Michael Scudiero

CS:3640 Final Paper

Part I – Research Question and Methodology

For my research project, I decided to measure the levels of plain text cookies on various websites, in order to analyze and see what cookies were most common, and what websites set the most plain text cookies in general. I found this to be a valid research question because most people do not think about what goes on behind the scenes when they open a web page, and plain text cookies, while not usually tracking cookies, are used by sites to maintain states of services like online shopping carts, login information, and other similar "memory" services that would simply be too expensive to maintain a database of. How this works is that the cookie is a small text file saved to a browser, that when a site that has saved that cookie is revisited, the cookie is accessed, and inputs the required data that was stored within the text file into the website, performing such tasks as automatic credential authentication, loading saved information such as a list of previous searches or a shopping cart, and other such tasks. They can also be used to track a user across sites by having the same cookie loaded on multiple sites, but that is not the focus of this research paper. The method I used to record my data was to write a Python web scraper program that first accepts a URL from the command line, then uses the Selenium web driver to load that URL, then gathers all the cookies from that webpage. I had the command line pipe the output to a text file so that it would be easier to read. I also had a second variant of this program that crawled to each site that was linked on the initial site, to see if the cookies remained the same on each linked site. I found this interesting because it allowed me to measure what information was being kept between various linked sites.

Part II – Data Gathering

To gather data, I went to a site that listed the top 20 visited sites in the United States in 2022, which was likely accurate because it had data up to December of that year. I decided to scrape the cookies from most of those sites that were listed, though some of the sites were pornography sites, which I did not measure due to the possibility of threats to my computer on those sites. The measured sites included widely trafficked sites like Google, Bing, YouTube, Yahoo, Facebook and other social media sites, Wikipedia and other informative sites, and Amazon and other shopping sites. What I found was that most of the cookies that I discovered on these sites were in some way related to the functionality of the site; for instance, Amazon had cookies that saved a user's login credentials so that they could use the site without having to log in every single time, and it also had multiple cookies that would store shopping cart data. Many sites, such as not only Amazon, but also Google and other search engines, had a specific cookie that stored data of the user's physical location or region, which the site likely derived from the user's IP address. This cookie could be used in multiple separate ways; on Google, it is likely used to localize the search responses so the user gets results from their local area, whereas on Amazon, it is likely used to keep the user's shipping area so that the site knows approximately how much to charge for shipping. These cookies remained the same throughout the linked sites, which, if they are related to the parent site's functionality, makes sense, since the child sites would have the same functions as their parent, almost certainly with additional functions added on.

Part III – Analysis

My analysis of the data I acquired is that, because there are many more cookies than I actually saw, I would conclude that the overwhelming majority of cookies are in fact used for tracking and advertising. This sadly makes a lot of sense, since the companies that run these websites need to make some sort of revenue in order to not only pay their employees, but also as an incentive to maintain their sites and services. There is a lot of money to be made advertising on the Internet, and personalized advertising is the best way to do that – a company does not want to spend money advertising to someone who is unlikely to be interested in their product.