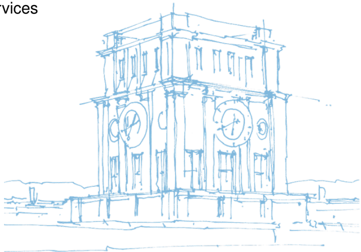
Chair of Network Architectures and Services
Department of Informatics
Technical University of Munich

HLOC: Hints-Based Geolocation Leveraging Multiple Measurement Frameworks

Quirin Scheitle, Oliver Gasser, Patrick Sattler, Georg Carle

TMA'17, Dublin
June 22, 2017

Chair of Network Architectures and Services
Department of Informatics
Technical University of Munich



Geolocating IP Addresses

Geolocation focuses:

- Human-centric, e.g. for businesses
- Structural mapping, e.g. of Internet routers

Geolocation approaches:

- Commercial databases
- Measurement-based algorithms

Geolocating IP Addresses

Geolocation focuses:

- Human-centric, e.g. for businesses
- Structural mapping, e.g. of Internet routers

Geolocation approaches:

- Commercial databases
- Measurement-based algorithms

Our goals:

- **Combine** ease-of-use of **databases** with accuracy of **measurement-based** approaches
- Focus on Internet **routers**

Related Work

Measurement-based:

- Large body of related work using latency, TTL, link-level topology, etc. for geolocation [6, 11, 12, 8, 4, 14, 13, 5, 9, 1]
- High barrier of entry through complex setup and calibration phase

DNS-based:

- RFC 1876: Store latitude and longitude in DNS [2] $\rightarrow$ rarely used
- DRoP [7]: Good results for ground-truth domains, no ready-to-use solution

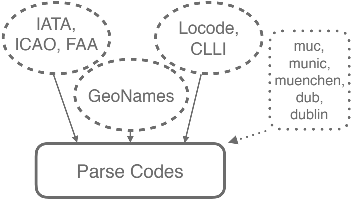
Database-based:

- Questionable accuracy of geolocation databases [3, 10]

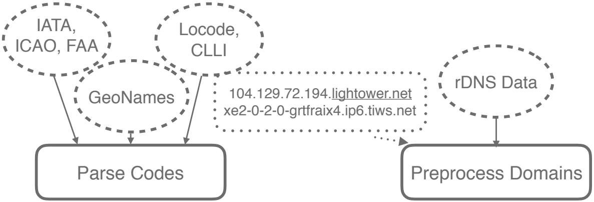
HLOC

- Geolocation based on hints in domain names
- Validation of geolocation hints using latency measurements
- Multi-level measurements
  - High-bandwidth scans
  - Globally distributed scans using RIPE Atlas
- Accuracy of dozens to hundreds of km $\rightarrow$ country-level
- Ready-to-use

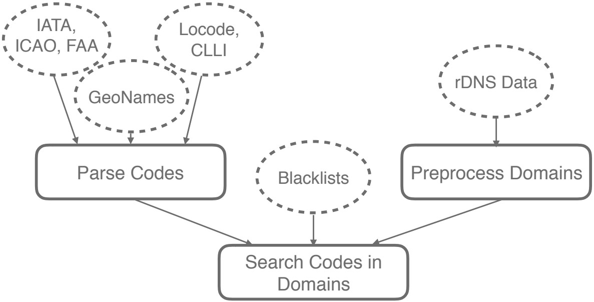
Approach


IATA,
ICAO, FAA
Locode,
CLLI
GeoNames
muc,
munic,
muenchen,
dub,
dublin
Parse Codes

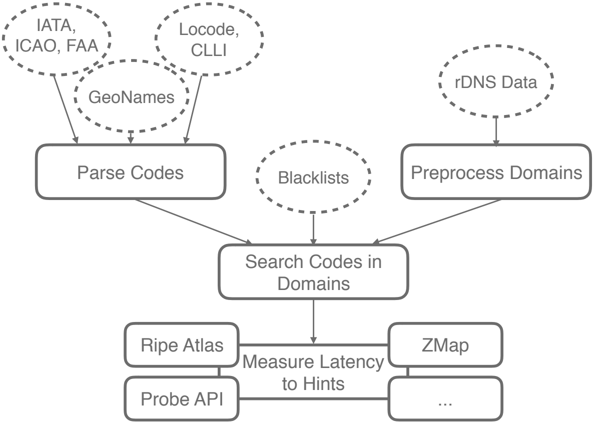
Approach


IATA,
ICAO, FAA
Locode,
CLLI
GeoNames
104.129.72.194.lightower.net
xe2-0-2-0-grtfraix4.ip6.tiws.net
rDNS Data
Parse Codes
Preprocess Domains

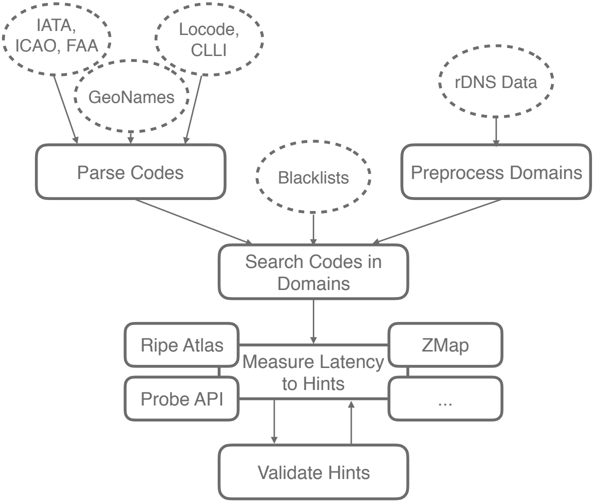
Approach


IATA, ICAO, FAA
Locode, CLLI
GeoNames
rDNS Data
Parse Codes
Blacklists
Preprocess Domains
Search Codes in Domains

Approach


IATA, ICAO, FAA
Locode, CLLI
GeoNames
rDNS Data
Parse Codes
Blacklists
Preprocess Domains
Search Codes in Domains
Ripe Atlas
ZMap
Measure Latency to Hints
Probe API
...

Approach


IATA,
ICAO, FAA
Locode,
CLLI
GeoNames
rDNS Data
Parse Codes
Blacklists
Preprocess Domains
Search Codes in
Domains
Ripe Atlas
ZMap
Measure Latency
to Hints
Probe API
...
Validate Hints

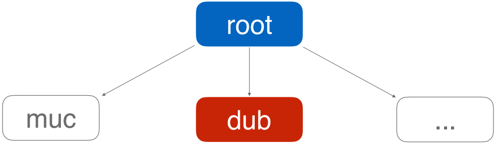
Challenges

- Fast search of location hints in domains
- Reduce number of unlikely matches
- Tailor to measurement limits

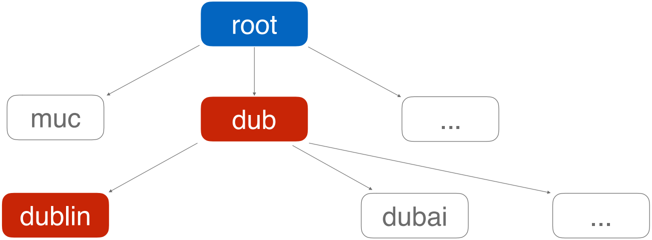
Challenges

- Fast search of location hints in domains → Trie
- Reduce number of unlikely matches → Blacklisting
- Tailor to measurement limits → Use multiple frameworks

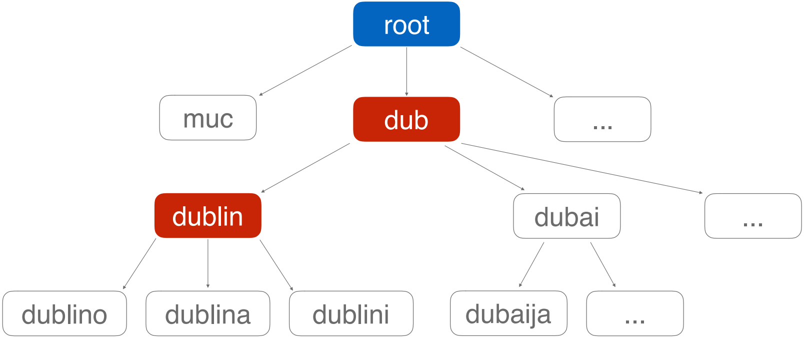
Fast Hint Search: Trie



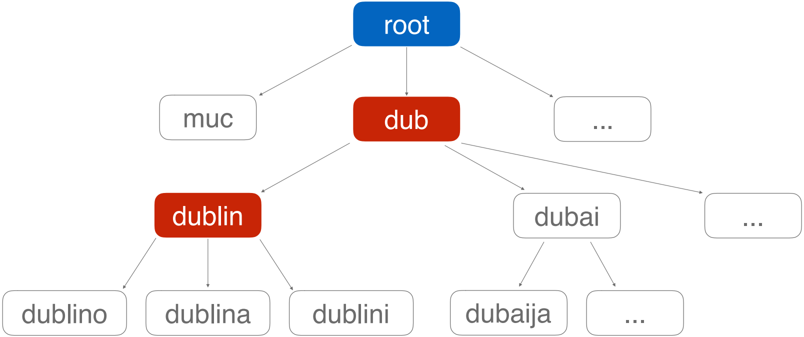
Fast Hint Search: Trie



Fast Hint Search: Trie



Fast Hint Search: Trie



→ Very fast lookup

Reduce Unlikely Matches: Blacklisting

Certain words in domains do not include a location

- Unnecessary increase of measurement duration

Reduce Unlikely Matches: Blacklisting

Certain words in domains do not include a location

- Unnecessary increase of measurement duration

Example:

`ae-0.facebook.amstnl02.nl.bb.gin.ntt.net`

Reduce Unlikely Matches: Blacklisting

Certain words in domains do not include a location

- Unnecessary increase of measurement duration

Example:

`ae-0.facebook.amstnl02.nl.bb.gin.ntt.net`

Reduce Unlikely Matches: Blacklisting

Certain words in domains do not include a location

- Unnecessary increase of measurement duration

Example:

`ae-0.facebook.amstnl02.nl.bb.gin.ntt.net`

- ams (IATA): Amsterdam, Netherlands (2.3 ms)
- ~~face~~ (ICAO): Ceres, South Africa
- ~~ace~~ (IATA): Lanzarote, Spain
- ~~ceb~~ (IATA): Lapu-Lapu City, Philippines
- ...

Reduce Unlikely Matches: Blacklisting

Certain words in domains do not include a location

- Unnecessary increase of measurement duration

Example:

`ae-0.facebook.amstnl02.nl.bb.gin.ntt.net`

- ams (IATA): Amsterdam, Netherlands (2.3 ms)
- ~~face~~ (ICAO): Ceres, South Africa
- ~~ace~~ (IATA): Lanzarote, Spain
- ~~ceb~~ (IATA): Lapu-Lapu City, Philippines
- ...

Publicly available blacklists on `Github`

- Crowdsourcing blacklists further improves measurement performance

Use Multiple Measurement Frameworks

Limitations in frameworks

- Parallel running measurements
- Requests per second

Use Multiple Measurement Frameworks

Limitations in frameworks

- Parallel running measurements
- Requests per second

Multi-level approach

1. Measure from high bandwidth servers in few locations
    - Pin-point hemisphere of location
    - e.g., dedicated servers with ZMap

Use Multiple Measurement Frameworks

Limitations in frameworks

- Parallel running measurements
- Requests per second

Multi-level approach

1. Measure from high bandwidth servers in few locations
   - Pin-point hemisphere of location
   - e.g., dedicated servers with ZMap
2. Measure from low bandwidth probes in many locations
   - Measurement close to hinted location
   - e.g., RIPE Atlas

Validation Concept

- Pick possible location match from right to left label
- Pick suitable probe $dist(probe, location) < x$
- Check validation threshold:

$$RTT(probe, host) < a + \frac{2 \cdot dist(probe, location)}{c \cdot c_0} \quad (1)$$

  - a is the maximal buffer time
  - $c \cdot c_0$ is the propagation speed in fiber optics
- If fulfilled, stop else repeat for the other location matches
- Our maximum error margin is 2900 km ($a = 9ms; x = 1000km$)

Measurement Example

- `cr-01.0v-00-04.anx32.nyc.us.anexia-it.com`

Measurement Example

- `cr-01.0v-00-04.anx32.nyc.us.anexia-it.com`
  - `nyc` (IATA): New York City, USA

Measurement Example

- `cr-01.0v-00-04.anx32.nyc.us.anexia-it.com`
  - `nyc` (IATA): New York City, USA
  - `anx` (IATA): Andenes, Norway

Measurement Example

- `cr-01.0v-00-04.anx32.nyc.us.anexia-it.com`
  - `nyc` (IATA): New York City, USA
  - `anx` (IATA): Andenes, Norway
- Select probe near suspected location

Measurement Example

- `cr-01.0v-00-04.anx32.nyc.us.anexia-it.com`
  - `nyc` (IATA): New York City, USA
  - `anx` (IATA): Andenes, Norway
- Select probe near suspected location
  - New York (Probe ID: 17736; distance: 0.84 km)

Measurement Example

- `cr-01.0v-00-04.anx32.nyc.us.anexia-it.com`
  - `nyc` (IATA): New York City, USA
  - `anx` (IATA): Andenes, Norway
- Select probe near suspected location
  - New York (Probe ID: 17736; distance: 0.84 km)
- Measure RTT from probe

Measurement Example

- `cr-01.0v-00-04.anx32.nyc.us.anexia-it.com`
  - `nyc` (IATA): New York City, USA → 1.3 ms
  - `anx` (IATA): Andenes, Norway
- Select probe near suspected location
  - New York (Probe ID: 17736; distance: 0.84 km)
- Measure RTT from probe
  - RTT(Probe(17736), "2001:2000:3080:c44::2") = 1.3 ms

Measurement Example

- `cr-01.0v-00-04.anx32.nyc.us.anexia-it.com`
  - `nyc` (IATA): New York City, USA → 1.3 ms
  - `anx` (IATA): Andenes, Norway
- Select probe near suspected location
  - 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 hints

Measurement Example

- `cr-01.0v-00-04.anx32.nyc.us.anexia-it.com`
  - `nyc` (IATA): New York City, USA → 1.3 ms
  - ~~`anx` (IATA): Andenes, Norway~~
- Select probe near suspected location
  - 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 hints

Measurement Example

- `cr-01.0v-00-04.anx32.nyc.us.anexia-it.com`
  - `nyc` (IATA): New York City, USA → 1.3 ms
  - ~~`anx` (IATA): Andenes, Norway~~
- Select probe near suspected location
  - 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 hints
- Validate RTT measurements using threshold

Measurement Example

- `cr-01.0v-00-04.anx32.nyc.us.anexia-it.com`
  - `nyc` (IATA): New York City, USA → 1.3 ms
  - ~~`anx` (IATA): Andenes, Norway~~
- Select probe near suspected location
  - 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 hints
- Validate RTT measurements using threshold

$$RTT(probe, host) < a + \frac{2 \cdot dist(probe, location)}{c \cdot c_0} \tag{2}$$

Measurement Example

- `cr-01.0v-00-04.anx32.nyc.us.anexia-it.com`
  - `nyc` (IATA): New York City, USA $\rightarrow$ 1.3 ms
  - ~~`anx` (IATA): Andenes, Norway~~
- Select probe near suspected location
  - 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 hints
- Validate RTT measurements using threshold

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

Measurement Example

- `cr-01.0v-00-04.anx32.nyc.us.anexia-it.com`
  - `nyc` (IATA): New York City, USA → 1.3 ms
  - ~~`anx` (IATA): Andenes, Norway~~
- Select probe near suspected location
  - 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 hints
- Validate RTT measurements using threshold

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

- Location confirmed ✓

Large-scale Measurements

- Conducted large-scale measurements to geolocate IPv4 and IPv6 routers

Large-scale Measurements

- Conducted large-scale measurements to geolocate IPv4 and IPv6 routers

# 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%)

Large-scale Measurements

- Conducted large-scale measurements to geolocate IPv4 and IPv6 routers

# 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%)

- Many falsified hints

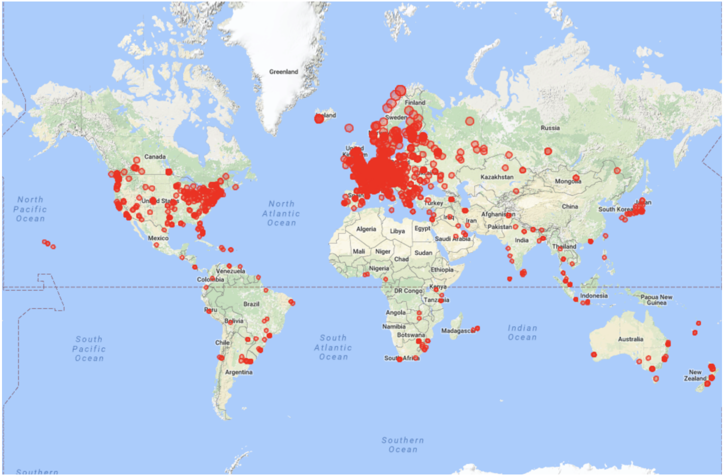
Large-scale Measurements

- Conducted large-scale measurements to geolocate IPv4 and IPv6 routers

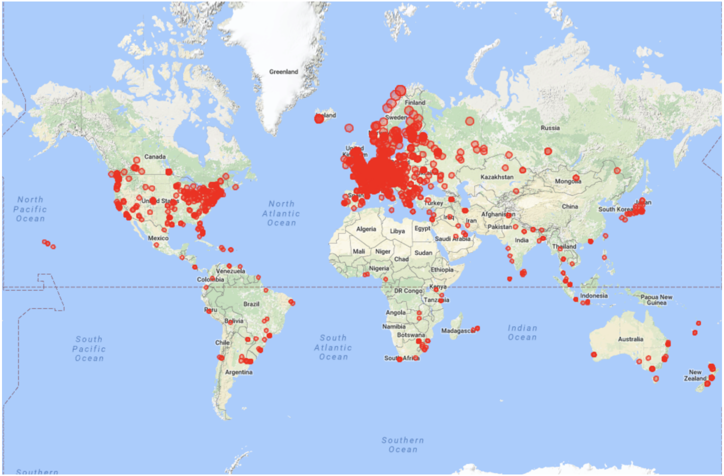
# IP addresses	IPv4	IPv6
Routers	2.5M	190k
– No Match	–1.0M	–7.2k
– Timeout	–431k	–151k
Responsive	961k (100%)	29k (100%)
All hints falsified	417k (**43.4%**)	7k (22.9%)
Hint verified	**45k** (4.7%)	**5k** (17.6%)
No hint verified	500k (52.0%)	17k (59.5%)

- Many falsified hints
- About 50k verified hints

RIPE Atlas Probe Coverage



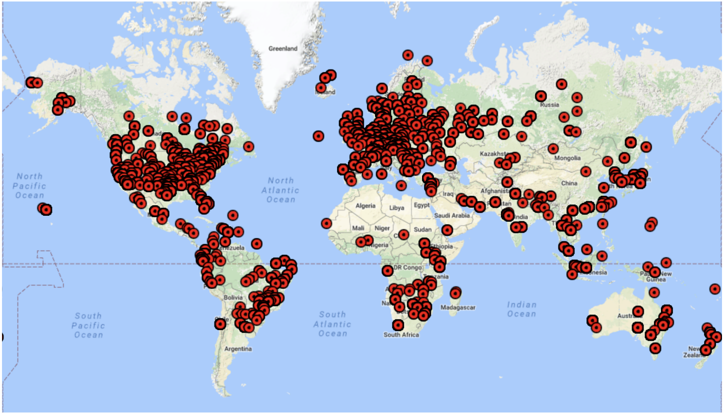
RIPE Atlas Probe Coverage



© Google Maps

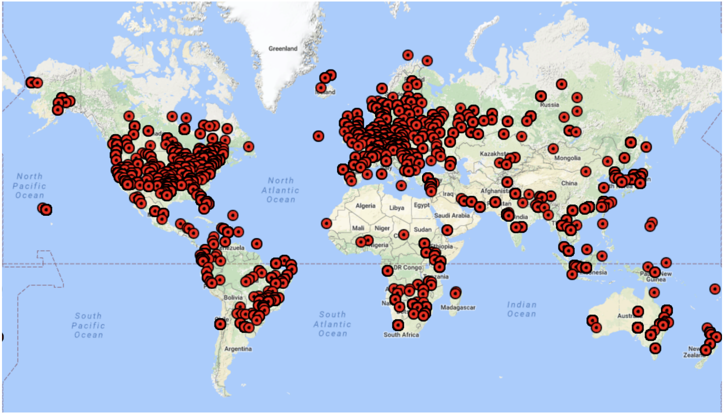
- Good coverage of Europe and USA
- Less coverage in Asia, Africa, and some parts of South America

IPv4 Locations of Validated Domains



© Google Maps

IPv4 Locations of Validated Domains



© Google Maps

- Similar coverage as RIPE Atlas probes

DRoP Comparison

- Goal: Compare our results with DRoP

DRoP Comparison

- Goal: Compare our results with DRoP
  - Reproduce the hint generator using DRoP rules
  - Evaluation on DRoP ground truth domains

DRoP Comparison

- Goal: Compare our results with DRoP
  - Reproduce the hint generator using DRoP rules
  - Evaluation on DRoP ground truth domains

- `cogentco.com`:

DRoP Comparison

- Goal: Compare our results with DRoP
  - Reproduce the hint generator using DRoP rules
  - Evaluation on DRoP ground truth domains

- `cogentco.com`:
  - 26% validated DRoP hints
  - 7% falsified DRoP hints

DRoP Comparison

- Goal: Compare our results with DRoP
  - Reproduce the hint generator using DRoP rules
  - Evaluation on DRoP ground truth domains

- `cogentco.com`:
  - 26% validated DRoP hints
  - 7% falsified DRoP hints
- `ntt.net`:

DRoP Comparison

- Goal: Compare our results with DRoP
  - Reproduce the hint generator using DRoP rules
  - Evaluation on DRoP ground truth domains

- `cogentco.com`:
  - 26% validated DRoP hints
  - 7% falsified DRoP hints
- `ntt.net`:
  - DRoP claims 96% of domains with location hint
  - Reproduction has 54% — HLOC 99%
  - NTT uses custom CLLI location hints (e.g., `londen`)

DRoP Comparison

- Goal: Compare our results with DRoP
  - Reproduce the hint generator using DRoP rules
  - Evaluation on DRoP ground truth domains

- `cogentco.com`:
  - 26% validated DRoP hints
  - 7% falsified DRoP hints
- `ntt.net`:
  - DRoP claims 96% of domains with location hint
  - Reproduction has 54% — HLOC 99%
  - NTT uses custom CLLI location hints (e.g., `londen`)

- `xe2-0-2-0-grtfraix4.ip6.tiws.net`
  - Validated in Frankfurt using HLOC
  - Complex pattern where DRoP would not match

Commercial Database Comparison

- How well do commercial databases work on geolocating routers?

Commercial Database Comparison

- How well do commercial databases work on geolocating routers?

	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%**

Commercial Database Comparison

- How well do commercial databases work on geolocating routers?

	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%**

- Falsified almost half of locations by most popular geolocation database

Result Summary and Next Steps

- Summarized

Result Summary and Next Steps

- Summarized
  - HLOC finds more locations by leveraging complex pattern matching
  - Commercial databases perform poorly on routers
  - IP-encoded domain names contain less locations

Result Summary and Next Steps

- Summarized
  - HLOC finds more locations by leveraging complex pattern matching
  - Commercial databases perform poorly on routers
  - IP-encoded domain names contain less locations
- Coming up

Result Summary and Next Steps

- Summarized
  - HLOC finds more locations by leveraging complex pattern matching
  - Commercial databases perform poorly on routers
  - IP-encoded domain names contain less locations

- Coming up
  - Improved probe selection
  - Direct integration into RIPE Atlas
  - Web service to geolocate hosts
  - Integration of additional measurement frameworks (e.g. ProbeAPI)

Key Contributions

- Geolocation focused on routers
- Multi-level measurement framework
- Configurable accuracy and error margins
- Source code and data available

Key Contributions

- Geolocation focused on routers
- Multi-level measurement framework
- Configurable accuracy and error margins
- Source code and data available

Questions?

Source code, blacklist, and data set: `https://github.com/tumi8/hloc`

Bibliography

[1] M. Calder, X. Fan, Z. Hu, E. Katz-Bassett, J. Heidemann, and R. Govindan.
Mapping the expansion of Google's serving infrastructure.
In ACM SIGCOMM Conference on Internet Measurement, 2013.

[2] C. Davis, P. Vixie, T. Goodwin, and I. Dickinson.
A Means for Expressing Location Information in the Domain Name System.
RFC 1876 (Experimental), Jan. 1996.

[3] B. Gueye, S. Uhlig, and S. Fdida.
Investigating the Imprecision of IP Block-Based Geolocation.
In Passive and Active Measurement, 2007.

[4] B. Gueye, A. Ziviani, M. Crovella, and S. Fdida.
Constraint-Based Geolocation of Internet Hosts.
IEEE/ACM Transactions On Networking, 2006.

[5] C. Guo, Y. Liu, W. Shen, H. J. Wang, Q. Yu, and Y. Zhang.
Mining the Web and the Internet for Accurate IP Address Geolocations.
In INFOCOM, 2009.

[6] Z. Hu, J. Heidemann, and Y. Pradkin.
Towards Geolocation of Millions of IP Addresses.
In ACM SIGCOMM Conference on Internet Measurement, 2012.

[7] B. Huffaker, M. Fomenkov, and k. c. Claffy.
DRoP: DNS-Based Router Positioning.
ACM SIGCOMM Computer Communication Review, 2014.

Bibliography

[8] E. Katz-Bassett et al.
Towards IP Geolocation Using Delay and Topology Measurements.
In ACM SIGCOMM Conference on Internet Measurement, 2006.

[9] V. N. Padmanabhan and L. Subramanian.
An Investigation of Geographic Mapping Techniques for Internet Hosts.
In ACM SIGCOMM Computer Communication Review. ACM, 2001.

[10] I. Poese, S. Uhlig, M. A. Kaafar, B. Donnet, and B. Gueye.
IP Geolocation Databases: Unreliable?
ACM SIGCOMM Computer Communication Review, 2011.

[11] Y. Wang, D. Burgener, M. Flores, A. Kuzmanovic, and C. Huang.
Towards Street-Level Client-Independent IP Geolocation.
In NSDI, 2011.

[12] B. Wong, I. Stoyanov, and E. G. Sirer.
Octant: A Comprehensive Framework for the Geolocalization of Internet Hosts.
In NSDI, 2007.

[13] K. Yoshida et al.
Inferring PoP-level ISP Topology through End-to-End Delay Measurement.
In Passive and Active Measurement, 2009.

[14] I. Youn, B. L. Mark, and D. Richards.
Statistical Geolocation of Internet Hosts.
In International Conference on Computer Communications and Networks. IEEE, 2009.

Which Code Sources are Valuable?

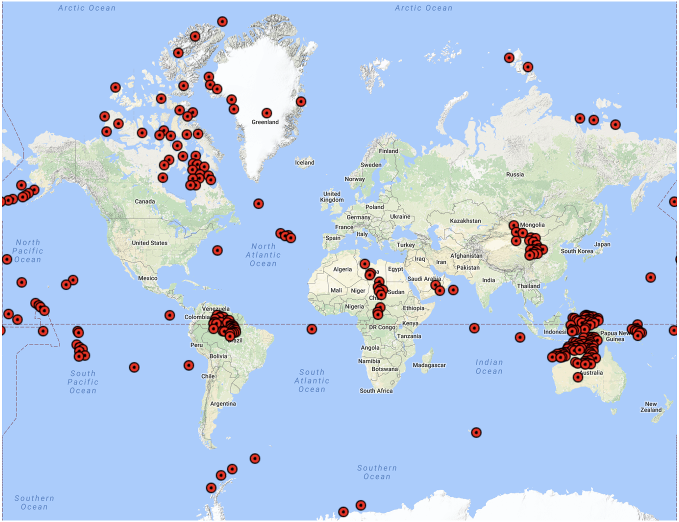
- Evaluate verified locations based on used location code source

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

- IATA, GeoNames and CLLI provide 99% of verified hints
- UN/Locode gives largest number of codes but negligible number of verified locations

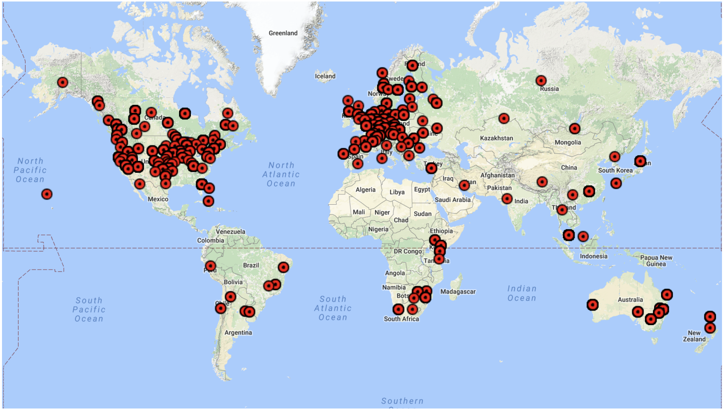
Backup Slides

Locations without RIPE Atlas Probe



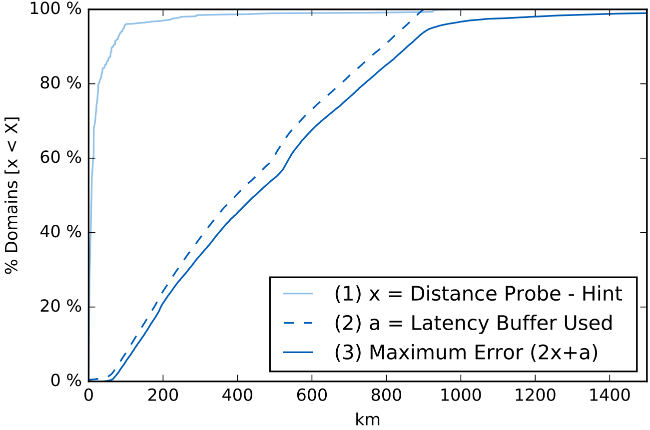
Backup Slides

IPv6 Locations of Validated Domains



Backup Slides

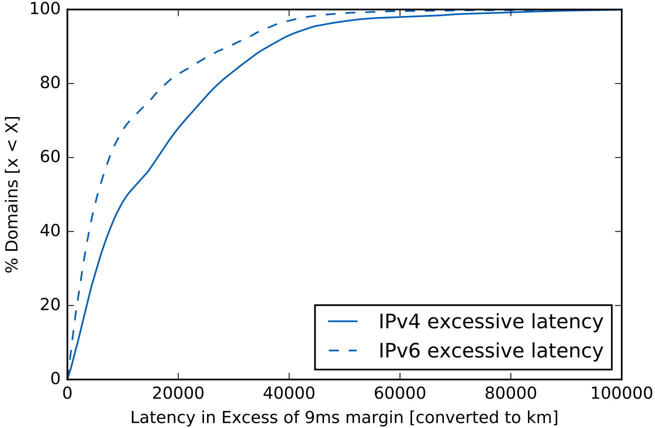
Verified: Error Margin Analysis



- 80% of distances under 25 km
- Used latency buffer and possible error increase linearly

Backup Slides

Not Verified: Sensitivity Analysis



- Excessive latency rises linearly

Domains with Encoded IP Addresses

- Encoded IP addresses in domain name
  - Point to automatically generated domain names
  - Assumption: Lower likelihood of included location in domain name
  - Goal: Find encoded IP addresses in domain names
- Deutsche Telekom domain name
  - `p4FE3C4A8.dip0.t-ipconnect.de`
  - `79.227.196.168`
  - Hexadecimally encoded IPv4 address
- Telus IPv6 domain name
  - `node-1w7jr9qi52esshkbkmpnz14yh.ipv6.telus.net`
  - `2001:569:71d6:2fff:4e8b:30ff:fe48:9e59`
  - Alphanumerically encoded IPv6 address
- Location match likelihood for IP-encoded domains
  - IPv4: Twice as low
  - IPv6: Ten times lower
- Pre-filter IP-encoded domains