FURMAN COMPUTER SCIENCE

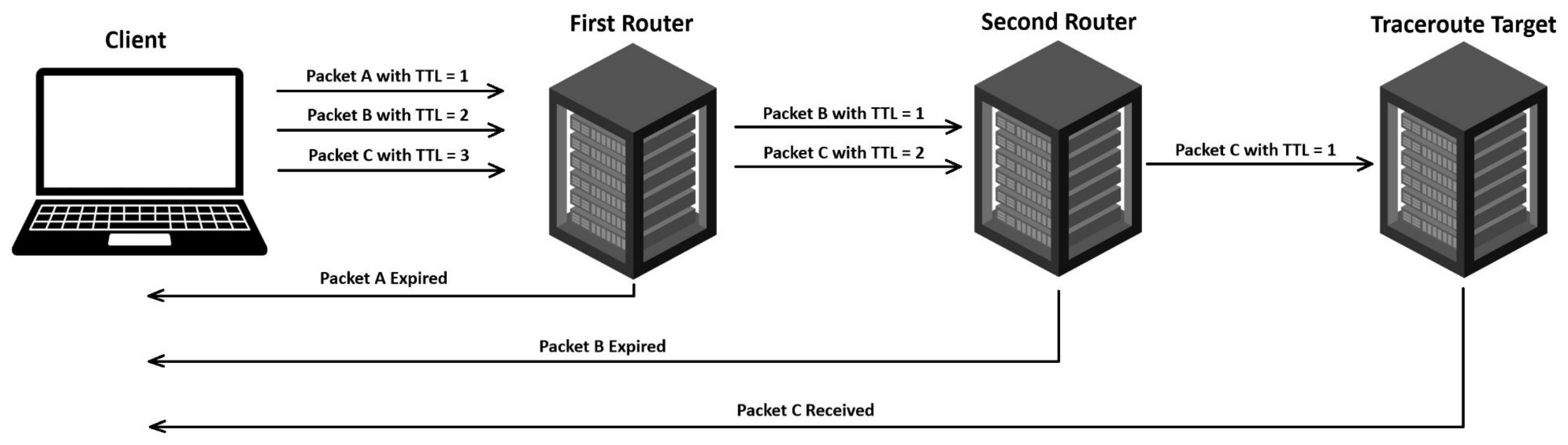
Visualizing Internet Traffic Using TCP Traceroute

Jack Crouse

ABSTRACT

To support today's digital world, there exists an extremely large network of infrastructure that orchestrates the routing of internet traffic to its destination. The purpose of this project was to come up with a way to visualize the path taken over the internet in as realistic and reliable a way as possible. To do this, a server located in New York City was obtained. This server hosts a simple website capable of finding the geographic path taken to a host. When a visitor provides it with a host, a TCP traceroute diagnostic test is performed to find the IP address of each stop taken along the way. These IP addresses are used to find the approximate geographic location of the stops, which can then be used to visualize the overall path taken. The server also performed routine traceroutes on hosts that are spread out across the world. The reason for this data collection was to learn how paths compare to one another and change over time.

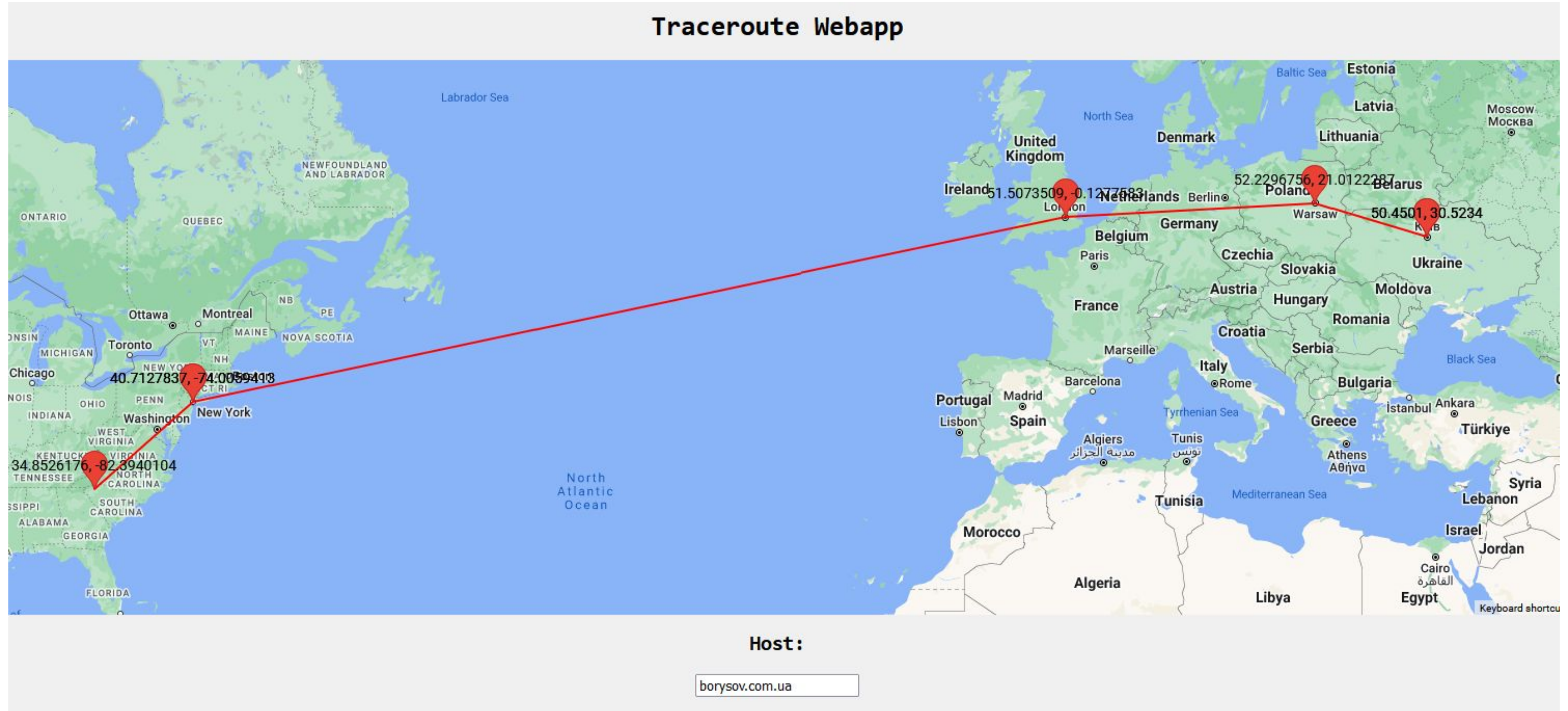
METHOD



TCP traceroute is a technique used to approximate the flow of traffic from point A to point B. The process utilizes a concept in networking called "time to live" or TTL, which starts as a number between 0-255. TTL indicates how long a packet can be handled by routers before it is discarded. When a router receives a packet, it decrements the TTL by 1 before forwarding it. If the packet's TTL reaches 0, the router sends an error message indicating that the packet has expired back to the packets source. TCP traceroute works by sending a series of crafted packets to the target destination. The first packet in the group has a TTL of 1. Each consecutive packet has its TTL incremented by 1, resulting in a bunch of packets being sent with TTL's in a series that increases by 1 (1, 2, 3…). Because the TTL causes these packets to expire, the device performing the traceroute can find out where the packets are routed by recording the source of each packets expiration error message along the way, until it receives a packet from the target destination. It is worth noting that traceroute is traditionally done using a different protocol called ICMP, but due to unsatisfactory results, TCP traceroute was used.

https://jacrouse.github.io

WEBSITE


Traceroute Webapp

Host:

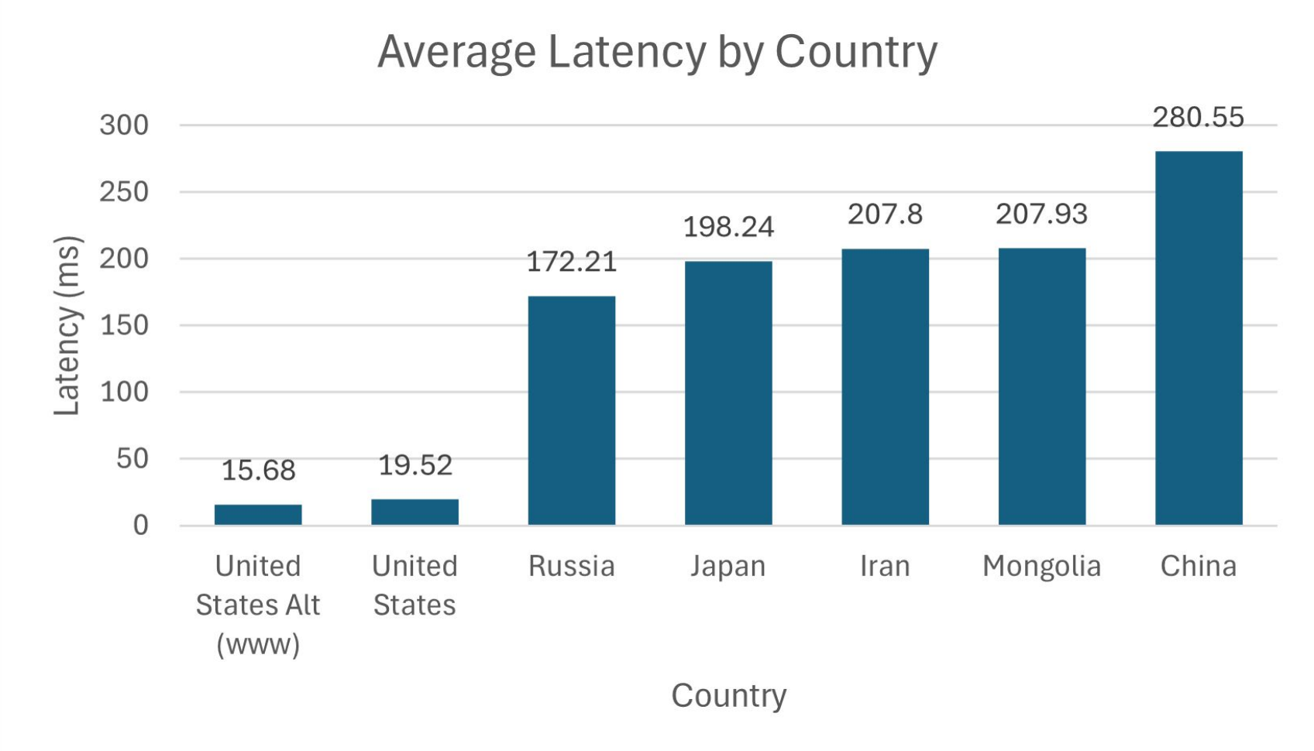
borysov.com.ua

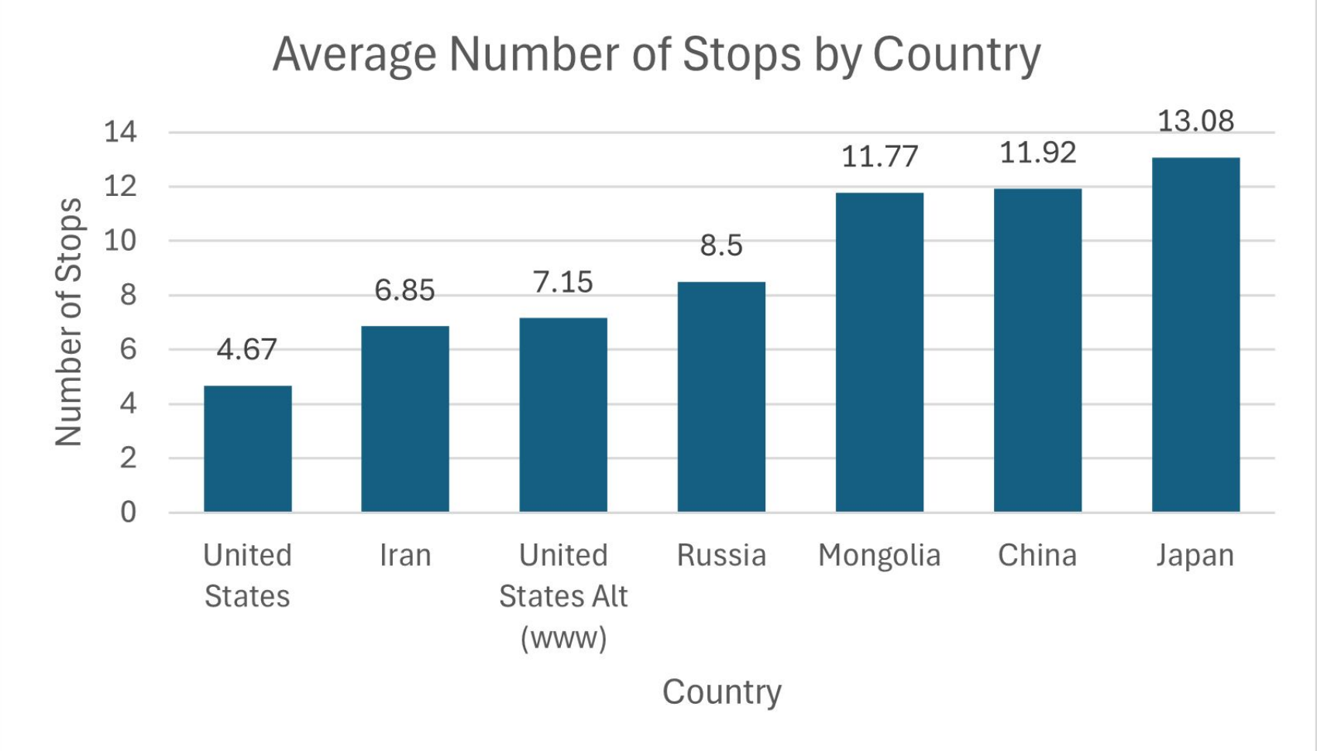
Above is an example of the output of the website when tracing the host "borysov.com.ua", which is a website located in Kyiv, Ukraine. The path taken was from Greenville, SC → New York City, NY → London, UK → Warsaw, Poland → Kyiv, Ukraine.

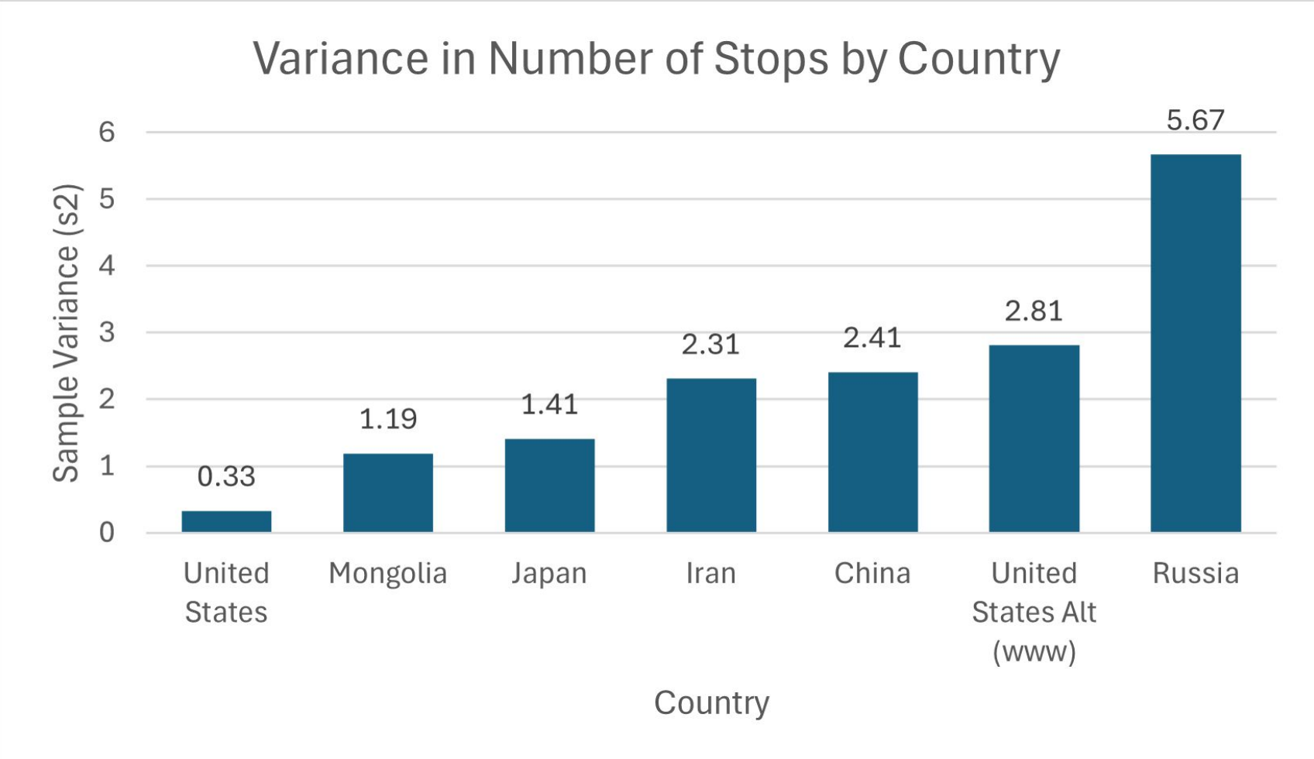
DATA COLLECTION AND ANALYSIS

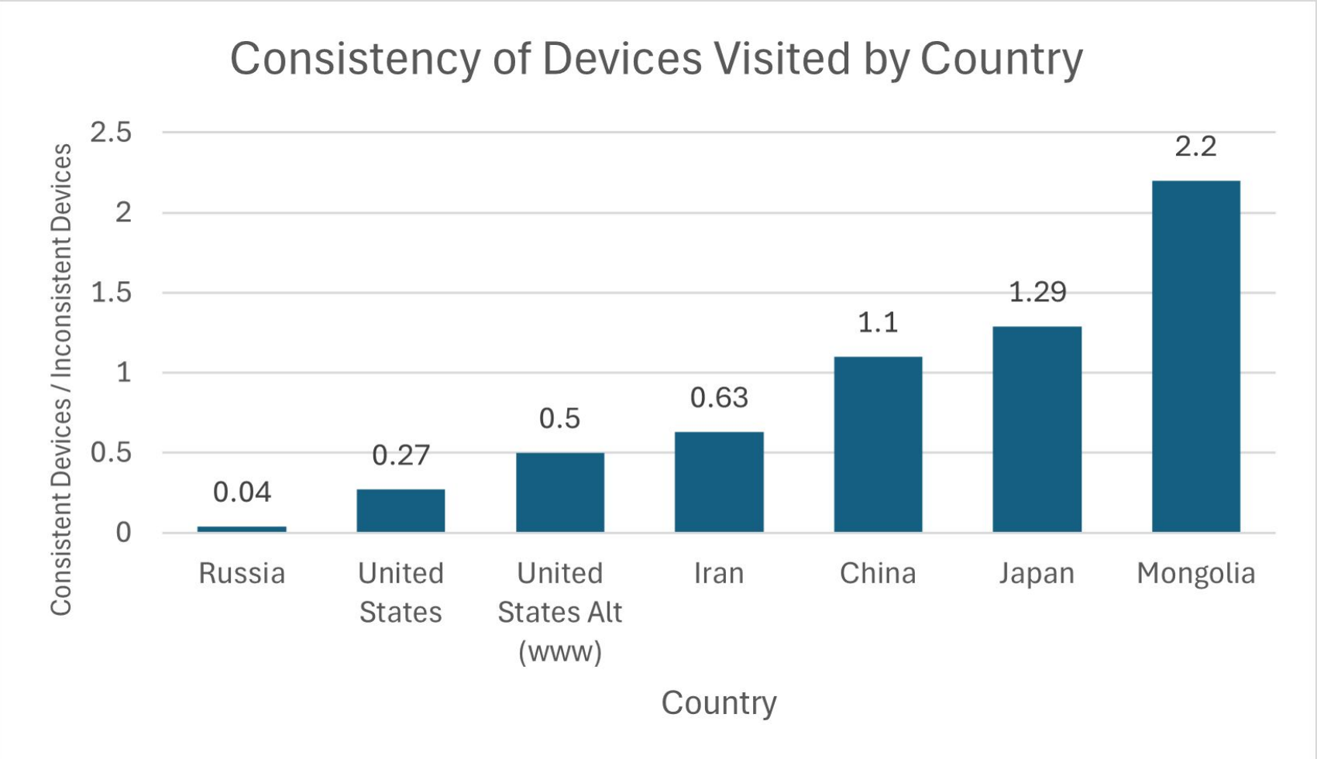
In addition to hosting the website, the server periodically performed and recorded traceroutes on several hosts around the world. The data consists of 91 traceroutes performed on 7 different hosts, with each scan being done every 2 hours. The websites were chosen based on their respective geographic locations.

Location	Hostname	Distance From Server (km)
Kansas City, MO, United States Alt (www)	www.redfishportorford.com	1,758.92
Los Angeles, CA, United States	redfishportorford.com	3,935.74
Moscow, Russia	mail.ru	7,510.31
Tehran, Iran	sbu.ac.ir	9,856.23
Ulaanbatar, Mongolia	news.mn	10,158.24
Tokyo, Japan	webtan.impress.co.jp	10,848.68
Shanghai, China	ftp.sjtu.edu.cn	11,859.49


Average Latency by Country


Average Number of Stops by Country


Variance in Number of Stops by Country

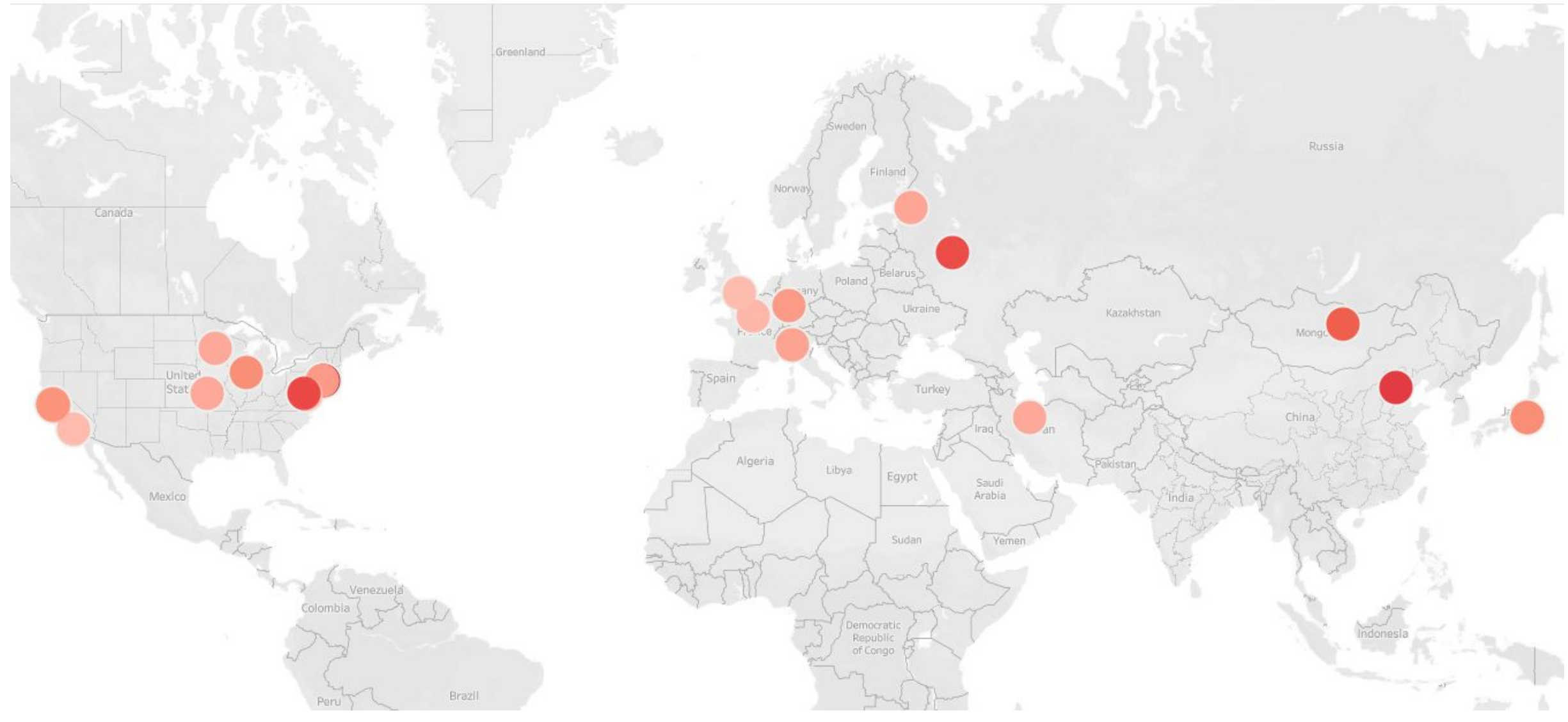

Consistency of Devices Visited by Country

RESULTS FROM WEBSITE

- Repeatedly tracing the same host can result in the exact same or very different paths, depending on the host. This could be happening for a multitude of reasons. One is load balancing, which is when DNS servers or routers manipulate the flow of traffic so that the workload is more evenly distributed. This helps prevent congestion that could occur due to networks becoming overwhelmed.
- Domestic and foreign high-traffic hosts will often times appear to exist conveniently closeby in the United States, despite this not being where their main servers reside. This is likely due to systemic caching of certain hosts through services like content delivery networks (CDNs), which is done to increase speed and availability for users. This phenomenon appears to be less likely to occur with smaller, low-traffic websites.

RESULTS FROM DATA

Frequency Map of Locations Visited During Traces



- Traffic sometimes takes seemingly nonsensical paths. The data collected recorded multiple instances of traffic bouncing around the United States before leaving the continent.
- Interestingly, the inclusion or exclusion of "www." before a website can result in a completely different path/destination. This could be due to DNS servers providing different information for the two hostnames, despite them both providing the same website. It could also be CDNs caching the same website in different locations.

CONCLUSION

Establishing meaningful patterns using the results of TCP traceroute is a difficult task. Factors such as the configuration of routers, load balancing, and redundant storage of content mean that the path taken at any given time is highly unpredictable. Performing research using more data, different techniques, or tools like Paris Traceroute and MTR would be useful in learning more about internet topology.