine, Aimee Tabor, & the other TRUST REU interns for their guidance and contributions to our research.

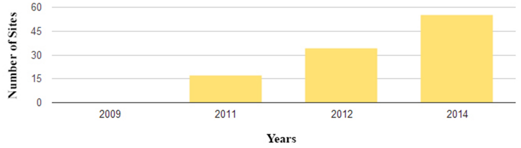
References

[1] Hoofnagle, Chris Jay, and Nathan Good. "Web Privacy Census." Available at SSRN (2012).
[2] Ayenson, Mika, Dietrich J. Wambach, Ashkan Soltani, Nathan Good, and Chris J. Hoofnagle. "Flash cookies and privacy II: Now with HTML5 and ETag respawning." Social Science Research Network (2011).

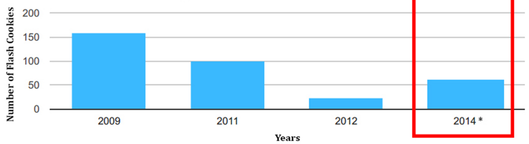
RESULTS

Compared to the data we collected for 2012, HTML5 local storage had a surprising increase in 2014. There was an unexpected increase in Flash cookies count for the top Global 100 Alexa ranked sites in the 2014 report. The top sites with increased Flash cookie count in 2014 were primarily Chinese and news websites. Therefore, it is likely that the high count of Flash cookies on Chinese sites was not captured in previous years when the sample size was the top 100 U.S. sites. Overall, after past years of decline in numbers, it is expected that Flash cookie counts will drop in future years.

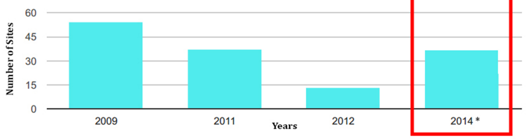
HTML5 Local Storage on Top 100 Sites



Flash Cookie Count on Top 100 Sites



Number of Top 100 Sites with Flash Cookies



*Shallow crawl: Approx. numbers compared with deep crawl

In addition to counting HTTP cookies, Flash cookies, and HTML5 local storage objects, we also collected data on the top trackers on different sets of sites. With these numbers and entity names, we were able to analyze which trackers are most concentrated on the top 100, 1,000, and 25,000 sites: doubleclick.net, scorecardresearch.com, google.com, godaddy.com and rubiconproject.com.

FUTURE WORK

Although HTTP cookies will continue to be the standard in tracking, as they carry the most compatibility with browsers, cookie counts cannot accurately represent to what extent users are being tracked. In quantifying the state of web privacy online, we examined other tracking techniques besides cookies. Fingerprinting has become more popular because it is a technique that gives trackers the ability to collect information without leaving a trace. It allows trackers to get a quick snapshot of the properties, technologies, and capabilities of a user's computer. Moving forward, reports on web privacy will need to measure tracking with multiple techniques.