Network systems: Challenge 1

By Sönke van Loh and Tjeerd Bakker

We couldn't have a TA sign of our work ("Just submit it on canvas!") so we added a little extra explanation but this did increase the page length a bit.

Findings while working through the Challange

A simple local site

What are we sending

- information about the clients machine and browser
- allowed languages
- connection type
- Host information

No ip address information -> not part of http because of different layer

Hiding identity

We cannot hide our ip. We can hide our machine information by redefining the header (User-Agent) and then sending it

A larger site

- Some of the domains are underlying structures of the website like images, json
- Some of the domain are adds
- Some of the domains are not understandable as they just link to more js scripts which are hard to read

BBC

- There are fewer obvious add sites here
- There is a site that pings the machine from time to time in order to see if one is still connected, this is disguised as an image
- There are a lot of underlying apis being accessed
- Only 6 requests go to `static.files.bbci.co.uk` (Where the articles are probably stored). The other 50 or so seem to go to external sites

Writing your own blocking program

It is hard to generally block ads and tracking since the websites are not stupid and dont put that in their names. Often when trying to block general statements ("ad") you also get alot of unwanted behaviour.

Creating an effective ad and tracking blocker:

For each website on the internet, go through each request, check if it is wanted/unwanted, if it is unwanted then block it.

- Advantages:
  - You will block ***all*** ads and tracking
- Disadvantages:
  - It takes a ***really*** long time, it is errorprone, and new websites are added daily (You are never done)

Find a list of domains that are known to serve ads or tracking, block all requests to those domains

- Advantage:
  - It's fast and easy to implement and will block most of the ads and tracking
- Disadvantage:
  - It will never be perfect, it may block wanted requests or let unwanted requests through

Our Solution

We tried both methods mentioned above. The first method gives a much better understanding of the web but also takes a lot more time. Some intresting patterns were found (some tracking tries to send an image). The method didn't work great, if you block one thing for one site then sometimes on another site something else stops working.

The second solution was much easier to implement but of course ignored the edge cases. The execution works as follows:

1. Find a list of domains that are known to serve ads and tracking (These lists were retrieved from Tjeerd's PiHole blocking list)
   - List of ad domains
   - List of tracking domains
2. Create a regular expression to get only the requested domain (~~jdldfnead9zi~~.adwebsite.com)
   - For a site like "subdomain.domain.extension", select everything before ".extension" until you encounter a dot (.)

    - **(([^.]*)(.(((com)|(net)|(de)|(nl)|(fi)|(be)|(jp)|(sg)|(uk)|(pt)|(eu)|(org)|(io)|(me)|(ru)|(dk)|(pl)|(lt)|(nz)|(au)|(cz)))))**
3. Pass any request through the RegEx, match the result against the list of known ad-serving domains, block the ones that match

Learning Outcomes

In this challenge one was supposed to learn about the workings of http. This includes the type of information and the extend of information that is shared when using web applications like websites.

insights into HTTP

http is one of the protocol layers commonly used in client-server applications. When examining http requests one can find numerous types of information beeing transmitted between the client and server. In order to examine these a simple js extension was used. When examining one could see that from the server commonly Host information and the type of request that is connected to a certain url are displayed. The information differed from request to request in more explicit things like security level information or cookies. From the user information about browser and operating system are transmitted.

How does the web work nowadays

Contrary to common believe when visiting a certain url one is not just getting that one url. Mostly there are big underlying structures which are activated when going to the specific website the url belongs too. For example every picture on these websites has its own url which is communicated via http aswell without the user taking notice of this. Adding to that there is stuff like frames in order to organize the websites, there is jason scripts that make the websites content and let the user interact with it and then there is stuff like tracking and ads. Most of the called urls when visiting a website belong to the same host, however also urls to external hosts are not uncommon. A prominent example of this would be googles ad service.

Security issues and solutions

As mentioned before there is user data transmitted in the http when excessing a site. This can be prevented by rerouting the http request and changing values in some of its headers like the one carrying the users info. However this cant deal with the fact that the server has access to things like ip-addresses which cant be changed this way as once they are not part of the http protocol and second they are strictly necessary in order to have communication between server and client. Then there is stuff like ads and tracking which one would like to get rid off aswell. This is possible via http as for example certain hosts can be blocked. However this often has drawbacks for functionality.