HLOC: Hints-Based Geolocation Leveraging Multiple Measurement Frameworks

Quirin Scheitle, Oliver Gasser, Patrick Sattler, Georg Carle

TMA’17, Dublin
June 22, 2017
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] → rarely used
- DRoP [7]: Good results for ground-truth domains, no ready-to-use solution

Database-based:

- Questionable accuracy of geolocation databases [3, 10]
• 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 → country-level
• Ready-to-use
Approach

Parse Codes
GeoNames
Locode, CLLI
IATA, ICAO, FAA

muc, munic, muenchen, dub, dublin

Parse Codes
Approach

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

- Parse Codes
  - GeoNames
  - Locode, CLLI
  - IATA, ICAO, FAA

- Search Codes in Domains
  - Blacklists
  - rDNS Data
  - Preprocess Domains

Q. Scheitle, O. Gasser, P. Sattler, G. Carle — HLOC: Hints-Based Geolocation Leveraging Multiple Measurement Frameworks
Approach

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

Q. Scheitle, O. Gasser, P. Sattler, G. Carle — HLOC: Hints-Based Geolocation Leveraging Multiple Measurement Frameworks
Approach

Parse Codes

Preprocess Domains

Search Codes in Domains

Measure Latency to Hints

Validate Hints

GeoNames

Locode, CLLI

rDNS Data

IATA, ICAO, FAA

Ripe Atlas

Probe API

ZMap

...
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

```
root
  muc
  dub
  ...
```

Very fast lookup
Fast Hint Search: Trie

- root
  - muc
  - dublin
  - dub
    - dubai
    - ...
  - ...

Very fast lookup
Fast Hint Search: Trie

```
root
   /   \
 m uc  \
   /   \
 d ub  \
   /   \
 d ub l i
   /   \
 d ub l in o  d ub l in a  d ub l in i
   /   \
   /   \
 d ub ai
   /   \
 d ub ai ja
```

Very fast lookup
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:
\texttt{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**.amstn102.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.amstn102.nl.bb.gin.ntt.net

- ams (IATA): Amsterdam, Netherlands (2.3 ms)
- faee (ICAO): Ceres, South Africa
- aee (IATA): Lanzarote, Spain
- eeb (IATA): Lapu-Lapu City, Philippines
- ...

Publicly available blacklists on Github

- Crowdsourcing blacklists further improves measurement performance
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 \( \text{dist}(\text{probe, location}) < x \)
- Check validation threshold:

\[
RTT(\text{probe, host}) < a + \frac{2 \cdot \text{dist}(\text{probe, location})}{c \cdot c_0}
\]  

- \( a \) 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. **anx**32. **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
  - \( RTT(\text{Probe}(17736), "2001:2000:3080:c44::2") = 1.3 \text{ ms} \)

- Eliminate impossible hints
  - Validate RTT measurements using threshold
    \[
    RTT(\text{probe}, \text{host}) < a + 2 \cdot \text{dist}(\text{probe}, \text{location}) + c \cdot c_0
    \]

- Location confirmed ✓
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
  - \( \text{RTT(Probe(17736), "2001:2000:3080:c44::2")} = 1.3 \text{ ms} \)

- Eliminate impossible hints

- Validate RTT measurements using threshold

\[
\text{RTT(probe, host)} < a + \frac{2 \cdot \text{dist(probe, location)}}{c \cdot c_0} \tag{2}
\]
Measurement Example

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

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

\[ 1.3 ms < 9 ms + \frac{2 \cdot 0.84 \text{ km}}{200 \frac{\text{km}}{\text{ms}}} \]  

- Location confirmed ✓
Large-scale Measurements

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

<table>
<thead>
<tr>
<th>IP addresses</th>
<th>IPv4</th>
<th>IPv6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Routers</td>
<td>2.5M</td>
<td>190k</td>
</tr>
<tr>
<td>– No Match</td>
<td>–1.0M</td>
<td>–7.2k</td>
</tr>
<tr>
<td>– Timeout</td>
<td>–431k</td>
<td>–151k</td>
</tr>
<tr>
<td>Responsive</td>
<td>961k</td>
<td>29k</td>
</tr>
<tr>
<td>All hints falsified</td>
<td>417k</td>
<td>(43.4%)</td>
</tr>
<tr>
<td>Hint verified</td>
<td>45k</td>
<td>(4.7%)</td>
</tr>
<tr>
<td>No hint verified</td>
<td>500k</td>
<td>(52.0%)</td>
</tr>
</tbody>
</table>
- Conducted large-scale measurements to geolocate IPv4 and IPv6 routers

<table>
<thead>
<tr>
<th># IP addresses</th>
<th>IPv4</th>
<th>IPv6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Routers</td>
<td>2.5M</td>
<td>190k</td>
</tr>
<tr>
<td>– No Match</td>
<td>-1.0M</td>
<td>-7.2k</td>
</tr>
<tr>
<td>– Timeout</td>
<td>-431k</td>
<td>-151k</td>
</tr>
<tr>
<td>Responsive</td>
<td>961k (100%)</td>
<td>29k (100%)</td>
</tr>
<tr>
<td>All hints falsified</td>
<td>417k (43.4%)</td>
<td>7k (22.9%)</td>
</tr>
<tr>
<td>Hint verified</td>
<td>45k (4.7%)</td>
<td>5k (17.6%)</td>
</tr>
<tr>
<td>No hint verified</td>
<td>500k (52.0%)</td>
<td>17k (59.5%)</td>
</tr>
</tbody>
</table>
Large-scale Measurements

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

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

- Many falsified hints
Large-scale Measurements

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

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

- Many falsified hints
- About 50k verified hints
RIPE Atlas Probe Coverage

© Google Maps
• Good coverage of Europe and USA
• Less coverage in Asia, Africa, and some parts of South America
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., lndon)
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?
How well do commercial databases work on geolocating routers?

<table>
<thead>
<tr>
<th></th>
<th>Same</th>
<th>Possible</th>
<th>Wrong</th>
<th>No Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>GeoLite</td>
<td>40.4%</td>
<td>15.6%</td>
<td><strong>44%</strong></td>
<td>-</td>
</tr>
<tr>
<td>ip2location</td>
<td><strong>76.6%</strong></td>
<td>11.3%</td>
<td><strong>12.1%</strong></td>
<td>-</td>
</tr>
<tr>
<td>DRoP</td>
<td>7.8%</td>
<td>0.1%</td>
<td>8.4%</td>
<td><strong>83.7%</strong></td>
</tr>
</tbody>
</table>
• How well do commercial databases work on geolocating routers?

<table>
<thead>
<tr>
<th></th>
<th>Same</th>
<th>Possible</th>
<th>Wrong</th>
<th>No Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>GeoLite</td>
<td>40.4%</td>
<td>15.6%</td>
<td>44%</td>
<td>-</td>
</tr>
<tr>
<td>ip2location</td>
<td>76.6%</td>
<td>11.3%</td>
<td>12.1%</td>
<td>-</td>
</tr>
<tr>
<td>DRoP</td>
<td>7.8%</td>
<td>0.1%</td>
<td>8.4%</td>
<td>83.7%</td>
</tr>
</tbody>
</table>

• 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

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

Questions?

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


An Investigation of Geographic Mapping Techniques for Internet Hosts.

IP Geolocation Databases: Unreliable?

Towards Street-Level Client-Independent IP Geolocation.
In NSDI, 2011.

In NSDI, 2007.

Inferring PoP-level ISP Topology through End-to-End Delay Measurement.

[14] I. Youn, B. L. Mark, and D. Richards.
Statistical Geolocation of Internet Hosts.
Backup Slides

Which Code Sources are Valuable?

- Evaluate verified locations based on used location code source

<table>
<thead>
<tr>
<th>Category</th>
<th>IATA</th>
<th>ICAO</th>
<th>FAA</th>
<th>UN/LO</th>
<th>GeoNames</th>
<th>CLLI</th>
</tr>
</thead>
<tbody>
<tr>
<td># Codes</td>
<td>8k</td>
<td>13k</td>
<td>20k</td>
<td>77k</td>
<td>32k</td>
<td>31k</td>
</tr>
<tr>
<td>Hints</td>
<td>4.5M</td>
<td>209k</td>
<td>472k</td>
<td>59k</td>
<td>215k</td>
<td>167k</td>
</tr>
<tr>
<td>Verified</td>
<td>32k</td>
<td>122</td>
<td>413</td>
<td>120</td>
<td>13k</td>
<td>5k</td>
</tr>
<tr>
<td>Verified (%)</td>
<td>.7%</td>
<td>&lt;.0%</td>
<td>.1%</td>
<td>&lt;.0%</td>
<td>5.9%</td>
<td>2.8%</td>
</tr>
</tbody>
</table>

- 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

© Google Maps
IPv6 Locations of Validated Domains

© Google Maps
• 80% of distances under 25 km
• Used latency buffer and possible error increase linearly
• Excessive latency rises linearly
Backup Slides

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
  - Alphanumerically encoded IPv6 address

- Location match likelihood for IP-encoded domains
  - IPv4: Twice as low
  - IPv6: Ten times lower

- Pre-filter IP-encoded domains