Amanda Anowi
Olivia Malcolmson
Carl Ekholm
Nyah Harrison

Analyzing Network Usage for students

Project Objective

Our research focuses on identifying the most commonly used network infrastructure across different geographic locations. Specifically, we aim to recognize dominant Internet Service Providers, Autonomous systems, and CDNs associated with specific websites. From this data, we will determine if there are any kinds of classifications or groupings we can make. The project motivation is for one, to allow students to be more aware of the digital services that comprise their digital footprint. Additionally, this research will provide valuable insights for any future research reliant on AS or name server information for commonly used websites.

Survey of Related Prior Artifacts

This research paper titled “Mobile Internet Usage - Network Traffic Measurement” speaks about phones and how their connection to the internet works. In this paper it talks about how a mobile phone will connect to a core network and this core network consists of two different domains. One that is packet switched and the other being circuit switched. This would be something that would be interesting to cover in our project if there is a way to gain access to the two different domains, as we do not know the domain of any app that we open up on our phones (page 16 of the paper).

Additionally, the research paper “Understanding Online Social Network Usage” focuses on tracking data usage specifically for online social networks (OSN’s). Specifically, they aim to answer the question of which features on an OSN due users interact with the most. Their methodology includes extracting clickstreams from passively monitor network traffic, followed by filtering such grouping clickstreams into categories. This research informs our project as it emphasizes the importance of analyzing network traffic to understand user behavior and

preferences. As we analyze data usage in our project, we may use similar filtering techniques to identify network usage patterns.

The nature of the geographic location information that our project seeks to obtain closely aligns with the classifications defined in a paper by Wang et al.[2] which differentiates locations by sources such as the provider, content, and server. Similarly, our team is interested in distinguishing the different locations of ISP and AS Organizations, their respective servers, and the CDNs used to support website queries. However, our work will not explicitly concern content locations in the same manner as we are not focused on websites with location-specific content.

Lastly, “Moving Beyond End-to-End Path Information to Optimize CDN Performance” by Rupa Krishnan is highly relevant to our project objective as our project aims to analyze network infrastructure usage patterns including CDNs. This paper directly addresses the optimization of CDN performance by analyzing latencies measured from servers in Google’s CDN to clients across the internet. The paper also analyzes the effectiveness of latency-based server selection in CDNs which aligns with our objective of identifying the most accessed services and optimizing network performance.

Project Plan:

Our project plan is divided into 3-4 general steps. The first step is accumulating the raw data we will use to research the geographic locations of CDN and name servers (along with the AS organization responsible for them) that responded to our queries, and the geographic paths our queries took. We intend on collecting a list of the most frequently visited websites Colgate students use within a 24-hour period. Each team member will be responsible for contributing to this list. The second step is where we will process and run our raw data through a series of Python scripts that will conduct traceroute and DNS lookup queries on these websites. As a team, we will outline the specific tasks of each script to achieve our intended objectives. This may include (but is not limited to) the general structure and goal of each script, what data structures we will use to store and pass information along to each other, and what external

support mechanisms we may need (e.g., setting up Docker containers, running traceroute, how we intend to perform DNS lookup queries, etc.). Afterwards, we expect to divide writing the scripts amongst ourselves. We expect to have approximately half the scripts by milestone 1, and finished and tested scripts by milestone 2. When all of our programs have been written and tested on small sets of data, we will likely allocate specific research objectives to team members. For instance, one individual may run traceroute to extract the paths for each website queried. This information may then be passed along to another team member who would extract the name servers. The specifications of how exactly this process may be divided amongst the team is subject to change though as we determine the structure of our scripts.

After we have the data for each research objective, we will consolidate to further analyze any trends or patterns. Within this process, we will begin working on constructing our artifacts as we determine the best ways to meaningfully present our work. We will likely populate our data into a pandas DataFrame or some other data structure that we can then use to produce graphs and examine our data visually.

Our Artifacts (our deliverables)

For our deliverables, we intend on presenting all of our data through maps and graphs. We intend to use multiple maps to show the geographic locations/areas where our responding AS organizations, CDN and name servers are stationed. These locations will be marked with icons that we will either organize by some color coding or scaling scheme to indicate how many queries made contact with those places. Additionally, we will also have maps that present the most common geographic routes our data and queries traversed. We also intend to present graphs categorizing our data along any relevant trends that we see based on the results of our research.

References and artifacts

[1] Krishnan, Rupa, et al. "Moving beyond end-to-end path information to optimize CDN performance." *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement*. 2009. https://research.google/pubs/moving-beyond-end-to-end-path-information-to-optimize-cdn-performance/

[2] Riikonen, A. 2009. Mobile Internet Usage - Network Traffic Measurements. https://www.researchgate.net/publication/258109957_Mobile_Internet_Usage_-_Network_Traffic_Measurements

[3] Chuang Wang, Xing Xie, Lee Wang, Yansheng Lu, and Wei-Ying Ma. 2005. Detecting geographic locations from web resources. In Proceedings of the 2005 workshop on Geographic information retrieval (GIR '05). Association for Computing Machinery, New York, NY, USA, 17–24. https://doi.org/10.1145/1096985.1096991

[4] Fabian Schneider, et al. “Understanding Online Social Network Usage from a Network Perspective.” Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement 2009, Chicago, Illinois, USA, November 4-6, 2009. https://www.researchgate.net/publication/221612053_Understanding_online_social_network_usage_from_a_network_perspective