Web Privacy Census: A Study of The Characteristics of High Cookie Websites

Sharon Pamela Santana
Department of Computer Science,
Smith College, Northampton MA, USA
ssantana@smith.edu

Chris Jay Hoofnagle
UC Berkeley School of Law
University of California, Berkeley, USA
correspondence to:
choofnagle@law.berkeley.edu

Nathaniel Good
UC Berkeley School of Information
University of California, Berkeley
ngood@ischool.berkeley.edu

ABSTRACT
The Web Privacy Census, a survey of the top 1,000 websites, finds that the average website has 8 cookies. However, a small number of them have hundreds or even thousands of cookies. This paper investigates why this may be so, and considers IP origins of sites, cookie names, trackers, website category, and the presence of online advertising. Most of the sites with high numbers cookies belong to a narrow set of categories, mainly News and Entertainment. Furthermore, websites with high numbers of cookies don't necessarily contain many advertisements, which was the expected. Nonetheless, certain categories of high cookie websites, such as News and Media, usually contain more online advertisement than other categories.

Keywords
Privacy Census, HTML, HTTP Cookies, Web Categorization, Online Advertisement

1. BACKGROUND
Three Web Privacy Censuses were conducted at the University of California Berkeley with the intention of exploring the current state of tracking and privacy on the web[1]. These censuses show that tracking on the web is becoming more common across the top rated sites. The reasons for tracking somebody on the web through the use of cookies vary. The main purpose, nonetheless, is because websites are stateless and can't remember the information of its users. The information stored in such cookies is also very useful to serve advertisements targeted to specific audiences online.

Since the Web Privacy Census started, there have been three website crawls and the fourth is currently in progress[2]. From May 2012 to April 2013, researchers found a significant increase in tracking mechanisms on the most popular sites.[1]

2. INTRODUCTION
The cookies being accounted for in this study are HTTP cookies from first and third party origins, as they are the most common type found. The domains examined in this report come from the data obtained in the three crawls mentioned above. Instead of looking at URLs ranked by popularity, however, this research focuses of he websites with the highest numbers of cookies. The aim of this study is to do quantitative analysis of the high cookie websites in order to examine the general characteristics of such sites, more specifically, to answer the question of why are there websites with such high numbers of HTTP cookies.

The difficulty in answering what seems to be a straightforward question is that there might be several factors influencing the number of cookies present in a website. Amount of advertisement served on a site, website category, location of website IP, and code structure of the website are only a few of the possible influencing factors. The latter reason could be due to recycling of unfamiliar code, which might contain cookies not accounted for. The reason supporting the first two elements is because some sites, known as "content farms", contain an excess of unorganized links, pictures, and advertisement. These components typically include vast numbers of cookies. Online advertisement, for example, uses cookies for tracking and ad targeting purposes.

3. METHODS

3.1 Identifying high cookie sites and most popular sites
Here we collect the sites with the most cookies (sites with between 95 and 469 cookies). These sites are the top 100 sites with the highest number of cookies from each of the crawls of the previous censuses.

3.2 Obtaining the categories and site origins

3.2.1 Categories
To determine the category information of the sites with the most cookies, a web filtering software called FortiGuard was used. A Python script was written to feed in the site names into FortiGuard and obtain the categorization of the sites in question. There are 78 possible categories offered by FortiGuard's URL Category Database, which are organized into six groups: Security Risk, Controversial, General Interest (Business), Potentially Liable, Bandwidth Consuming, General Interest (Personal).

3.2.2 Locations
The script used to categorize the websites using FortiGuard also collected information about the country in which the IP of the server is located by using a process similar to the one used to obtain the category information. This information was collected to get an idea of what the countries hosting high cookie sites are.

3.2.3 Python Categorizing Script
The python script used to categorize the websites made use of FortiGuard's URL category database. It uses BeautifulSoup, which is an html parser, in order to extract the category information.

[1] The Web Privacy Censuses referenced were completed during the months of May 17 2012, October 24 2012 and April 18 2013.

[2] A web crawler was used to visit the top rated websites from Quantcast.com and collect the cookie information. Websites were visited on two modes: shallow crawl (the home page of domain only) and deep crawl (6 random links of domain).

The script first reads a file containing 100 URLs. The list of 100 is split into 10 lists of 10 to comply with FortiGuard's query limitations of 10 sites per minute. Each of the 10 lists go through a data retriever, which takes a single URL at a time and concatenates it with the string "http://www.fortiguard.com/ip_rep/index.php?data=" in order to retrieve the html file containing the data for that specific website. After the HTML file is obtained, it is parsed by creating a HTMLParser object. For each of these HTML files, the category information is obtained from reading the data in between the only "<h3>" tags found in each of the HTML files. To get the country information, the script looks for the tables of class "large" and selects the second table that gets returned. Lastly, it extracts the country information from the table.

3.2.3.1 Script Limitations

Since FortiGuard has a query limit of 10 sites per minute and 200 queries per hour, each one of the 10 lists has to be selected manually after every minute to avoid getting the IP of the computer running the script temporarily blocked. This script can be fully automated by slowing down the script so that it feeds the lists into FortiGuard after every minute.

3.3 Obtaining the most popular cookie names and trackers

The web crawls of May 2012, October 2012 and April 2013 collected general information about the cookies found on these sites such as the names, values, paths, total HTTP cookie count, Flash cookie count, hosts, trackers, etc.… and stored them into SQL format tables. For this project, a series of python scripts were written to extract and manipulate the data from those sequel tables. These scripts allowed for the collection of the most popular cookie names and trackers on the sites with the most cookies in each of the crawls. To achieve this, the ids of the top 100 sites with the most cookies from each of the crawls were matched to the ids on the cookie tables. This allowed for only the cookies belonging to the domains in question to be retrieved.

4. RESULTS AND DISCUSSION

4.1 Online Advertising

Advertisement could be defined as the activity (usually persuasive) of attracting public attention to a product, business, service, or information and it is usually paid for when promotions are made thought a print, broadcast, or electronic media. In online advertisement, advertisers buy inventory from publishers (the websites hosting the ads). Inventory is the space available on a website for ad displaying or serving purposes. Online ads make use of a lot of cookies, especially targeting ads, which tailor ads to specific audiences. Ads also use cookies to make sure that the ads served are not presented multiple times to the same user.

There are various important elements to advertising that should be paid attention to because the crowd to which the ads might be tailored to might be very diverse or specific. The sites with high cookies were suspected to have a lot of advertisements because online advertisements typically require many cookies on the sites serving them. This, however, turned out not to be the case necessarily. After a close manual inspection of 35 sites from the shallow crawl from 4-18-13 and 35 from the shallow crawl of 10-24-12, the sites with high cookies didn't have that many more ads in them when compared to regular sites. There were sites with no advertisements at all, and the most advertisements found in a site from the two lists previously mentioned was 30. The only possible trend found was a slightly higher number of ads in the categories of News and Media and Entertainment domains.

4.2 Categories

The categories with the highest frequencies examined from the top 100 high cookie sites across the crawls turned out to be similar as shown by table 1.

The most popular website categories were News and Media and Entertainment. There were a total of 37 out of 78 possible categories found in the 600 hundred high cookie domains that were looked at. The most prominent categories from the top 100 popular domains in the previous censuses were Search Engines and Portals and Web Hosting.

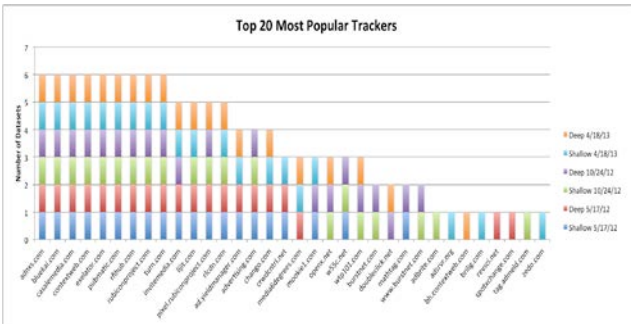
Table 1. Most popular categories on high cookie sites

Crawl	Categories	Frequency
Shallow 10/24/2012	News and Media Entertainment	25/100 22/100
Deep 10/24/2012	Education Entertainment	22/100 25/100
Shallow 5/17/2012	News and Media Entertainment	32/100 18/100
Deep 5/17/2012	Education Content Servers	29/100 20/100
Shallow 4/18/2013	News and Media File Sharing and Storage	57/100 10/100
Deep 4/18/2013	Entertainment File Sharing and Storage	20/100 15/100

4.3 Prominent Trackers in high cookie sites

Even though knowing the most prominent trackers do not help determine why some sites have many cookies, it is still interesting to examine this for comparison purposes. Python scripts were run in order to identify the most popular cookie names and trackers. All of the trackers with counts in the top 20 in all of the crawls are included in image 1 below.

Image 2. Top 20 trackers of top 100 high cookie sites



Ad.yieldmanager.com was the top 1 in 4 out of 6 crawls and bluekai.com was within the top 5 of all of the crawls. Furthermore, some of the trackers predominant in the top ranked sites such as doubleclick.net, rubiconproject.com, and pubmatic.com were within the top 20 trackers in high cookie sites. On the other hand, there were also trackers such as scorecardsearch.com, quantserve.com, atdmt.com, and

addthis.com, which were found in the top ranked sites, but not in the top 20 trackers of the high cookie sites. Maybe the most popular or well-known trackers are not present in the top high cookie sites because the ranks of the high cookie sites are typically around the 500th, which is not very good.

4.4 Popular cookie names in high cookie sites

Before getting into the results of the most frequent cookies used in the high cookie sites, it is important to first understand what a cookie is and what it does as well. A cookie is a text file that gets downloaded to the visitor's machine after visiting a website with the purpose of helping the browser navigate through a particular web. These cookies get stored in the browser's folder or subfolder and can be deleted by the user.

Since there are different types of information that websites collect with regards to its user's online activity, there are various types of cookies available to obtain the specific data needed. The names of HTTP cookies could be the same across sites, regardless who the tracker company setting the cookie is, but they usually change. Most HTTP cookies have values, which is where the users' information gets stored. The most common values used are name, value, expiration date, path, domain, and secure connection parameter. Due to the nature of the information stored in HTTP cookies, nevertheless, they can't be used to store viruses or malicious files. Storage space also limits the ability for cookies to be used with malevolent intentions. As more technologies evolve though, new types of cookies become available and allow more data to be stored. This allows for more efficient online tracking methods to emerge, but it compromises' users privacy more as a result.

4.5 Location results

Since Quantcast.com rates the top sites within the United States, most of the websites IPs from the crawls originated in the U.S. The second country with the most IP addresses was the United Kingdom and the third was Europe.

In the top 97 ranked sites from Quantcast.com, the United States was hosting 96 IP addresses, which is to be expected, given that Quantcast.com rates the popular sites in the U.S. only.

5. FUTURE WORK

A thorough examination of the code of these websites, in conjunction to all of the elements already discussed above, should be done in order to more directly answer the question of why there are websites with incredibly high numbers of cookies. The information from the code of these high cookie sites would allow for a more concrete answer to this research question because if a lot of these sites contain recycled code then it's cookies may also be recycled without the programmer being aware of that. As a result, some of these high cookie sites might contain many embedded cookies that are unwanted or needed, which might be part of this high cookie issue.

6. CONCLUSION

The most prominent categories were actually similar, which supports one of the believed reasons that explain the high numbers of cookies on a site: some website categories are more likely to have more cookies than others. In this case, Education, News and Media, and Entertainment encompass by far most of the categories of the high cookie sites studied in this project. In addition, these categories turned out to be not as malicious or adult content bearing as they were expected to be. Moreover, websites with high numbers of cookies were not necessarily the ones with the most online advertisement; nonetheless, the categories of News and Media did tend to have more ads than any of the other categories of sites considered. Furthermore, the only similarity between the cookie names identified in the sites with the most cookies and the cookie names in the top ranked sites were the Google analytics cookies, the rest of the cookie names were different for the most part as predicted. Most of the trackers of high cookie sites and top popular sites also differ. This might support the suggestion that the common trackers and cookies used in the top sites are not the same as the common trackers in high cookie domains.

7. AWKNOWLEDGEMENTS

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).

Thanks to FortiGuard for providing the information in its URL Category Database that allowed the web categorization to be done in this project.

Aimee Tabor for her support and guidance throughout the TRUST REU

8. REFERENCES

[1] Chris Jay Hoofnagle & Nathan Good, *The Web Privacy Census*, October 2012. Retrieved July 29, 2013 from http://law.berkeley.edu/privacycensus.htm