A Look at Router Geolocation in Public and Commercial Databases

Manaf Gharaibeh¹, Anant Shah¹, Bradley Huffaker², Han Zhang¹, Roya Ensafi³, Christos Papadopoulos¹

¹Colorado State University  ²CAIDA / UC San Diego  ³University of Michigan

Internet Measurement Conference
London, United Kingdom
November 1–3, 2017
IP geolocation

- IP geolocation maps an IP address to a physical real-world location
The edge vs the center

• Most of the money and commercial interest is in the edge
  – users
  – content

• So geolocation databases focus most effort on edge
The edge vs. the center

• Many important research questions focus on the center
  – Censorship
  – Geographic stretch
  – Ownership

• How accurate are the geolocation databases for the center?
Motivation

• Router geolocation is used in network research:
  – BGP route visualization and detection of BGP threats
  – Detection of routing paths that experience international detours
  – Studying censorship and monitoring

• Geolocation databases (geo-DBs) accuracy for infrastructure addresses
  – Geo-DBs accuracy evaluation is dominated by the results over end-host addresses
  – Researchers are left unsure about the geo-DBs accuracy over infrastructure addresses such as routers
Goals

• Quantify geo-DBs **consistency and coverage** for router geolocation

• Quantify expected **accuracy** for router geolocation
  – Identify which geo-DBs perform better and **where** (regional evaluation)
# Geo-DBs in this study

<table>
<thead>
<tr>
<th>Free</th>
<th>Commercial</th>
</tr>
</thead>
<tbody>
<tr>
<td>IP2Location DB11.Lite</td>
<td>Digital Envoy NetAcuity</td>
</tr>
<tr>
<td>MaxMind GeoLite</td>
<td>MaxMind GeoIP2</td>
</tr>
</tbody>
</table>

- **Netacuity**: CAIDA has agreement for free access
- **GeoIP2**: purchased access at full price
## Validation datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Ark-topo-router</th>
<th>Ground Truth</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Source/method</strong></td>
<td></td>
<td><strong>Ground Truth</strong></td>
</tr>
<tr>
<td><strong>IP addresses count</strong></td>
<td>1.64M  0.69M (city consistency)</td>
<td>CAIDA DNS Dataset*  Location hints ground truth rules  RIPE Atlas traceroute built-in measurement / RTT-based</td>
</tr>
<tr>
<td><strong>Used to study</strong></td>
<td>Coverage &amp; Consistency</td>
<td>Accuracy</td>
</tr>
</tbody>
</table>

* Macroscopic Internet Topology Data Kit (ITDK)  
+ IPs with city-level coordinates in all geo-DBs
DNS-based (accuracy validation)

- Some operators encode geographic hints into some DNS names
- Operator provided geographic heuristics for 7 domains*

\[...\langle \text{airport code}\rangle\d*\text{.atlas.cogentco.com}\]

be1273.ccr41.lax04.atlas.cogentco.com Los Angeles, US
be3257.ccr41.iad02.atlas.cogentco.com Washington, US
te0-7-0-1.rcr21.b054208-1.lhr01.atlas.cogentco.com London, UK

<table>
<thead>
<tr>
<th>Domain</th>
<th>belwue.de</th>
<th>cogentco.com</th>
<th>digitalwest.net</th>
<th>ntt.net</th>
<th>peak10.net</th>
<th>seabone.net</th>
<th>pnap.net</th>
</tr>
</thead>
<tbody>
<tr>
<td>IP address count</td>
<td>23</td>
<td>6,462</td>
<td>29</td>
<td>2,331</td>
<td>170</td>
<td>1,405</td>
<td>1,437</td>
</tr>
</tbody>
</table>

RTT-proximity (accuracy validation)

- Leverage RIPE Atlas built-in traceroute measurements data
  - From May 25th, 2016
  - Find all IP hops within 0.5ms threshold from monitor/probe
    - IP is within 50 km from the probe
  - Associate each IP with probe
  - Filter incorrectly geolocated probes
  - 4,838 addresses satisfy the RTT threshold
Incorrectly geolocated Atlas probes
Probes with default country-level coordinates

- Typically near the geographic center of a country
- Indicate lack of specific city-level location
- E.g., The United States: 38 00 N, 97 00 W
- Out of 1,387 probes associated with our 0.5ms threshold data
  - 19 probes have default country coordinates
  - Associated with 109 IP addresses
  - All are omitted from the dataset
Incorrectly geolocated Atlas probes
RTT-nearby probes with very different locations

- **Insight:** RTT-nearby probes should also be near each other
- 495 RTT-proximity addresses have RTT-nearby groups of 2 or more probes
  - Only 12 addresses (2.4%) have RTT-nearby probes with **inconsistent** locations.
  - 4 have prominent location inconsistencies.
  - 8 have relatively small disagreements (< 128 km)
- Overall, 223 probes are part of one or more RTT-nearby groups
  - Only 5 probes (2.2%) are disqualified (along with 13 interface addresses associated with them in the dataset)
  - The final RTT-proximity dataset has 4,838 addresses
Methodology

- **40 km city radius**
  - Distance between database coordinates for the same city

- **Coverage**
  - IP has an answer at the given level

- **Consistency (geo-DB vs itself)**
  - All the router’s IPs has the same country
  - All the router’s IPs are within a city radius

- **Accuracy (geo-DBs vs ground truth)**
  - IP address has the same country as GT
  - IP address is within city radius of the GT
Ark-topo-router (coverage validation)

<table>
<thead>
<tr>
<th>Geo-DB</th>
<th>IP2Location-Lite</th>
<th>NetAcuity</th>
<th>MaxMind-GeoLite</th>
<th>MaxMind-Paid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Country</td>
<td>~100%</td>
<td>~100%</td>
<td>99.3%</td>
<td>99.3%</td>
</tr>
<tr>
<td>City</td>
<td>99.9%</td>
<td>99.9%</td>
<td>43%</td>
<td>61.6%</td>
</tr>
</tbody>
</table>

- **Country level**
  - All databases provided country level geolocations for **all IP**

- **City level**
  - IP2Location-Lite and Netacuity provided **almost 100% coverage**
  - MaxMind-GeoLite covers **43%**, paid improves to **61%**
Ark-topo-router (cross consistency)

- Country-level (1.64M IPs)
  - **Pairwise:** > 97% for any two geo-DBs
  - **All 4 geo-DBs:** 95.8%

- City-level (0.69M IPs)
  - The 2 MaxMind DBs **disagree on 11.4%** of the IPs
  - Different vendors disagree on at least **29%** of the IPs

![CDF chart showing IP2Loc vs NetA, MM_G vs MM_P, MM_P vs NetA, IP2L vs MM_P]

40 km city-radius

**ACM Internet Measurement Conference 2017**

15
Quantifying Geo-DBs accuracy
Using ground truth data (DNS-based + RTT-proximity)

- **Country-level**
  - IP2Location-Lite and MaxMind DBs are comparable: **77.5% to 78.6% accuracy**
  - NetAcuity: **89.4%**
- **City-level (40 km city radius)**
  - IP2Location-Lite: **lowest** accuracy
  - MaxMind-Paid vs. MaxMind-GeoLite:
    - 41.3% for paid
    - 30.4% for geolite
  - NetAcuity highest with **73% accuracy**
Geo-DBs regional accuracy
City-level breakdown by RIR

- **ARIN** does poorly across all geo-DBs
- Much of **NetAcuity**’s advantage comes in **ARIN**
- **LACNIC** and **AFRINIC** under sampled
Summary - country-level

- **Good coverage** for all databases
- IP2Location-Lite and MaxMind have similar accuracy (77.5% to 78.6%)
- NetAcuity **highest accuracy** (89.4%)
Summary - city-level

- IP2Location-Lite:
  - High coverage (99.9%), but low accuracy (36%)

- MaxMind-GeoLite vs. MaxMind-Paid (what you pay for):
  - Large coverage increase from 43% to 61%
  - Moderate accuracy increase from 47% to 52%
  - Poor ARIN accuracy 35% and 40%

- NetAcuity:
  - High coverage (99.9%) and higher accuracy (73%)
  - Better ARIN accuracy (69%)
Conclusions

• All geo-DBs have room to improve their router geolocation accuracy at both country- and city-level
• Researchers need to be aware of the geo-DBs inaccuracies and their impact on their research results

Our ground truth dataset is available via IMPACT: https://www.impactcybertrust.org/dataset_view?idDataset=792
Additional Slides
Regional and topological distribution
DNS-based and RTT-proximity sets

<table>
<thead>
<tr>
<th>Ground truth</th>
<th>IP count</th>
<th>Countries</th>
<th>Unique coordinates</th>
<th>ARIN</th>
<th>APNIC</th>
<th>AFRINIC</th>
<th>LACNIC</th>
<th>RIPENCC</th>
<th>Transit ASes</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNS-based</td>
<td>11,857</td>
<td>53</td>
<td>238</td>
<td>9,588</td>
<td>560</td>
<td>0</td>
<td>0</td>
<td>1,709</td>
<td>99.9%</td>
</tr>
<tr>
<td>RTT-proximity</td>
<td>4,838</td>
<td>118</td>
<td>1,347</td>
<td>1,123</td>
<td>372</td>
<td>131</td>
<td>52</td>
<td>3,160</td>
<td>74.5%</td>
</tr>
</tbody>
</table>
DNS-based data correctness
Agreement with latency measurement data

• Vs. the RTT-proximity ground truth
  – 109 common addresses
  – 105 addresses agree within 10 km and 4 addresses agree within 43 km

• Vs. a second RTT-proximity dataset**
  – A set of routers within 1ms RTT threshold from Atlas probes (collected on April 2017)
  – 384 addresses are common with our DNS-based dataset
  – 355 addresses (92.45%) agree within 100 km (337 addresses (87.8%) agree within 40 km)
  – 19 addresses are likely reassigned to hosts at different locations (as recent rDNS records show)
    • No conflict with the DNS-based data
  – Remaining 10 addresses disagreements might be a result of stale hostnames, or few incorrect Atlas probes locations

How often IP addresses move?

• For the 11,857 DNS-based addresses
• Between May 2016 and September 2017
• Hostnames changed for 24% of the addresses
  – Not all hostnames changes indicate location changes
  – Only 30.8% have different location (7.4% of all DNS-based addresses in about 16 months)
Regional (RIR) accuracy

Country-level

- AFRINIC and LACNIC are under sampled
- NetAcuity is the most accurate in all regions
- IP2Location-Lite, MaxMind DBs are comparable at country-level
Country-level accuracy
Individual countries

• Graph shows top 20 countries in ground truth (number of addresses)

• Geo-DBs accuracy varies greatly from one country to another

• NetAcuity is the most consistent: at least 74% in all countries
Low city-level accuracy at ARIN
MaxMind-Paid as a case study

- 2,793 ARIN addresses are not in the US
  - 1,955 of them (70%) are geolocated to the US
    - 519 of the 1,955 addresses have city-level geolocation
    - 504 out the 519 have disagreements > 1,000 km with ground truth
  - Possible fallback to registry information

- 3,897 ARIN addresses are located in the US with city-level information
  - 2,267 (58.2%) have geolocation error > 40 km
    - 91% of them have block-level (/24 block or larger) locations
    - Compared to 78% of the correctly geolocated addresses at city-level
  - Block-level location assignments can be responsible for large geolocation errors (previous work)
  - We didn’t investigate blocks co-locality in this work