HLOC: Hints based Geolocation Leveraging Multiple Measurement Frameworks*

Quirin Scheilte, Oliver Gasser, Patrick Sattler, Georg Carle

Traffic Measurement & Analysis Conference 2017

Presentation by Manasvini Sethuraman

*Borrowed diagrams/tables from Oliver Gasser's original ppt

Background

- IP Geolocation is the task of mapping an IP address to a physical location
- Useful in understanding network topology and mapping the Internet
- Businesses locate people through their device location

Background contd.

- How can we do IP geolocation?
  - From **commercial databases**: ip2location, MaxMind GeoLite
  - **Measurement based**: Infer location from latency, time to live and other parameters by triangulation
  - **DNS based**: Use DNS name entries to understand topology [1]
- Why is this not sufficient?
  - DNS names are sometime re-used, and rule-based methods are narrow
  - Latency measurements lead to partially incorrect mappings at city level
    - Precise measurements will require probes to be close to the target routers, not always possible
  - Databases can be inaccurate
    - e.g., routers geolocated in Antarctica

[1] Huffaker et al, SIGCOMM 2014, DRoP: DNS-Based Router Positioning

Motivation

- Approaches to IP geolocation
  - Latency measurement tools => hard to set up, and use
  - Commercial databases => inaccurate for internet nodes like routers
- Paper proposes a framework for IP geolocation that is ready-to-use and accurate
- Combines information from several sources in order to create a novel system that can verify the accuracy of entries in existing databases
- Code is open source

How does HLOC work?

- Use **geolocation hints** from domain names
- Validate geolocation hints by making **latency measurements**
- Accuracy at country level

Tools used in the project

- **ZMap scan**
  - ZMap is a network scanning tool
  - It randomly samples the IP address space and sends out packets to those addresses
  - Records the responses
- **RIPE Atlas**
  - Set of servers that act as probes which can measure reachability to other nodes
  - Measure RTT from the probe to a node in order to estimate the distance between the node and the probe

Datasets- IP Addresses

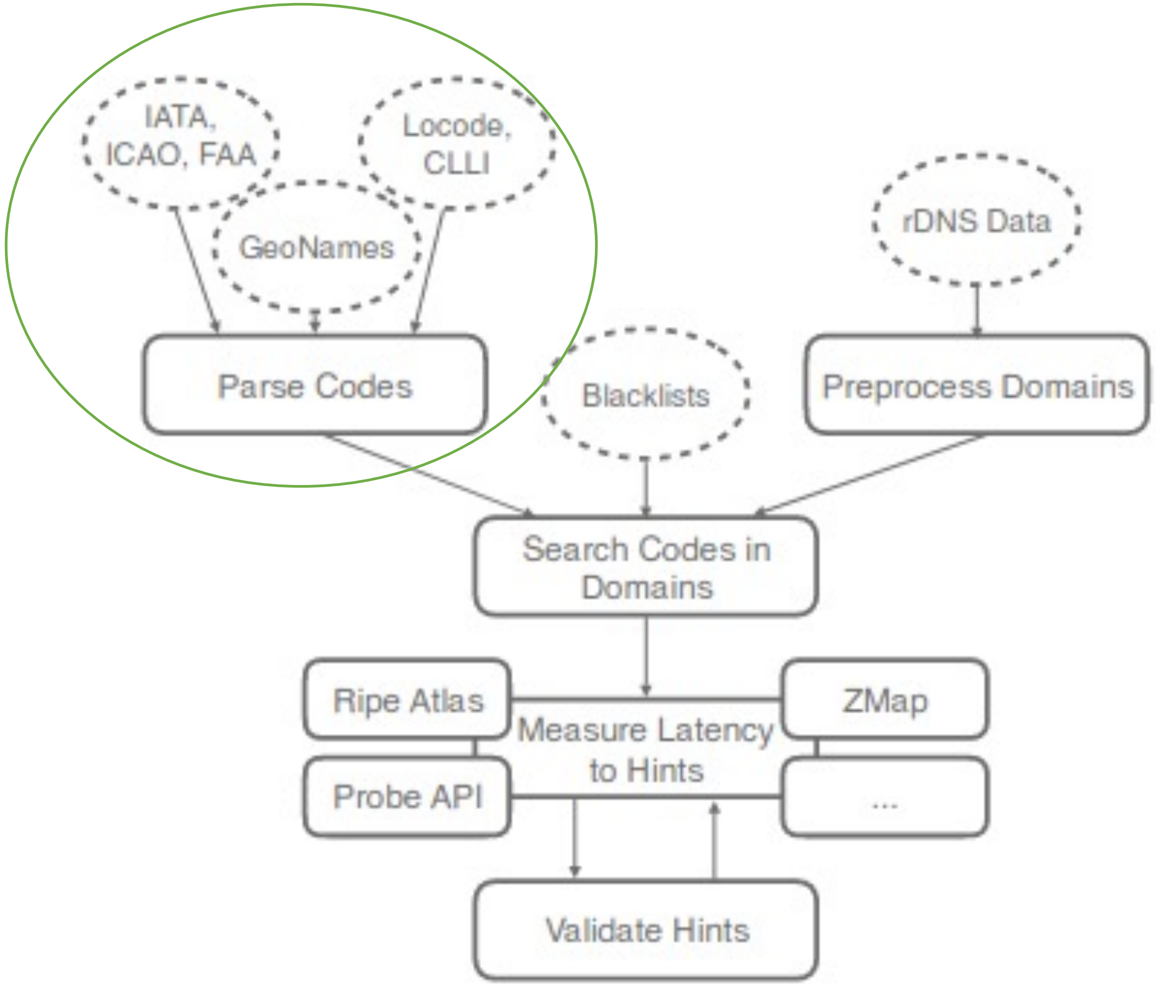
- CAIDA ITDK IPv4 traces
    - Identify router IP addresses
- Project SONAR
    - Scans IP addresses and looks for open ports
    - Extracts names that represent DNS records
    - Analyzes names from HTML links (from HTTP traffic studies)
    - RDNS (reverse DNS) files
- Paper uses rDNS file to get domain names for hosts in the CAIDA dataset

Datasets- Geolocation

- Airport codes
    - e.g., Houston, TX <=> HOU
- UN/Location codes
    - e.g., Houston, TX <=> US HOU
- CLLI codes (Telco data)
    - e.g., Houston, TX <=> HSTNTX~~MOCG0~~
- GeoNames
    - e.g.; Houston, TX <=> Hiustonas

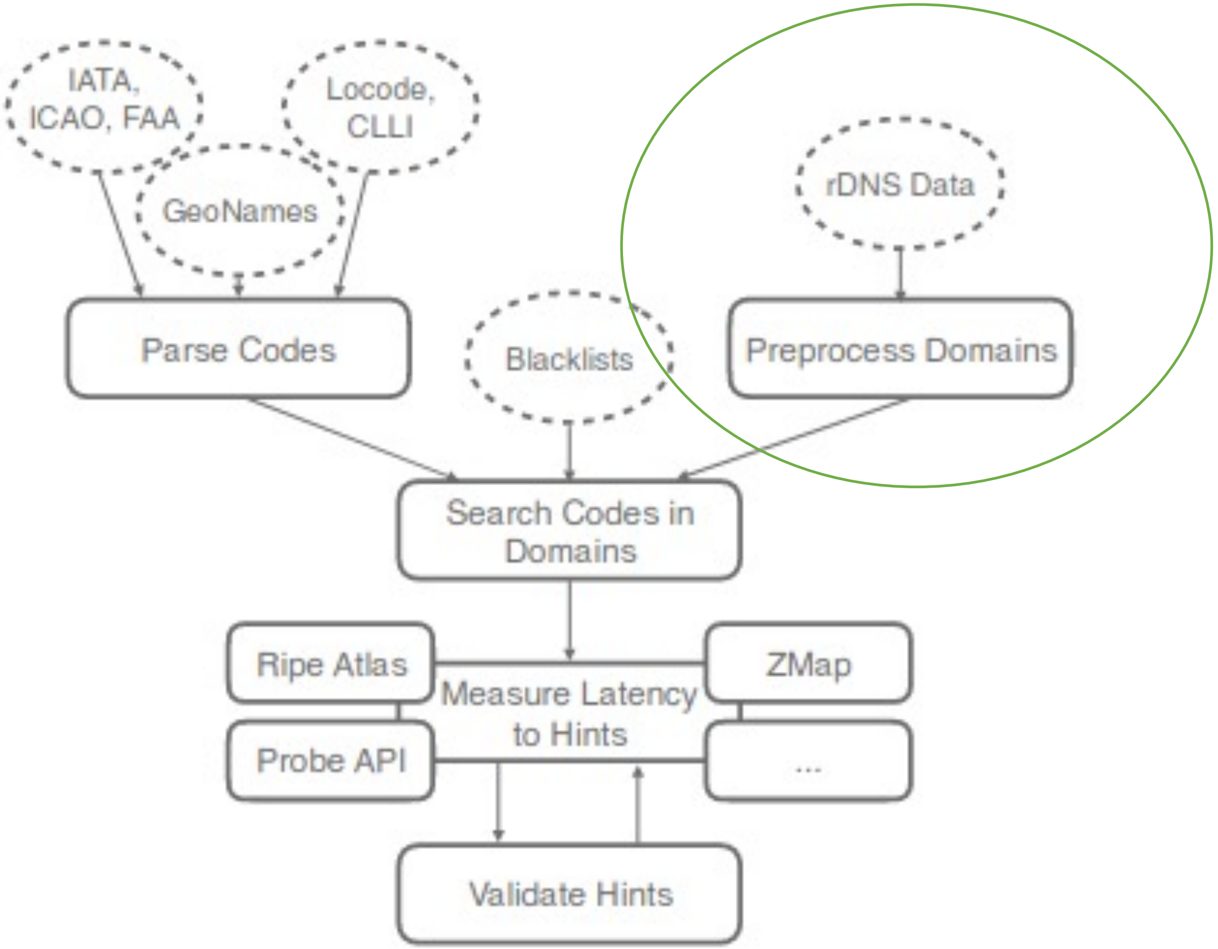
HLOC components- Parse Codes

- **Parse codes** creates a mapping of location codes to a single location entry using several datasets like IATA, GeoNames etc. (total 448K location codes and 5474 locations)



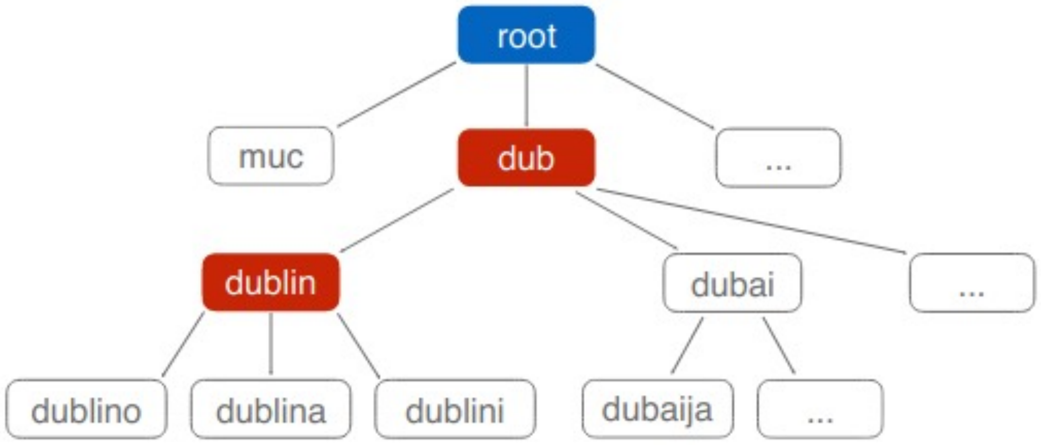
HLOC components – Preprocess Domains

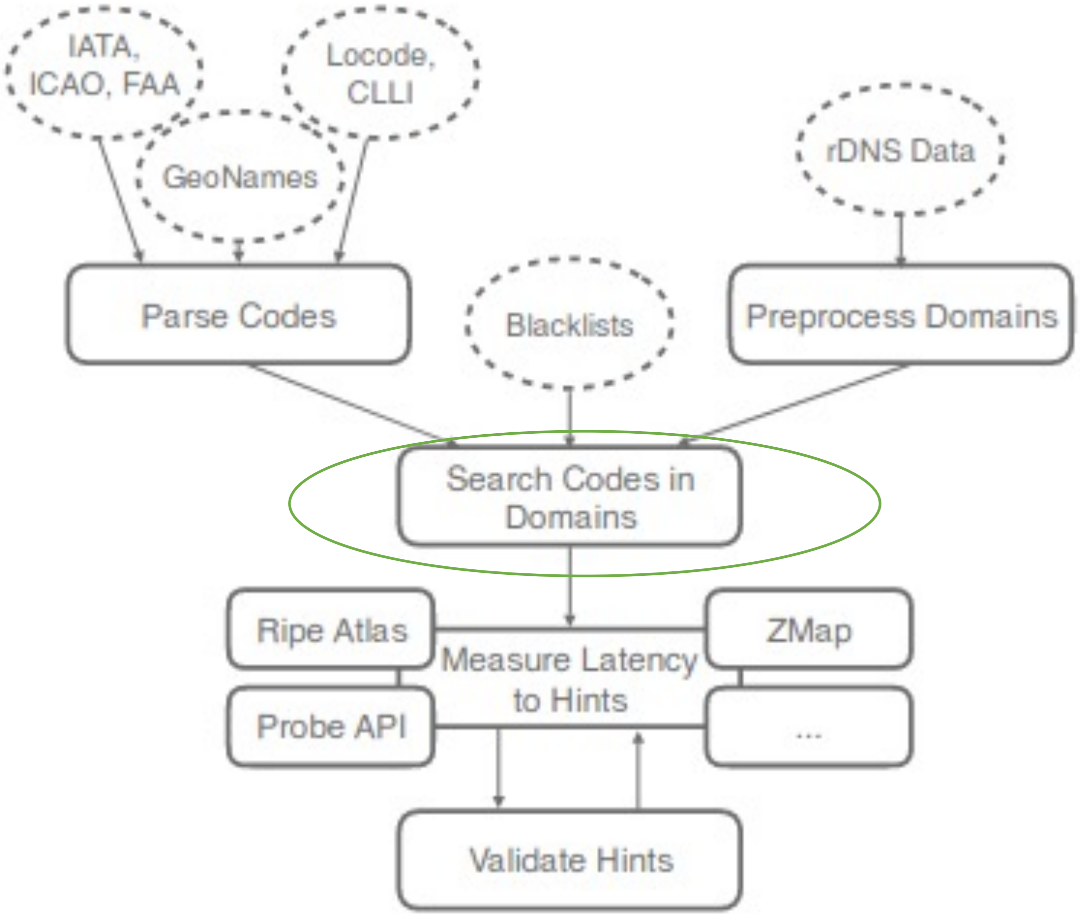
- **Preprocess domains** takes list of IP addresses (CAIDA ITDK dataset) and maps them to domain names using rDNS queries on SONAR dataset.
- Filter invalid domains (e.g.; *.local*)
- Total 1.6M domain names



HLOC components – Search Domains

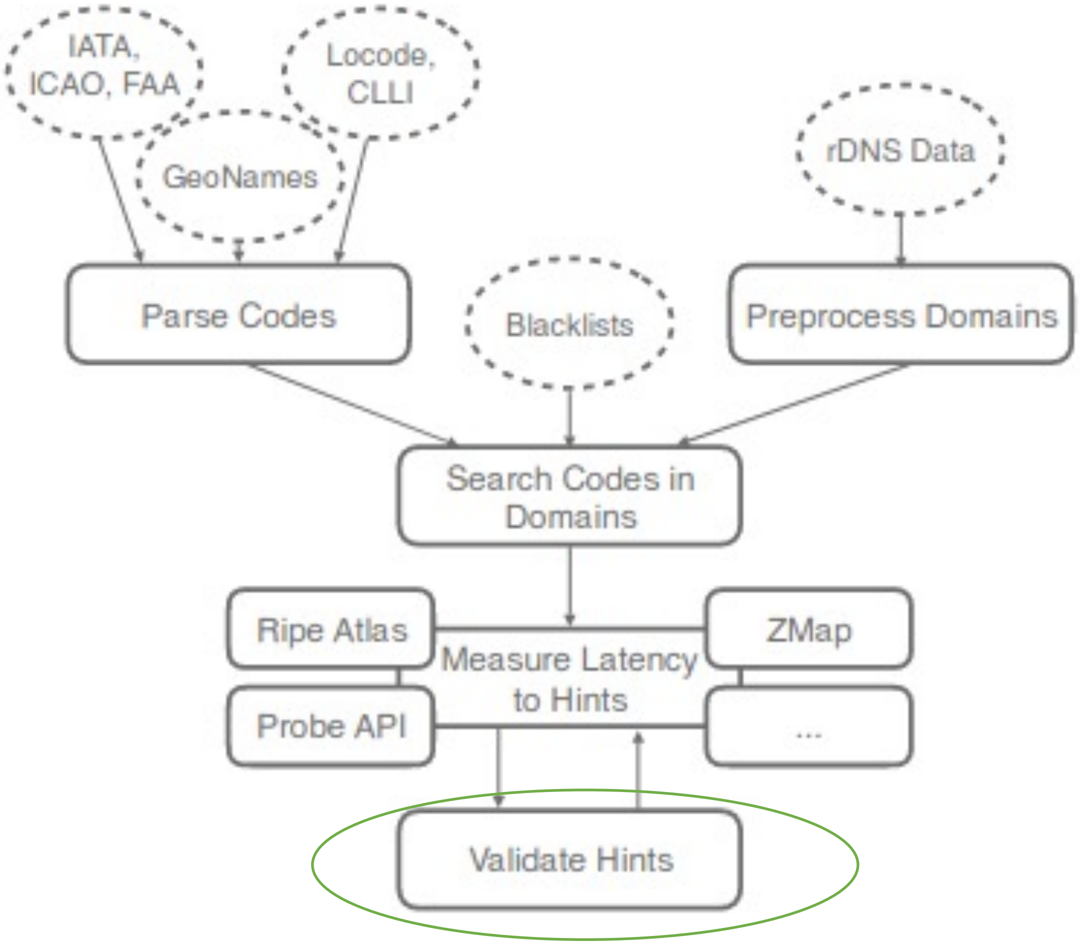
- **Search domains** organizes location codes in a prefix tree and match domains against prefix tree
- Matching needs to be fast
- Matches of domain names to prefix tree are called hints
- e.g., search for 'dublin'





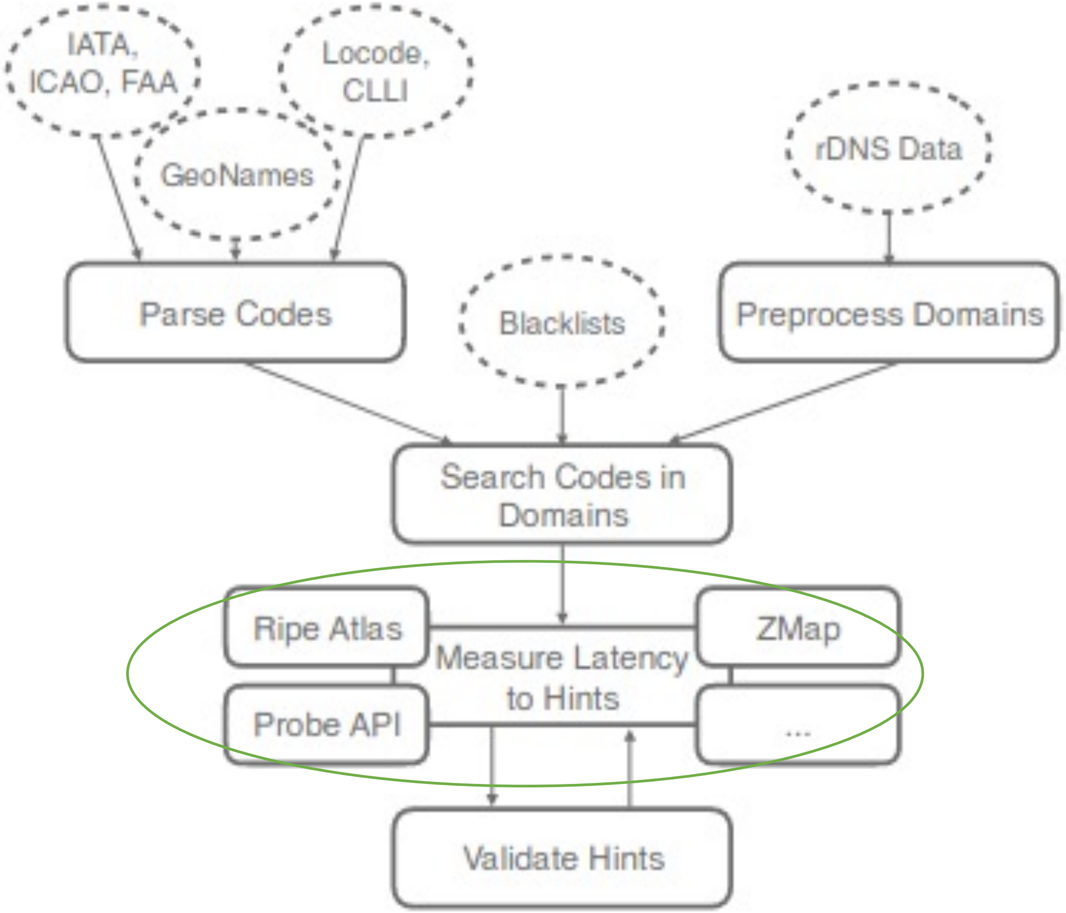
HLOC components – Validate hints

- **Validate hints** is called for every location hint for a domain.
- It eliminates possible locations based on speed of light constraints (RTT).
- Repeat validation until a matching prefix is found
- Distance between probe p and host h is defined as:
    - ***d(p, h) < x***
- Round trip time (RTT) constraint:
- ***RTT(p,h) < a + (2*d(p,h))/(c*c0)***
- *c0* is the speed of light and *c* is the refractive index of optical fiber (c=2/3) and *a* is latency buffer
- In the experiments, **a=9 ms and x is 1000 Km**



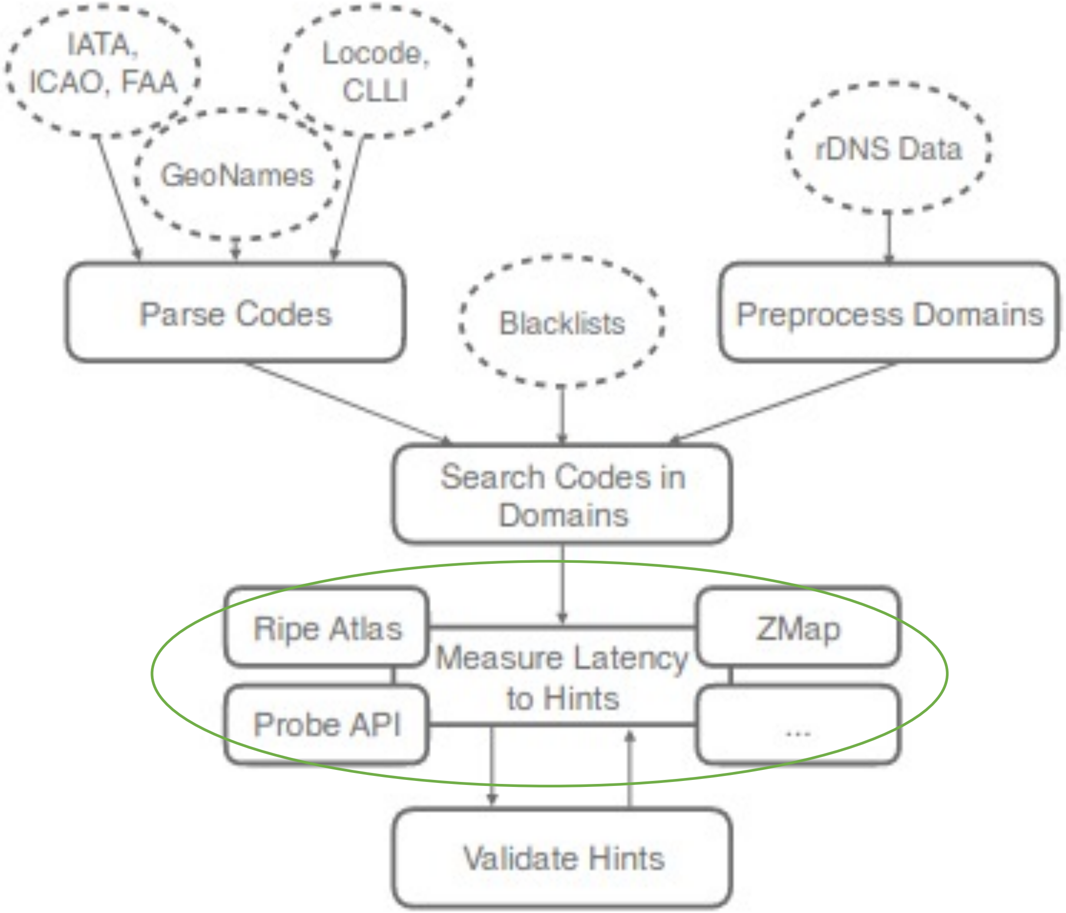
HLOC Components- measurement & challenges

- **Abundance of matches**
  - Prefix tree yields 20 matches per domain name
  - Remove locations with < 100k inhabitants
  - Create blacklists (e.g, prefix matches which do not signify location)
  - ae-0.facebook.amstnl02.nl.bb.gin.ntt.net
    - Keep ams (IATA): Amsterdam, Netherlands (2.3 ms)
    - Remove ceb (IATA): Not a location
    - Remove face (ICAO): Ceres, South Africa
    - Remove ace (IATA): Lanzarote, Spain
  - Remove top level and second level matches, e.g., *.fr*



HLOC Components- measurement & challenges

- **Validation runtime**
  - Use **ZMap** scans to filter unresponsive IPs using ICMP echo messages
  - ZMap scans done from Dallas, Frankfurt and Singapore=> helps in identifying hemisphere for IP address
  - ZMap scans require high bandwidth
  - **RIPE Atlas** allows only a certain number of measurements per time interval
  - low bandwidth, more accurate latency measurements



Putting it all together

- Suppose we want to geolocate cr-01.0v-00-04.anx32.nyc.us.anexia-it.com
- Identify location prefix matches
  - anx (IATA): Andenes, Norway
  - nyc (IATA): New York City, USA
- Select probe near these locations
  - Andenes (Probe ID: 20229; location: Skien, distance: 990 km)
- Measure RTT from probe
  - RTT(Probe(20229), "2001:2000:3080:c44::2") = 25 ms
  - ***RTT(p,h) < a + (2*d(p,h))/(c*c0)***
    - 9ms + 2.1000/(200) = 9 + 10 = 19ms
  - 25 ms > 19 ms

Putting it all together

- Suppose we want to geolocate cr-01.0v-00-04.anx32.nyc.us.anexia-it.com
- Identify location prefix matches
    - ~~anx (IATA): Andenes, Norway~~
    - nyc (IATA): New York City, USA

Putting it all together

- Suppose we want to geolocate cr-01.0v-00-04.anx32.nyc.us.anexia-it.com
- Identify location prefix matches
  - ~~anx (IATA): Andenes, Norway~~
  - nyc (IATA): New York City, USA
- Select probe near these locations
  - New York (Probe ID: 17736; distance: 0.84 km)
- Measure RTT from probe
  - RTT(Probe(17736), "2001:2000:3080:c44::2") = 1.3 ms
- Eliminate impossible location hints
- Validate location hints using RTT constraints

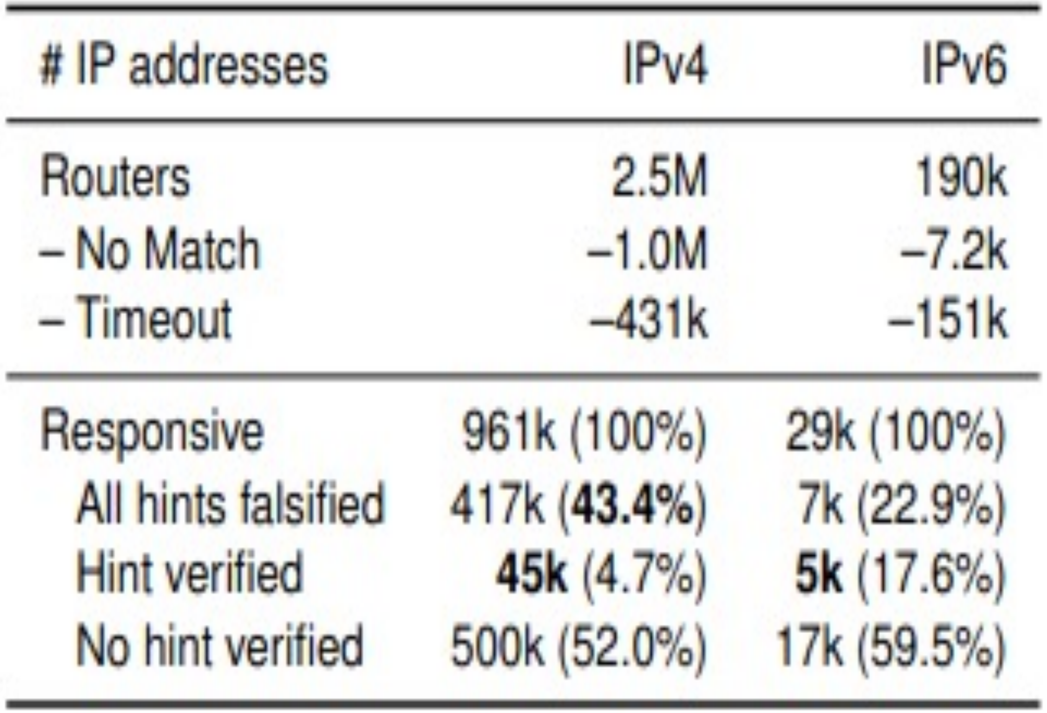
$$1.3ms < 9ms + \frac{2 \cdot 0.84km}{200\frac{km}{ms}}$$

Measurement Results

- 41% IPv4 and 4% IPv6 addresses filtered out since they do not have location prefix
- 2% IPs in blacklist (from previous studies) or not announced on BGP border router
- 28% IPv4 and 78% IPv6 addresses unresponsive in ZMap scan
  - Large portion of IPv6 addresses correspond to home routers

Measurement Results

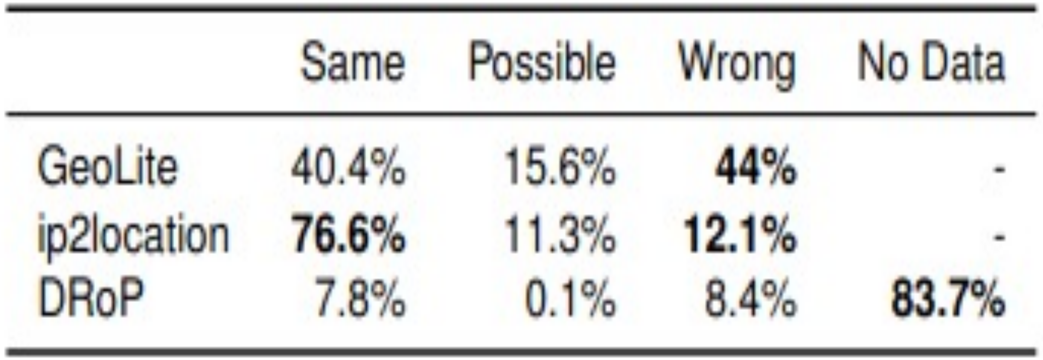
- **Hint verified**: location prefix exists and latency measurement within threshold for one location hint. Many of the IPs which were verified hints were present in other measurement datasets, and are important in the internet topology
- **All hints falsified**: All location hints for the domain name were invalidated via latency constraints (Zmap).
- **No hint verified**: all hints were not falsified. I.e, either no probe was nearby or the latency was too high.



# IP addresses	IPv4	IPv6
Routers	2.5M	190k
– No Match	–1.0M	–7.2k
– Timeout	–431k	–151k
Responsive	961k (100%)	29k (100%)
All hints falsified	417k (**43.4%**)	7k (22.9%)
Hint verified	**45k** (4.7%)	**5k** (17.6%)
No hint verified	500k (52.0%)	17k (59.5%)

Comparison with DRoP, GeoLite and ip2location

- DRoP is a system for geolocating IP addresses based on rules for domain names
- GeoLite and ip2location are commercial databases
- Compare addresses verified by HLOC against values in other systems
- Possible=> the reported location in the database is within latency bounds
- Wrong => location reported in the database violates speed of light constraints
- Paper does NOT claim that HLOC is more right than others but only points out that there could be inconsistencies in databases.



	Same	Possible	Wrong	No Data
GeoLite	40.4%	15.6%	**44%**	-
ip2location	**76.6%**	11.3%	**12.1%**	-
DRoP	7.8%	0.1%	8.4%	**83.7%**

Things I liked about the paper

- The ideas are clear
- Evaluation is easy to follow
- Paper does not claim to be a source of truth
- Has an extensive discussion section
- Code + datasets are easy to obtain

Other Questions

- What other methods exist for IP geolocation apart from measuring RTT that can be plugged into HLOC?
    - RIPE Atlas hosts are deployed in far fewer numbers outside of Europe/North America
    - Is there a better way to geolocate IPs outside of Europe/North America

Questions/Limitations

- What happens if the wrong prefix is selected from the Search Domains step?
  - This is problematic since the paper filters from 20 domains to ~1.3 based on their criteria
  - Probe selection is based on prefix match, so it is possible to produce a larger latency estimate
  - Livadariu et al [2] show that this is indeed an issue and HLOC picks the wrong prefix
- We know that ICMP messages are ignored by most hosts
  - Can we do better by using the technique presented in the proxies paper?
  - Why would/wouldn't it work?
- Would it help to use multiple probes?
  - How would it impact the number of verified/unverified hosts?

[2] Livadariu et al, ANRW 2020, On the Accuracy of Country-Level IP Geolocation

Thank you for your attention!