Rahul Padhi - 919975938

Arun Khanijau - 919496132

11/25/24

Project 2 Part 2 Report

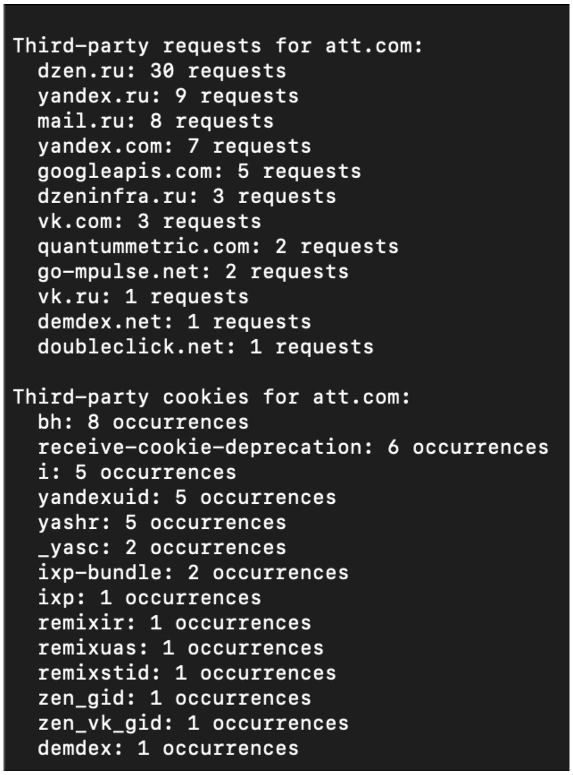
Part 2: Web Crawling and HAR file analysis

Running the “scan_har_files.py” program will go through a directory of har files and list out the third-party domains and third-party cookies accessed by each site. At the end, it will list the 10 most commonly accessed third-party domains and the 10 most common third-party cookies.

Link to ChatGPT session for improving code:

https://chatgpt.com/share/67455413-7668-8007-a058-1b40aaba1894

The following example image shows the results for att.com:



```
Third-party requests for att.com:
  dzen.ru: 30 requests
  yandex.ru: 9 requests
  mail.ru: 8 requests
  yandex.com: 7 requests
  googleapis.com: 5 requests
  dzeninfra.ru: 3 requests
  vk.com: 3 requests
  quantummetric.com: 2 requests
  go-mpulse.net: 2 requests
  vk.ru: 1 requests
  demdex.net: 1 requests
  doubleclick.net: 1 requests

Third-party cookies for att.com:
  bh: 8 occurrences
  receive-cookie-deprecation: 6 occurrences
  i: 5 occurrences
  yandexuid: 5 occurrences
  yashr: 5 occurrences
  _yasc: 2 occurrences
  ixp-bundle: 2 occurrences
  ixp: 1 occurrences
  remixir: 1 occurrences
  remixuas: 1 occurrences
  remixstid: 1 occurrences
  zen_gid: 1 occurrences
  zen_vk_gid: 1 occurrences
  demdex: 1 occurrences
```

Running the program will show the results for all the HAR files in the specified directory.

After analyzing 1000 sites from the list, the ten most seen third parties across all sites are:

1. doubleclick.net: 1554 requests
2. google.com: 1459 requests
3. media-amazon.com: 1434 requests
4. googletagmanager.com: 1316 requests
5. gstatic.com: 1139 requests
6. googlesyndication.com: 995 requests
7. googleapis.com: 995 requests
8. microsoft.com: 912 requests
9. cookielaw.org: 809 requests
10. ifood.com.br: 798 requests

The ten most seen cookies are:

1. receive-cookie-deprecation: 672 occurrences
    - This cookie is from a domain owned by Criteo and is used for advertising.
2. __cf_bm: 429 occurrences
    - This cookie comes from a Nvidia domain and is used by Cloudflare for bot management.
3. sca: 346 occurrences
    - This cookie is used by various sites for advertising tracking.
4. audit_p: 210 occurrences
    - This cookie comes from the Rubicon Project and is used for advertising.

5. audit: 210 occurrences
    - This cookie also comes from the Rubicon Project and is used for advertising.
6. bcookie: 196 occurrences
    - This is a cookie from Mircosoft and is used for targeted advertising.
7. _cfuvid: 154 occurrences
    - This cookie is used by Cloudflare to help distinguish users with the same IP addresses in order to enforce rate-limiting rules.
8. khaos: 150 occurrences
    - This cookie is used by the Rubicon Project for targeted advertising. It carries information about how the end user uses the website.
9. khaos_p: 150 occurrences
    - This cookie is also used by the Rubicon Project for targeted advertising
10. __cfruid: 139 occurrences
    - This cookie is used by Cloudflare to identify trusted web activity.