HLOC: Hints-Based Geolocation Leveraging Multiple Measurement Frameworks

Quirin Scheitle, Oliver Gasser, Patrick Sattler, Georg Carle
Chair of Network Architectures and Services
Technical University of Munich (TUM)
Email: {scheitle,gasser,sattler,carle}@net.in.tum.de

Internet Measurement Reading Group @ LINCS

IP geolocation

132.227.123.8

↓

48° 50' 47.148'' N

2° 21' 16.596'' E

In practice: city or country

IP geolocation — Why?

- Advertisement
- Content licensing
- Content personalization (e.g. weather, news, language…)
- Content delivery networks
- …
- Internet research (mapping, routing, security…)

Commercial services

- IP2Location, MaxMind, NetAcuity…
- Opaque geolocation mechanisms
  - Impossible to reproduce their results
- Paid or limited free version
- Works well for end-user IPs, less so for **infrastructure** IPs
  - Gouel, Matthieu, et al. "IP Geolocation Database Stability and Implications for Network Research." *2021 Network Traffic Measurement and Analysis Conference (TMA)*.

MaxMind self-reported end-user coverage

	Correctly Resolved	Incorrectly Resolved	Unresolved
GeoLite2 City *(free service)*	41%	55%	4%
GeoIP2 City	46%	48%	6%
GeoIP2 Precision City Service	48%	47%	6%

MaxMind GeoIP2 accuracy, France, 10km radius, excluding cellular networks
https://www.maxmind.com/en/geoip2-city-accuracy-comparison

Measurement-based techniques

- Nearest-neighbor
  - Measure latency from multiple vantage points, assign target location to the closest vantage point
- Multilateration
  - Measure latency from multiple vantage points, find intersections of the speed-of-light circles
  - CBG (Constraint-Based Geolocation)
- Multilateration with topological information
  - TBG (Topology-Based Geolocation)

⚠️ Vantage points distribution, measurement cost, probe filtering…

Data-based techniques

- Social graph
- Web page content
- WHOIS database
- …
- **Reverse DNS records**

Reverse DNS records

$$8.8.4.4 \Rightarrow 4.4.8.8.\text{in-addr.arpa} \underset{\text{PTR}}{\Rightarrow} \text{dns.google}$$

Reverse DNS records

154.54.36.130	be2334.ccr42.**par**01.atlas.cogentco.com
193.51.181.170	gi8-7-**rennes**-rtr-021.noc.renater.fr
99.162.80.168	99-162-80-168.lightspeed.**irvnca**.sbcglobal.net.
4.4.119.193	et-4-0-0-0.bar4.**SaltLakeCity**1.Level3.net.

Dataset: https://opendata.rapid7.com/sonar.rdns_v2/

Geolocation hints

- IATA & ICAO airport codes
  - Paris Charles de Gaulle Airport: CDG, LFPG
- CLLI (Common Language Location Identifier)
  - Houston, Texas: HSTNTX
- UN/LOCODE (United Nations Location Code)
  - Berlin, Germany: DEBER
- Raw or partial city names
  - Irvine, irvn...

Ambiguities

be2334.**ccr**42.**par**01.**atlas**.cogentco.com

Airport code for Concord, CA

Paris, France or one of the 20 towns named Paris in the USA?

Salas Atlas, Spain

HLOC framework

1. Map code to cities (100km aggregation)
2. Extract location hints from reverse DNS names
3. Verify or falsify hints based on delay constraints

Match reduction

- Ignore cities below 100k inhabitants
- Ignore common words
  - tel (telecom), cpe (customer premises equipment)
  - Internet, Linux, static…
- Ignore ambiguous codes
  - lin ⇒ Milan (IATA), Illinois, Carolina, Dublin
- Ignore top and second-level domains
  - .com, cogentco.com

Hint validation

1. Falsify hint based on speed-of-light violation
   - Measure the latency from a landmark towards the target
   - If it is inferior to the minimal latency (at 0.66*c) towards the hint location, falsify the hint
2. Verify hint based on pin-point measurements
   - Find a RIPE Atlas probe close (< 1000km) to the hint location
   - Measure the latency between the probe and the hint location
   - If it is inferior to a tight bound on the latency (twice the distance at 0.66*c + 9ms), verify the hint

Evaluation — Dataset

- CAIDA ITDK (Internet Topology Data Kit)
  - IP addresses of routers (aliases) with their associated reverse DNS
  - 2.5M IPv4 router IPs and 146k IPv6 router IPs (in 2017)
- IPv4 filtering
  - - 14k invalid domains
  - - 1M domains with no matches
  - - 465k unresponsive addresses
  - ⇒ 961k remaining addresses/domains pairs
- IPv6 filtering
  - ⇒ 29k remaining addresses/domains pairs

Evaluation — DRoP, GeoLite, ip2location

TABLE V: Evaluation of location decisions by databases and DRoP against HLOC measurements: ip2location more accurate than GeoLite, DRoP frequently with "no data". All information-based approaches with a significant number of wrong decisions.

	HLOC		GeoLite			ip2location			DRoP			
	Location Dec.	n	Same	Possible	Wrong	Same	Poss.	Wrong	Same	Poss.	Wrong	No data
IPv4	Verified	45k	40.4%	15.6%	**44.0%**	**76.6%**	11.3%	**12.1%**	7.8%	0.1%	8.4%	**83.7%**
	All falsified	417k	n/a[1]	100%	0%	n/a	100%	0%	n/a	n/a	2.2%	97.8%
	No verified	499k	n/a	96.1%	3.9%	n/a	98.8%	1.2%	n/a	10.5%	4.1%	85.4%
	Timeout	465k	n/a	100%	n/a[2]	n/a	100%	n/a	n/a	26.4%	n/a	73.6%
IPv6	Verified	5k		—		25.7%	10.6%	**63.6%**	33.7%	1.0%	1.8%	63.5%
	No verified	17k		—		n/a	74.2%	**23.9%**	n/a	25.5%	3.3%	71.2%

1: With no verified HLOC match, other approaches can not have the same match. 2: With HLOC timeout, it is not possible to evaluate other approaches.

Evaluation — DRoP ground truth

TABLE VI: For DRoP's ground truth domains, we show performance for (a) DRoP's reported performance, (b) our reproduction of DRoP and its validation against latency measurements, and (c) HLOC-generated hints and their latency validation.

Domain	DRoP 2014 [18]				DRoP 2016 Reproduction					HLOC			
	n	Type	Match	TP[1]	n	Match	TP[1]	Ver.[2]	Fals.[2]	Match	TP[1]	Ver.[2]	Fals.[2]
belwue.de	161	City	52%	86%	53	64%	65%	22	1	94%	64%	32	5
cogentco.com	13,129	IATA	90%	99%	9,475	95%	26%	2,381	**628**	99%	23%	2,144	295
digitalwest.net	111	IATA	49%	100%	47	49%	26%	6	0	100%	15%	7	2
ntt.net	2,584	CLLI	96%	100%	3,125	54%	37%	**622**	5	99%	30%	**937**	148
peak10.net	115	IATA	100%	100%	199	99%	9%	18	0	100%	9%	18	0

1: % of matches that are true positives 2: Total count of verified or falsified matches. "possible" and "time out" results not displayed.

Hints contribution

TABLE IV: IATA, GeoNames and CLLI codes provide 99% of verified hints.

Category	IATA	ICAO	FAA	UN/LO	GeoNames	CLLI
# Codes	8k	13k	20k	77k	32k	31k
Hints (100%)	4.5M	209k	472k	59k	215k	167k
Verified	**32k**	122	413	120	**13k**	**5k**
Verified (%)	.7%	$<$ .0%	.1%	$<$ .0%	**5.9%**	**2.8%**

Takeaways

- Reverse DNS information is valuable, **when present and containing geolocation information**
  - ~60% of the interfaces in Diamond-Miner traceroutes have a reverse DNS name
- Reverse DNS information can be outdated or wrong, it should be verified with latency measurements

HLOC — Advantages

- The code is provided (https://github.com/tumi8/hloc)
- The code works
- It uses public datasets and public measurement platforms

HLOC — Limitations

- Anycast
  - Unlikely for routers
- Reverse DNS coverage
- Routing detours
  - Choose probes closer to AS
- Aggregation of cities in a 100km radius
- Ignore location hints for location < 100k inhabitants (only ~300 cities in the US)
- Validated only against DRoP and commercial databases on ITDK

Other works

Dan, Ovidiu, Vaibhav Parikh, and Brian D. Davison. "IP Geolocation through Reverse DNS." *ACM Transactions on Internet Technology (TOIT)* 22.1 (2021): 1-29.

- Use a ground truth of 67 million IP addresses obtained from Bing search logs where users opted-in to provide the device location
- Doesn’t perform active measurements
- Claims to outperform other reverse DNS based techniques
  - Hard to reproduce since the dataset is private and the code doesn’t work out-of-the-box (hardcoded paths, no documentation)
  - https://github.com/microsoft/ReverseDNSGeolocation

Discussion

Thanks for your attention :-)