IP Geolocation with Two-Tier Neural Network

Hao Jiang[1] Yaoqing Liu[2] Jeanna N. Matthews[2]

[1]Department of Computer Science
The University of Chicago

[2]Department of Computer Science
Clarkson University

Outline

Background

System Design

Data Collection

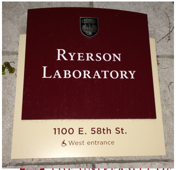
Evaluation

Conclusion

What is IP Geolocation?

128.135.100.110

41.79N
87.60W



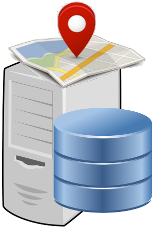
What can we do with IP Geolocation?

Credit card Fraud

CDN

Online Ads

Previous work - IPGeo Database



MAXMIND

IP2LOCATION™

IPligence

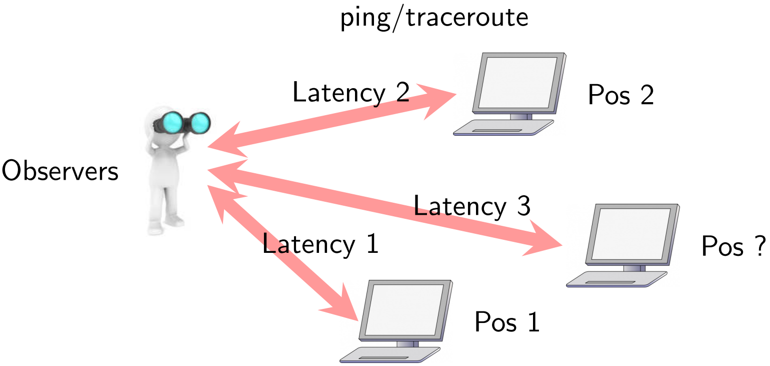
Pros

- Easy to use

Cons

- Less accurate (City level)
- Not up to date (Periodic update)

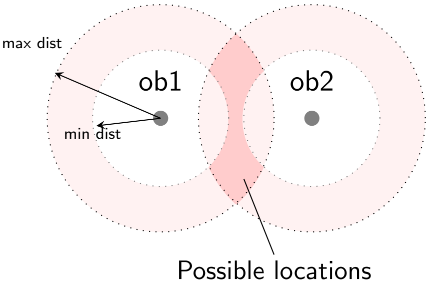
Previous work - Measuring network delay



$$\theta = train(Latency_{1,2}, Pos_{1,2})$$

$$Pos_3 = predict(Latency_3, \theta)$$

Previous work - Build a model

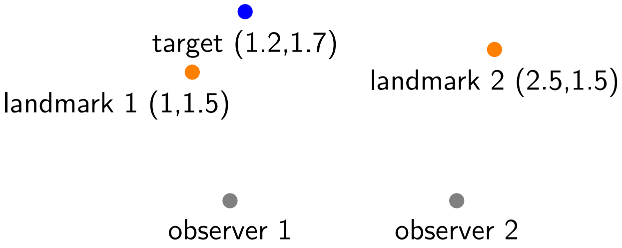


Design a simple model (mostly based on triangulation) and calculate the parameters. [GZCF06, KBJK+06, WSS07, DPCS12]

Accuracy: ∼10km median error

Such a model requires a lot of assumptions, which are not necessarily true. (E.g., is there a linear relationship between latency and geographic distance?)

Previous work - Find nearby landmarks



Find the landmark that has the most similar observation results with the target. [WBF+11]

Accuracy: ∼1km median error

Accuracy is greatly relied on the density of the landmarks.

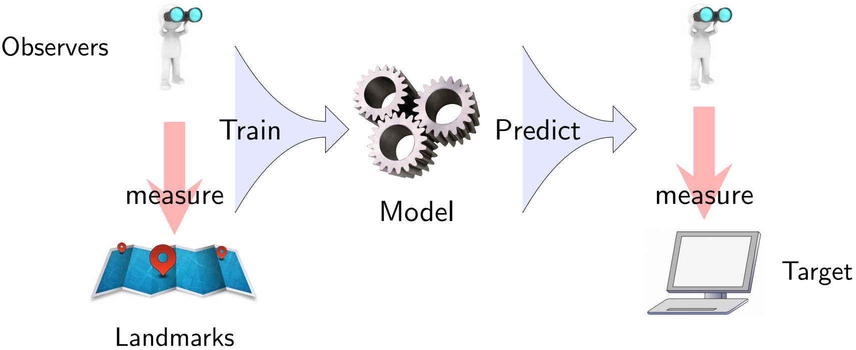
Hard to maintain a large group of landmarks.

Previous work - What do we learn

- Physically adjacent nodes have similar measurements
- Network topology is simpler in a local area than in a larger area

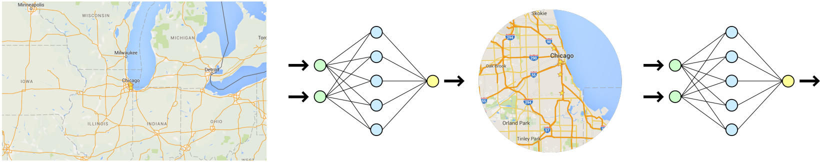
Outline

Design Idea



Our method employs machine learning technique to solve the problem. Instead of "choose" a model, we collect latency data from landmarks with known locations and train a model, then use this model to predict the location of unknown targets.

Two-Tier Neural Network



Region estimation Location estimation

Intuition:
Measurement from adjacent landmarks can yield a better estimation result.

Make a rough estimation with all landmarks, locate the region the target resides in. Then use only the landmarks in that region to do a more accurate prediction.

Outline

Background

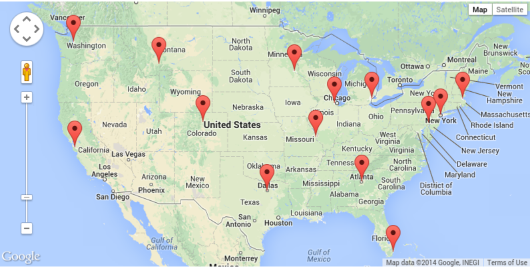
System Design

Data Collection

Evaluation

Conclusion

Data Collection - Observers



Ripe Atlas Probes

Ripe Atlas Anchors

We choose 14 anchors from Ripe Atlas Network as observers. These observers covers most area of US continental evenly. They are able to send ping/traceroute requests to arbitrary IP addresses.

Data Collection - Datasets

A large enough landmark dataset is crucial to the accuracy of our method. Our dataset consists of landmarks from three data sources

- Ripe Atlas Probes
- University Webservers
- City Government Webservers

Data Collection

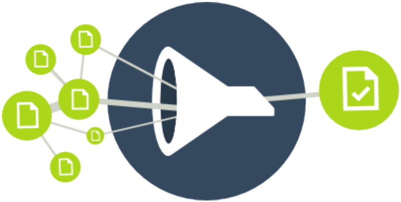
University Dataset

- Get a U.S. university list from Wikipedia
- Use Google search API to obtain the geographic location and its website
- Use `host` command to obtain corresponding IP address

City Dataset

- Get a U.S. city and population list from government website
- Choose the top 50 cities of each state ordered by population in descending order
- Use Google search API to obtain the geographic location and its website
- Use `host` command to obtain corresponding IP address

Data Collection - Filtering



We filter out invalid data using various methods

- Look for popular virtual host providers (Amazon, GoDaddy, Rackspace, etc.)
- Look for owners that own multiple IP addresses (through whois)
- Cross-validation using GeoIP database

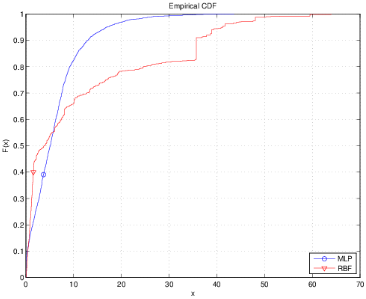
Data Collection - Result

Category	Raw	Valid	Reachable
Ripe Atlas Probes	637	637	429
University Websites	2170	1858	826
City Government Websites	2880	740	292
Total	5687	3235	1547

Table: Landmark Detail (Raw: All landmark candidates. Valid: Landmarks after filtering and cross-validation. Reachable: Landmarks that respond to `ping`)

Outline

Evaluation - Error Distribution



;

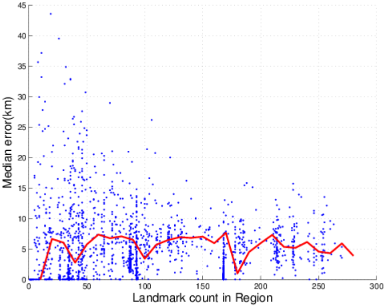
Error Distribution of the estimation result

We compare the performance of two popular neural network types: Multi-Layer Perceptron (MLP) and Radial-Basis Function (RBF)

Accuracy:

- Over 80% estimations have a error within 10km
- MLP has a overall better performance than RBF

Evaluation - Accuracy related to number of landmarks



;

MLP Error related to Landmark Density

- 3.7km in regions with > 100 landmarks
- 6km in regions with $< 50 landmarks$
- Error decreases when landmark density increases

Outline

Conclusion

Our Contribution:

- A novel method for IP Geolocation
- Achieved similar accuracy with state-of-the art with a fixed amount of landmarks

Future Work

- **Mobile client**
  In this research, our data source contains only wired network nodes. Mobile network, especially cellular network clients may have different properties that is not represented in our dataset.

  Contribution: High   Complexity: High
- **Region**
  Our method assumes two geographically adjacent IP addresses will be adjacent on network topology. While this has been justified by our research result on U.S. territory, we are interested in expanding the testing in regions such as Europe and Asia.

  Contribution: High   Complexity: Medium

Question?

References I

Ziqian Dong, Rohan D.W. Perera, Rajarathnam Chandramouli, and K.P. Subbalakshmi.
Network measurement based modeling and optimization for {IP} geolocation.
Computer Networks, 56(1):85 – 98, 2012.

Bamba Gueye, Artur Ziviani, Mark Crovella, and Serge Fdida.
Constraint-based Geolocation of Internet Hosts.
IEEE/ACM Trans. Netw., 14(6):1219–1232, December 2006.

Ethan Katz-Bassett, John P. John, Arvind Krishnamurthy, David Wetherall, Thomas Anderson, and Yatin Chawathe.
Towards IP Geolocation Using Delay and Topology Measurements.
In *IMC '06*, pages 71–84, New York, NY, USA, 2006. ACM.

References II

Yong Wang, Daniel Burgener, Marcel Flores, Aleksandar Kuzmanovic, and Cheng Huang.
Towards Street-level Client-independent IP Geolocation.
In *NSDI '11*, pages 27–27, Berkeley, CA, USA, 2011. USENIX Association.

Bernard Wong, Ivan Stoyanov, and Emin Gün Sirer.
Octant: A Comprehensive Framework for the Geolocalization of Internet Hosts.
In *NSDI '07*, pages 23–23, Berkeley, CA, USA, 2007. USENIX Association.