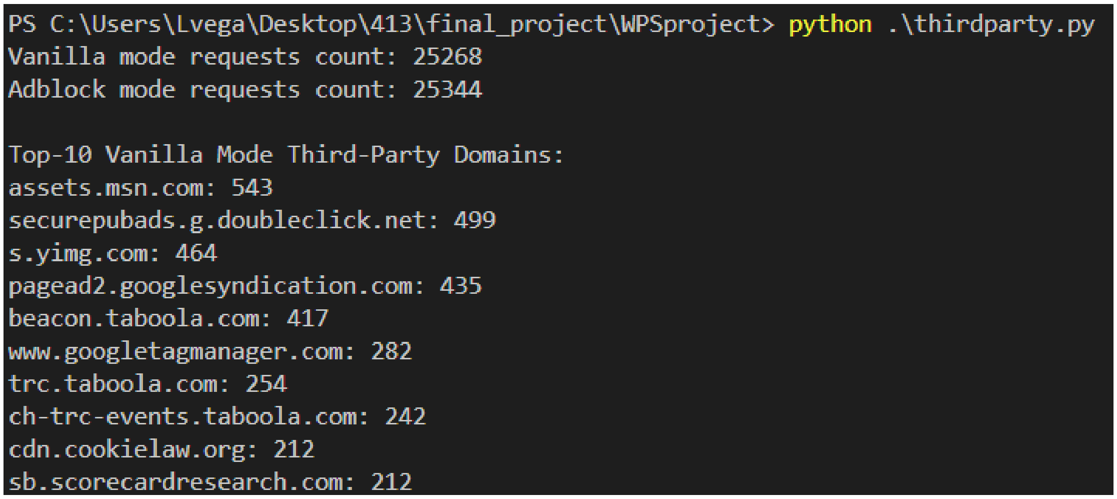
Final Project: Measuring the Prevalence of Online Tracking

Task 1

We followed the instructions and created scripts to extract the more valuable data from the resulting json files. There are 4 scripts, each one corresponds to a task and gives different graphs depending on the attribute we want to analyze. To run them, please refer to the included readme file.

Task 2

The script automatically lists the top 10 vanilla mode third party domains in the console output.



```
PS C:\Users\Lvega\Desktop\413\final_project\WPSproject> python .\thirdparty.py
Vanilla mode requests count: 25268
Adblock mode requests count: 25344

Top-10 Vanilla Mode Third-Party Domains:
assets.msn.com: 543
securepubads.g.doubleclick.net: 499
s.yimg.com: 464
pagead2.googlesyndication.com: 435
beacon.taboola.com: 417
www.googletagmanager.com: 282
trc.taboola.com: 254
ch-trc-events.taboola.com: 242
cdn.cookielaw.org: 212
sb.scorecardresearch.com: 212
```

Figure 1: Top-10 Vanilla Mode Third-Party Domains and Overall Requests Counts

A closer inspection to these domains (and a quick Google search for each one) gives us a closer look at how they are being used in these sites:

assets.msn.com: This is a manager for assets on Microsoft's MSN network, it most likely means the website is requesting the MSN portal to load some images, scripts, videos, or other resources. This also provides performance analytics and other insights pertaining to the content they deliver.
securepubads.g.doubleclick.net: We explored the purpose of this domain in class a little. This domain is in charge of ad delivery across the internet. It also facilitates the management of ads, their placements and their performance on a site. Since it serves so well for data collection, media-heavy sites will see a lot of requests
s.yimg.com: This is the domain belonging to yahoo that serves resources similar to Microsoft's domain we saw at first.
pagead2.googlesyndication.com: This is another google service that tracks advertisements analytics and other performance data for marketing purposes.

beacon.taboola.com: This one is a bit of a different type of tracker, you can see that the domain has the word beacon in it, indicating this type of tracker checks what content the user is interacting with. This is also used to track user activity and ad performance. Also managed by a marketing company that is a little less well knows (Taboola)
www.googletagmanager.com: Google Tag Manager is a tag management system that allows you to set up and manage tags on your site without changing your website's code.
trc.taboola.com, ch-trc-events.taboola.com: These are part of the Taboola ecosystem
cdn.cookielaw.org: A cookie management system that allows for users to and hosts to access the cookies of a site and modify them
sb.scorecardresearch.com: From their website, “a leading global market research effort that studies and reports on Internet trends and behavior. ScorecardResearch conducts research by collecting Internet web browsing data and then uses that data to help show how people use the Internet, what they like about it, and what they don’t.”

After running the thirdparty.py file, we will see this graph showing the frequency in which third parties are seen.

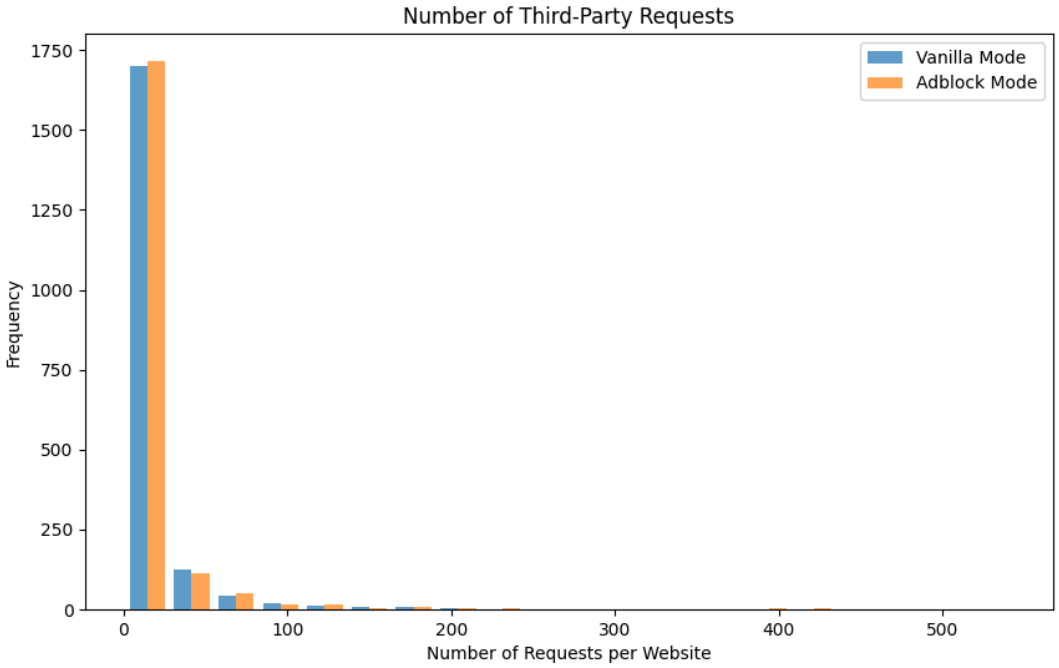


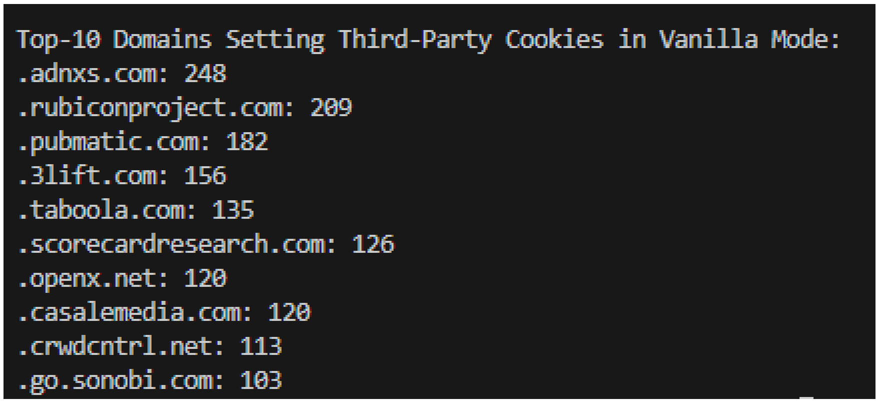
Figure 2: Distribution of third-party requests from vanilla and adblocker mode

These two distributions seem very similar. In that first bar we can see the frequency of adblock mode sites is slightly higher than the one in vanilla mode, and strangely enough we can see there are a bit more requests in adblock mode overall.

In general, there are under 100 requests per website with a very high frequency in each, their frequency decreases when we check the higher number of requests. It seems the adblock mode added some extra requests that cause the numbers to be the opposite as we would expect. Adblocker mode should be getting some of those requests out of our distribution. Nonetheless, this graph gives us a good idea of how the list of websites we crawled are being used for advertisements.

Task 3

When discussing the cookies that are being set in these websites, we have these top 10 domaines that are being used to set the cookies in the sites we crawled. A quick google search for each one gives us some insight on how they are used.



```
Top-10 Domains Setting Third-Party Cookies in Vanilla Mode:
.adnxs.com: 248
.rubiconproject.com: 209
.pubmatic.com: 182
.3lift.com: 156
.taboola.com: 135
.scorecardresearch.com: 126
.openx.net: 120
.casalemedia.com: 120
.crwdcntrl.net: 113
.go.sonobi.com: 103
```

Figure 3: Top-10 Domains Setting Third-Party Cookies in Vanilla Mode:

.adnxs.com (From an article from the Guardian): Plug into other advertising serving platforms, such as Google's Doubleclick, and "data aggregators", such as Quantcast, which provide behavioural targeting.
.rubiconproject.com: This is the name of a digital advertisements company that most likely works to manage cookies of clients in several sites
.pubmatic.com: Provides sell-side, real-time programmatic ad transaction advertising software.
.3lift.com: Allows for marketing analytics even in cookie-constrained environments. It enables audience targeting through data activation technologies
.taboola.com, .scorecardresearch.com: See task 1 for some research on these
.openx.net: Cookies from OpenX track user activity to deliver personalized ads based on interests and demographics

.casalemedia.com: It uses cookies to track users' web behavior and deliver personalized ads. It helps advertisers reach targeted audiences across multiple platforms.
.crwdcntrl.net: Lotame is a data management platform that lets marketers, agencies and publishers harness audience data to make smarter marketing, product and business decisions.
.go.sonobi.com: Sonobi is an advertising technology company that provides a header bidding solution, allowing publishers to maximize ad revenue. Their cookies help track users' interests and behavior for targeted ad delivery across various sites.

All of these as you see are helper companies that allow advertisers to better reach their audiences and manage the information of their users.

In the graph we see the same behavior we witnessed in the previous task where the adblock mode seems to be a bit higher in some instances which skews the analysis of the overall performance of the adblocker.

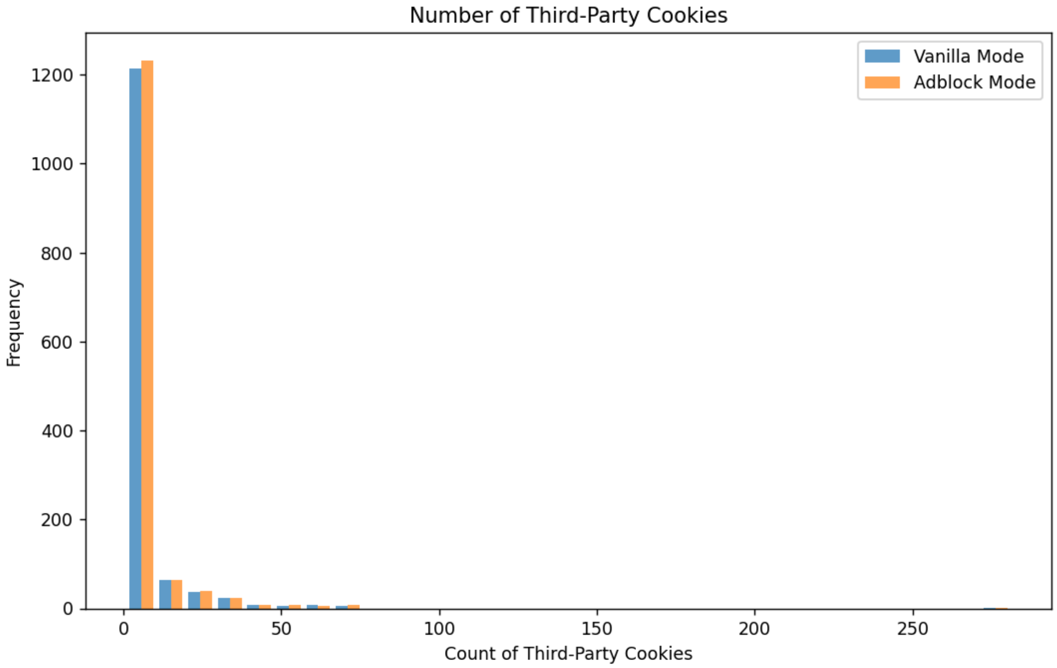


Figure 4: Distribution of Third-Party Cookies from Vanilla and Adblocker Mode

The two distributions do not differ greatly and like in our first task most of the cookies are high in frequency but low in count. In reality this is concurrent with the general hypothesis that there are few sites that set and manage the cookies for advertisers.

<u>Cookie content:</u>

.adnxs.com (WPSproject\vanilla_data\ajc.com_4d76.json):

```
"set-cookie":"XANDR_PANID=I74Xage358erj5o29yX35faVWWMyvfIMEI_5dlqRO5A
36CNqvyRPQAuDtOG0GIPUDor-2PLAvsQxgttxmy2OY4-9eDmCdf0YAv8RGU7m8n
Q.; SameSite=None; Path=/; Max-Age=7776000; Expires=Wed, 12-Mar-2025
19:58:16 GMT; Domain=.adnxs.com; Secure;
Partitioned\nreceive-cookie-deprecation=1; SameSite=None; Path=/;
Max-Age=314496000; Expires=Thu, 30-Nov-2034 19:58:16 GMT;
Domain=.adnxs.com; Secure; HttpOnly;
Partitioned\nuuid2=5213143099403789933; SameSite=None; Path=/;
Max-Age=7776000; Expires=Wed, 12-Mar-2025 19:58:16 GMT;
Domain=.adnxs.com; Secure; HttpOnly",
```

The first part of the cookie contains a user id type of string. The cookie is not restricted to one site as seen by the same site attribute set to none. It is available to all the directory pages from the site given by the path attribute. Since the domain attribute is set to .adnxs.com it will be available to that site and all its subdomains. The other attributes are to make sure the cookie information is transported securely. The other highlights of the cookie are the security against cross-scripting attacks by making it http only and the receive-cookie-deprecation attribute which is used to signal the deprecation of third-party cookies in Chrome browsers.

.rubiconproject.com (WPSproject\vanilla_data\ajc.com_4d76.json):

```
"set-cookie":"khaos=M4LQVLDT-24-1SNP; Domain=.rubiconproject.com; Path=/;
Expires=Fri, 12-Dec-2025 20:01:13 GMT; Max-Age=31536000; SameSite=None;
Secure\naudit=1|tcR/wBEzWcKA0JbQik6AZvazqvChu469akN+8AnfzNXCbQhPjIsMkT+
E+GmX6EjusT2cgpgigr3z8yenK/o+f83XpEqTIAKQRw6xunrQy4Eijy0RC4Zd8SKPLREL
hl3x0A+VO7RH1E0=; Domain=.rubiconproject.com; Path=/; Expires=Fri, 12-Dec-2025
20:01:13 GMT; Max-Age=31536000; SameSite=None; Secure",
```

The rest of these cookies will have a similar formatting where we can find a user id that will help keep track of this user across different sites. Here the id is stored under that khaos attribute and you can see it differs greatly from the previous one. Some remarkable aspects of this one is that audit attribute which is encrypted and most likely is storing important information for tracking like clicks or page visits that are important for the original site to obtain

.pubmatic.com (WPSproject\vanilla_data\app.com_d2b1.json):

```
"set-cookie": "SyncRTB4=1735171200%3A220; domain=pubmatic.com; path=/;
max-age=7776000; SameSite=None;
secure;\nipc=159706^https%3A%2F%2Fwww.app.com%2Fpbd%2Fsetuid%3Fbidd
er%3Dpubmatic%26gdpr%3D%26gdpr_consent%3D%26f%3Di%26uid%3D%23PM
UID^1^0; domain=pubmatic.com; path=/; max-age=3; SameSite=None;
secure;\npi=159706:2; domain=pubmatic.com; path=/; max-age=86400;
SameSite=None;
secure;\nKADUSERCOOKIE=03B969D1-B57D-45BA-8FAF-9FC324D38EA2;
domain=pubmatic.com; path=/; max-age=31536000; SameSite=None;
secure;\nchkChromeAb67Sec=1; domain=pubmatic.com; path=/;
max-age=7776000; SameSite=None; secure;",
```

Once again we have a user id, here the notable item is the ipc and that first attribute we see that most likely represents a session id. The user id makes this a unique string that will be linked to the user with more information than some of the other cookies we will see here. We also see the use of a browser check in this cookie (look for “Chrome” in the string)

.3lift.com (WPSproject\vanilla_data\azcentral.com_902e.json):

```
"set-cookie": "tluidp=3782794062811195484213; Path=/; Domain=.3lift.com;
Max-Age=7776000; Expires=Wed, 12 Mar 2025 19:56:40 GMT; Secure;
SameSite=None; Partitioned;\ntluid=3782794062811195484213; Max-Age=7776000;
Expires=Wed, 12 Mar 2025 19:56:40 GMT; Path=/; Domain=.3lift.com; Secure;
SameSite=None",
```

We have a unique identifier for the user here as well. Nothing upstanding here, but we can note that all of these cookies are set to be set for a long time. The Max_Age attribute is set to something around 5 to 10 years. Which makes clearing your cache and cookies super important.

.taboola.com (WPSproject\vanilla_data\app.com_d2b1.json):

```
"set-cookie": "t_gid=e27feecc-d52e-41c1-be0f-5e119ac928b2-tucte54c637;Version=1;Path=/;Domain=.taboola.com;Expires=Fri, 12-Dec-2025 19:59:51 GMT;Max-Age=31536000;Secure;SameSite=None\nt_pt_gid=e27feecc-d52e-41c1-be0f-5e119ac928b2-tucte54c637;Version=1;Path=/;Domain=.taboola.com;Expires=Fri, 12-Dec-2025 19:59:51 GMT;Max-Age=31536000;Secure;SameSite=None;Partitioned\nreceive-cookie-deprecation=1;Path=/;Domain=.taboola.com;Expires=Fri, 12-Dec-2025 19:59:51 GMT;Secure;HttpOnly;SameSite=None;Partitioned\ntaboola_session_id=v2_e71130474f8a87e9995d16b314d31fc1_e27feecc-d52e-41c1-be0f-5e119ac928b2-tucte54c637_1734033591_1734033591_CIi3jgYQ0qI9GPCU0-O7MiABKAEw4QE4kaQOQJ-FD0jMzNkDUOEEWABgAGjGo52xrKzMo-8BcAGAAQA;Version=1;Path=/gannettcompany-app/;Domain=.taboola.com;Secure;SameSite=None",
```

We have the unique identifier yet again. Something notable here is this other identifier t_pt_gid which likely is tracking a session id. Here we see the receive-cookie-deprecation attribute that we hadn' seen before but it is part of Chrome's scheme to deprecate cookie usage to make sure functionality is intact

.scorecardresearch.com (WPSproject\vanilla_data\app.com_d2b1.json):

```
"set-cookie": "UID=169a50c0c5be95a5506f41d1734033594; SameSite=None; Secure; domain=.scorecardresearch.com; path=/; max-age=33696000\nXID=169a50c0c5be95a5506f41d1734033594; SameSite=None; Secure; Partitioned; domain=.scorecardresearch.com; path=/; max-age=33696000",
```

We have that user id and an XID that can that session ID

.openx.net (WPSproject\vanilla_data\app.com_d2b1.json):

```
"set-cookie": "i=ce425abd-e406-41a2-8184-8b1b93bbfebd|1734033591; max-age=31536000; domain=openx.net; path=/; secure; SameSite=None",
```

This one is smaller but still contains that user ID and this one seems to combine the session ID right at the end.

.casalemedia.com (WPSproject\vanilla_data\app.com_d2b1.json):

```
"set-cookie": "CMID=Z1tAz8AoIlMAAHNJCBFvEwAA; Path=/;
Domain=casalemedia.com; Expires=Fri, 12 Dec 2025 20:00:15 GMT;
Max-Age=31536000; Secure; SameSite=None\nCMPS=2034; Path=/;
Domain=casalemedia.com; Expires=Wed, 12 Mar 2025 20:00:15 GMT;
Max-Age=7776000; Secure; SameSite=None\nCMPRO=2034; Path=/;
Domain=casalemedia.com; Expires=Wed, 12 Mar 2025 20:00:15 GMT;
Max-Age=7776000; Secure; SameSite=None",
```

.crwdcntrl.net (WPSproject\vanilla_data\app.com_d2b1.json):

```
"set-cookie": "_cc_dc=0;Path=/;Domain=.crwdcntrl.net;Expires=Mon,
08-Sep-2025 19:09:00
GMT;SameSite=None;Secure\n_cc_id=d18ceecdfcb615030f312f46dc9b7cae;Path=/;D
omain=.crwdcntrl.net;Expires=Mon, 08-Sep-2025 19:09:00
GMT;SameSite=None;Secure\n_cc_cc=\"ACZ4nGNQSDG0SE5NTU5JS04yMzQ1MDZIMzY0SjM
xS0m2TDJPTkxlAIL0aLu%2FDAgAAHkOC8c%3D\";Version=1;Path=/;Domain=.crwdcntrl
.net;Expires=Mon, 08-Sep-2025 19:09:00
GMT;Max-Age=23328000;SameSite=None;Secure\n_cc_aud=\"ABR4nGNgYGBIj7b7ywAHA
BgaAf4%3D\";Version=1;Path=/;Domain=.crwdcntrl.net;Expires=Mon,
08-Sep-2025 19:09:00 GMT;Max-Age=23328000;SameSite=None;Secure",
```

We have two springs here that look identifiable and they are also set to be there for a long time. We can see this cookie will be available in other sites as well thanks to that Same_Site attribute too.

.go.sonobi.com (WPSproject\vanilla_data\azcentral.com_902e.json):

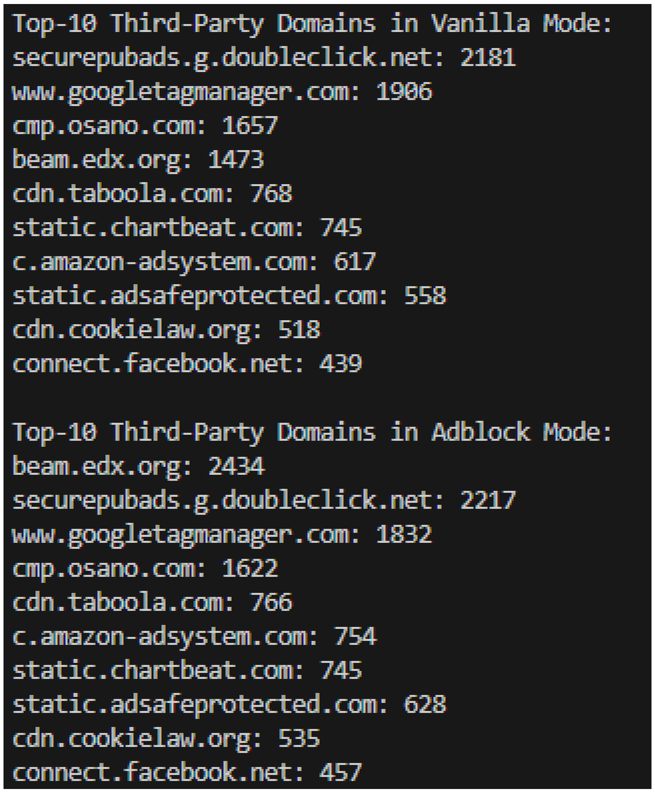
```
          "set-cookie": "__uis=2201423f-7c8a-4345-ba90-2321c742982d;
expires=Fri, 12 Dec 2025 19:52:26 GMT; domain=.go.sonobi.com; path=/;
secure; SameSite=None\nHAPLB8G=s86214|Z1s+/; path=/;
domain=.go.sonobi.com; SameSite=none; Secure",
```

We have another user id here and a string that differentiates the session as well.

Overall all cookies keep their information safe by partitioning and giving access to everything possible in the target website as well as theirs. They also set the expiration time to be super long for prolonged access. A lot of the providers of these products also are developing solutions to have more resilient cookies in hostile environments (looking for loopholes in browser and website implementation to stay within user reach)

Task 4

I added some extra information to print on this task:



```
Top-10 Third-Party Domains in Vanilla Mode:
securepubads.g.doubleclick.net: 2181
www.googletagmanager.com: 1906
cmp.osano.com: 1657
beam.edx.org: 1473
cdn.taboola.com: 768
static.chartbeat.com: 745
c.amazon-adsystem.com: 617
static.adsafeprotected.com: 558
cdn.cookielaw.org: 518
connect.facebook.net: 439

Top-10 Third-Party Domains in Adblock Mode:
beam.edx.org: 2434
securepubads.g.doubleclick.net: 2217
www.googletagmanager.com: 1832
cmp.osano.com: 1622
cdn.taboola.com: 766
c.amazon-adsystem.com: 754
static.chartbeat.com: 745
static.adsafeprotected.com: 628
cdn.cookielaw.org: 535
connect.facebook.net: 457
```

Figure 5: API calls from vanilla and adblock data

We can see a lot of APIs from big companies are being called on these websites:

securepubads.g.doubleclick.net: We explored this is class, this is used to display ads and it tracks user behavior to deliver personalized ads and manage ad delivery on websites.
cdn.taboola.com: After encountering this one before a couple of times, this one is a cooke manager that had recommendations for user and gives related items and articles to users, also used for delivering personalized content and ads
c.amazon-adsystem.com: This is the ad platform that amazon offer, similar to Google and Facebook, Amazon's system tracks user for as targeting
static.chartbeat.com: This platform provides data and analytics to global publishers as well
static.adsafeprotected.com: This domain offers a lot of advertisements to big websites like youtube, yahoo, and google
beam.edx.org: edX is an online learning platform. This domain can serve cookies related to tracking user progress in courses or personalizing learning content. It can also be user to market other courses that relate to what the user is looking for in other sites.

cmp.osano.com: Osano is a consent management platform (CMP). It manages user consent for cookies and other tracking technologies.
connect.facebook.net: Facebook's domain handles functions like social sharing, social login, and ad targeting. It tracks users across sites for personalized Facebook ads and social interactions.
cdn.cookielaw.org: is a service that provides tools for websites to comply with cookie consent laws, enabling websites to manage and obtain user consent for cookies.
www.googletagmanager.com: Google's tag management system that help websites manage and deploy cookies and tags (scripts) for tracking, analytics, and other user experience and advertisement purposes

The graph here is a little misleading because that first bar for adblocker seems a bit higher than the vanilla data but when we see the console output there are less requests in adblocker mode which is what is expected. Some sites have higher APIs and this would be interesting to investigate how the adblocker is being defeated by some sites but works very well for others.

The two distributions are similar and we can see again a big decrease in count of API calls but a high frequency for the first 100ish counts of API calls.

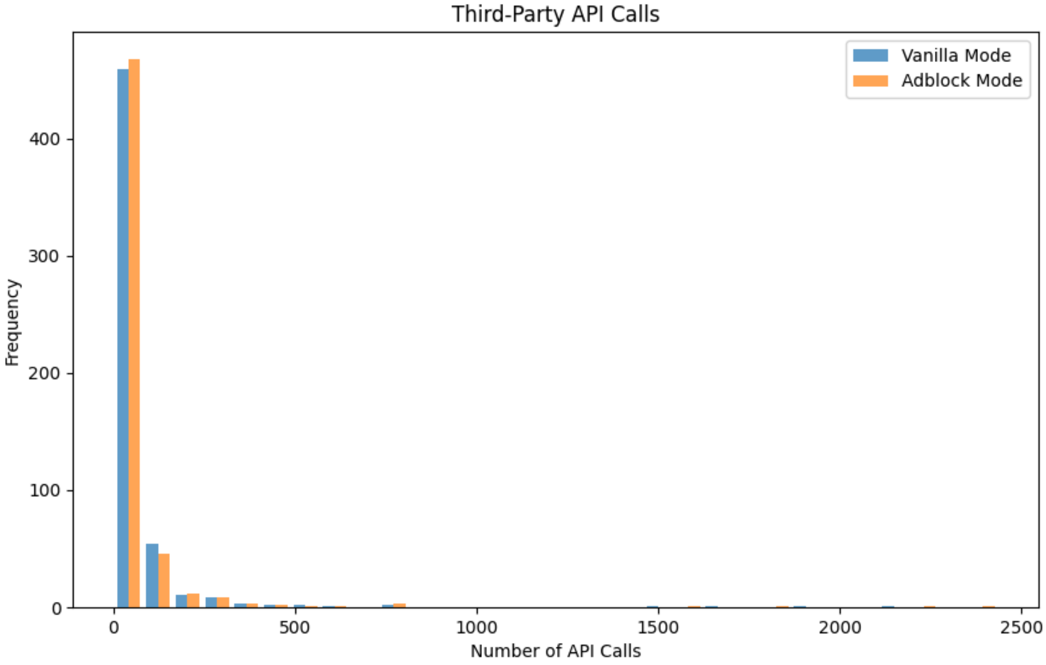
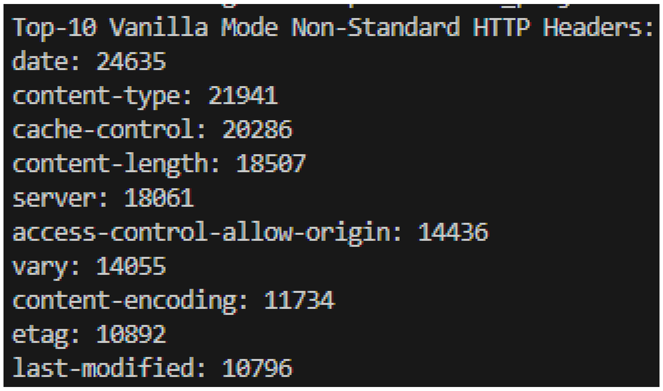


Figure 6: Distribution if API calls from vanilla mode and adblock mode

Task 5

For this task we ran the script we created and it provided us with a list of non standard http headers:



```
Top-10 Vanilla Mode Non-Standard HTTP Headers:
date: 24635
content-type: 21941
cache-control: 20286
content-length: 18507
server: 18061
access-control-allow-origin: 14436
vary: 14055
content-encoding: 11734
etag: 10892
last-modified: 10796
```

Figure 7: Top-10 Vanilla Mode Non-Standard HTTP Headers

We can see that the headers are accessing a lot of web permissions that would make it easier to create fingerprints for tracking and modify cookies and understand the security measures a user may have in place against trackers.

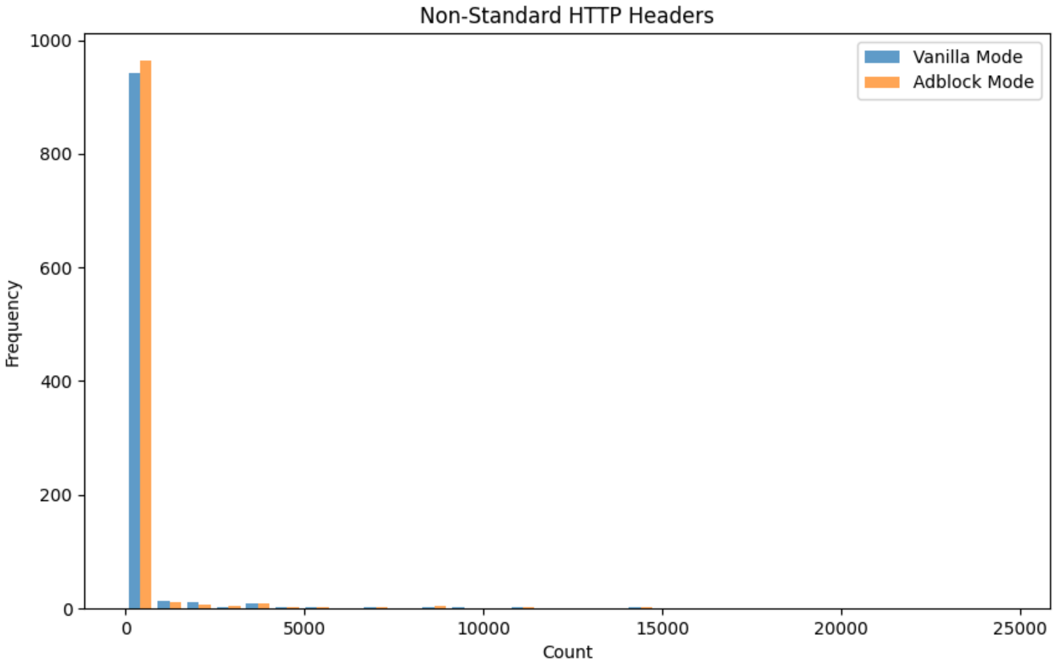


Figure 8: Distribution of non-standard http headers from vanilla mode and adblock mode

All the distributions seem very very similar which allows us to see the correlation between all of these accesses for tracking purposes. We have in this distribution that in the adblock mode there seems to be more http non-standard headers at a higher frequency in the websites we crawled. This is at par with the other distributions we saw where the frequency is very high in comparison with the counts of the element we were looking at (cookies, API calls, and third party calls). It is interesting to see that even after implementing security measures to prevent tracking, it seems not to be nearly as effective as we believe. It is also interesting to see why this is the case, after what we have discussed in class with machine learning models and resilient methods for blocking and creating the trackers, it becomes imperative we recognize how this presents in real life. This project is therefore important to recognize how the methods we studied in class are being used to penetrate and alternate the way consumers interact with items on the internet.

The analysis of third-party cookies, non-standard HTTP headers, and API calls reveals key insights into how websites track and manage user data, even with ad-blocking measures in place. Despite ad-blockers reducing the overall number of third-party cookies and some tracking mechanisms, they do not entirely eliminate them. This highlights the adaptive nature of tracking technologies and the need for more robust privacy tools which we clearly saw and studied in class. Overall there need to be systems in place that help users understand and keep privacy while online.

References

- What web beacons are and how they are used on websites and in e-mail | Securelist
- Introduction to Tag Manager - Tag Manager Help
- Taboola - Wikipedia
- Cookie Consent | Products | OneTrust
- What is Native Advertising? | Taboola
- Home Page - Scorecard Research
- Adnxs (AppNexus): What is it and what does it do? | Cookies and web tracking | The Guardian
- Magnite Inc - Wikipedia
- Pubmatic - Wikipedia
- Lotame (crwdcntrl.net) | Better
- Our Story • Sonobi
-