Ayaan Puri: 920893614
Vikram Penumarti: 920928592

Part 1:

File location:

- part1_dns/dnsclientAyaan_Puri_920893614_Vikram_Penumarti_920928592.py

In this part, we implemented a DNS client using the socket API. We constructed a DNS query manually, sent it to a public DNS resolver, parsed the reply, extracted A records for tmz.com, and then measured RTT values for both the DNS lookup and an HTTP connection to the IP address.

Constructing the DNS Request

We first built the DNS request packet. This included:

- A 12-byte DNS header (transaction ID, Standard query flags)
- The Question section (encoded domain name)

The request was made using basic Python operations and struct.pack().

Sending the DNS Query

We used a UDP socket to communicate with the public DNS resolver 1.1.1.1 and recorded the start time before sending the packet. Once the response came, we recorded the end time to calculate the DNS RTT.

Parsing the DNS Response

To parse the DNS reply:

- Unpacked the 12-byte response header and checked the RCODE. ● Skipped the Question section by reading the encoded domain name. ● Processed each resource record by reading its TYPE, CLASS, TTL, and RDLENGTH.
- For A records, we converted the RDATA field into an IPv4 address.

The resolver returned two valid A records for tmz.com.

HTTP Connection RTT

We used a TCP socket to connect to port 80 on the server, after getting the first IP address. The time taken for the TCP connect() call was measured as the RTT between our machine and the tmz.com web server.

Results

- DNS resolver RTT: ~26–29 ms
- Returned A records:
  - 13.248.160.137
  - 76.223.34.124
- TCP connect RTT to tmz.com server: ~23 ms

Part 2:

File locations:

- part2/crawl.py
- part2/analyze.py

Overview

In Part 2, we built a web crawler which visited the top 100 websites listed in the top-1m.csv file and collected HAR files containing all HTTP(S) network traffic generated. We used Selenium + BrowserMob Proxy to intercept and record HTTP requests and responses. We then analyzed these HAR files to identify requests made to third party domains and third party cookies.

Third-Party Request Analysis

From all 100 collected HAR files, our script extracted all requests and computed: total third-party requests per site, number of unique third-party domains contacted per site, and global frequency of each third-party domain.

Top 10 Most Common Third-Party Domains:

microsoft.com: 965
office.net: 955
spotifycdn.com: 266
media-amazon.com: 249
google.com: 182
googlesyndication.com: 148
rbxcdn.com: 147
tiktokcdn-us.com: 116
githubassets.com: 109
awsstatic.com: 103

A majority of top third-party domains belong to large platforms (Microsoft, Google, Amazon, TikTok, GitHub). Even sites unrelated to Microsoft often load Microsoft analytics, login, or CDN assets. Advertising domains such as googlesyndication.com appear frequently even on non-Google sites.

Third-Party Cookie Analysis

Our script scanned all HAR entries and recorded both request and response cookies.

Top 10 Most Common Third-Party Cookies:

receive-cookie-deprecation: 55
- Associated with Google and used to collect information on user behavior on multiple websites in order to optimize the relevance of advertisements on the website.

bh: 44

TiPMix: 31
- Used to track the load balancing of Azure cloud services to ensure user requests are distributed effectively.

x-ms-routing-name: 31
- Used in Azure to route a user to a deployment slot which is used for load balancing and to route traffic to deployment slots during deployments.

t_gid: 30
- Could be related to performance and optimization

t_pt_gid: 30
- Could be related to performance and optimization

audit: 29

audit_p: 29

uid: 28
- Provides a uniquely assigned, machine-generated user ID and gathers data about activity on the website. Data may be sent to a 3rd party for analysis and reporting.

khaos: 27
- Carries out information about how the end user uses the website and any advertising that the end user may have seen before visiting the said website.

Cookies from Microsoft and Google appear across unrelated sites. Advertising domains often set tens or hundreds of unique cookies per site.