Olivia Malcolmson

Nyah Harrison

Carl Ekholm

Amanda Anowi

Progress Report

Current Progress:

Since the project proposal, we have collected a list of 24 unique domains that we as a group use daily. The majority of our project will center on analyzing data from these domains. For milestone 1, we focus on extracting the data we will need for later analysis (milestone 2). Given that our analysis is centered on finding patterns/graphing/identifying what services receive the most traffic from our daily use, we've identified three types of data that could yield useful insights: CDNs, AS paths, and geographic location. We ran paris traceroutes on each of the domains in our collection, and saved the output in a folder TracerouteOutput. We created a script (traceASPaths.py) that analyzes paris traceroute output on each of our domains of interest to parse for a list of AS organizations traversed. Additionally, we have a script that goes through the destination IP in each traceroute file, and retrieves the AS name, then checks if the IP is in our list of CDNs. If it is, then we add it to a file, cdns.txt. We can analyze this data later on to see what CDN's receive the most traffic, location of these CDNs, etc. Lastly, we wrote a script (get_geolocation.py) to again process the traceroute files, extract their IP addresses, and retrieve geolocation information for each IP address. We use a token received from ipinfo.io to make queries for the geographic location, and we save all of this information to an output file (geolocations.txt).

Future Tasks:

Are there other CDN's accessed when we load a webpage? In order to find out, we will experiment by extracting data using the chrome dev tool to see what other third parties help load the contents of a domain's webpage. Websites will oftentimes query other websites in order to get particular assets, so seeing if CDN's are used will allow for more data we can use for analysis.

Additionally we hope to write code to create graphs and groupings. Our aim is to find patterns. What locations/CDNs/Ases are getting the most traffic? Is there a reason or pattern we can identify? We may potentially need to make more queries and collect more data. But the extraction we have done so far provides a solid base. We also plan to explore some more research papers to identify data we can collect - like Matteo Dell'Amico's "Lean on me: Mining internet Services Dependencies from Large scale DNS Data." In the end, we hope we can make posters to put around Colgate so students can get an idea of what happens when they use these domains - since a lot of the domains in our data set are also largely used by Colgate students, e.g moodle, gradescope, dine on campus, etc.

How do our future tasks align with our original project plan:

In our proposal, we focused on identifying the most commonly used domains among Colgate students and figuring out what goes on in the background when we log on to those domains or go on those websites. With our now collected CDN, Geolocation, and AS data, we can begin to craft a picture of this behind-the-scenes. Now we can move on to find patterns like which ASes are traversed through the most often in order to access our favorite domains, the locations that a lot of the traffic is routed to, etc.