Assignment 4: Web Tracking

This project is due on **November 12, 2021** at **5 pm**. You may work with a partner on this assignment and submit one project per team. You may NOT work with the same partner you have worked with on a previous assignment. Submit your solutions electronically.

Background

Internet tracking involves the use of many web technologies, including cookies, local data storage, and browser fingerprinting, to build profiles of users' Internet habits. These techniques allow data aggregation companies to make sophisticated inferences about the interests and personalities of individuals, even if they do not know these individuals' exact identities.

A result of web tracking is that different people may have very different experiences on the web based on how sites personalize content and advertising to their profiles. While occasionally convenient (e.g. some personalized ads), this type of personalization can also cause and perpetuate biases or create "content bubbles."

For example, real estate sites may choose to display ads only to individuals who are inferred to be from majority groups, online loan quotes may be cheaper for people inferred to be wealthy and at lower risk of default, and articles about fringe theories may be more prominently displayed to individuals who are already primed for conspiracies.

One goal of Internet privacy researchers is to measure the prevalence of web trackers and identify especially egregious third-party tracking domains. In this assignment, you will perform these measurement tasks to discover the scale of cookie-based web tracking.

Objectives

- Use a clean environment to collect first- and third-party cookies
- Plot cookie prevalence and identify prolific tracking domains

Provided Files

- `Readme.pdf`: This file.
- `Solutions.txt`: File for your written responses to open-ended questions.

Instructions

Your task is to measure the number and distribution of cookies that someone like you might encounter during a typical session of web browsing.

In order to get a non-biased measurement, you will use a clean install of a browser that is not associated with any of your existing profiles. Read and follow the below instructions to complete the assignment:

1. Download and install the Opera browser from `https://www.opera.com/`. If you already use Opera as your day-to-day browser, contact Prof. Apthorpe for alternatives.

2. Open Opera and find the Settings/Preferences page (this will be in a different place whether you are on Mac or Windows, but it should be similar to other applications).

   (a) Ensure that "Block ads and surf the web up to three times faster" is turned **OFF**

   (b) Ensure that "Block trackers" is turned **OFF**

3. Using Opera, browse the web for *at least 1 hour*. Yes, this really is homework! Just make sure to follow these guidelines:

   (a) **Do not** create any new accounts on any websites

   (b) **Do not** log in to any of your existing accounts (this measurement should be separate from any existing profiles you may have).

   (c) **Do not** spend too long on any one website or domain. While you may sometimes spend more than an hour on a site during your normal use of the Internet (e.g. while watching a movie), this won't give you a realistic measurement of the trackers someone might encounter while browsing more actively.

   (d) **Do** choose the most permissive cookies option (e.g. "allow all cookies") if any individual websites you visit ask.

   (e) **Important!** Keep a text document `domains.txt` with a list of the *domains* of *all* the websites you visit. For example, if you were to visit `https://www.colgate.edu/academics` and `https://www.colgate.edu/student-life/athletics-recreation`, you would add `colgate.edu` to your list. You will use this list to distinguish between first- and third-party cookies.

4. Return to the Opera Settings/Preferences page. Go to "Advanced > Privacy & security > See all cookies and site data." This should display a list of all the cookies and other local information stored while you were browsing. Marvel at what you collected. While this view is useful for inspecting cookies manually, it is not that convenient for programmatic analysis. **Do NOT** clear these cookies.

5. Download a version of the cookies that will be easy to process programmatically:
   (a) Go to "About Opera" to find the directory that the browser uses to save local files. On MAC, probably `C:\Users\Your_User_Name\AppData\Roaming\Opera Software\Opera Stable`
   On Windows, probably `C:\Documents and Settings\[username]\Application Data\Opera\Opera\`
   (b) Open this folder on your hard drive and copy the `Cookies` file into the folder you downloaded with the provided files for this assignment.
   (c) This `Cookies` file is a binary SQLite table, so you will need to convert it to JSON for easy processing:
      i. Download and install the SQLite Browser: `https://sqlitebrowser.org/`.
      ii. Open the SQLite browser, then go to "File > Open Database Read Only" and select the `Cookies` file you copied in (b) above.
      iii. Go to "File > Export >Table to JSON...", select the "cookies" table (keeping "Pretty print" checked), and save the "cookies.json" file to the assignment directory.
6. Write a program in any language that analyzes `cookies.json` and `domains.txt` to create two bar plots:
   (a) `firstparty.png`: Counts of **first-party** cookies by domain
   (b) `thirdparty.png`: Counts of **third-party** cookies by domain

   These charts should be similar to the examples in the `example_plots` folder, although the specific colors and other stylistic formatting is up to you. If you decide to use Python, I recommend the `json` library for parsing JSON and the `matplotlib` library for plotting.
7. Answer the questions in `Questions.txt`.

Deliverables

Upload the following files to Moodle:

- `cookies.json` (5%)
- `domains.txt` (5%)
- Bar plots `firstparty.png` and `thirdparty.png` (30%)
- Your program to parse `cookies.json` and `domains.txt` and create the bar plots (30%)
- Completed `Questions.txt` (30%)

Extra Credit "Bug Bounty"

If you find a bug anywhere in this assignment, please inform Prof. Apthorpe. The first student (or partners) to find any particular bug will be given a small amount of extra credit. This is an incentive to start the assignment early and will help make the course better for students in future years.