ECE46300/54700 Intro to Computer Communication Networks

Project Title: IP Geolocation using Traceroute and IP Interpolation

Partners: Marcel McCormick, Karthik Banakar, Shree Wooradi, Josue Gutierrez

1. Introduction

Geolocation is a term used to describe the methods used to determine the location of someone or something on the Earth. One of the methods for geolocation is the use of IP addresses. There are a multitude of problems in our world in which accurate geolocation would prove to be useful. Emergency services, such as police and fire departments, as well as emergency medical and rescue services, are a great example. This would include ambulances, fire departments, police, and rescue services that pertain to securing the safety of a civilian. A study states that geolocation technology can lower the response time of an emergency service arriving to whoever is in need. This is known because based on this study, in 10% of emergency calls, a person requesting a rescue service did not know their exact location. Being able to identify the position of a caller in an emergency could result in time trying to figure out the location being spent on getting to the actual person. This situation's setting is based on phone calls being made, however, there are situations where the use of IP addresses to pinpoint a location would work alternatively in place of phone calls [1]. Moreover, another issue that geolocation can help resolve is fraud detection and prevention. The idea is that financial institutions, banking companies, or any e-commerce business that conducts online transactions can use geolocation data and trace the location based on the IP address. Then this technology could identify login locations, and transaction authorization, then based on this information, companies can make the assumption if the action is authentic [2]. Lastly, geolocation can be used as a tool in the cyber security world. Due to attacks in the digital and real world being a very real danger, geolocation based on IP addresses can be used to trace those responsible. This could include the cybersecurity in a company, an individual, or even national security. As mentioned in the referenced IEEE article, by using geolocation, an investigation showed that there was a lot of hacking activity whose target was the United States. Cyber security firm Mandiant was able to trace this activity to an office building in Shanghai with the help of this geolocation technology [3]. It is because of the prevalence of the aforementioned issues that plague the world that our team aims to provide an accurate way of determining the geolocation based on IP addresses, in conjunction with, IP interpolation methodology which utilizes known data points to derive location coordinates. Now when discussing the objective of our project, it is important to state we needed to find related work to make sure we could reproduce similar results from the initial paper assigned to us. **[Josue]**

2. Related Work

When researching the work outside of the paper given to us, we were able to find several related articles. When researching the related work, the team had to create a database of IP addresses which the team could use to find the geolocation, and while creating the database of IP ranges using the ipinfo(use it in a reference) it was found that, not all IPs can be used or assigned by the Internet Service Providers. The **Internet Assigned Numbers Authority (IANA)** is responsible for distributing large sets of IP addresses to various regions and Internet Service Providers. There are currently five Regional Internet Registries (RIR) [4]:

1. **AFRINIC**-African region
2. **APNIC**-Asia Pacific region
3. **ARIN**-North America and several Caribbean and North Atlantic islands
4. **LACNIC**-Latin America and the Caribbean
5. **RIPE NCC-**Europe, the Middle East, and parts of Central Asia

These RIRs are responsible for distributing the IP's to the local ISP's.

There exist different classes of IP addresses (Classes A-E) which are used for different purposes. **Class A** is for extensive communications which can accommodate a large number of network hosts such as universities and telecommunications networks where the address ranges from 10.0.0.0 to 10.255.255.255. **Class B** is used for medium and large-sized networks, namely, small businesses and home networks where the address ranges from 172.16.0.0 to 172.31.255.255. **Class C** addresses are used in small local area networks (LANs) where the address ranges from 192.168.0.0 to 192.168.255.255. **Class D** only works as a multicast service, for things such as streaming audio and video for applications like Netflix, YouTube, Spotify. Hosts are unable to use this class. Lastly, **Class E** addresses reserved for research and development purposes are unavailable for general use [8]. **[Shree Ganesh]**

Many valuable studies have been done in the realm of geolocation. One study in particular completed by Shichang Ding, Xiangyang Luo, Meijuan Yin, Yan Liu, Fenlin Liu in 2015 proposed a method for finding geolocation based on the use of well-known subnetworks. In the study, the team used the covariances and variances of network delay and distance as a means of finding the delay-distance correlation of a network. Using the value obtained from these calculations allowed for their team to classify whether a subnetwork was strongly or weakly linked. If a subnetwork was strongly linked then it would be used in helping determine the geolocation of IPs in the other remaining weakly linked networks. Their team found that other geolocation techniques saw significant improvement overall when employed with their new found subnetwork technique [6] Another study completed by a team from the Zhengzhou Science and Technology Institute, aimed at being able to accurately determine the region that an IP address resides

in. The team in this research acknowledges the weakness that delay-distance correlation from the aforementioned 2015 research project may hold and proposed instead to use the features that are intrinsic to the mapping of paths between IP addresses. They were able to find in their research that using the path features provided high accuracy in determining the region of an IP address [7]. While both of these methods have different approaches and utilize different geolocation techniques they both shine light on the fact that using different geolocation techniques. **[Marcel]**

3. Implementation and Results

For this project, we have used a select amount of **Class A** and **Class B** addresses to find the geolocation. The **Class A** addresses include IP addresses of various universities located in different parts of the United States while the **Class B** addresses include the IP range taken from one of the team members IP who is currently subscribed to Comcast.

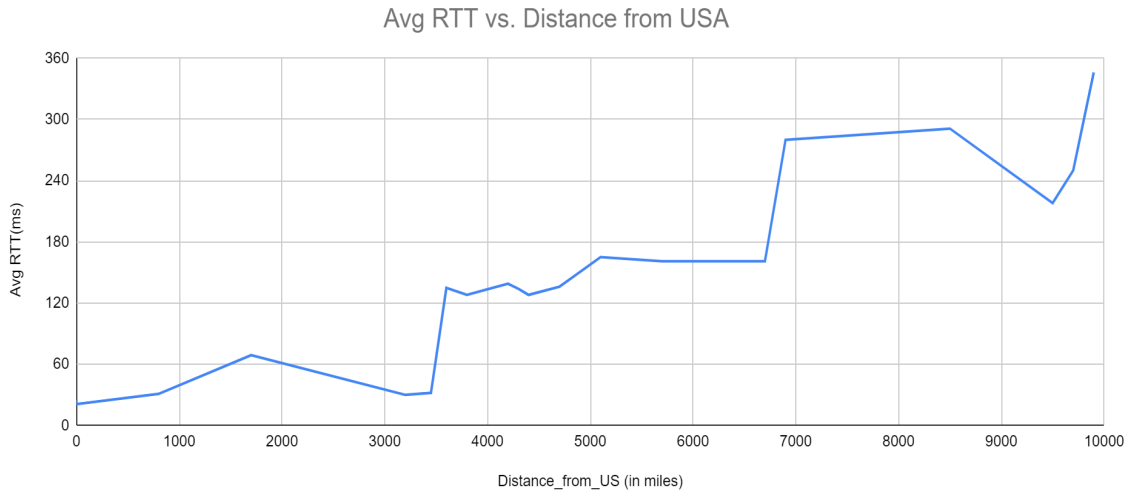
3.1 Method 1: Using Traceroute on Class A IP's and finding the country of the IP address

In this method we are using the traceroute command on 20 IP addresses of various universities. The collected IP addresses are almost equidistant with each other. The **`tracert`** and **`ping`** command in the Windows command prompt to trace the route of a packet to an IP address over multiple hops through various servers, as shown in Figure 1. The journey of the packet begins from the first hop which is usually the local DNS server and subsequently to the DNS lookup tables of the ISP. The default number of hops for any given IP address is 30 hops, but the ping packets can also reach their destination in less than 30 Hops For every hop, three Round Trip Times are recorded for three different packets.

```
C:\Users\PNW_checkout>tracert purdue.edu

Tracing route to purdue.edu [128.210.7.200]
over a maximum of 30 hops:

  1     7 ms     6 ms     7 ms  10.0.0.1
  2    20 ms    17 ms    14 ms  96.120.28.1
  3    21 ms    12 ms    16 ms  po-303-1216-rur02.hammond.in.chicago.comcast.net [68.85.181.189]
  4    20 ms    33 ms    18 ms  po-2-rur01.hammond.in.chicago.comcast.net [162.151.184.53]
  5    34 ms    59 ms   123 ms  be-171-ar01.area4.il.chicago.comcast.net [68.85.183.165]
  6    73 ms     *        *     be-32241-cs24.350ecermak.il.ibone.comcast.net [96.110.40.61]
  7    17 ms    23 ms    16 ms  be-2411-pe11.350ecermak.il.ibone.comcast.net [96.110.33.206]
  8    18 ms    19 ms    19 ms  96-87-9-106-static.hfc.comcastbusiness.net [96.87.9.106]
  9     *        *        *     Request timed out.
 10     *        *        *     Request timed out.
 11     *        *        *     Request timed out.
 12    25 ms    18 ms    18 ms  et-2-2-4.1235.rtr.ll.indiana.gigapop.net [64.57.21.174]
 13    22 ms    20 ms    22 ms  lamb-20-c7710-01-ptp-po103-891.tcom.purdue.edu [192.5.40.185]
 14     *        *        *     Request timed out.
 15     *        *        *     Request timed out.
 16     *        *        *     Request timed out.
 17     *        *        *     Request timed out.
 18     *        *        *     Request timed out.
 19    37 ms    28 ms    25 ms  128.210.7.200

Trace complete.
```

Fig. 1 Example of **`tracert`** command

With the help of the **`ping`** command, the average RTT for that IP address can be found. For each of the 20 IP's we have performed **`tracert`** and ping which gave the route of the packet and average RTT. If the geolocation of the IP address is outside the United States with every 500-1000 miles increase in distance, the Average RTT time increases which proves that the packet is traveling a large distance to reach its destination. For every 500-1000 miles, a different RTT was found, Using Graph 1, we can determine the country or neighboring country to which the IP belongs to. The IP's can be collected using the nslookup [9] tool available online.



Graph 1. Average RTT vs Distance from US

The main motive of using the **`tracert`** command is to find the IP geolocation of various **Class A** IPs. With the help of **`tracert`** command the geolocation of **Class A** IP's can be even more precisely estimated by using the total round trip time of the packet sent from the user's computer and IP interpolation techniques. If the RTT can be determined multiple times during a day, then the accuracy can also be improved. The accuracy can also be improved by performing tracert on more IP's located in neighboring cities, states and countries. The **`tracert`** command takes a lot of time for every given IP and hence it is almost impossible to **`tracert`** every possible IP combination hence we are using IP interpolation, which is implemented in Method 2. **[Shree Ganesh]**

Our team split our approach into a multistep process. The first step of the process being to obtain a large ground truth seed. The dataset utilized in this study was obtained from ipap.is [10], a platform that utilizes public IP addresses to populate its database. After obtaining the large amount of IP data, the next step was to perform IP interpolation on sets of IP addresses.

3.2 Method 2: The IP Linear Interpolation[5] Method using Class B IP's

The IP interpolation attempt initially refers to the dataset containing Indiana's IP addresses. This technique makes use of IP addresses and their corresponding locations to give an estimated location of the unknown IP address. Lets understand this through an example, consider an unknown IP 65.175.34.127 which is located at Hammond. The python code developed for this project searches among the known IP addresses in the database. The linear interpolation formula [12] is represented as:

```
Est_Value=Strt_Value+(Rel_Position×(End_Value-Strt_Value))
```

In our case, the "Start Value" and "End Value" represent the latitudes or longitudes of the start and end points of the IP address range. The "Relative Position" is the position of the unknown IP address within the range, normalized to a value between 0 and 1.
For latitude (or longitude) estimation:

```
Est_Lat=Strt_Lat+(Rel_Position×(End_Lat-Strt_Lat))
Est_Lon=Strt_Lon+(Rel_Position×(End_Lon-Strt_Lon))
```

Where,

Est_Lat	: Estimated Latitude value
Rel_Position	: Relative Position of unknown IP address
Strt_Lat	: Start value of latitude
End_Lat	: end value of longitude
Strt_Lon	: start value of longitude
End_Lon	: end value of longitude

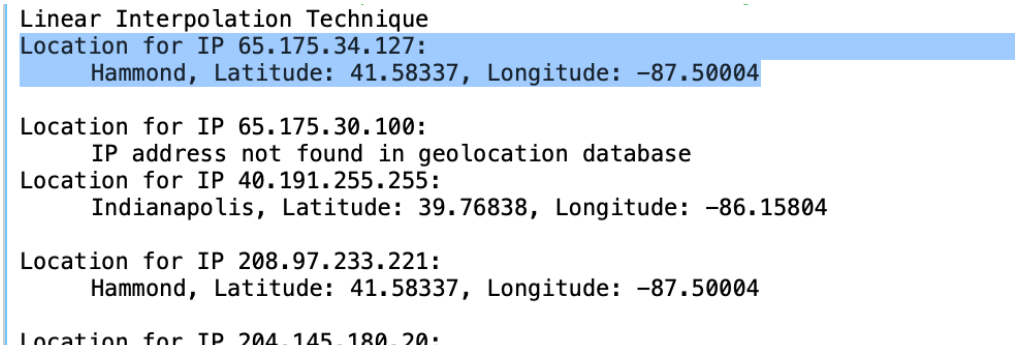
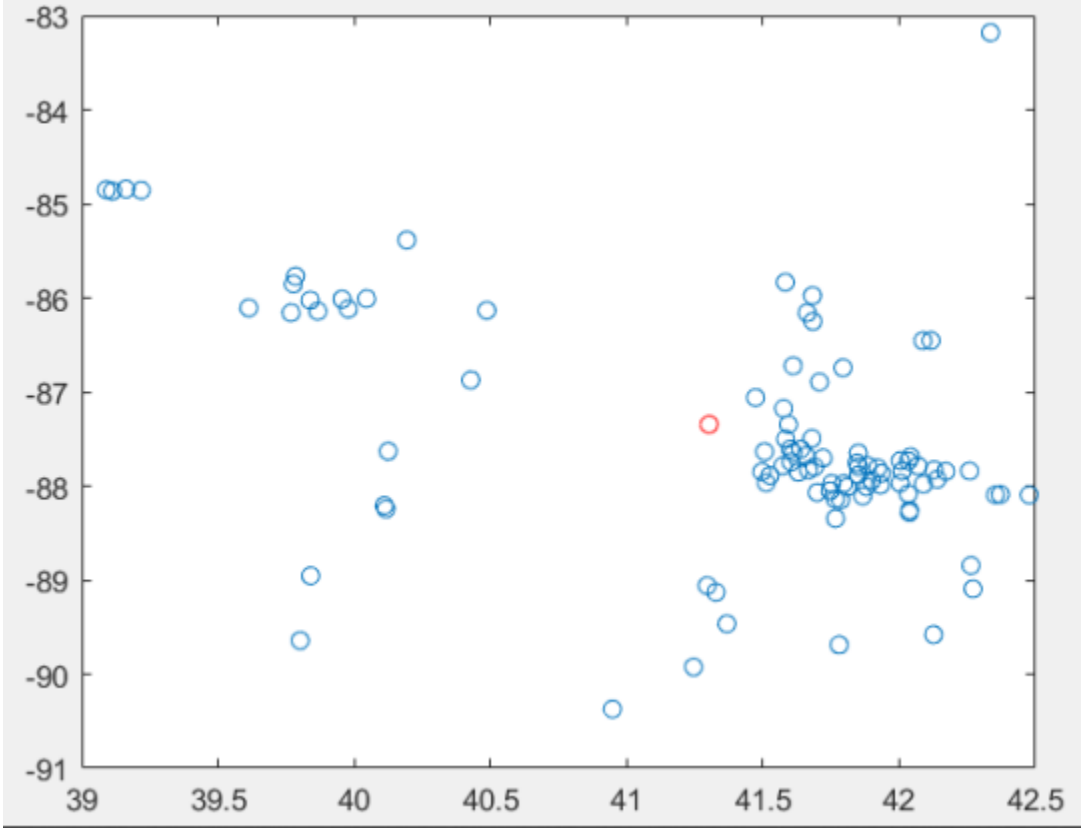


Fig 2: Output of shell terminal showing the location of IP 65.175.34.127

Upon finding a match at 65.175.34.120 (lower IP) and 65.175.34.127, linear interpolation is applied to make an estimate of the corresponding latitude and longitude. In this case the value displayed will be 'Hammond, Latitude: 41.58337, Longitude: -87.50004'. A set of IP addresses was then input to MATLAB to study further concepts. **[Karthik]**

The IP interpolation was performed using MATLAB. The process was intuitive in how to properly interpolate the data points. The IP addresses were loaded into MATLAB and the

centroid of the set of IP addresses was calculated. The obtained centroid in Graph 2, is the value which was then mapped to a latitude and longitude coordinate on the Earth and was taken as the “true” spot where the entire set of IPs lied.



Graph 2: Latitude vs Longitude and Centroid

After implementing both methods, Method 2 was found to be more accurate compared to method 1. Although Method 1 can be improved if more countries are added to the dataset. Method 2 currently uses ipinfo api which is considered one of the accurate databases for finding ip address geolocation, with the help of centroid we can determine the city where the IP address is located. Similarly in method 1 we can determine the country. It was found through the analysis of the aforementioned procedure that a large set of IPs have the same IP geolocation which are owned by organizations. ISP’s are responsible for assigning these IP’s to various organizations. Similarly IANA is responsible for assigning the ranges to ISP’s.

3.3 Comparing two methods to find the IP geolocation

The first method uses collection IP addresses in a particular region which is later mapped to a city whereas the second method uses the traceroute command to find the country where the IP address is located. From Figure 1, it can be observed that the ping packet travels to the local DNS server first and it reaches to comcast servers which contains

lookup tables for IP's. Usually it's the same route for most of the IP's located outside the United States. From Graph 1, it can be observed that the slope is not a straight line. The main reason behind this is due to usage of different types of transmission media. The fluctuations occur due to losses in the material. Few countries still depend on cable networks and few countries have adapted the usage of optical fiber hence ping fluctuations are observed at the time of sending packets to a larger distance. **[Shree Ganesh]**

The linear interpolation is simple to implement and computationally efficient, making it highly suitable for real time applications where spontaneous estimations are crucial. This makes it slightly better than k-nearest neighbor which requires much computation. Linear interpolation utilizes only the two nearest data points, while K nearest neighbors requires analyzing and comparing multiple neighbors, making it less efficient [11]. However, it assumes a linear relationship between IP addresses and locations within a range, which may not always hold true. The accuracy of the method relies on the quality and granularity of the underlying IP address database. In conclusion, IP linear interpolation shows valuable responses in effectively estimating the IP address locations, thereby offering pace for various real-time applications. Nevertheless, researchers and practitioners should be mindful of the resources involved and the limitations of acquiring up-to-date IP addresses. **[Karthik]**

4. Conclusion

Per the time constraint of the project, the team was able to provide an adequate solution for how IP interpolation can be used to assist in finding the geolocation of IP addresses. In the future, the team could work to obtain a larger truth seed (known IP locations) in order to better perform interpolation. For example, if multiple centroids were used instead of just one to interpolate the IP addresses, an even better estimation would be obtained for the geolocation. Furthermore, a larger set of IPs for testing would be helpful in consolidating the findings. It is important to note that there is no one geolocation-detecting technique which is perfect. Some techniques may be more or less accurate than others due to many uncontrollable factors, such as weather , internet stability, etc. But all do provide some improvement in determining the geolocation more than that of just solely using an IP address. **[Marcel]**

5. Contributions

Discuss contributions of each member in project implementation and writing the report.

Contributions and efforts to this report and project were provided equally by each member of the team. Shree Ganesh was responsible for **Method 1,** Gathering and creating IP address databases for various ranges, for the report Shree Ganesh wrote Related Works until classes of IP addresses, Implementation, results and future work of **Method 1,** Gathering and creating IP address databases for various ranges. For the

project, Marcel was responsible for providing the MATLAB code to map the IP addresses to specific coordinates and performing IP interpolation to find the centroid of the ground truth seed of IPs for **Method 2**. For the report, Marcel wrote the latter half of the Related Work Section, part of the Implementation and Results Section, as well as the Conclusion Section. Karthik was responsible for developing the Python code to perform linear interpolation on the gathered IP addresses from earlier sections. He then conducted an analysis of the results, provided contributions to the results section, and implemented aspects of method 2. Josue's responsible for gathering the open source IP address databases from various sources including ipapis [10] which was used in **Method 2** and wrote the **Introduction.**

References:

1. M. Weinlich, P. Kurz, M. B. Blau, F. Walcher and S. Piatek. "Significant Acceleration of Emergency Response Using Smartphone Geolocation Data and a Worldwide Emergency Call Support System." PLoS ONE 13 (5): pp. 1–10, May. 2018, doi:10.1371/journal.pone.0196336.
2. P. P. Vishwakarma, A. K. Tripathy and S. Vemuru, "An empiric path towards fraud detection and protection for NFC-enabled mobile payment system." Telkomnika, 17(5), pp. 2313–2320, Oct. 2019, 10.12928/TELKOMNIKA.v17i5.12290
3. R. Koch, M. Golling, L. Stiemert and G. D. Rodosek, "Using Geolocation for the Strategic Preincident Preparation of an IT Forensics Analysis," in IEEE Systems Journal, vol. 10, no. 4, pp. 1338-1349, Dec. 2016, doi: 10.1109/JSYST.2015.2389518.
4. Dan, Ovidiu, Vaibhav Parikh, and Brian D. Davison. "Improving IP geolocation using query logs." *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. 2016.
5. Olufunke Baruwa, Brian Haberman, & Ramanou Biaou. (2023, August 31). *A short guide to IP addressing*. www.internetsociety.org/resources/deploy360/2015/short-guide-ip-addressing
6. S. Ding, X. Luo, M. Yin, Y. Liu and F. Liu, "An IP geolocation method based on rich-connected sub-networks," *2015 17th International Conference on Advanced Communication Technology (ICACT)*, PyeongChang, Korea (South), 2015, pp. 176-181, doi: 10.1109/ICACT.2015.7224779.
7. J. Chen, F. Liu, T. Wang, X. Luo, F. Zhao and G. Zhu, "Towards region-level IP geolocation based on the path feature," *2015 17th International Conference on Advanced Communication Technology (ICACT)*, PyeongChang, Korea (South), 2015, pp. 468-471, doi: 10.1109/ICACT.2015.7224839.
8. *5 classes of ipv4 addresses [class A, B, C, D and E]*. (n.d.). Retrieved December 8, 2023, from www.meridianoutpost.com/resources/articles/IP-classes.php

9. *DNS lookup*. (n.d.). NsLookup.Io. Retrieved December 8, 2023, from www.nslookup.io
10. ipapi.is. (n.d.). *ipapi.is*. Geolocation. Retrieved December 8, 2023, from www.ipapi.is/geolocation.html
11. Y. Hamed, A. Shafie, Z. B. Mustaffa and N. R. B. Idris, "An application of K-Nearest Neighbor interpolation on calibrating corrosion measurements collected by two non-destructive techniques," 2015 IEEE 3rd International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA), Kuala Lumpur, Malaysia, 2015, pp. 1-5, doi: 10.1109/ICSIMA.2015.7559030.
12. Lloyd N. Trefethen, *Finite Difference and Spectral Methods for Ordinary and Partial Differential Equations*, unpublished text, 1996, available at people.maths.ox.ac.uk/trefethen/pdetext.html