

Home » Berkeley Center for Law and Technology » Research » Privacy at BCLT » Web Privacy
 Census » Literature Review

LITERATURE REVIEW

Web privacy measurement is a nascent field, with significant contributions developed by academic computer scientists and others interested in discovering tracking vectors and quantifying them. At [Web Privacy Measurement 2012](#), leaders in the field will attempted to formalize these efforts.

The Electronic Privacy Information Center made the earliest attempts to enumerate privacy practices in a systematic fashion. In June 1997, it released [Surfer Beware: Personal Privacy and the Internet](#), a survey of the top 100 websites. Only 17 of the top 100 websites had privacy policies. Twenty-three sites used cookies, although it appears that EPIC used a “surface crawl” to detect those cookies, meaning that it only visited the homepage of the site and did not click other links. By 2009, Soltani et al. found cookies on 98 of the top 100 sites, and by 2011, Ayenson et al. found cookies on all 100 most popular sites (see discussion below).

In [Surfer Beware II: Notice is Not Enough](#), published in June 1998, EPIC surveyed websites of companies that had recently joined the Direct Marketing Association. At the time, the Direct Marketing Association (DMA) had committed to basic privacy protections, including notice and an ability for consumers to opt out. EPIC found that there were 76 new members of the DMA, but only 40 had websites. Of those 40, all collected personal information. Only eight of the sites had a privacy policy

The Federal Trade Commission conducted the first large-scale privacy measurement study in [Privacy Online: A Report to Congress](#). Released in June 1998, the Commission studied the privacy practices of 1,402 websites, using a sophisticated sample procedure to ensure that a variety of consumer-oriented websites were studied (health, retail, financial, sites directed to children, and the most popular websites). The FTC found that, “the vast majority of Web sites -- upward of 85% -- collect personal information from consumers. Few of the sites -- only 14% in the Commission’s random sample of commercial Web sites -- provide any notice with respect to their information practices, and fewer still -- approximately 2% -- provide notice by means of a comprehensive privacy policy.”

In EPIC’s [Surfer Beware III: Privacy Policies without Privacy Protection](#), the group surveyed the practices of 100 e-commerce sites. This was the most comprehensive, but last of the EPIC surveys. It evaluated sites for compliance with a full range of fair information practices, such as whether the site collected personal information, whether the site linked to a privacy policy, whether the site had agreed to a seal program, and whether users had access and correction rights for personal information. Eighty-six of the sites used cookies, 18 lacked privacy policies, and 35 had some form of network advertiser active on the site. The text of the report makes it clear that EPIC evaluated both the privacy politics of these sites and tested them to see whether they were setting cookies. However, it is unclear whether EPIC performed a surface crawl of just the homepage or a deeper crawl that explored more of the site.

In May 2000, the Federal Trade Commission released a [survey](#) of sites that detected third party cookies. In its study, the FTC drew from two groups of websites: those with over 39,000 visits a month and a second sample of popular sites (91 of the top 100). The FTC found that, “57% of the sites in the Random Sample and 78% of the sites in the Most Popular Group allow the placement of cookies by third parties.... The majority of the third-party cookies in the Random Sample and in the Most Popular Group are from network advertising companies that engage in online profiling.”

In a multiple-year [study](#) of 1,200 websites, Bala Krishnamurthy and Craig Wills found increasing collection of information about users from an increasingly concentrated group of tracking companies. Krishnamurthy and Wills describe what we call “DNS aliasing” in their paper (this was also described in their 2006 paper), a practice where, “...what appeared to be a server in one organization (e.g. [w88.go.com](#)) was actually a DNS CNAME alias to a server ([go.com.112.207.net](#)) in another organization (Omniture).” They found a massive increase in such aliasing: “...the percentage of first-party servers with multiple top third-party domains has risen from 24% in Oct’05 to 52% in Sep’08...This increase is significant because it shows that now for a majority of these first-party servers, users are being tracked by two and more third-party entities.” It is also significant because through DNS aliasing, tracking companies can present cookies to users directly as first parties, thereby circumventing third party cookie blocking.

Through decoding aliased domains, Krishnamurthy and Wills found that third party trackers were becoming more concentrated. Sampling from five periods, concentration grew from 40% in October 2005 to 70% in September 2008. Further, they found that, “The overall share of the top-five families: Google, Omniture, Microsoft, Yahoo and AOL extends to more than 75% of our core test set with Google alone having a penetration of nearly 60%.”

In June 2009, Gomez et al. published the [KnowPrivacy report](#). The report focused on several areas of consumer privacy, and featured a large-scale crawl of sites based upon data from Ghostery. Google-owned

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

trackers were present on over 88% of a sample of 393,829 distinct domains. Further, in a survey of the top 100 sites, Google Analytics appeared on 81 of them.

In August 2009, Soltani et al. [demonstrated](#) that popular websites were using “Flash cookies” to track users. Some advertisers had adopted this technology because it allowed persistent tracking even where users had taken steps to avoid web profiling. Soltani et al. also demonstrated “respawning” on top sites with Flash technology. This allowed sites to reinstate HTTP cookies deleted by a user, making tracking more resistant to users’ privacy-seeking behaviors. In a survey of the top 100 sites according to Quantcast, Soltani et al. found 3602 cookies set on 98 of the top 100 sites. They also found 281 Flash Cookies set on 54 of the top 100 sites.

In July 2010, Julia Angwin, Tom McGinty, and Ashkan Soltani of the Wall Street Journal [reported](#) that in a scan of the top 50 sites, 3,180 “tracking files” (this comprised HTTP cookies, Flash cookies, and web beacons) were detected. Twelve sites set over 100 each.

In 2010, Michael Coates [surveyed](#) the top 1,000 websites in order to determine how many were using HTTPS. Coates sent a basic HTTPS request to these sites, and they responded with 559 cookies. Coates’ method appeared to not collect any third party cookies.

Flash cookies have become a major focus of research. In 2001, McDonald and Cranor of Carnegie Mellon investigated the presence of Flash cookies on websites. They [found](#) a dramatic decline from the Soltani et al. investigation in 2009. McDonald and Cranor found Flash cookies on only 20 of the top 100 sites. They were also careful to attempt to determine whether Flash cookie values were unique or not—six of the top 100 sites had Flash cookies that were not unique, and thus probably not used to track individuals.

Krishnamurthy et al. have made significant contributions to the study of privacy “leakage.” In a [study](#) of websites that required registration, they found that a majority of the popular sites they analyzed “directly leak sensitive and identifiable information to third-party aggregators.” The problem they identified was widespread: “56% of the 120 popular sites in our study (75% if we include userids) directly leak sensitive and identifiable information to third-party aggregators.”

In July 2011, Stanford Law/Computer Science graduate student Jonathan Mayer released “FourthParty,” an “open-source platform for measuring dynamic web content.” Mayer has posted the raw data from web crawls made with the platform, and has released two reports flowing from the system. In the [first](#), Mayer tested how members of the Network Advertising Initiative (NAI) interpret opt outs. The NAI considers the scope of opt out rights to pertain only to targeting ads, not to tracking. Thus, if a consumer opts out, NAI members may still track them. Mayer found that half of the NAI members tested (N=64) still used tracking cookies after an opt out was expressed.

In the [second](#), Mayer found that in developing FourthParty, he detected “browser history stealing.” This is a practice where a website, “exploits link styling to learn a user’s web browsing history. The approach is simple: to test whether the user has visited a link, add it to a page and check how it’s styled.”

In August 2011, Ayenson et al. [surveyed](#) the top 100 web sites, simulating a user session by clicking on 10 random links on each site. Cookies were detected on all top 100 sites. The group found 5,675 cookies, 4,615 of which were set by third parties. Six-hundred third-party hosts were detected. Google-controlled cookies were present on 97 of the top 100 sites, including popular government websites.

Ayenson et al. found that 17 sites were using HTML5 local storage, and seven of those sites had HTML5 local storage and HTTP cookies with matching values. Flash cookies were present on 37 of the top 100 sites.

In October 2011, Jonathan Mayer tested signup and interaction on 185 of the Quantcast top 250 sites. He found 113 of the sample leaked userids or usernames to third parties.

REVERSE TIMELINE

Study	Year	Major Finding	Sample Size
Mayer	2011	Most popular websites were “leaking” usernames and userids to third parties.	185 of the Quantcast top 250
Ayenson et al.	2011	5675 HTTP cookies detected, 4615 of which were third party. 37 sites with 100 Flash cookies detected. All top websites had cookies.	Top 100 sites, 10-click user session simulated
Mayer	2011	Network Advertising Initiative members continued to use tracking cookies after opt out	64 of the Network Advertising Initiative Members
Krishnamurthy & Wills	2011	Majority of popular websites with registration leaking personal data to third parties	Array of popular websites that required registration
McDonald & Cranor	2011	Flash cookies present on 20 of top 100 sites	Surface crawl of homepages of top

			100 sites
Coates	2010	559 first party cookies detected	Limited HTTPS request to top 1,000 sites
Angwin et al. (Wall Street Journal What They Know)	2010	3,180 tracking mechanisms detected. Only one site lacked cookies.	Top 50 sites, 20-click user session simulated
Gomez et al. (KnowPrivacy Report)	2009	Google-owned web beacons were present on 88% of a large sample of websites	393,829 unique domains
Soltani et al.	2009	3602 HTTP cookies detected, 281 Flash cookies detected. 98 of the top 100 sites had cookies.	Top 100 sites, 10-click user session simulated
Krishnamurthy et al.	2009	Large increase in DNS aliasing; penetration of major third party trackers to 75% of sample sites	1,200 sites scanned over four years
FTC	2000	57% of the sites in the Random Sample and 78% of the sites in the Most Popular Group set cookies.	Random sample of 335 sites and 91 of top 100 sites
EPIC Surfer Beware III	1999	86 used cookies.	100 ecommerce sites
FTC Privacy Online	1998	Most websites collect personal info, but only 14% have privacy notices	1,400
EPIC Surfer Beware II	1998	Few of the newest DMA members had privacy policies	New DMA members
EPIC Surfer Beware I	1997	Homepages of 23 sites used cookies	Top 100

References

Julia Angwin, The Web's New Gold Mine: Your Secrets, A Journal investigation finds that one of the fastest-growing businesses on the Internet is the business of spying on consumers, Wall Street Journal, Jul. 30, 2010, available at <http://online.wsj.com/article/SB10001424052748703940904575395073512989404.html>.

Ayenson, Mika, Wambach, Dietrich James, Soltani, Ashkan, Good, Nathan and Hoofnagle, Chris Jay, Flash Cookies and Privacy II: Now with HTML5 and ETag Respawning (July 29, 2011) available at: <http://ssrn.com/abstract=1898390>.

Michael Coates, A Study of HTTPOnly and Secure Cookie Flags for the Top 1000 Websites, Dec. 28, 2010, available at <http://michael-coates.blogspot.com/2010/12/study-of-httponly-and-secure-cookie.html>.

Electronic Privacy Information Center, Surfer Beware: Personal Privacy and the Internet, Jun. 1997, available at <https://epic.org/reports/surfer-beware.html>.

Electronic Privacy Information Center, Surfer Beware II: Notice is Not Enough, Jun. 1998, available at <https://epic.org/reports/surfer-beware2.html>.

Electronic Privacy Information Center, Surfer Beware III: Privacy Policies without Privacy Protection, Dec. 1999, available at <https://epic.org/reports/surfer-beware3.html>.

Federal Trade Commission, Privacy Online: A Report to Congress, Jun. 1998 <http://www.ftc.gov/reports/privacy3/toc.shtm>.

Federal Trade Commission, Privacy Online: Fair Information Practices In the Electronic Marketplace: A Report to Congress, May 2000, available at <http://www.ftc.gov/reports/privacy2000/privacy2000.pdf>.

Joshua Gomez, Travis Pinnick, and Ashkan Soltani, KnowPrivacy (Jun. 1, 2009), available at http://www.knowprivacy.org/report/KnowPrivacy_Final_Report.pdf.

Krishnamurthy, B., & Wills, C., Privacy diffusion on the web: A longitudinal perspective, Proceedings of the 18th ACM international conference on World wide web (2009)(p. 541–550), available at <http://portal.acm.org/citation.cfm?id=1526782>.

Krishnamurthy, B., Naryshkin, K., & Wills, C. E., Privacy leakage vs. Protection measures: the growing disconnect, presented at W2SP 2011: WEB 2.0 SECURITY AND PRIVACY 2011 (2011), available at <http://www.cs.wpi.edu/~cew/papers/w2sp11.pdf>.

Jonathan Mayer, FourthParty, available at <http://fourthparty.info/>.

Jonathan Mayer, Tracking the Trackers: Early Results, Jul. 12, 2011, available at <http://cyberlaw.stanford.edu/node/6694>.

Jonathan Mayer, Tracking the Trackers: To Catch a History Thief, Jul. 19, 2011, available at <http://cyberlaw.stanford.edu/node/6695>.

Jonathan Mayer, Tracking the trackers: Where everybody knows your username, Oct. 11, 2011, available at <http://cyberlaw.stanford.edu/node/6740>.

McDonald, A. M., & Cranor, L. F., A Survey of the Use of Adobe Flash Local Shared Objects to Respawn HTTP Cookies, CMU-CyLab-11-001 (2011), available at <http://www.casos.cs.cmu.edu/publications/papers/CMUCyLab11001.pdf>.

Ashkan Soltani, Shannon Canty, Quentin Mayo, Lauren Thomas, and Chris Jay Hoofnagle, Flash Cookies and Privacy, Aug. 10, 2009, available at: <http://ssrn.com/abstract=1446862>, accepted for publication at AAAI Spring Symposium on Intelligent Information Privacy Management, CodeX: The Stanford Center of Computers and Law.

UC Berkeley School of Law

215 Boalt Hall
Berkeley, CA 94720-7200
510-642-1741

[Directions](#)
[Feedback](#)
[UC Berkeley](#)

[For Students](#)
[For Faculty & Staff](#)
[For Employers](#)

[Job Openings](#)
[Social Media](#)
[About This Site](#)