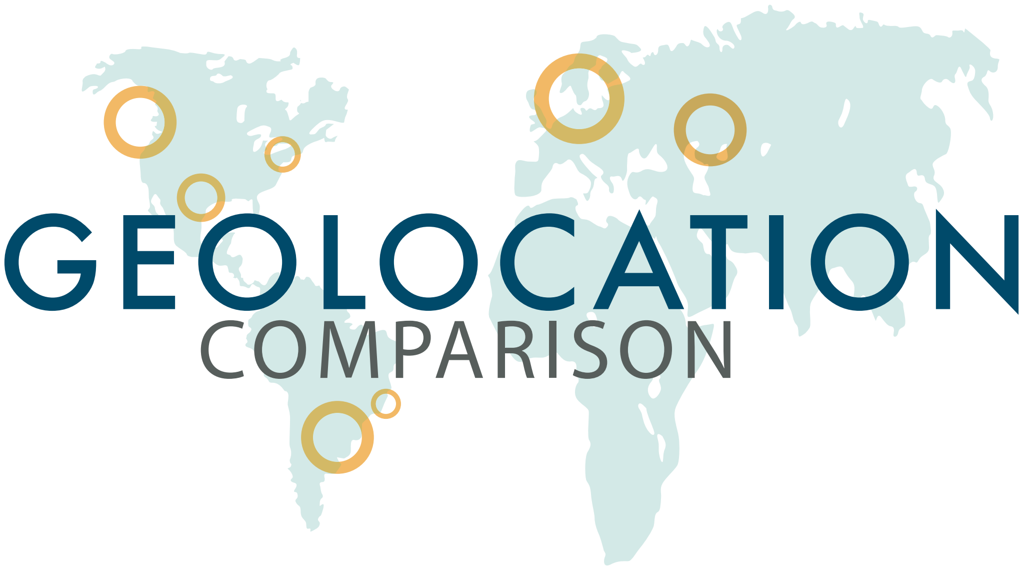


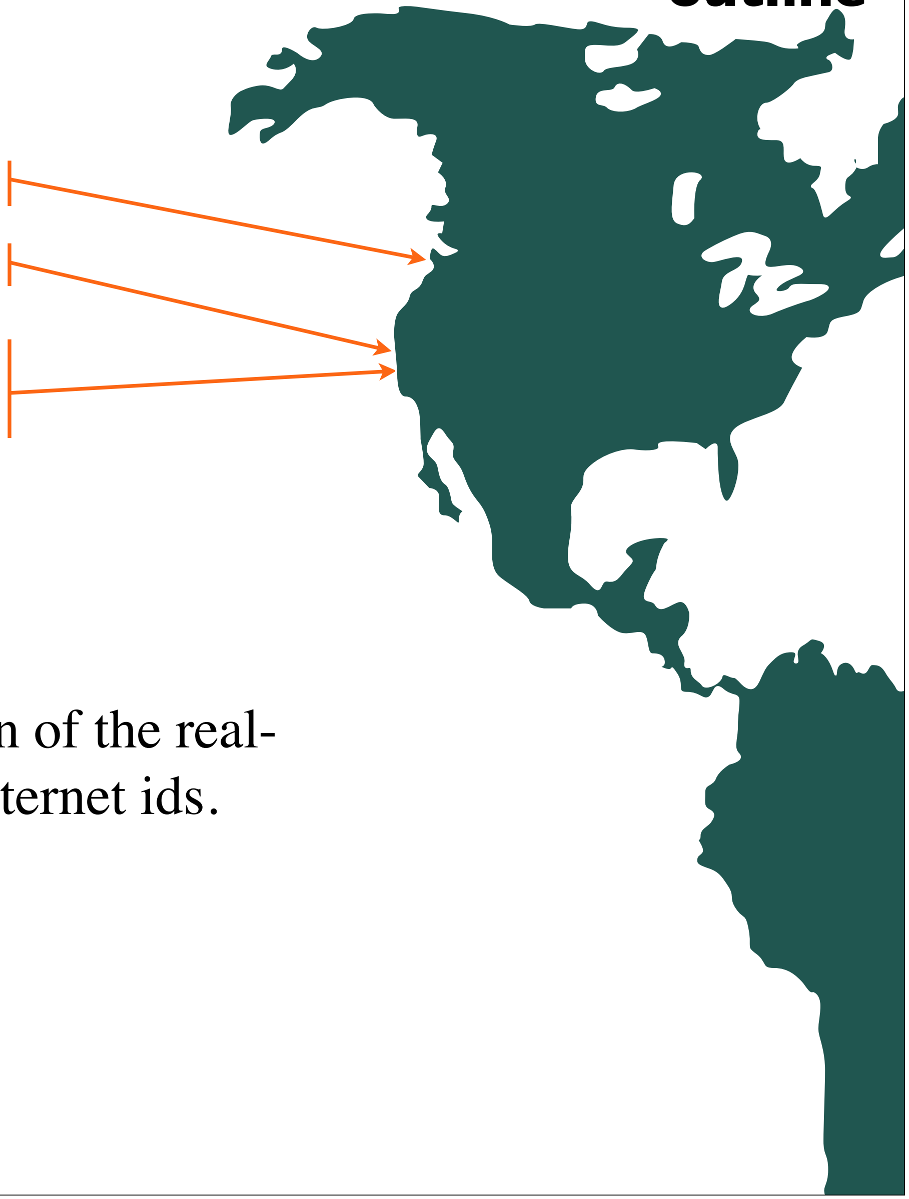
CAIDA's Geolocation Database Comparison

Bradley Huffaker
bradley@caida.org

CAIDA
University of California
at San Diego, La Jolla, CA

Network Mapping and Measurement Conference 2011

Geolocation



sinet-1-lo-jmb-702.lsanca.pacificwave.net (207.231.240.135)

hpr-lax-hpr--sdsc-10ge.cenic.net (137.164.26.33)

dolphin.sdsc.edu (132.249.31.17)
piranha.sdsc.edu (198.17.46.8)
pinot-g1-0-0 (192.172.226.1)

Geolocation is the identification of the real-world geographic location of Internet ids.

Outline

- datasets
- methodology
- analysis
- conclusion

Database types

- Databases
- Address Breakdown
  - Organization Type
  - Regional Internet Registry
- Ground Truth

Databases

Table 1: Geolocation service provider database statistics.

Database	cost[1]	date	addr[2]	blocks	countries	cities	lat,long
RIR$_f$	-	2010.10.31	100.0%	105,380	229	-	-
Software77$_f$	-	2010.12.01	99.5%	105,334	229	-	-
HostIP$_f$	-	2010.10.04	15.9%	780,287	216	-	23,906
IPligence	$	2010.10.06	98.0%	3,155,821	234	-	56,004
Cyscape	$$	2010.08.31	96.8%	54,639	234	-	-
MaxMind GeoIP	$$$	2010.12.01	100.0%	5,774,006	239	128,368	130,707
MaxMind Lite$_f$	-	2010.11.01	100.0%	3,536,604	239	113,216	115,982
IPInfoDB$_f$[3]	-	2010.12.01	100.0%	3,533,709	228	113,209	115,950
Digital Envoy	$$$$	2010.12.02	100.0%	6,082,327	241	33,247	33,195

Indented databases are derived from the database in the row above.

$_f$ marks the free datasets

[1] cost of unlimited geolocation: $ = $1-$300 $$ = $300-$900 $$$ = $900-$1800 $$$$ = $1800+

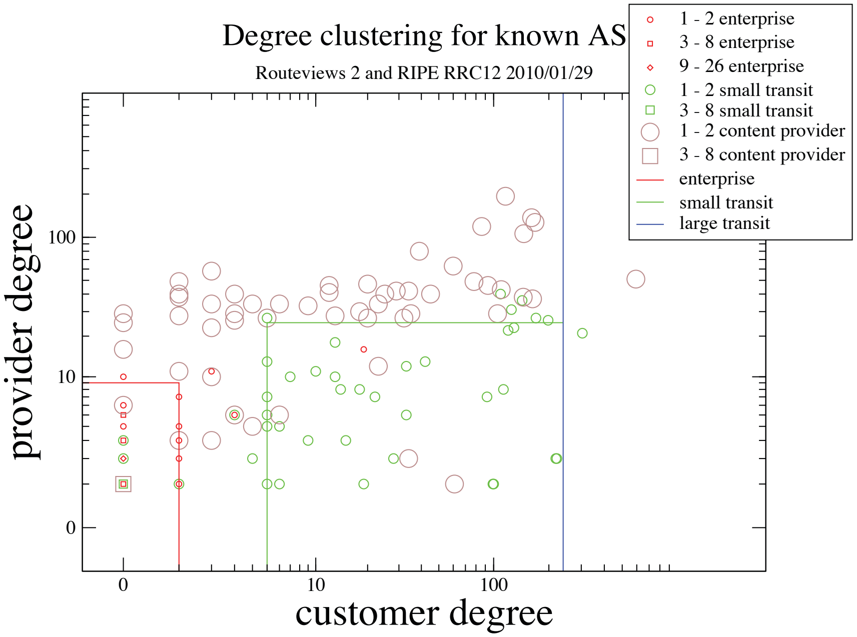
[2] out of RIR delegated addresses

[3] IPInfoDB is almost indistinguishable from MaxMind Lite and is not individually displayed in the rest of the paper.

Organization Type

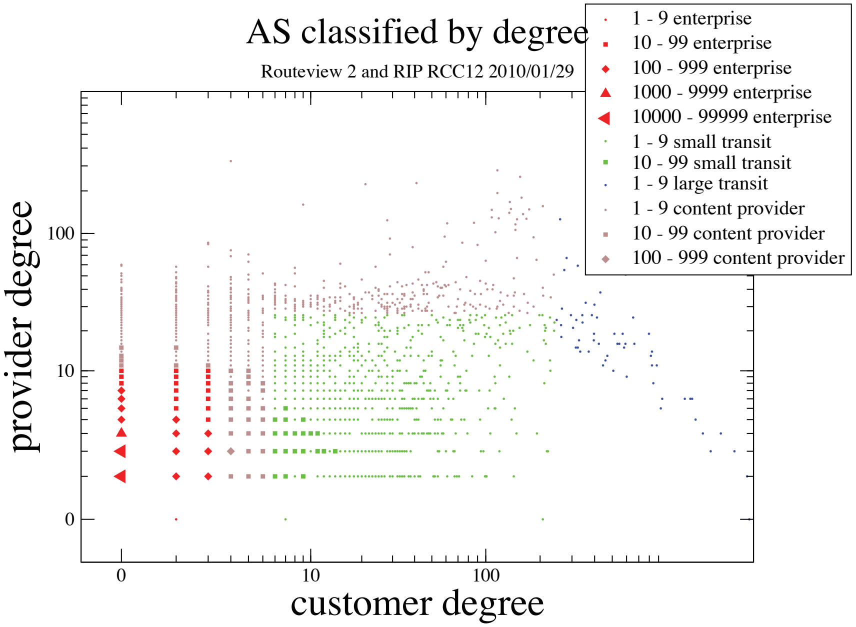
- **Enterprise Customers** (**EC**): typically organizations, universities and companies at the edge, comprised of mostly users
- **Content/Access/Hosting Providers** (**CAHP**): also at the edge, but typically provide content and/or Internet access
- **Small Transit Providers** (**STP**): provide transit to smaller ASes, in addition to content and access services, but purchase transit from a larger Transit Provider
- **Large Transit Providers** (**LTP**): same services as the STP, but have sufficient coverage that they do not need to pay for transit

Org. Classification



Degree clustering for known AS

Routeviews 2 and RIPE RRC12 2010/01/29



AS classified by degree

Routeview 2 and RIP RCC12 2010/01/29

Density plot for 50 ASes classified by hand, with manually drawn bounding boxes that separate most ASes of a given type into their own class.

Using bounding boxes from hand-classified data set on full AS set.

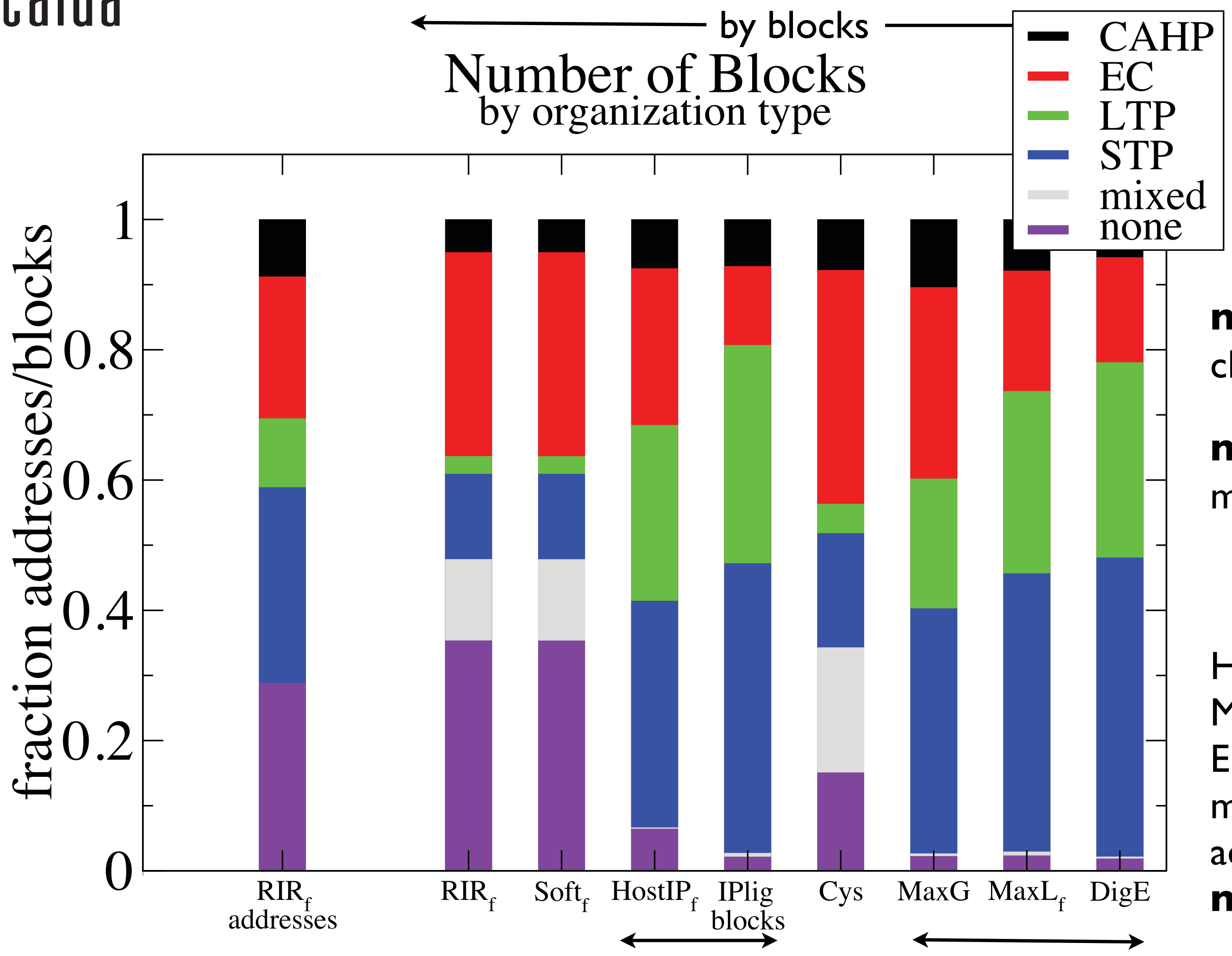
AS links relationships from CAIDA's as-rank.caida.org

Org. Breakdown

GEOLOCATION COMPARISON

datasets

by blocks



none: addresses not classified, not in BGP table

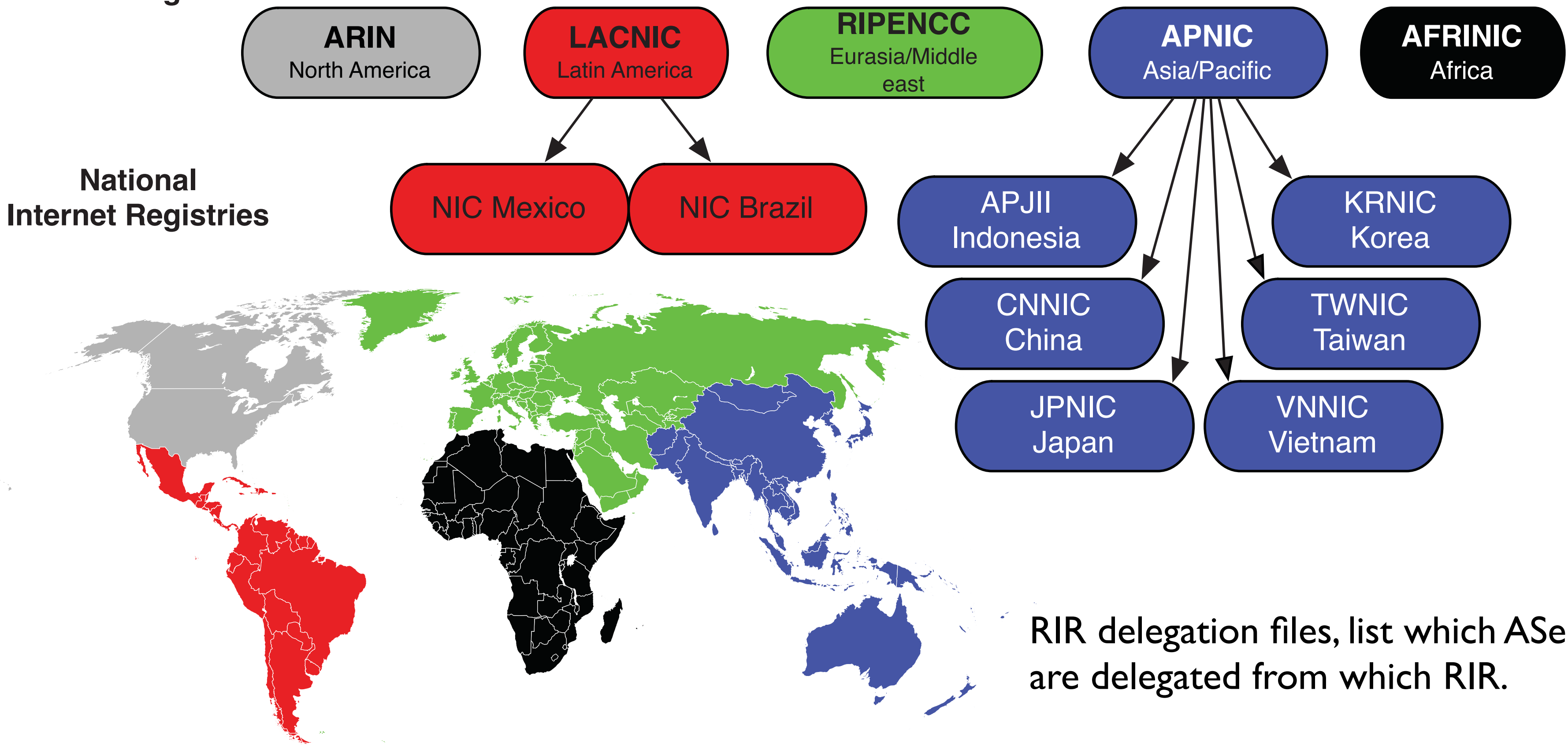
mixed: blocks covered by multiple classification

HostIP, IPligence, Maxmind, and Digital Envoy have less then 6%, most less then 3%, of address blocks in **none** or **mixed**

Region Classification

Regional
Internet Registries

National
Internet Registries



RIR delegation files, list which ASes are delegated from which RIR.

Ground Truth

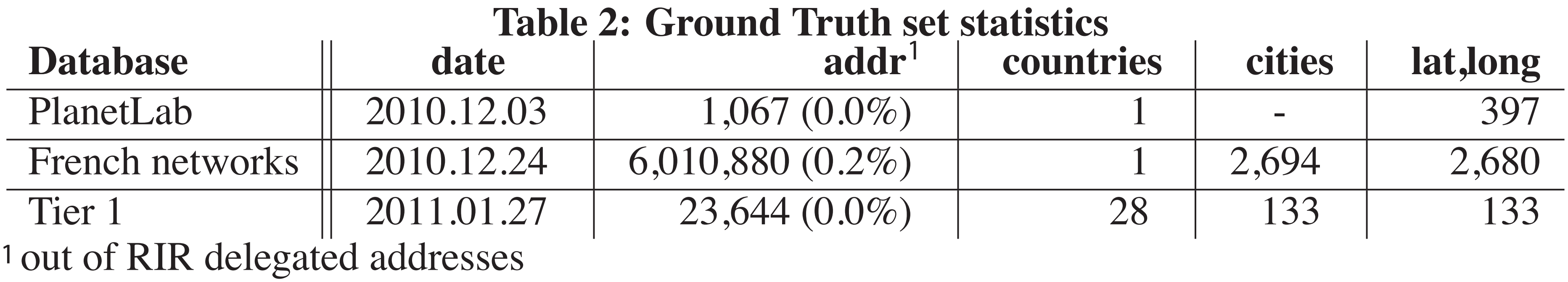


Table 2: Ground Truth set statistics

Database	date	addr[1]	countries	cities	lat,long
PlanetLab	2010.12.03	1,067 (0.0%)	1	-	397
French networks	2010.12.24	6,010,880 (0.2%)	1	2,694	2,680
Tier 1	2011.01.27	23,644 (0.0%)	28	133	133

[1] out of RIR delegated addresses

PlanetLab is a globally distributed set of computers available as a testbed for computer networking and distributed systems research.

French networks FreeNet's list of SDSL networks by geographic region.

US Tier 1 a large US transit provider.

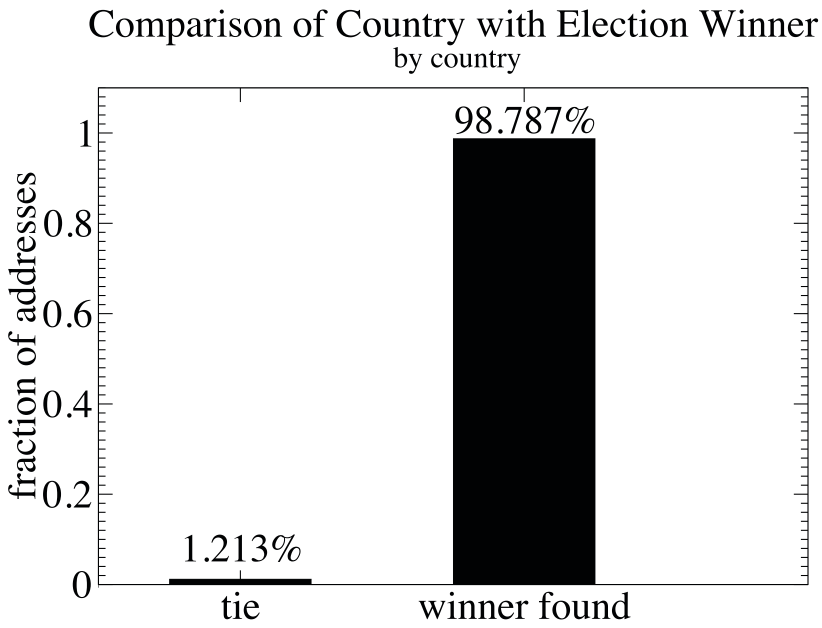
Methodology

- Country Election
  - country agreement with the majority of databases
- Coordinate lat/long Election
  - distance from coordinate cluster derived from majority of databases
- Ground Truth
  - distance from ground truth location

Methodology

country election

- Country Election
  - An election is held across all databases; country with most votes wins.
  - Databases agree or disagree with winner.
  - RIR, IPInfoDB, and MaxMind Lite not included in election



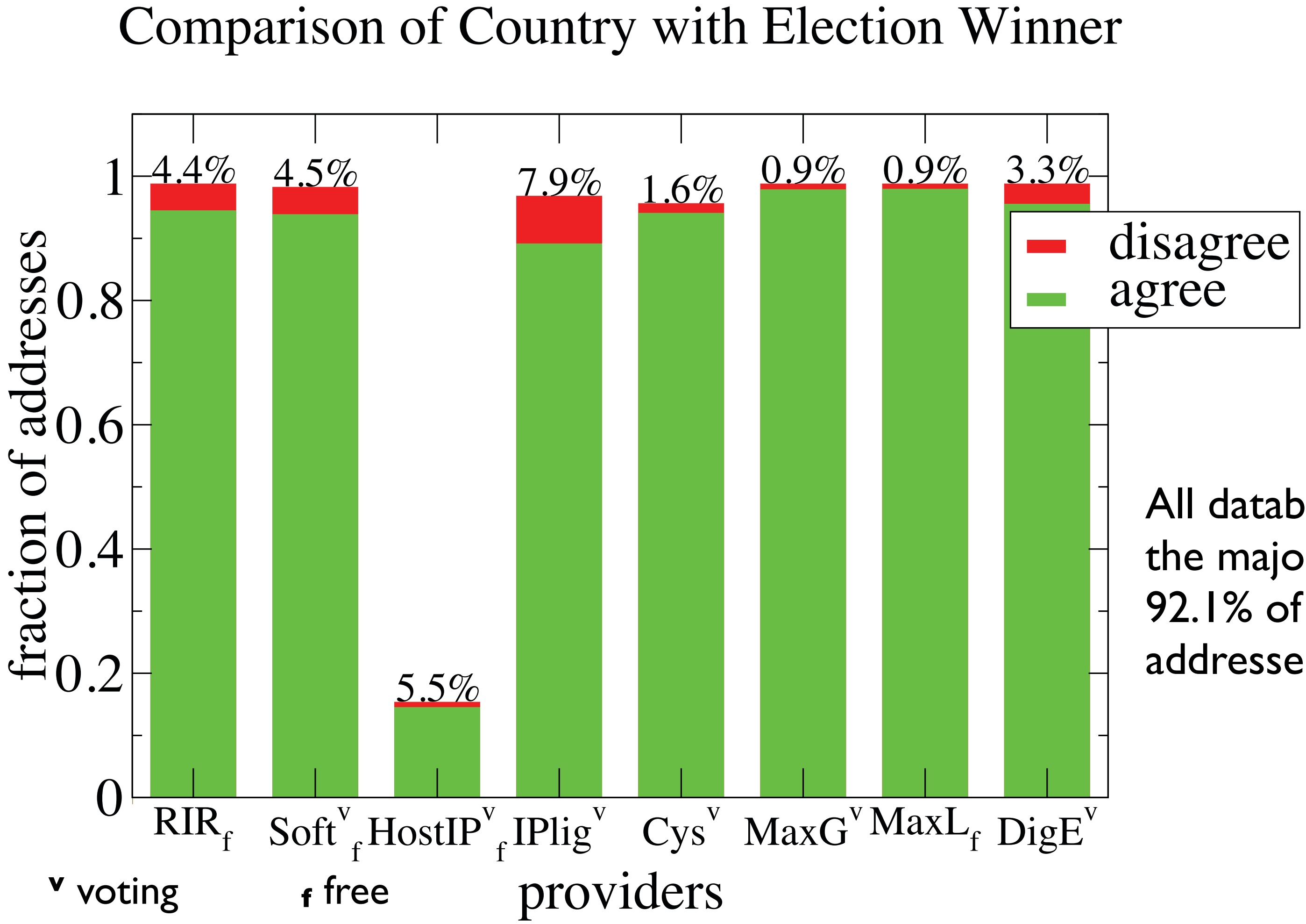
Election held for an IPv4 address if

- **tie**: top two countries have same vote count
- **winner found:** one country got more votes then any other country

Comparison to Winner

GEOLOCATION COMPARISON

country election



^{v} voting $_f$ free

All databases agree with the majority for at least 92.1% of RIR-delegated addresses.

Pairwise Comparison

GEOLOCATION COMPARISON

country election

The column reports the percentage of addresses for which the row database had an answer that matched the column's database.

	RIR	Soft	HostIP	IPlig	Cys	MaxG	MaxL	DigE	avg[1]
RIR$_f$	A -	99.9	88.9	87.2	93.6	94.1	94.2	91.8	93.3
Software77$_f^v$	99.4	-	88.8	86.6	93.0	93.5	93.6	91.2	91.1
HostIP$_f$ B	14.1	14.2	-	13.6	15.4	14.4	14.4	14.9	14.5
IPligence^v C	85.4	85.3	83.8	-	89.3	89.5	89.6	86.2	87.6
Cyscape^v	90.7	90.6	94.2	88.3	-	93.2	93.3	95.7	92.0
MaxMind GeoIP^v	94.1	94.0	90.9	91.4	96.2	D -	99.8	94.9	94.1
MaxMind Lite$_f$	94.2	94.1	91.0	91.5	96.3	99.8	-	94.9 E	95.3
Digital Envoy^v	91.8	91.7	93.9	87.9	98.8	94.9	94.9	-	93.3
average[1]	92.3	90.4	90.3	88.6	94.3	92.8	94.3	92.0	-

A. Software77 almost undistinguishable from RIR delegation file

B. HostIP has low agreement with other databases, because it lacks full coverage

C. IPligence has largest disagreement with other databases

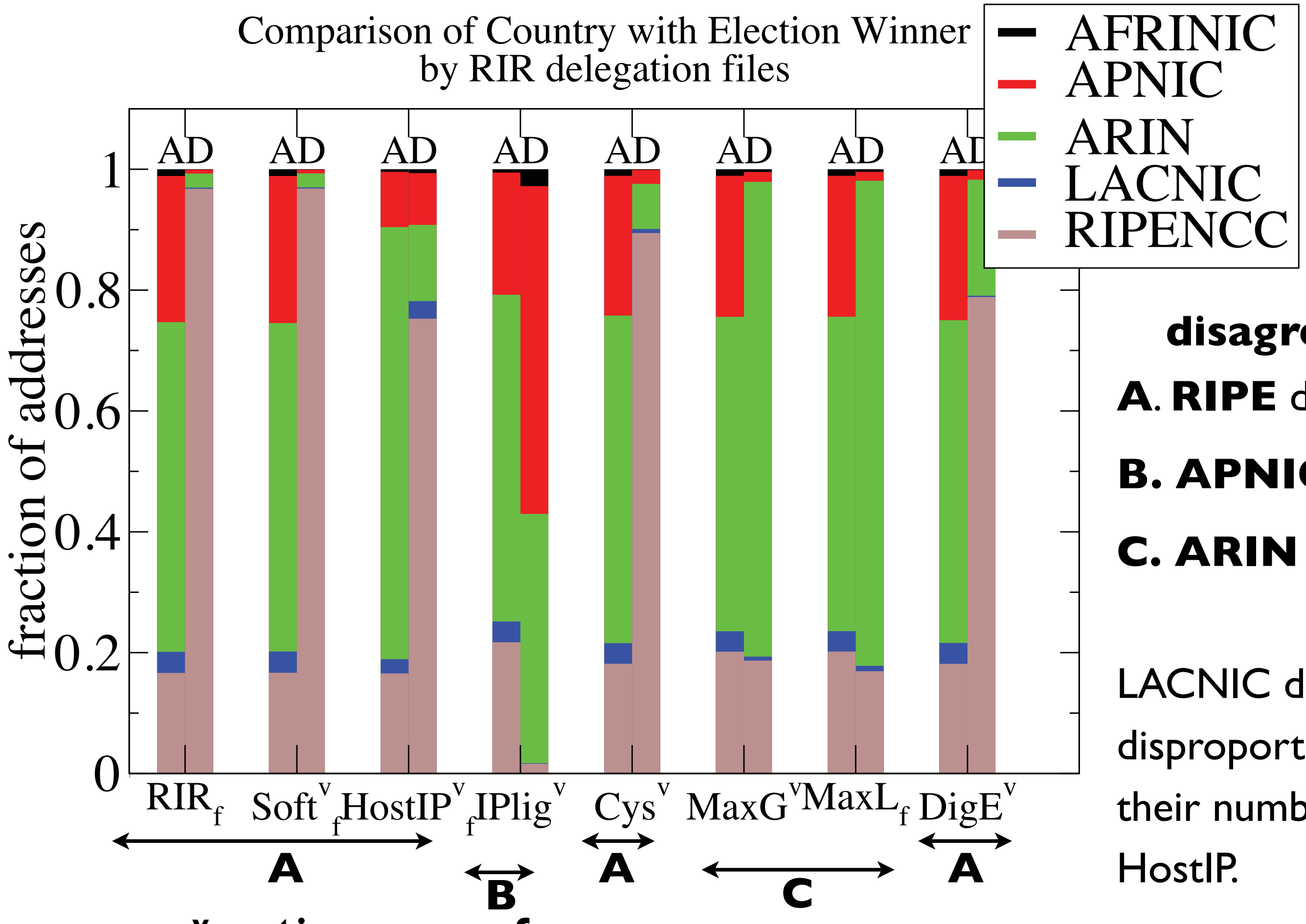
D. MaxMind Lite and MaxMind GeoIP agree on 99.8%

E. MaxMind Lite had greatest overall agreement with majority of databases: 95.4%

Regional Breakdown

GEOLOCATION COMPARISON

country election



v **voting**

f **free**

disagreement breakdown

A. RIPE dominates disagreements

B. APNIC dominates disagreements

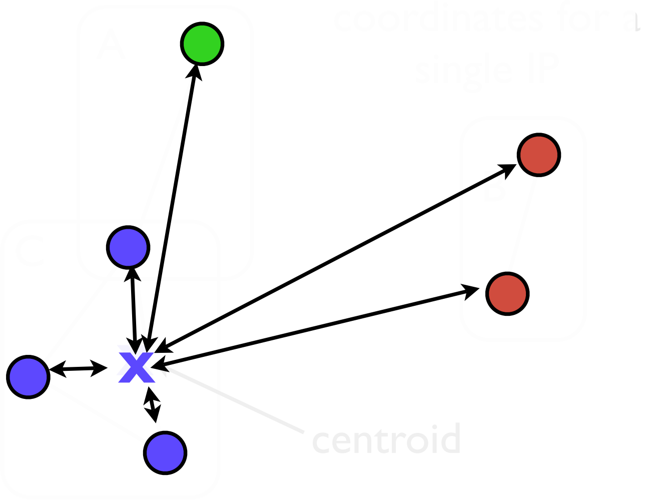
C. ARIN dominates disagreements

LACNIC disagreements are disproportionately few (compared to their number of addresses) except in HostIP.

Winning Cluster Centroid

coordinate election

- cluster coordinates into clicks, all coordinates within given threshold
- cluster gets one vote per member, members get multiple votes
- winning cluster has most votes: **C**
- calculate centroid of winning cluster

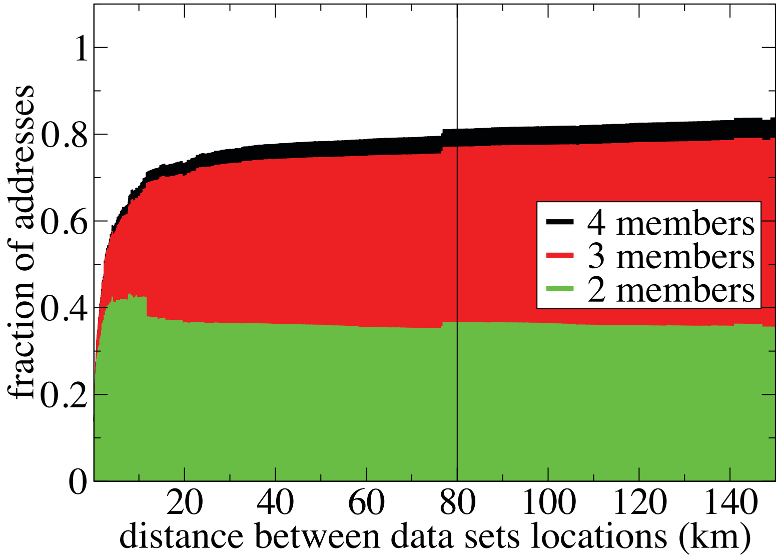


distance measured from winning cluster centroid

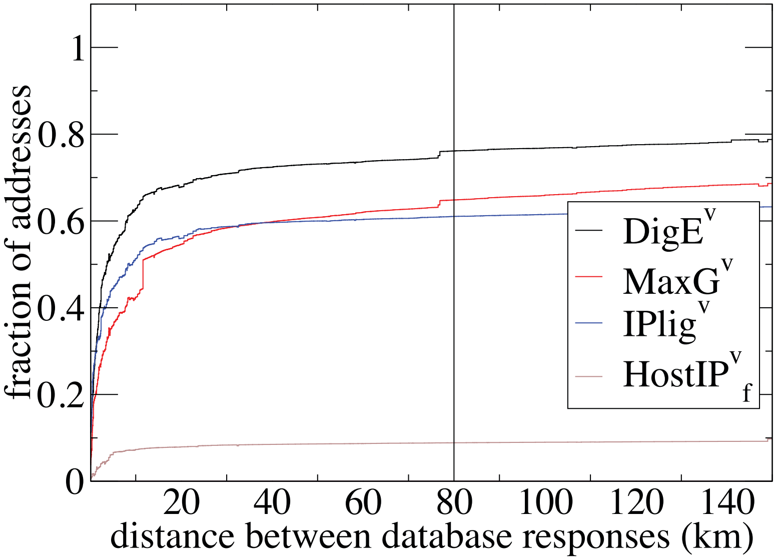
Effects of Thresholds

coordinate election

Number of Addresses in Winning Cluster by number of members in cluster



Number of Addresses where Database is Part of Winning Cluster



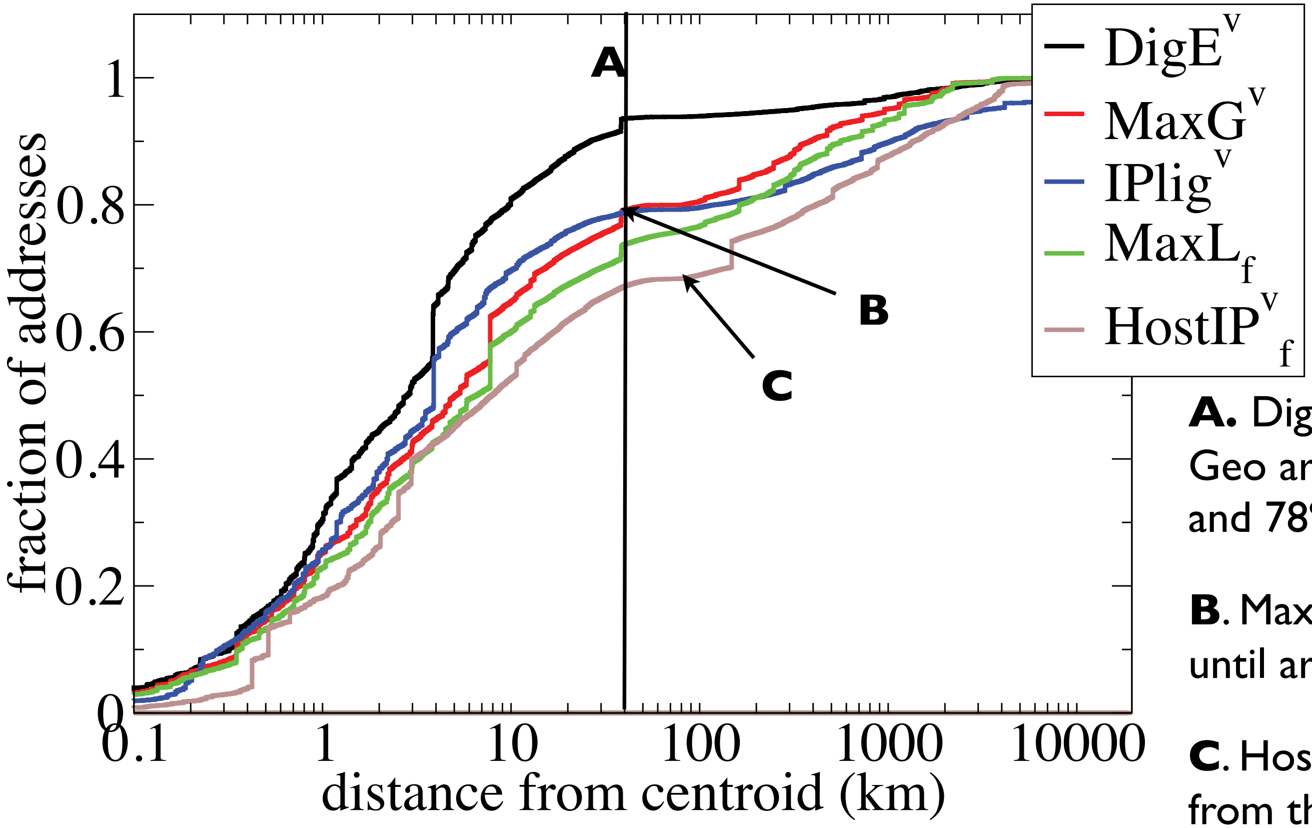
Typical city diameter suggests a threshold over 20km.
An 80 km threshold maximizes the number of two-member clusters

Centriod Distances

GEOLOCATION COMPARISON

coordinate election

Distance from Winning Cluster's Centroid



A. Digital Envoy and MaxMind Geo are within 40 km for 93% and 78% respectively

B. MaxMind Geo trails IPligence until around 33 km

C. HostIP is always furthest from the centroid caida

Pairwise Comparison

GEOLOCATION COMPARISON

coordinate election

distance between coordinates: 25 % / 50% / 75%

	HostIP	IPlig	MaxG	MaxL	DigE
HostIP$_f$	- A	3 / 134 / 1160	9 / 216 / 1140	10 / 248 / 1220	20 / 511 / 2360
IPligence	3 / 134 / 1160	-	4 / 85 / 722	4 / 88 / 722	1 / 9 / 721
MaxMind GeoIP	9 / 216 / 1140	4 / 85 / 722	B -	0 / 0 / 0	2 / 15 / 318
MaxMind Lite$_f$	10 / 248 / 1220	4 / 88 / 722	0 / 0 / 0	-	2 / 19 / 377
Digital Envoy	20 / 511 / 2360	1 / 9 / 721	2 / 15 / 318	2 / 19 / 377	C -

color key 0-49 50-149 150-449 450-1049 1050-

A. HostIP is furthest from all database

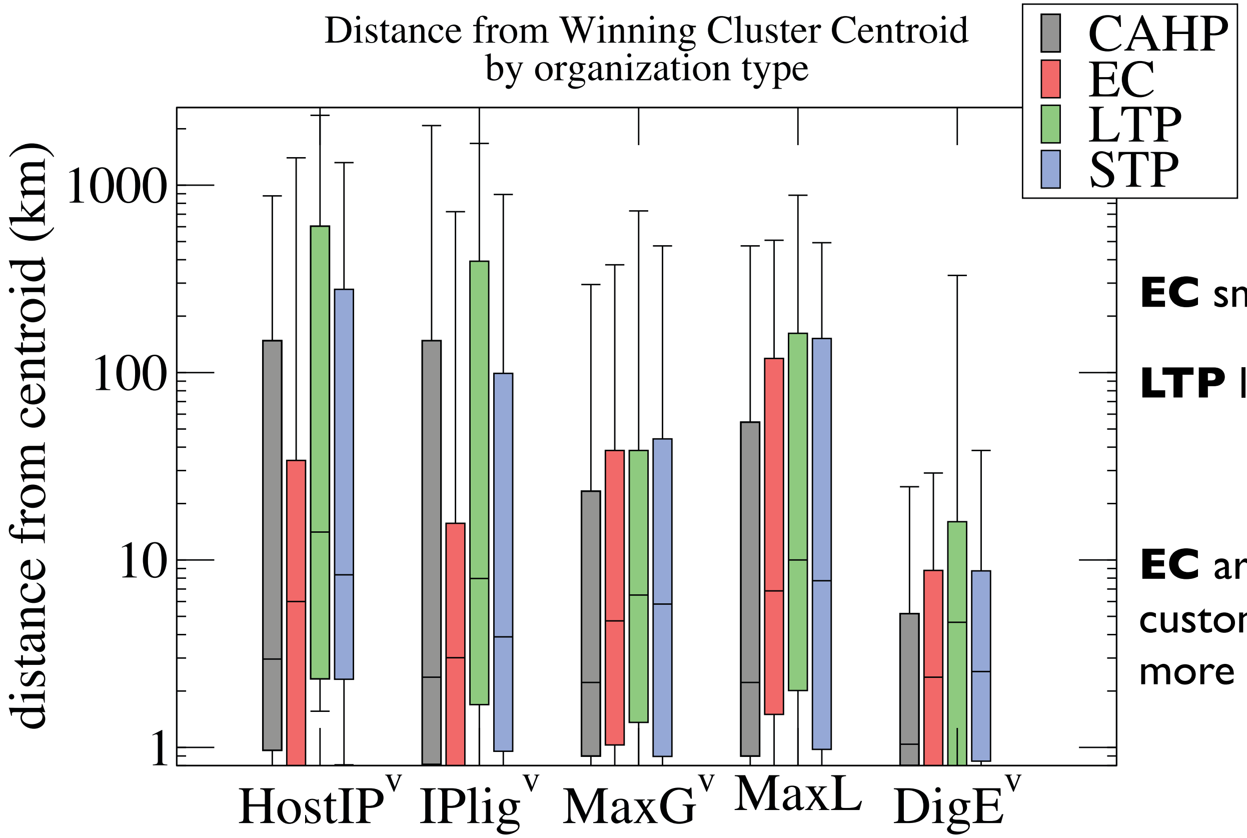
B. MaxMind Lite and MaxMind Geo are within 0 km for 75% of addresses.

C. For 50% of addresses, Digital Envoy is within 19 km of MaxMind Lite and 15 km of MaxMind Geo

Distance by Org

GEOLOCATION COMPARISON

coordinate election



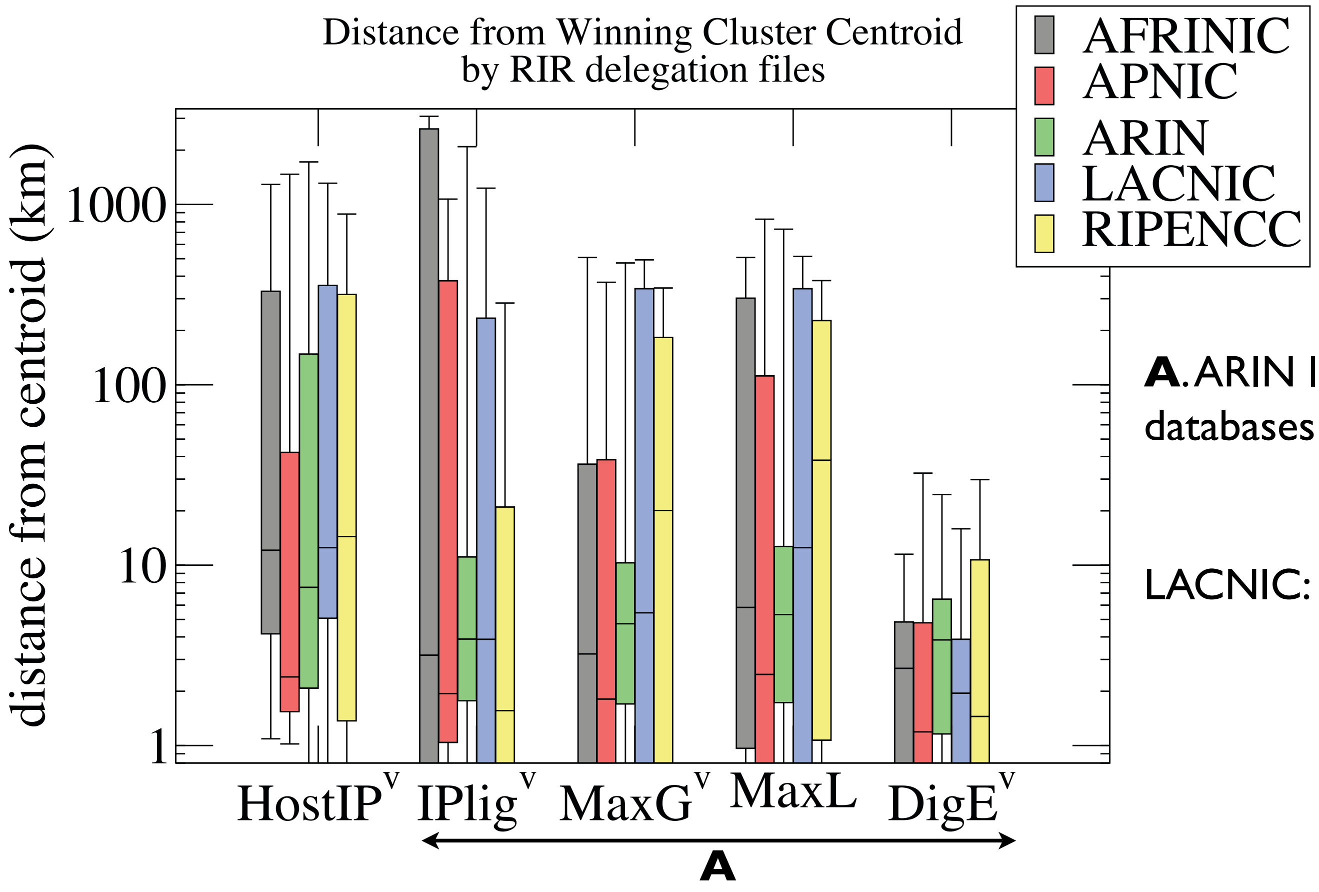
EC smallest variance

LTP largest variance

EC are edge organizations with lots of customer addresses, while **LTP** have more infrastructure addresses

Distance by Region

coordinate election



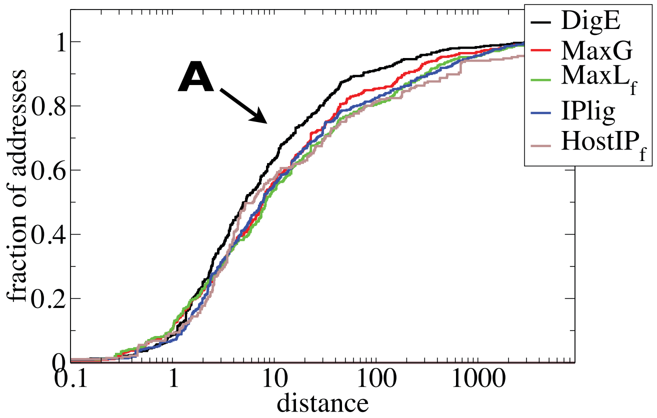
A. ARIN lowest variance across four databases (focused on US market?)

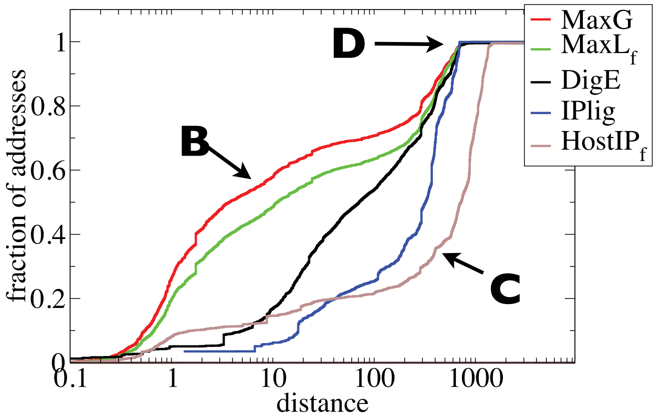
LACNIC: overall highest variance

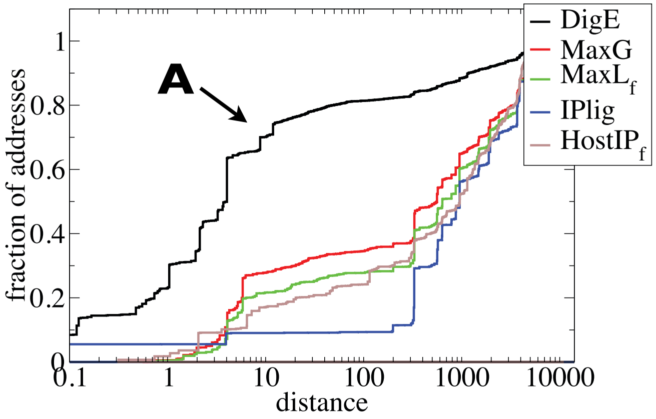
Distance to ground truth

GEOLOCATION COMPARISON

ground truth


Comparison with Planetlab's Groundtruth


Comparison with French Nets' Groundtruth


Comparison with Tier's Groundtruth

A. Digital Envoy has shortest median distance from truth for PlanetLab and Tier 1 IPs

B. MaxMind has shortest median distance from truth for French Network IPs

C. HostIP geolocated most of the French Network IPs in Germany

D. 717km approximate radius of France

Conclusion

- **country election**
  - Databases agree with the majority for 92%~99% of RIR-delegated IPv4 addresses
  - Databases agreed with the majority more then they did in pairwise comparison
  - For many databases, RIPE NCC's address were the source of most disagreements
- **coordinate election**
  - Digital Envoy and MaxMind Geo are within 40 km for 93% and 78% of addresses
- **ground truth**
  - Digital Envoy had shortest median distance to the Tier 1 and Planet Lab IPs
  - MaxMind Geo had shortest median distance to the French Network IPs.