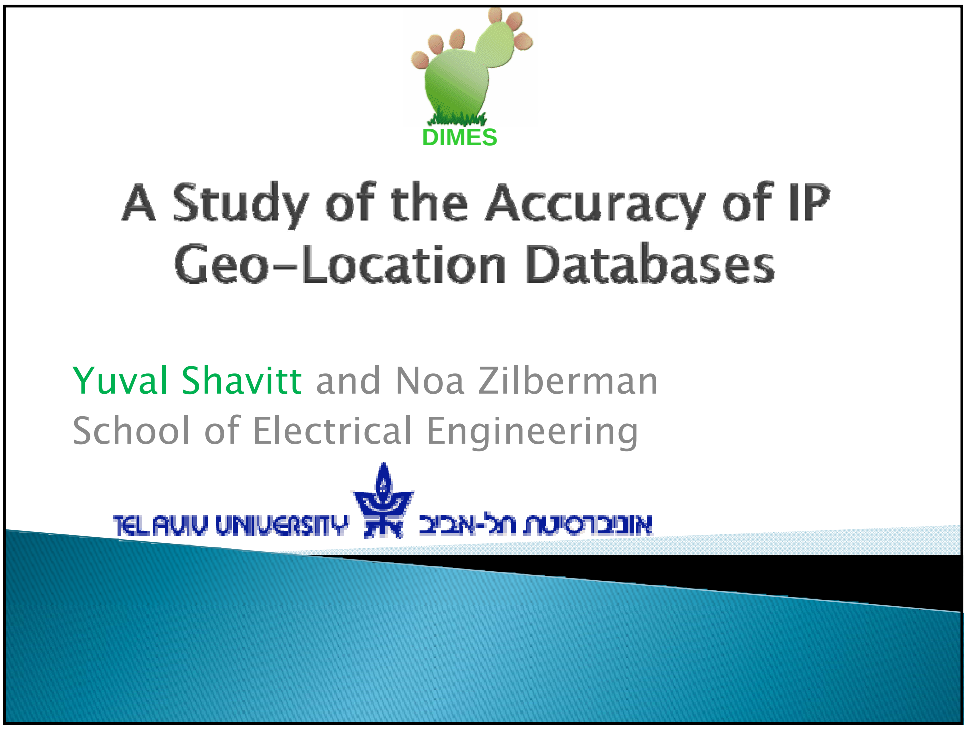
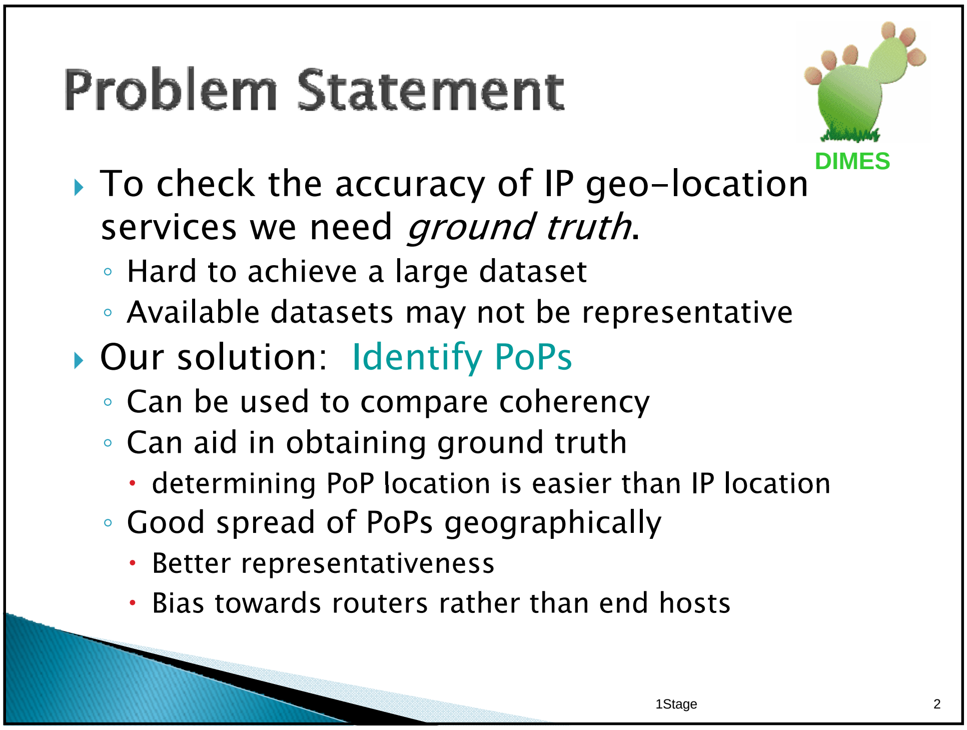


DIMES

A Study of the Accuracy of IP Geo-Location Databases

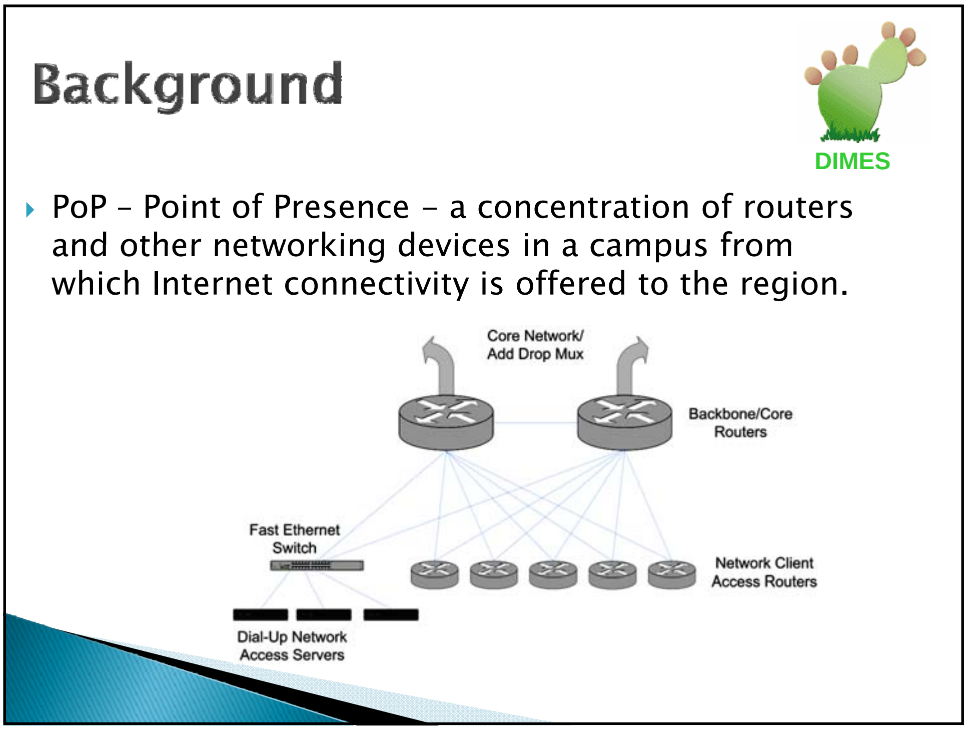
Yuval Shavitt and Noa Zilberman
School of Electrical Engineering
TEL AVIV UNIVERSITY אוניברסיטת תל-אביב



Problem Statement

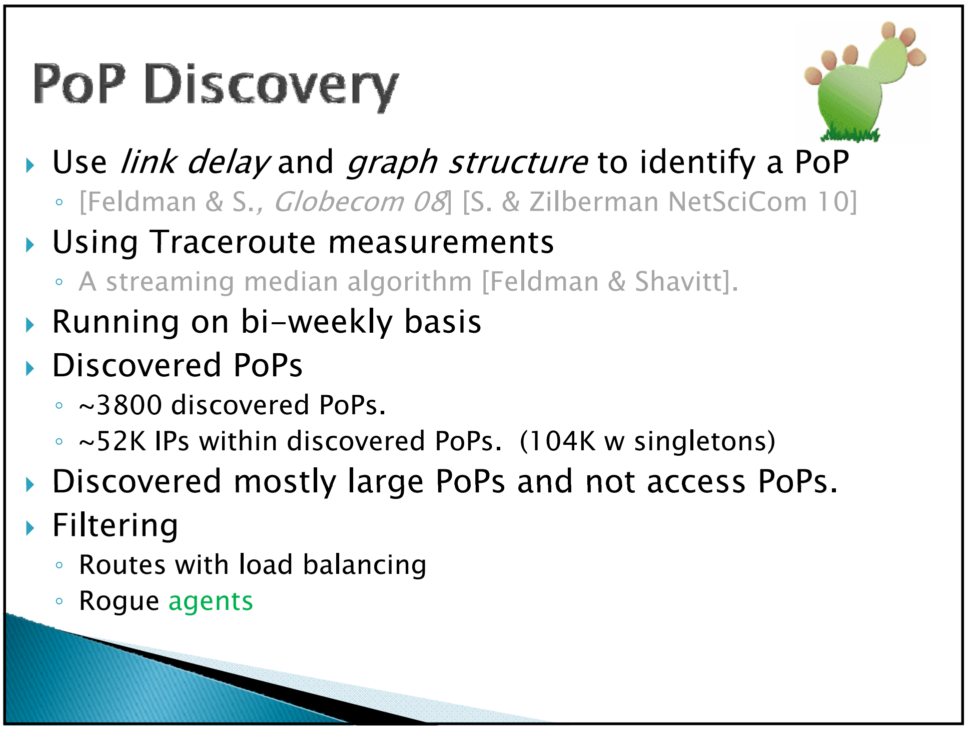
DIMES

- To check the accuracy of IP geo-location services we need *ground truth*.
  - Hard to achieve a large dataset
  - Available datasets may not be representative
- Our solution: Identify PoPs
  - Can be used to compare coherency
  - Can aid in obtaining ground truth
    - determining PoP location is easier than IP location
  - Good spread of PoPs geographically
    - Better representativeness
    - Bias towards routers rather than end hosts



Background

- PoP – Point of Presence – a concentration of routers and other networking devices in a campus from which Internet connectivity is offered to the region.

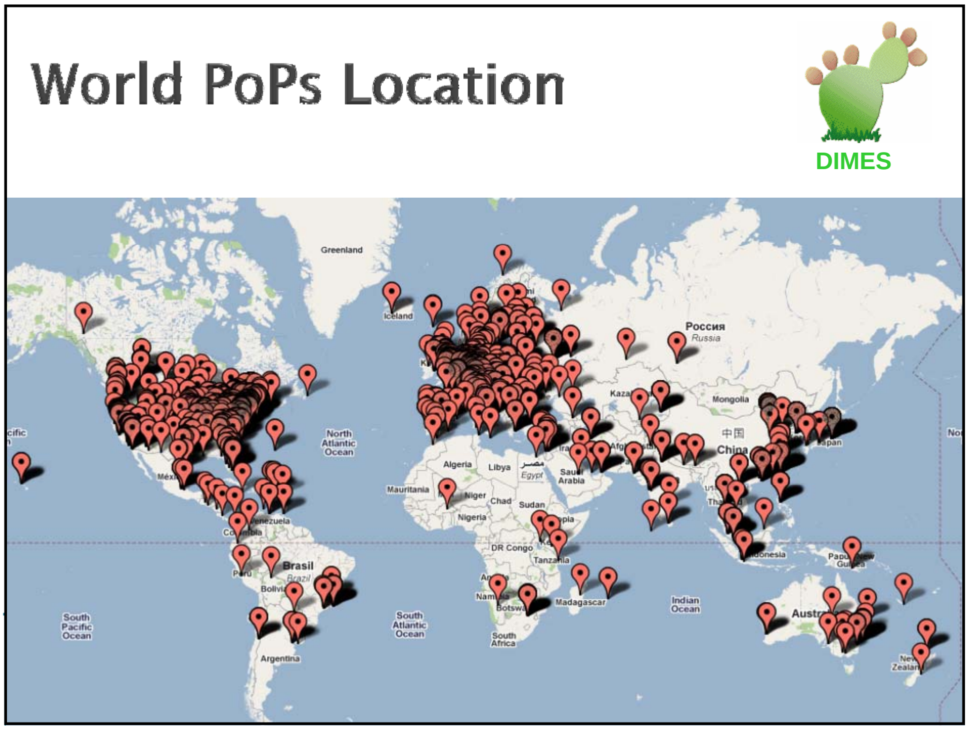


PoP Discovery

- Use *link delay* and *graph structure* to identify a PoP
  - [Feldman & S., *Globecom 08*] [S. & Zilberman NetSciCom 10]
- Using Traceroute measurements
  - A streaming median algorithm [Feldman & Shavitt].
- Running on bi-weekly basis
- Discovered PoPs
  - ~3800 discovered PoPs.
  - ~52K IPs within discovered PoPs. (104K w singletons)
- Discovered mostly large PoPs and not access PoPs.
- Filtering
  - Routes with load balancing
  - Rogue agents

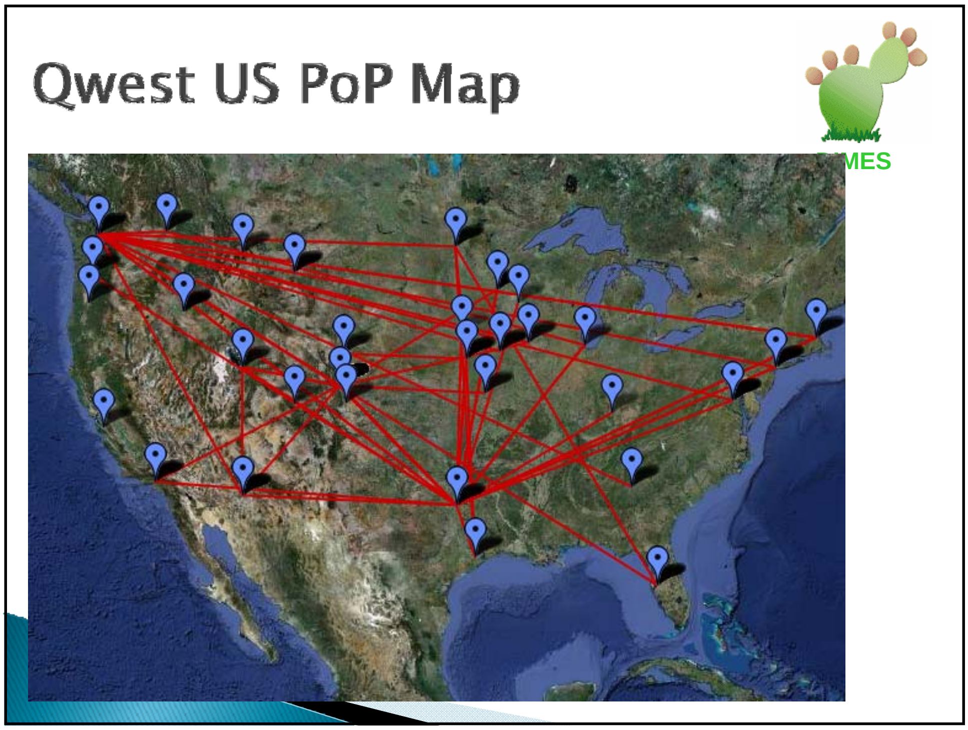
World PoPs Location

DIMES



Qwest US PoP Map

DIMES



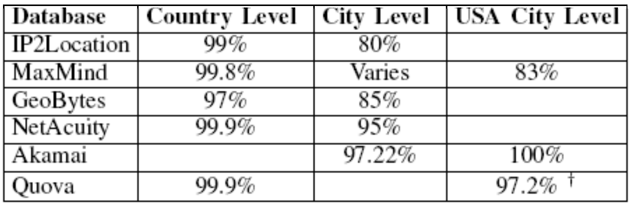
Evaluation of Geolocation Databases

DIMES

- Seven databases were used for the evaluation.
  - NetAcuity (Digital Element) – High end
  - GeoBytes
  - GeoIP (MaxMind)
  - IPligence Max
  - IP2Location
  - HostIP.info – Free service
  - Spotter – Research tool
- Dataset: DIMES measurements, March 2010
  - 52K IP addresses (+ 52K singletons IP addresses)
  - 3800 PoPs

Vendor Reported Accuracy

DIMES



Database	Country Level	City Level	USA City Level
IP2Location	99%	80%	
MaxMind	99.8%	Varies	83%
GeoBytes	97%	85%	
NetAcuity	99.9%	95%	
Akamai		97.22%	100%
Quova	99.9%		97.2% †

TABLE I
GEOLOCATION DATABASE ACCURACY AS REPORTED BY VENDOR

†US state accuracy

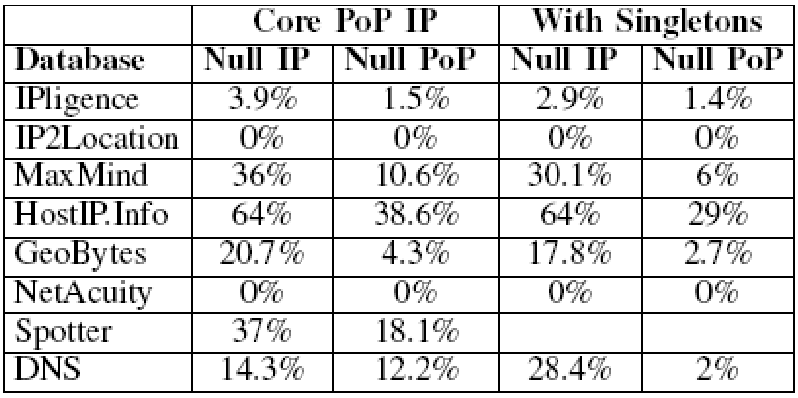
Evaluation methods

DIMES

- Null Replies
- Agreement within a database – coherency
- "Ground Truth" location
- Comparison Between databases
  - Similarity
  - By majority Vote
- Database anomalies

Null Replies

DIMES



	Core PoP IP		With Singletons	
Database	Null IP	Null PoP	Null IP	Null PoP
IPligence	3.9%	1.5%	2.9%	1.4%
IP2Location	0%	0%	0%	0%
MaxMind	36%	10.6%	30.1%	6%
HostIP.Info	64%	38.6%	64%	29%
GeoBytes	20.7%	4.3%	17.8%	2.7%
NetAcuity	0%	0%	0%	0%
Spotter	37%	18.1%		
DNS	14.3%	12.2%	28.4%	2%

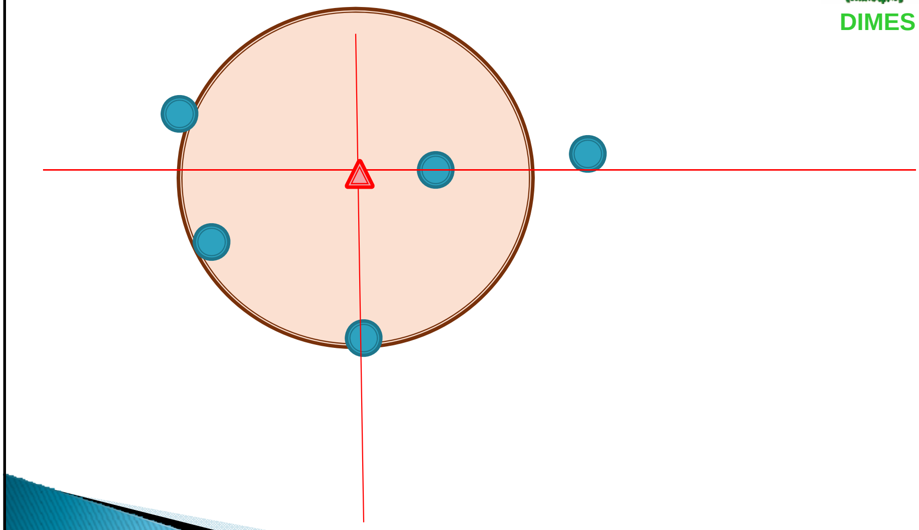
PoP Location

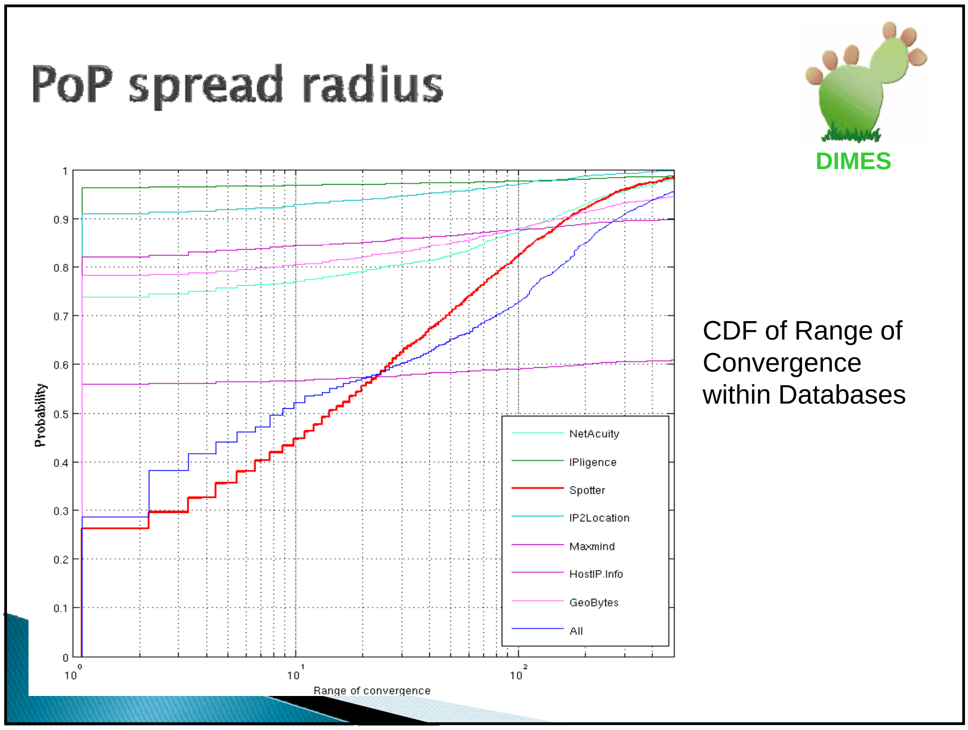
DIMES

- For each IP in the PoP (*N* IPs), each database (*M*) get a vote on the geo-location
  - Number of votes *N*•*M*
- Using the votes we define the PoP *location* and *convergence radius*

PoP Location and ‘Convergence’

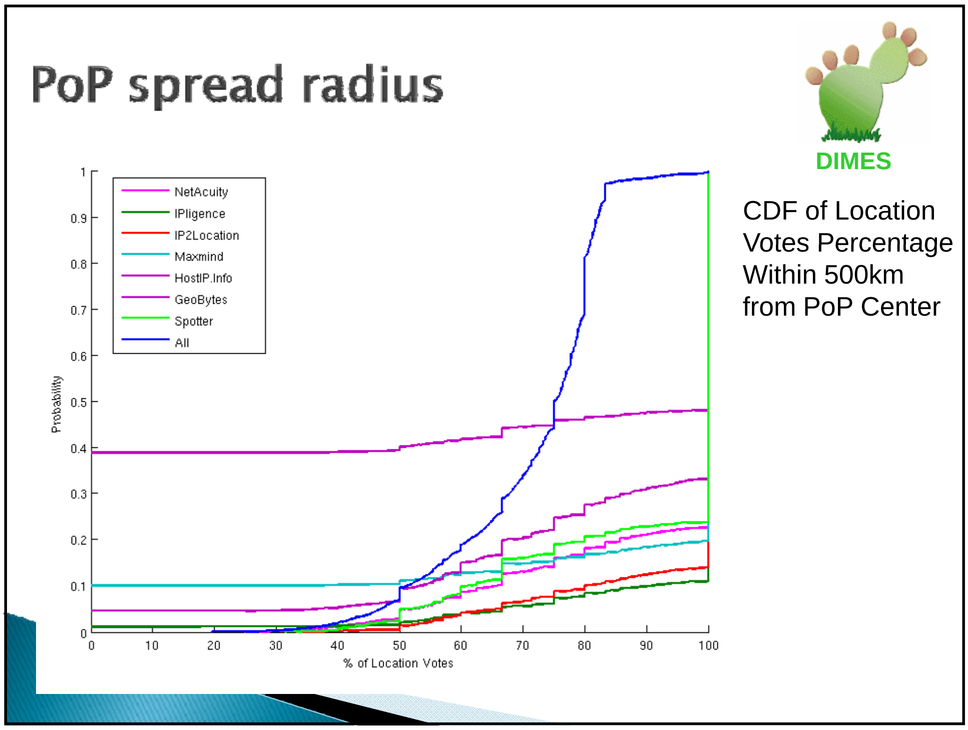
DIMES





PoP spread radius

CDF of Range of Convergence within Databases



PoP spread radius

CDF of Location Votes Percentage Within 500km from PoP Center

Ground Truth evaluation

DIMES

- Using CAIDA's 25K "Ground Truth" IP addresses
  - January-2010 database, based on DNS & ISP collaboration
  - In the results, city range considered at 100km range

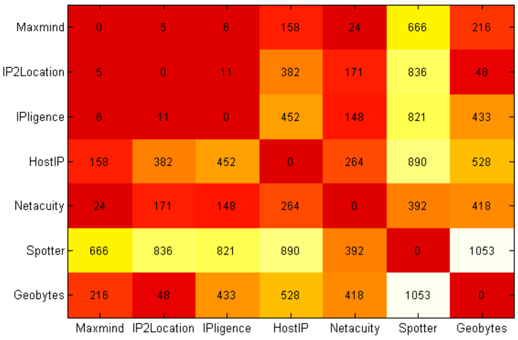
Database	IP hits	Country Match	City Match
Geobytes	67.3%	80.1%	26.5%
HostIP.Info	28.1%	89.0%	17.9%
IP2Location	100%	76.0%	13.3%
IPligence	100%	76%	0.7%
Netacuity	67.9%	96.9%	79.1
Spotter	54.1%	---	27.8

10.1K wrongly located in Washington DC

20.5K wrongly located in Washington DC

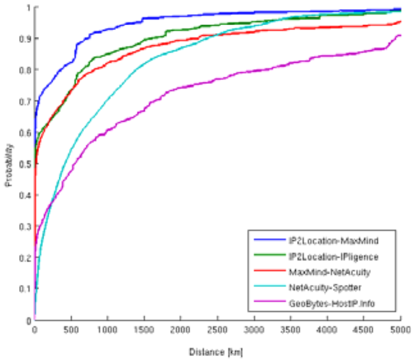
Correlation among Databases

DIMES

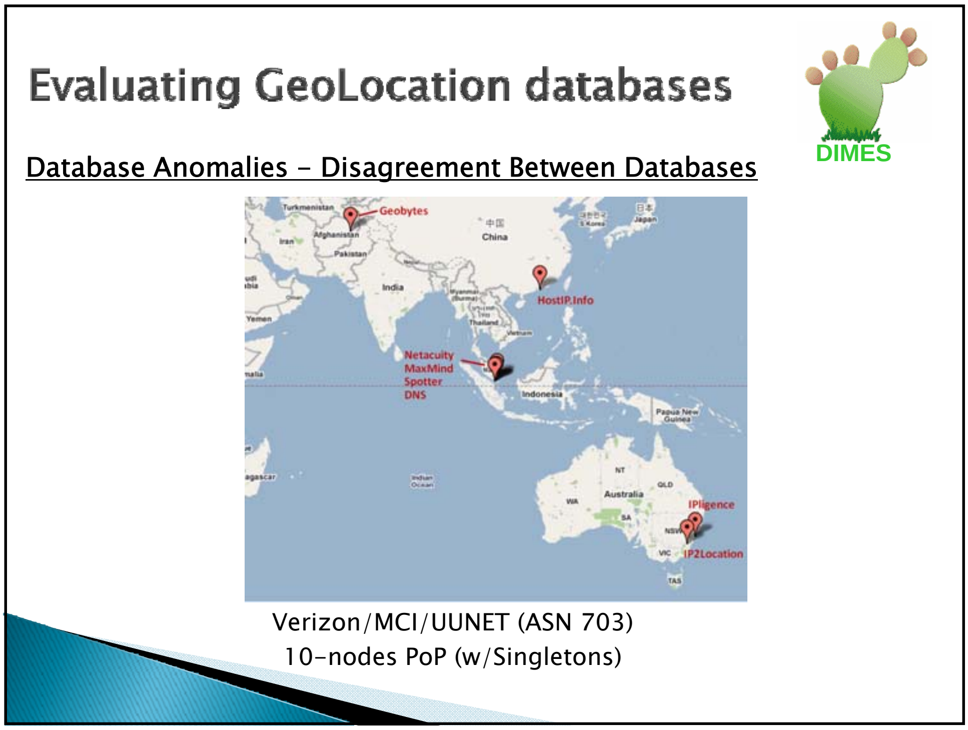


	Maxmind	IP2Location	IPligence	HostIP	Netacuity	Spotter	Geobytes
Maxmind	0	5	6	158	24	666	216
IP2Location	5	0	11	382	171	836	48
IPligence	6	11	0	452	148	821	433
HostIP	158	382	452	0	264	890	528
Netacuity	24	171	148	264	0	392	418
Spotter	666	836	821	890	392	0	1053
Geobytes	216	48	433	528	418	1053	0

Heatmap – Median distance between databases



CDF- distances between databases

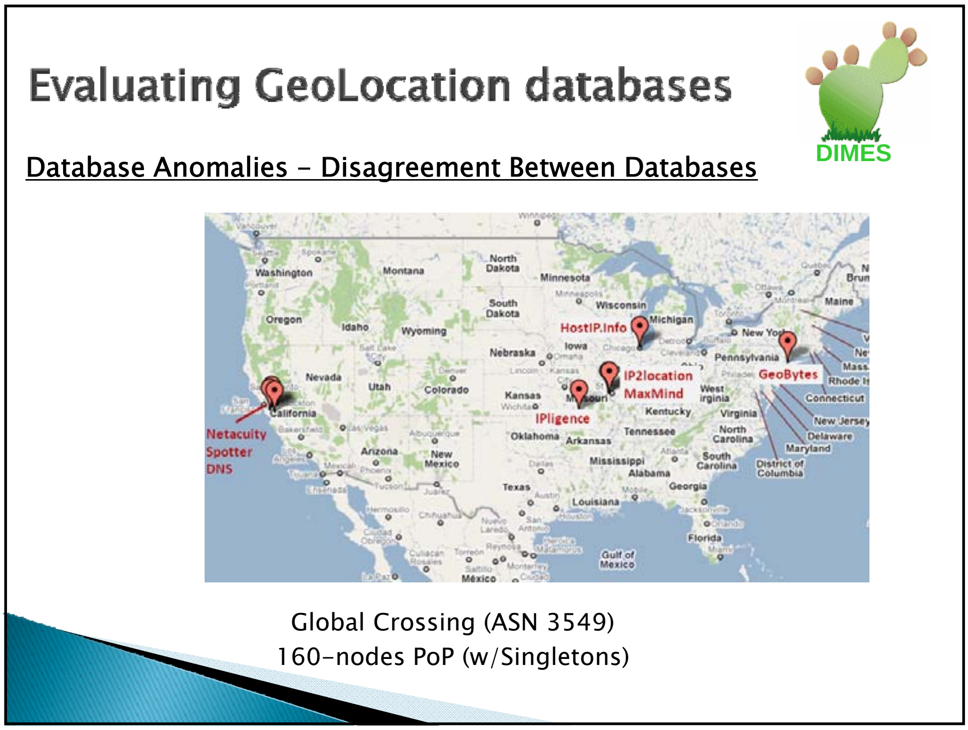


Evaluating GeoLocation databases

DIMES

Database Anomalies – Disagreement Between Databases

Verizon/MCI/UUNET (ASN 703)
10-nodes PoP (w/Singletons)



Evaluating GeoLocation databases

DIMES

Database Anomalies – Disagreement Between Databases

Global Crossing (ASN 3549)
160-nodes PoP (w/Singletons)

Database Anomalies – False Location Replies

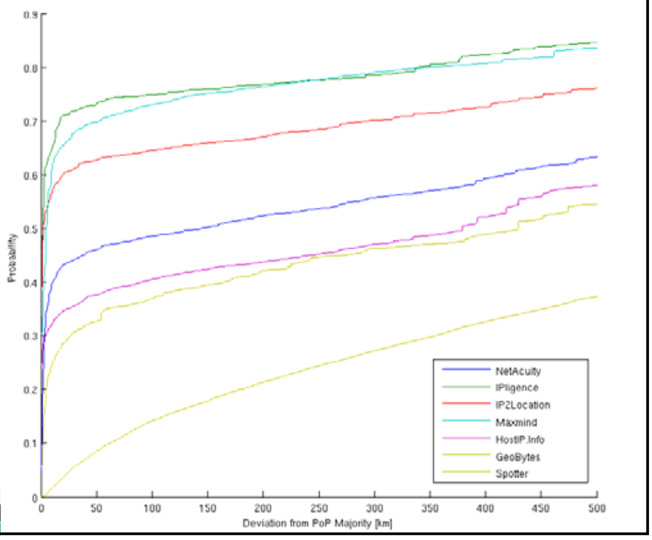
Qwest as an example

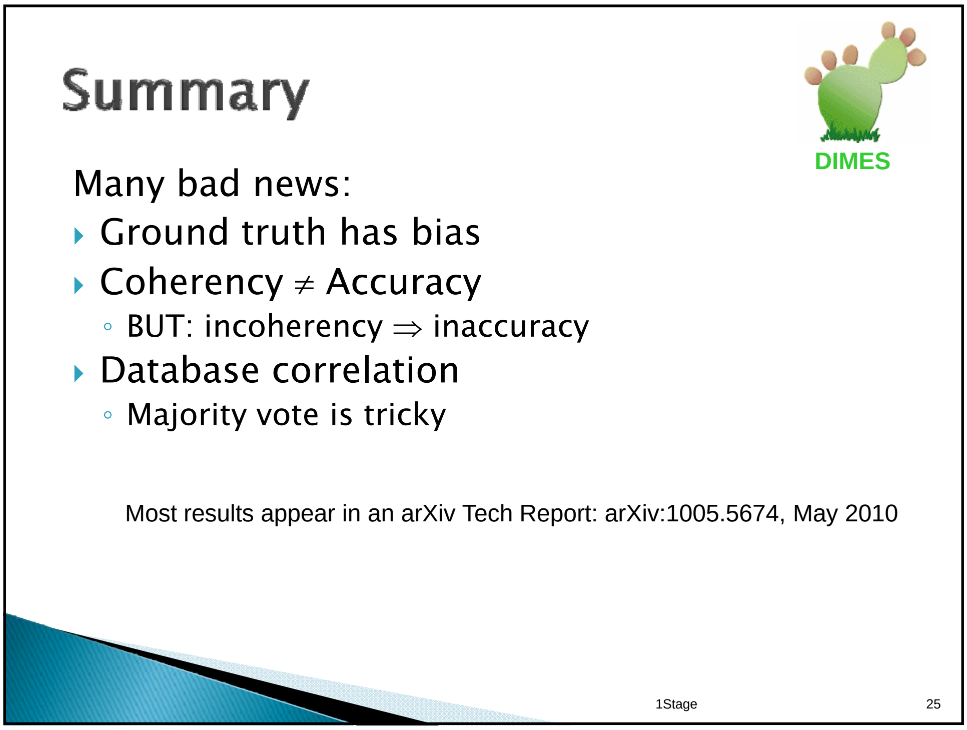
- 70 PoPs were discovered by the algorithm
- MaxMind assigned the PoPs to 55 different locations
- HostIP.Info assigned the PoPs to 46 different locations
- IP2Location assigned the PoPs to 35 different locations
- IPligence located the PoPs in only one distinct location;
  - All the PoPs were placed in Denver, where Qwest HQ are located.
  - Out of 20291 Qwest entries in IPligence, 20252 are located in Denver.
- MaxMind had the same problem as IPligence in their May-2009 DB, but it was fixed in July-2009 DB.

Agreement Between Databases – By Majority Vote

CDF of Database Location Deviation From PoP Median.

Long tail.





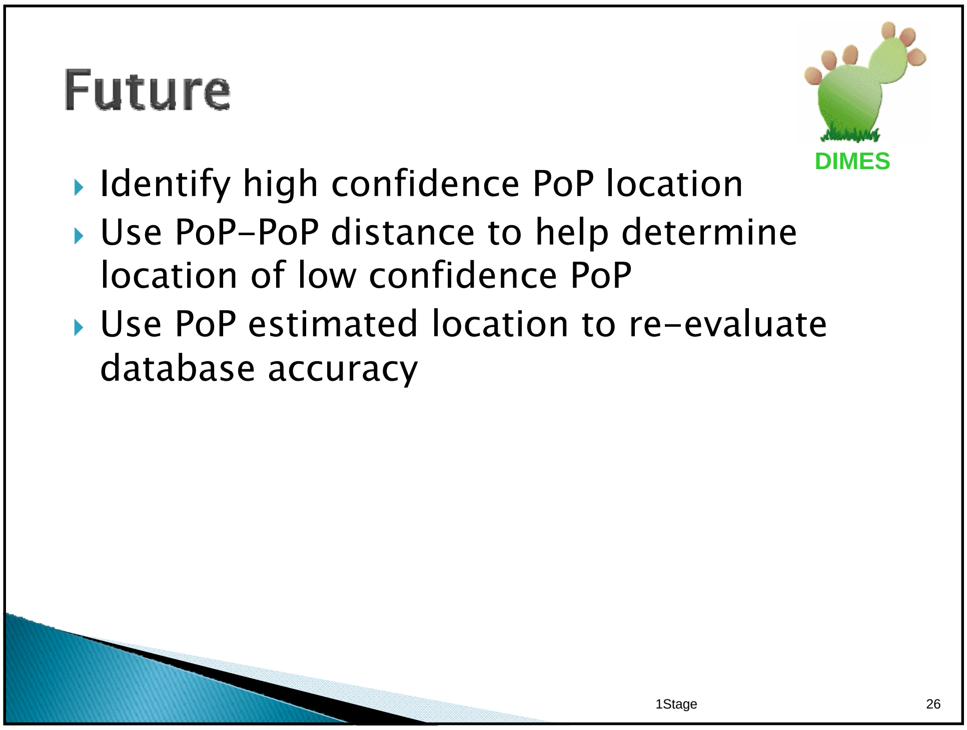
Summary

DIMES

Many bad news:

- Ground truth has bias
- Coherency ≠ Accuracy
  - BUT: incoherency ⇒ inaccuracy
- Database correlation
  - Majority vote is tricky

Most results appear in an arXiv Tech Report: arXiv:1005.5674, May 2010



Future

DIMES

- Identify high confidence PoP location
- Use PoP-PoP distance to help determine location of low confidence PoP
- Use PoP estimated location to re-evaluate database accuracy