A Look at Router Geolocation in Public and Commercial Databases

Manaf Gharaibeh[1], Anant Shah[1],
Bradley Huffaker[2], Han Zhang[1], Roya Ensafi[3], Christos Papadopoulos[1]

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




IIJ Presentation

Tokyo, Japan

November 21, 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 **coverage and consistency** for **router geolocation**

- Quantify expected **accuracy** for router geolocation
  - Identify which geo-DBs perform better and **where** (**regional evaluation**)

Geo-DBs in this study

Free	Commercial
IP2Location DB11.Lite	Digital Envoy NetAcuity*
MaxMind GeoLite	MaxMind GeoIP2^{+}

- ***Netacuity**: CAIDA has agreement for free access
- **$^{+}$GeoIP2**: purchased access at full price

Validation datasets

Dataset	Ark-topo-router	Ground Truth	
		DNS-based	RTT-proximity
Source/method	CAIDA Router Topology*	CAIDA DNS Dataset* Location hints ground truth rules	RIPE Atlas traceroute built-in measurement / RTT-based
IP addresses count	1.64M 0.69M (city consitency) +	11,857	4,838
Used to study	Coverage & Consistency	Accuracy	

* Macroscopic Internet Topology Data Kit (ITDK)
http://www.caida.org/data/internet-topology-data-kit/

\+ IPs with city-level coordinates in all geo-DBs

DNS-based (accuracy validation)

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

…<**airport code**>\d*.atlas.cogentco.com

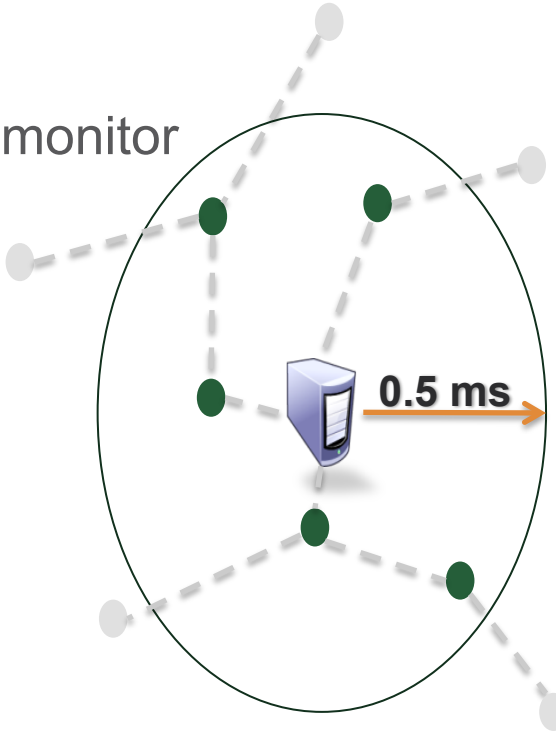
be1273.ccr41.**lax**04.atlas.cogentco.com	Los Angeles, US
be3257.ccr41.**iad**02.atlas.cogentco.com	Washington, US
te0-7-0-1.rcr21.b054208-1.**lhr**01.atlas.cogentco.com	London, UK

Domain	belwue.de	cogentco.com	digitalwest.net	ntt.net	peak10.net	seabone.net	pnap.net
IP address count	23	6,462	29	2,331	170	1,405	1,437

*Huffaker et al., DRoP: DNS-based Router Positioning. ACM SIGCOMM Computer Communication Review 44, 3 (2014), 5–13.

RTT-proximity (accuracy validation)

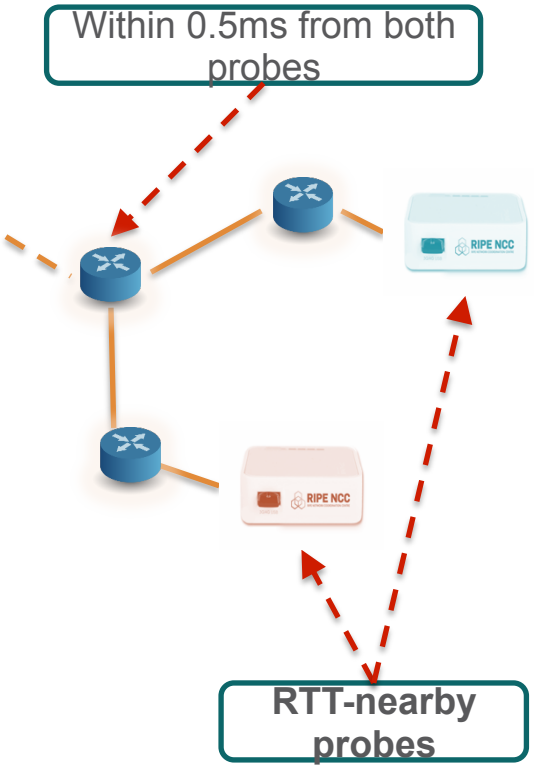
- Leverage RIPE Atlas built-in traceroute measurements data
  - From May 25th, 2016
  - Find all IP hops **within 0.5ms** threshold from monitor
    - IP are then within **50 km** from the probe
  - Associate each IP with **monitor**
  - Filter incorrectly relocated probes
  - **4,838 addresses** satisfy the RTT threshold



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)



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 with in a **city radius**
- **Accuracy (geo-DBs vs ground truth)**
  - **IP address** has the **same country** as GT
  - **IP address** is with in **city radius** of the GT (Geoname coordinates)

Ark-topo-router (coverage validation)

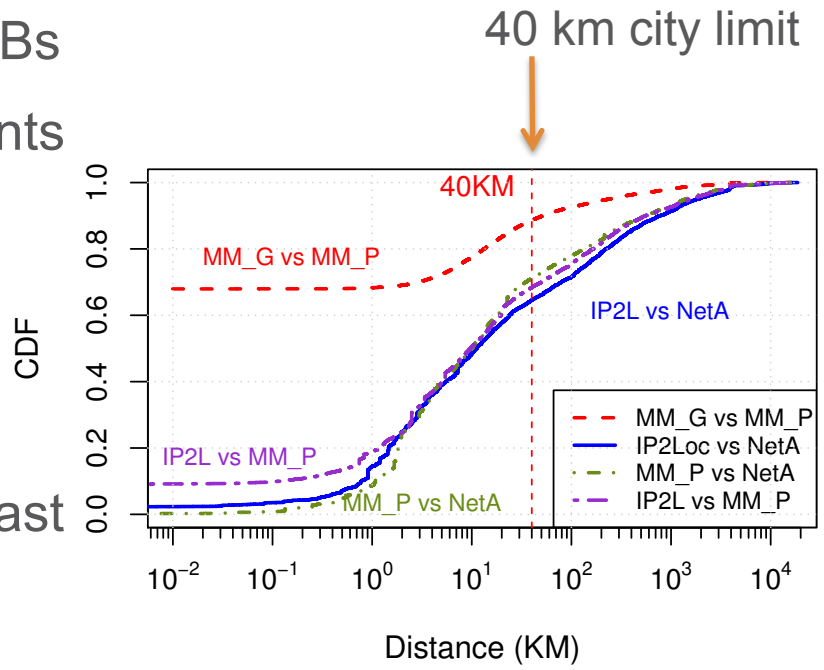
Geo-DB	IP2Location-Lite	NetAcuity	MaxMind-GeoLite	MaxMind-Paid
Country	~100%	~100%	99.3%	99.3%
City	99.9%	99.9%	**43%**	**61.6%**

- 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%**

* Macroscopic Internet Topology Data Kit (ITDK)
http://www.caida.org/data/internet-topology-data-kit/

Ark-topo-router (cross consistency)

- Country-level (1.64M IPs)
  - **Pairwise > 97%** for any two geo-DBs
  - **95.8% for all 4 geo-DBs** agreements
- City-level (0.69M IPs)
  - The 2 MaxMind DBs **disagree on 11.4%** of IPs
  - Different vendors disagree on at least **29%** of IPs

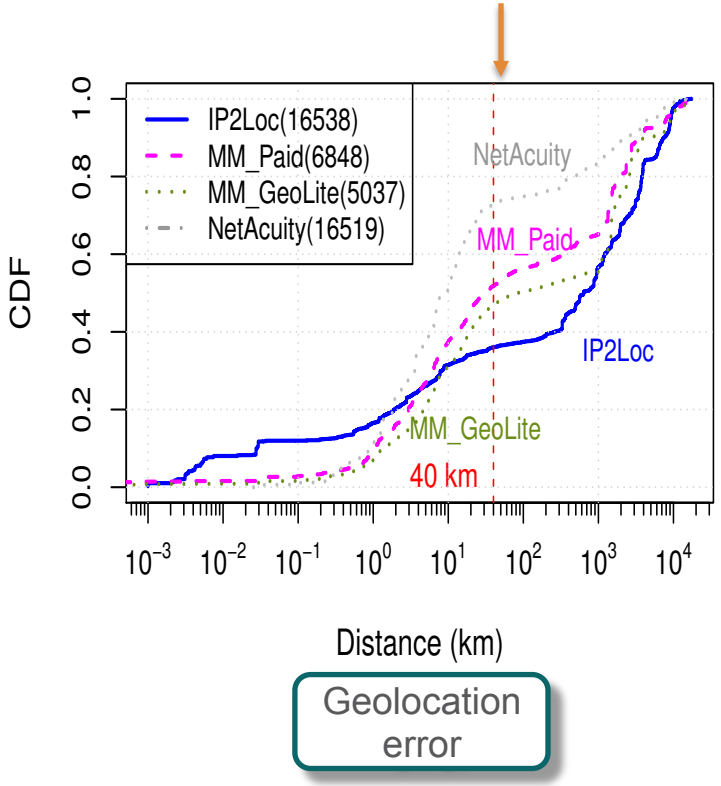


Quantifying Geo-DBs accuracy

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:
    - **30.4%** for geolite
    - **41.3%** for paid
  - NetAcuity highest with **73%** accuracy

40 km, is a knee for both NetAcuity a& MM

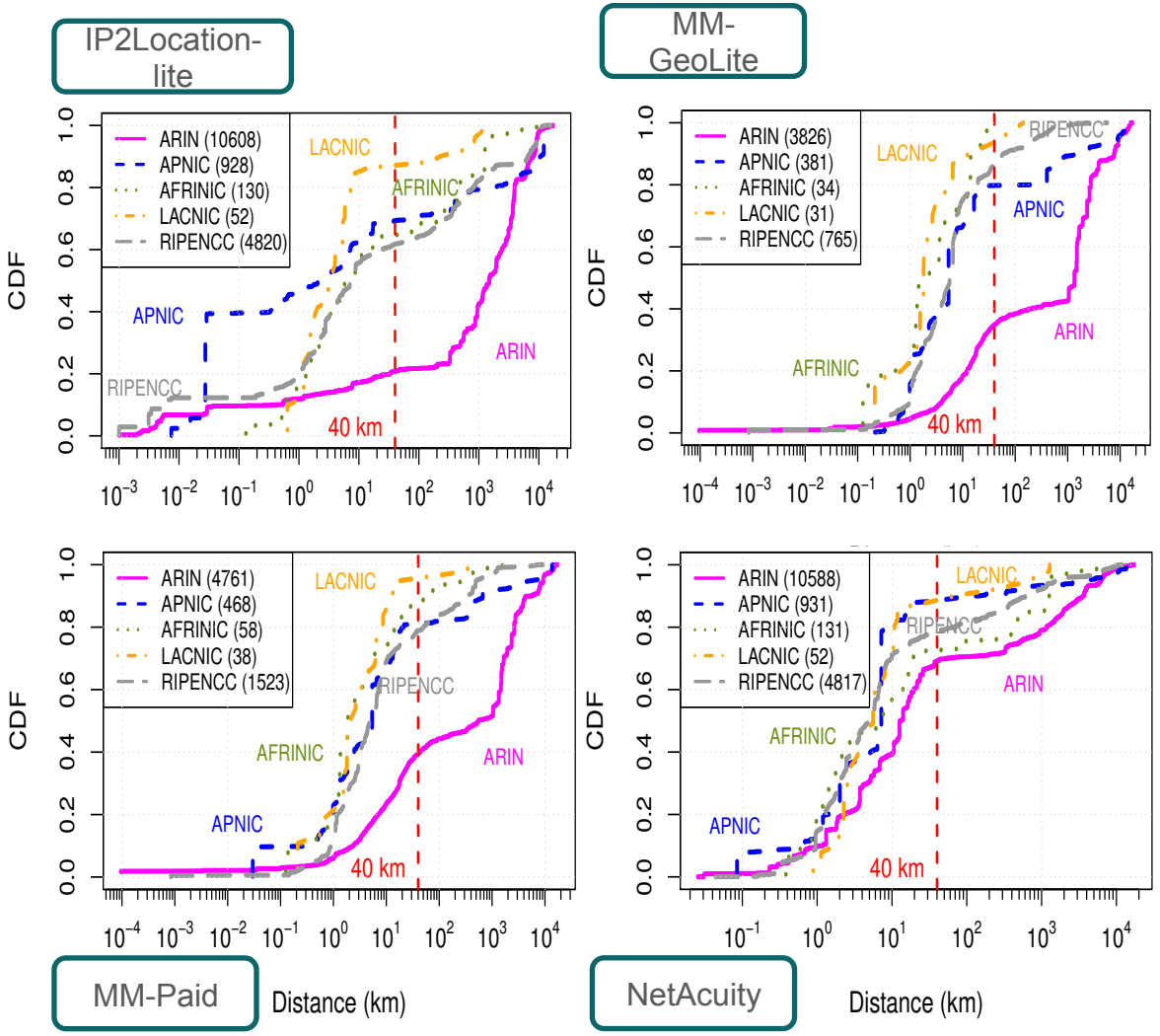


Geolocation error

Geo-DBs **regional** accuracy

City-level breakdown by RIR

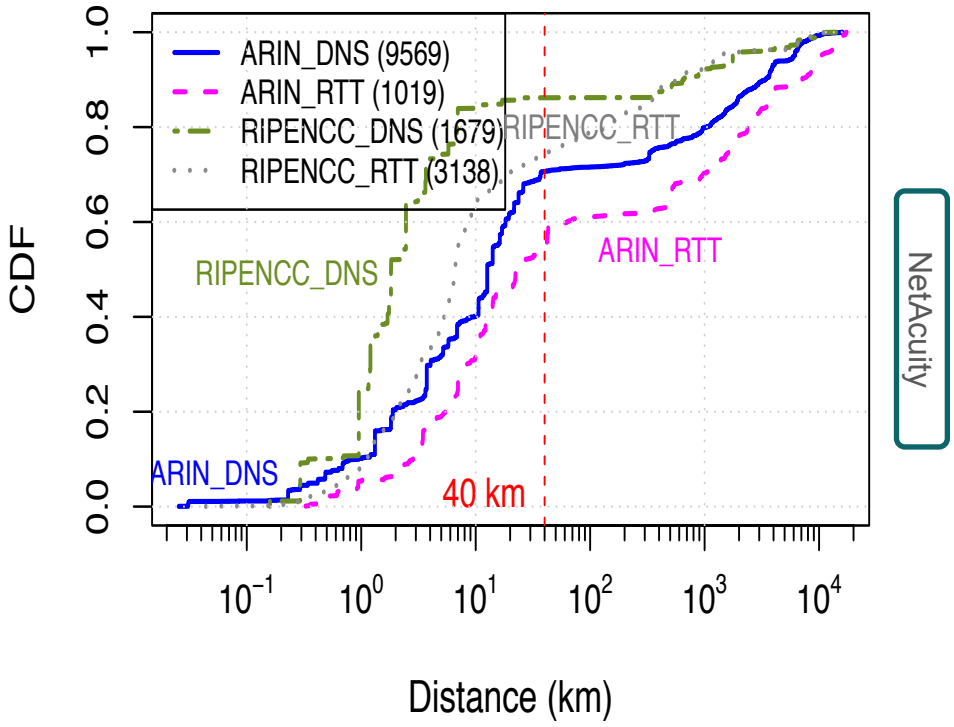
- **ARIN** does does **poorly** across all geo-DBs
- Much of **NetAcuity's advantage** comes in **ARIN**
- LACNIC and AFRINIC under sampled

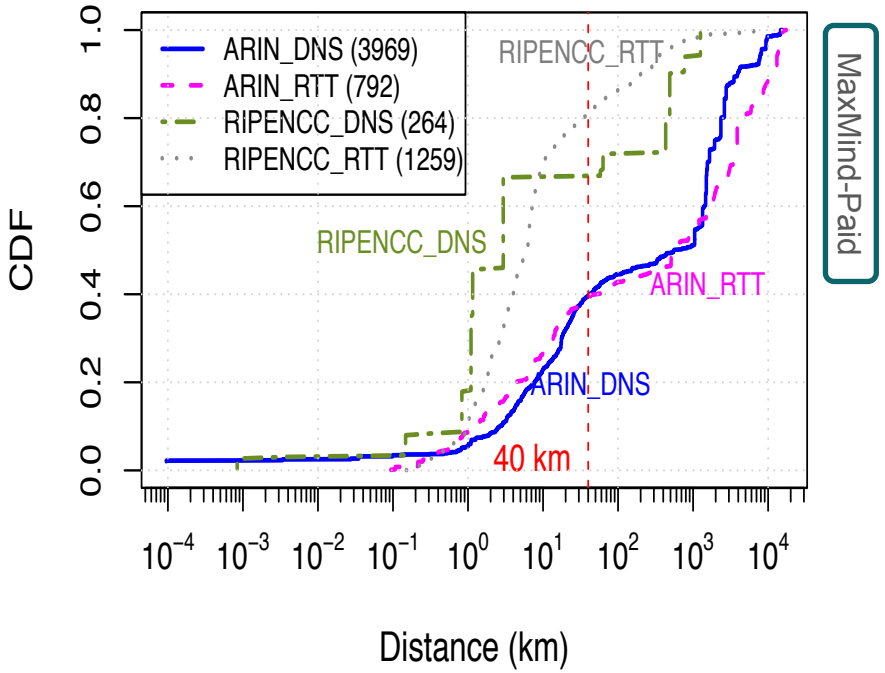


Do databases take advantage of location hints?

Results vs. DNS-based set and vs. RTT-bases set (city-level)

- NetAcuity results over DNS-based data are somewhat better than its results over RTT-proximity data





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 **highest accuracy** (73%)
  - **Better ARIN accuracy** (69%)

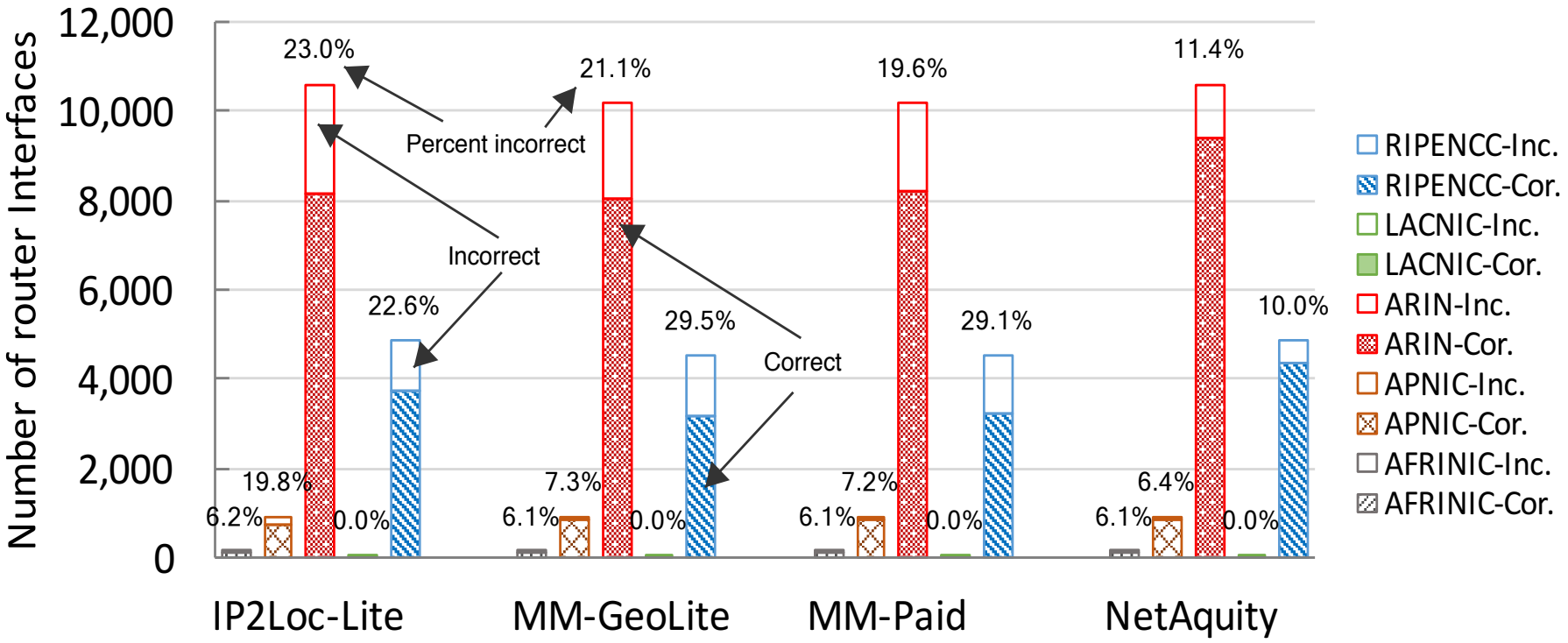
Conclusions

- 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

Backup Slides

Regional (RIR) accuracy



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

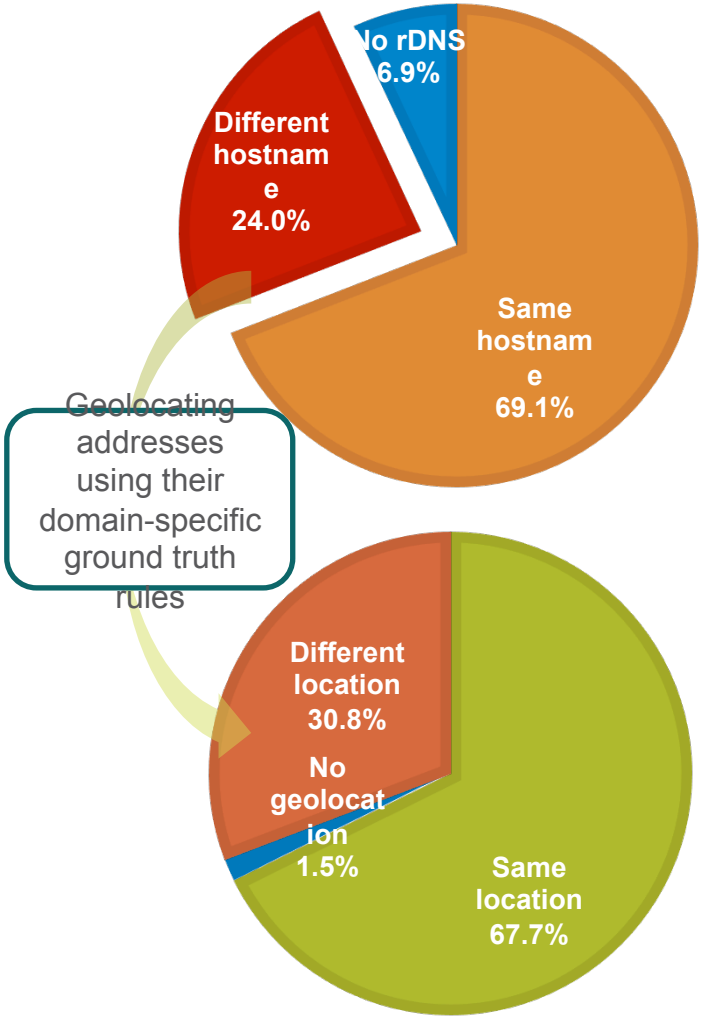
Incorrectly geolocated Atlas probes

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

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**)



Recommendations for researchers

- NetAcuity has the best combination of coverage, accuracy across all regions
  - **We recommend NetAcuity to geolocate routers if geo-DBs is the only option available**
- IP2Location-Lite overall accuracy is too low
  - **We do not recommend it**
- MaxMind DBs are doing bad in ARIN, good in other regions, but have very low city-level coverage
  - **We do not recommend them if high city-level accuracy and coverage are required**
  - **We recommend the paid version over the public one (better city-level coverage and accuracy)**
- All geo-DBs show less accuracy for ARIN addresses
  - **We recommend users to be extra careful when geolocating ARIN**

Incorrectly Geolocated Atlas Probes

- Second method is based on the insight:
  - Multiple probes near the same router 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.
  - The 8 remaining addresses have relatively small disagreements (< 128 km)
  - Overall, we have 223 different probes that are part of one or more RTT-nearby groups

DNS-based set

- We decode location hints in routers' hostnames
  - Use **domain-specific rules from 7 ground truth domains***
  - Rules are obtained from the domains operators
- Performing reverse DNS lookups to the Ark-topo-router addresses
  - 905K addresses have hostnames (55%)
  - About 13.5K belong to the 7 ground truth domains
  - **11,857 addresses are geolocated using the ground truth rules:**

Domain	belwue.de	cogentco.com	digitalwest.net	ntt.net	peak10.net	seabone.net	pnap.net
IP address count	23	6,462	29	2,331	170	1,405	1,437

*Huffaker et al., DRoP: DNS-based Router Positioning. ACM SIGCOMM Computer Communication Review 44, 3 (2014), 5–13.

DNS-based data correctness

Agreement with latency measurement data

- Our RTT-proximity ground truth
  - **109 common addresses**
  - **105 addresses agree within 10 km and 4 addresses agree within 43 km**
- Using 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

**Giotsas et al. 2016, The Remote Peering Debate: Boarding the Remote Peering Ecosystem (PAPER 73)

Regional and topological distribution

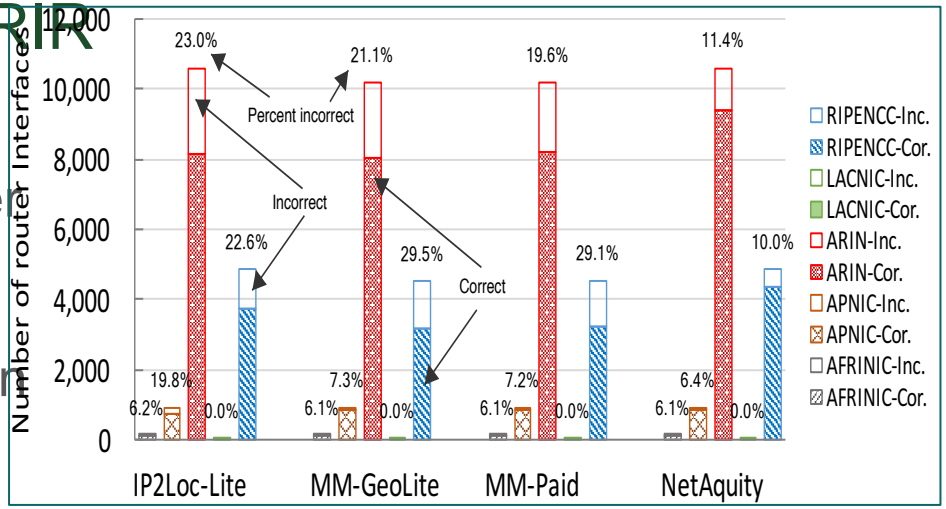
DNS-based and RTT-proximity sets

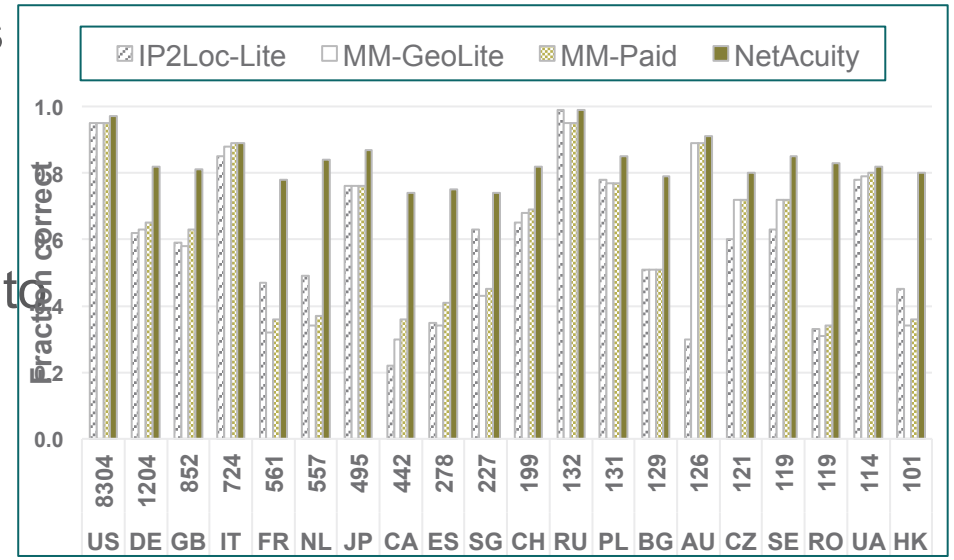
Ground truth	IP count	Countries	Unique coordinates	ARIN	APNIC	AFRINIC	LACNIC	RIPENCC	Transit ASes
DNS-based	11,857	53	238	9,588	560	0	0	1,709	**99.9%**
RTT-proximity	4,838	118	1,347	1,123	372	131	52	3,160	**74.5%**

Quantifying Geo-DBs **regional** accuracy

Country-level breakdown by RIR

- AFRINIC and LACNIC are under sampled
- NetAcuity is the most accurate in all regions
- IP2Location-Lite, MaxMind DBs are comparable at country-level
- However, geo-DBs accuracy varies greatly from one country to another (as the bottom graph shows)

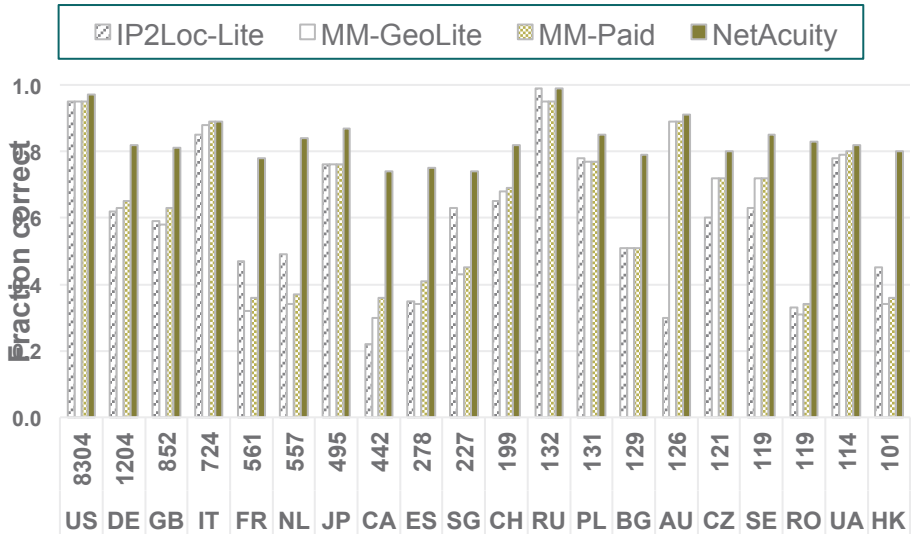




Quantifying Geo-DBs **regional** accuracy

Country-level breakdown by RIR

- 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