WEB PRIVACY CENSUS

Our goal is to define and quantify vectors for tracking consumers on the internet. By doing this, using consistent methods over time, we will be able to make empirical statements about the state of internet tracking and privacy.

Introduction

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. This assumption may be true, as online advertising and metrics companies have developed more sophisticated ways to track and identify individuals online. This has been substantiated in the academic literature, and in the popular press through an influential news series, "What they Know," by Wall Street Journal reporters.

WEB PRIVACY CENSUS LINKS

- June 2012 Census
- Web Privacy Census Literature Review
- Web Privacy Census Methods
- Privacy Census in the News

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 literature review discusses this project and its context more fully. Key to this project is our methods, which we apply consistently over time.

This effort was developed and executed in partnership with Abine, Inc. Abine has been our technical collaborator and resource partner, helping us develop a reliable method for web crawling and analysis of tracking vectors. This project is supported by and builds upon prior research in collaboration with the National Science Foundation Team for Research in Ubiquitous Secure Technology.

Results and Discussion

This report contains data from our most recent crawl, conducted on 10/24/12, and compares it to the results of our June 2012 Web Privacy Census.

We conduct two different crawls—a shallow one where our test browser just visits the homepage of a site, and a deep crawl where our browser visits six links on a site.

We found cookies on all popular websites (by "popular websites," we mean the top 100 most popular according to Quantcast). Historically, there has been a large upswing in cookies on popular websites. When we first measured cookies in 2009, we found 3,602 cookies on popular websites, and in 2011, we found 5,675.

Here we found statistically significant upticks in tracking mechanisms from just five months ago: more popular sites are using more cookies. We found a total of 6,485 cookies on the top 100 websites; the vast majority of these cookies are from third party domains.

Berkeley Center for Law and Technology

About
Research
Law & Tech Research Portal
Privacy at BCLT
Faculty Casebooks
Tech-Academics-Policy
Students
Events
Past Events
Sponsors
Contact

Deep Crawl - Most Popular 100 Sites (six links deep)			
crawl date	5/17/12	10/24/12	trend*
Total HTTP Cookies	5,795	6,485	up ↑
Total HTTP Cookies: First Party	932	992	
Total HTTP Cookies: Third Party	4,863	5,493	up ↑
Total Flash Cookies	23	17	
Total Flash LSO: First Party	8	6	
Total Flash LSO: Third Party	15	11	
Total Session Cookies	301	259	
Total HTML5 LSO	34	38	

*We only indicate trends that are statistically significant at the .05 level or stronger.

Key Tracking Metrics - Most Popular 100 Sites			
crawl date	5/17/12	10/24/12	trend
Do all popular sites have cookies?	Yes	Yes	
Sites with 100 or more cookies	21	21	
Sites with 150 or more cookies	6	11	
Percentage of cookies set by a third party host	84%	84.7%	
Number of third party hosts	446	457	
Number of top websites with a Google presence	78	74	
Number of sites with Flash cookies	13	11	
Number of sites with HTML5 storage	34	38	

We are observing an overall downward trend in the use of Flash cookies. In 2011, 37 sites used Flash cookies. In our May 2012 crawl, 13 were, and now just 11 use Flash cookies. Websites may be changing strategies here by adopting HTML5 local storage. In 2011, when we first surveyed local storage, we found only 17 sites using HTML5. Our May 2012 crawl found 34, and now 38 sites are using HTML5 local storage.

Top Trackers - Most Popular 100 Sites	
5/17/12	10/24/12
doubleclick.net(73)	doubleclick.net(69)
scorecardresearch.com(58)	scorecardresearch.com(54)
adnxs.com(48)	bluekai.com(41)
quantserve.com(47)	atdmt.com(40)
ad.yieldmanager.com(42)	adnxs.com(40)

Google's DoubleClick leads the top trackers statistic in all three crawls.

Trackers Setting the Most Cookies - Most Popular 100 Sites	
5/17/12	10/24/12
Bluekai(321 cookies)	bluekai.com (328 cookies)
Rubiconproject.com(192)	Rubiconproject.com(242)
Adnxs.com(169)	rfihub.com(213)
Advertising.com(169)	advertising.com(211)
Pubmatic.com(164)	doubleclick.net(151)

The most frequently appearing cookie keys were: "__utma," "__utmb", "__utmc", "__utmz", and "UID." Many of these keys are commonly associated with unique user tracking and Google Analytics. For instance, __utma is used by Google for identifying unique visitors.
Our shallow crawl data indicates that by merely visiting the homepage of the most popular sites, perhaps without even receiving a privacy policy, thousands of cookies are installed.

Shallow Crawl - Most Popular 100 Sites			
crawl date	5/17/12	10/24/12	trend
Total HTTP Cookies	2616	3152	up ↑
Total HTTP Cookies: First Party	729	828	up ↑
Total HTTP Cookies: Third Party	1887	2324	up ↑
Total Flash Cookies	6	7	
Total Flash LSO: First Party	3	2	
Total Flash LSO: Third Party	3	5	
Total Session Cookies	236	257	
Total HTML5 LSO	27	34	

Top 1,000 Websites

We observed increased presence of trackers in our crawl of the top 1,000 websites as well. The total number of first and third party cookies placed on computers was up significantly.

Deep Crawl - Most Popular 1,000 Websites			
crawl date	5/17/12	10/24/12	trend
Total HTTP Cookies	62,755	65,381	up ↑
Total HTTP Cookies: First Party	8,302	8,658	up ↑
Total HTTP Cookies: Third Party	54,453	56,723	up ↑
Average HTTP Cookies: First Party	8.32	8.69	
Average HTTP Cookies: Third Party	54.61	56.95	
Total Flash Cookies	176	181	
Total Flash LSO: First Party	44	41	
Total Flash LSO: Third Party	132	140	
Total Session Cookies	2,767	2,448	down ↓
Total HTML5 LSO	311	318	

Key tracking metrics remains level among the top 1,000 websites.

Key Tracking Metrics - Most Popular 1,000 Websites			
crawl date	5/17/12	10/24/12	trend
Percentage of sites with cookies	97.4%	97.9%	
Sites with 100 or more cookies	191	198	
Sites with 150 or more cookies	117	114	
Sites with 150 or more cookies	87%	86%	
Number of sites with a Google presence	712	733	
Number of sites with Flash cookies	110	97	
Number of sites with HTML5	311	318	

The trackers present in the top 1,000 sites are consistent with those predominating the top 100.

Most Prevalent Trackers - Most Popular 1,000 Sites	
5/17/12	10/24/12
Doubleclick.net(685 sites)	Doubleclick.net(681 sites)
Scorecardresearch.com(489)	Scorecardresearch.com(475)
Adnxs.com(404)	Adnxs.com(439)
Quantserve.com(445)	Quantserve.com(409)
Admt.com(385)	Admt.com(391)

Trackers Setting the Most Cookies - Most Popular 1,000 Sites	
5/17/12	10/24/12
Bluekai(2,906 cookies)	Bluekai(2,562 cookies)
Rubiconproject.com(2,049)	Rubiconproject.com(2,470)
Pubmatic.com(1,673)	rfihub.com(2005)
Doubleclick.net(1,539)	Pubmatic.com(1941)
Adnxs.com(1,505)	Adnxs.com(1555)

The most frequently appearing cookie keys were: "__utmb," "__utma," "__utmc," "__utmz," and "UID"

Top 25,000 Websites

Our crawl of the top 25,000 websites is shallow—we only visit the homepage of these websites. The goal was to get a basic understanding of cookie counts for a wider range of sites to develop an understanding of trackers in the long tail.

Shallow Crawl - Most Popular 25,000 Sites			
crawl date	5/17/12	10/24/12	trend
Total HTTP Cookies	442047	476492	up ↑
Total HTTP Cookies: First Party	108,044	111,069	up ↑
Total HTTP Cookies: Third Party	334,003	365,423	up ↑

Total Flash Cookies	441	454	
Total Flash LSO: First Party	136	115	
Total Flash LSO: Third Party	305	339	
Total Session Cookies	33,404	33,918	up ↑
Total HTML5 LSO	2,417	2,758	up ↑

We saw an increase in the number of sites that placed 150 or more cookies.

Key Tracking Metrics - Most Popular 25,000 Websites			
crawl date	5/17/12	10/24/12	trend
Percentage of sites with cookies	87%	87%	up ↑
Sites with100 or more cookies	730	771	
Sites with 150 or more cookies	133	267	up ↑
Percentage of cookies set by a third party host	76%	76%	
Number of sites with a Google presence	8,993	9252	
Number of sites with Flash cookies	344	351	
Number of sites with HTML5	2417	2758	up ↑

Most Prevalent Trackers - Most Popular 25,000 Sites	
5/17/12	10/24/12
Doubleclick.net(8,554 sites)	Doubleclick.net(8,855 sites)
Quantserve.com(4,817)	Scorecardresearch.com(4,759 sites)
Scorecardresearch.com(4,565)	Quantserve.com(4,653 sites)
Adnxs.com(3,249)	Adnxs.com(4,557 sites)
Twitter.com(2,475)	Invitemedia.com(3,318 sites)

Trackers Setting the Most Cookies - Most Popular 25,000 Sites	
5/17/12	10/24/12
Bluekai(18,142 cookies)	Doubleclick.net(17,690 cookies)
Doubleclick.net(16,832)	Bluekai(17,158 cookies)
Adnxs.com(9,540)	Adnxs.com(12,611 cookies)
Scorecardresearch.com(9,402)	Addthis.com(11,603 cookies)
Casalemedia.com(9,392)	Rubiconproject.com(10,056 cookies)

The most frequently appearing cookie keys were: "__utmb," "__utma," "__utmc," "__utmz," "UID."

Conclusion

In this first update to our original June 2012 Web Privacy Census, we observed statistically signifiant increases in the amount of tracking on all three of our samples--the top 100, 1,000, and top 25,000 websites. Flash cookies use is declining among the most popular websites, and HTML5 local storage is rising across all three groups.

Sponsors

ABINE the online privacy company
Good Research
TROST

How to cite this report:

Chris Jay Hoofnagle & Nathan Good, *The Web Privacy Census* , October 2012, available at http://law.berkeley.edu/privacycensus.htm

Berkeley, CA 94720-7200
510-642-1741

Feedback
UC Berkeley
For Faculty & Staff
For Employers
Social Media
About This Site