Do You Know How Far Your Data Goes? – Following Network Traffic of Sites Commonly Used by Colgate Students

Amanda Anowi, Carl Ekholm, Nyah Harrison, Olivia Malcolmson

Project Overview

Our capstone project primarily focused on what happens when looking up a website you use in your everyday life, such as moodle, instagram, linkedin, etc. Specifically, we focused on the geographical locations of popular domains' network infrastructures, as well as the CDNs and ASes (CDN stands for Content Delivery Network and is a network of interconnected servers: AS stands for Autonomous System and is a group of networks with a routing policy). The motivation behind this project was that we wanted students at Colgate to be more aware of the digital services that make up their digital footprint, as there are many moving parts operating in the background that you would normally not know about or expect, such as the other domains that websites rely on in order to get all of their artifacts that make up their website.

Process

To conduct our research, each team member was responsible for coming up with their own most commonly used domains, which resulted in a total of 40 domains. From this we then filtered for unique domains and analyzed them using Google Chrome's Developer Tools to track network activity. Within the Network tab, we logged all related domains that were queried by the original domain. We exported this data to .har files (JSON formatted file of all network interactions between the web browser and the site) and extracted the URLs of the domains that were being queried. This provided in total around 2000 supporting domains that could then be used for further analysis.

Next, we ran various scripts on both the original and supporting domains to track the following information:

- Commonly used Content Delivery Networks (CDNs) and Autonomous Systems (ASes)
- Network activity path, specifically capturing data on intermediate routers
- Geographic locations of intermediate and end servers

Our original domains:

moodle.colgate.edu
colgate.edu
portal.colgate.edu
google.com
instagram.com
netflix.com
amazon.com
dineoncampus.com
gradescope.com
gmail.com
chat.openai.com
youtube.com
linkedin.com
docs.google.com
open.spotify.com
x.com
buff.163.com
di.se
docs.google.com
www.notion.so
github.com
drive.google.com
pandas.pydata.org

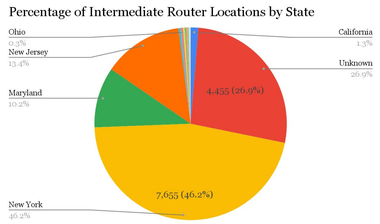
Case Study

DineonCampus is a platform that many Colgate students use to check daily dining hall menus. When you visit the Dine on Campus website, additional domains (tracked by the chrome dev tool) are used to load its features.. A good amount of these supporting domains retrieve their data exclusively from the popular CDN provider, Cloudflare.

The location Dine on Campus frequents the most when loading its data is Secaucus, New Jersey. But overall, It routes its data between three different locations: San Francisco, California; Secaucus, New Jersey; and New York City, New York.

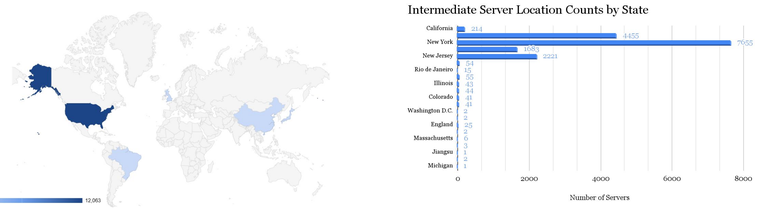
Intermediate Servers

Unsurprisingly, the geographic locations of almost half of the intermediate routers were in New York. This aligns with our expectations as the location where we were querying from is in New York.

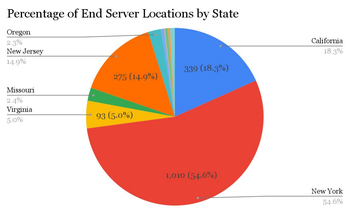


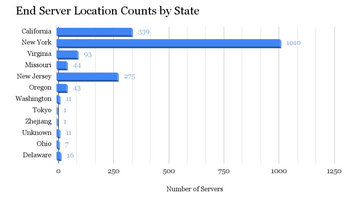
Looking beyond the U.S., there were actually a handful of servers in other countries such as China, Japan, England, and Brazil.

When manually checking the responses to IP address queries that returned unknown, many revealed a 'bogon' tag, indicating that the particular packet in transit may not be from the reserved address range it claims to be.



End Servers



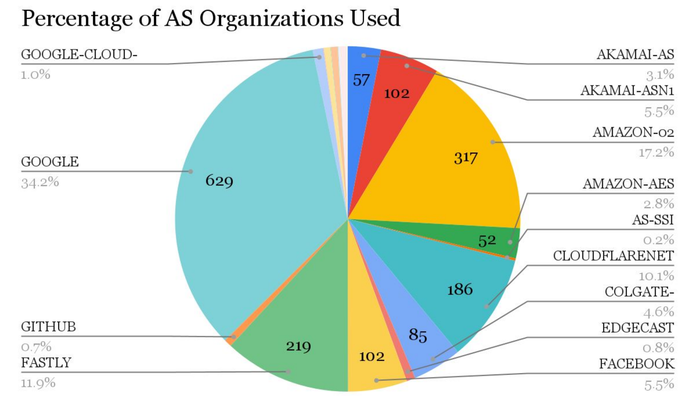


The range of end server locations was a bit more limited in comparison to intermediate routers, with only a handful being located outside of the U.S. Within the U.S., New York was the state with the highest number of servers. Interestingly, California was the second.

When it comes to optimizing network activity, it makes sense that users would be routed to the nearest available servers. However, regional availability can depend on multiple factors like resources or population. NYC had the highest number of end servers compared to other U.S. cities like San Francisco (relative to our data). NYC's large population makes the destination an ideal server location.

Autonomous Systems

Approximately a fifth of our original domains are services owned by Google, making Google representative of almost a third of all tracked AS organizations used by our domains. Amazon was the second largest, used by Amazon, gradescope, Netflix, buff.163.com, Colgate's Moodle page, and di.se, to name a few. Tracking autonomous system usage was also not solely restricted to the original domains but included querying those used by supporting domains.



CDNS

When we browse platforms like Spotify, a lot of the data we see actually comes from Content Delivery Networks (CDNs), not directly from the website's own servers. CDNs store copies of website content in various places around the world to make it load faster for users.

In our study, we noticed that many of the websites we looked at were using CDNs, with Fastly and CloudFlarenet being the most popular ones. This means that a large number of websites we visit route their data through these CDNs to speed up loading times and improve our browsing experience.

Very few sites these days are strictly text. Loading in multimedia content such as images or videos is often delegated to CDNs for optimal load-balancing.