IP GEOLOCATION

Using Traceroute Location Propagation and IP Range Location Interpolation

Ovidiu Dan, Vaibhav Parikh, Brian. D. Davison

Bing LEHIGH UNIVERSITY

Presentation Outline

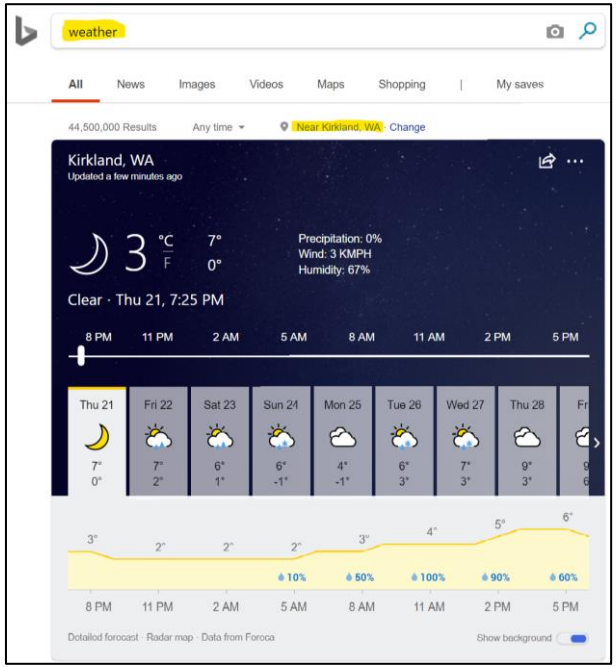
- What is IP Geolocation
- Commercial Geolocation Services
- Motivation and Goal
- Privacy and Datasets
- IP Colocation
- IP Interpolation
- Traceroute Location Propagation
- Related Work (Appendix)

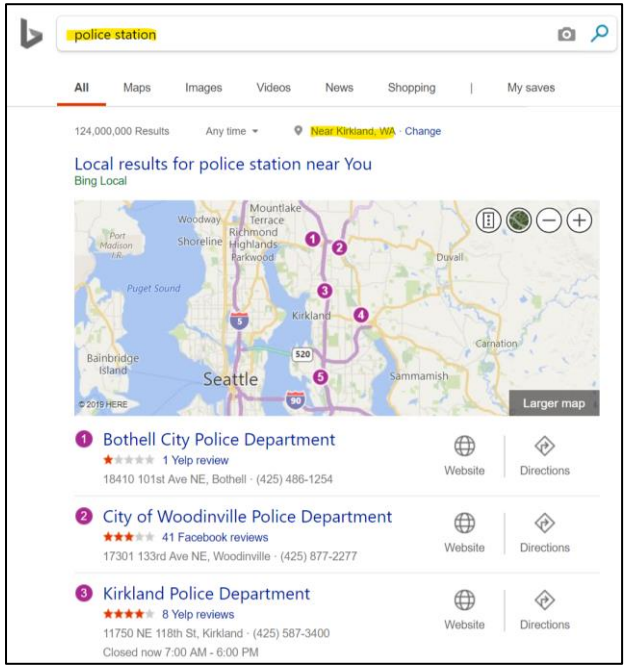
1.

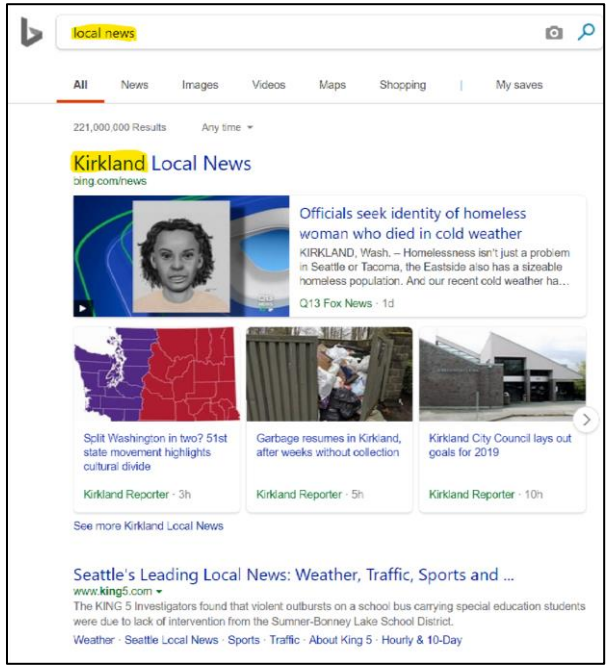
WHAT IS IP GEOLOCATION

Background, Motivation, Research Questions, Contributions

IP Geolocation Examples

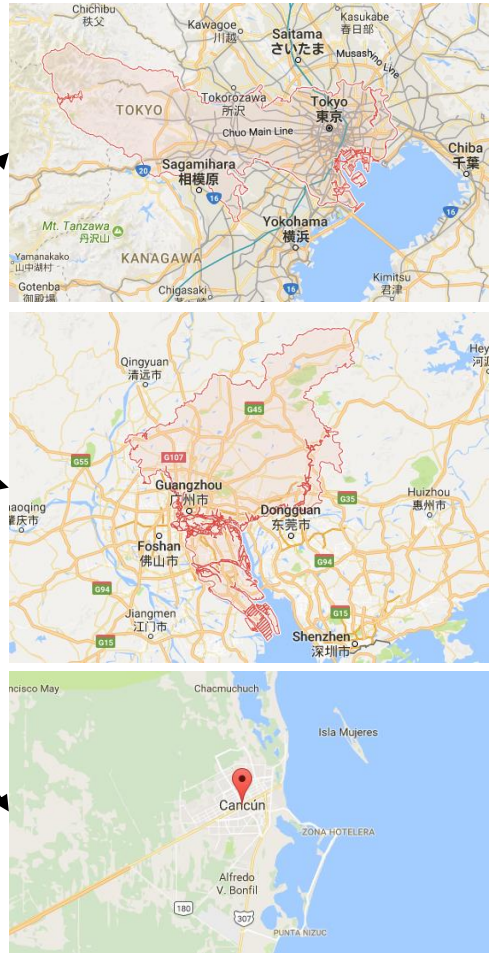






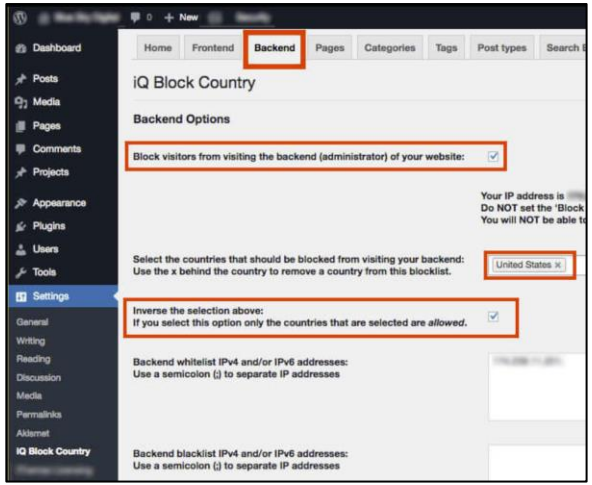
Geolocation Databases

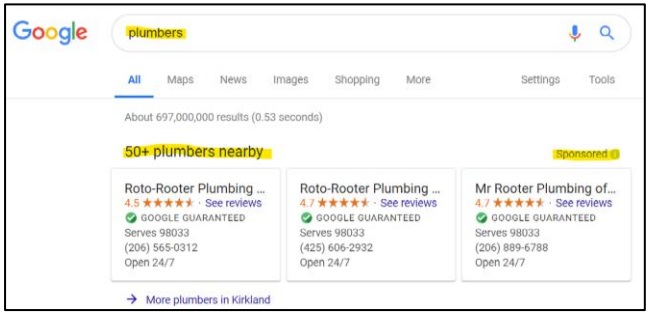
StartIP	EndIP	Country	State	City
1.0.16.0	1.0.16.255	JP	Tokyo	Tokyo
1.0.48.0	1.0.48.255	CN	Guangdong	Guangzhou
187.153.184.0	187.153.184.255	MX	Quintana roo	Cancun

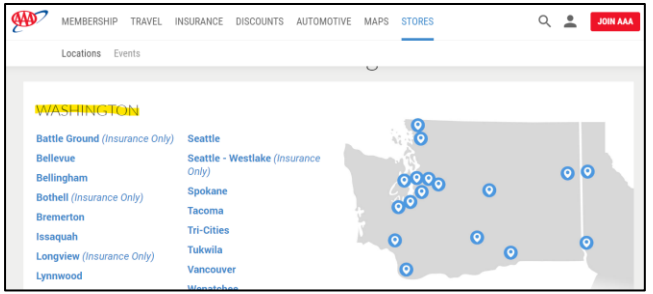


IP Geolocation Use Cases

- Search Engine Personalization
- Content Delivery Networks
- Credit card fraud protection
- Click fraud protection
- Advertising network targeting
- Cybercrime investigations
- Location based content licensing
- E-commerce
- Organizations with regional offices
- Automatic Language Selection
- Network Security
- Academic Research
  - Personalization
  - Cohort modeling







2.

COMMERCIAL GEOLOCATION

Current commercial location services and their accuracy

Commercial Geolocation Services

- **Akamai EdgeScape** (since 2000)
- **Neustar IP Intelligence** (since 2000)
- **IP2Location** (since 2001)
- **IPligence** (since 2006)
- **MaxMind** (since 2002)
- **Digital Element's NetAcuity** (since 1999)

Data Sources

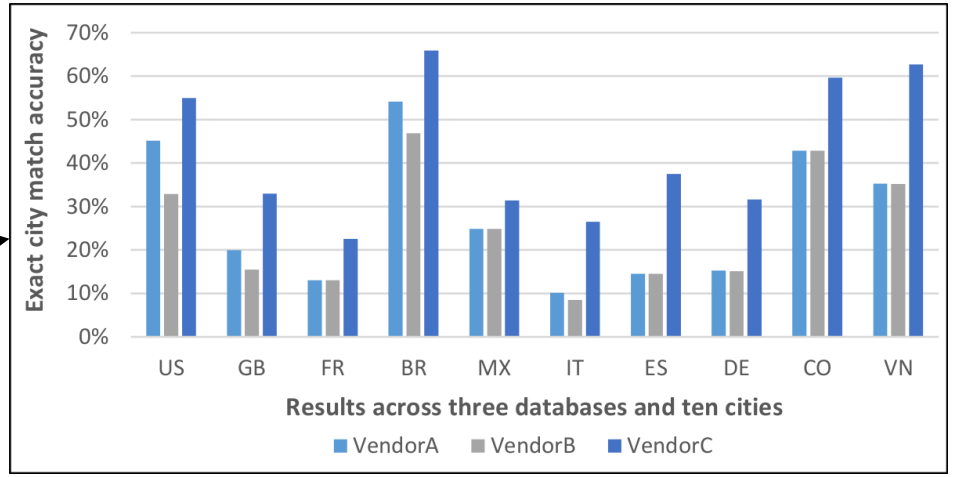
- **Data and techniques are proprietary**
  - Direct partnerships with ISPs
  - Mobile phone traffic
  - Data from weather, news apps
  - Crowdsourcing
  - Network delay (ping)
  - Network topology (traceroute)
  - WHOIS databases
  - Reverse DNS hostnames

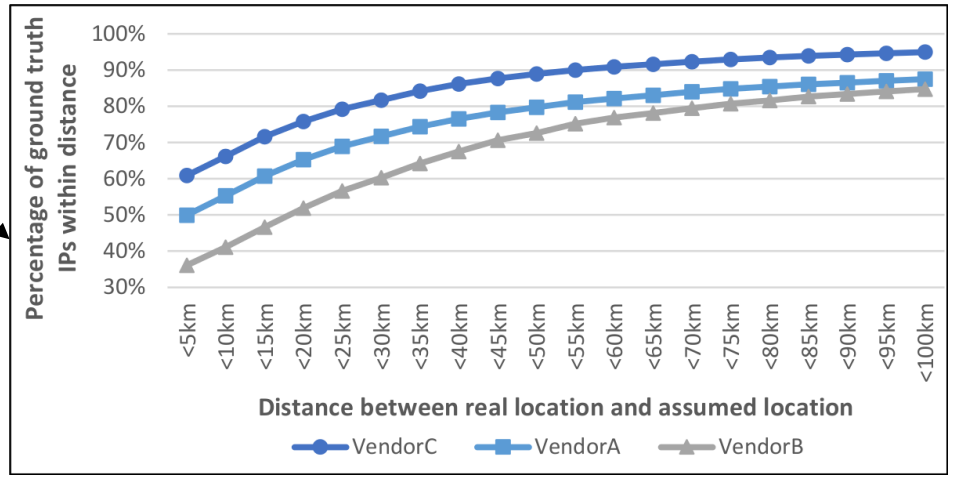
Self-reported accuracy

- MaxMind and IP2Location reveal city-level accuracy
- Numbers self-reported, not audited
- **Maxmind**
  - Mean unweighted accuracy for 100 countries is 76.5% at 50 km
  - USA: accuracy of 86% at 50 km
- **IP2Location**
  - Mean unweighted accuracy 77% for error smaller than 50 kilometers
  - USA: accuracy of 99.52% at 50 km

Accuracy of commercial geolocation services

- Ground truth: 8.4 million IPs
- Three commercial services
- Exact city accuracy
- Error distance in USA





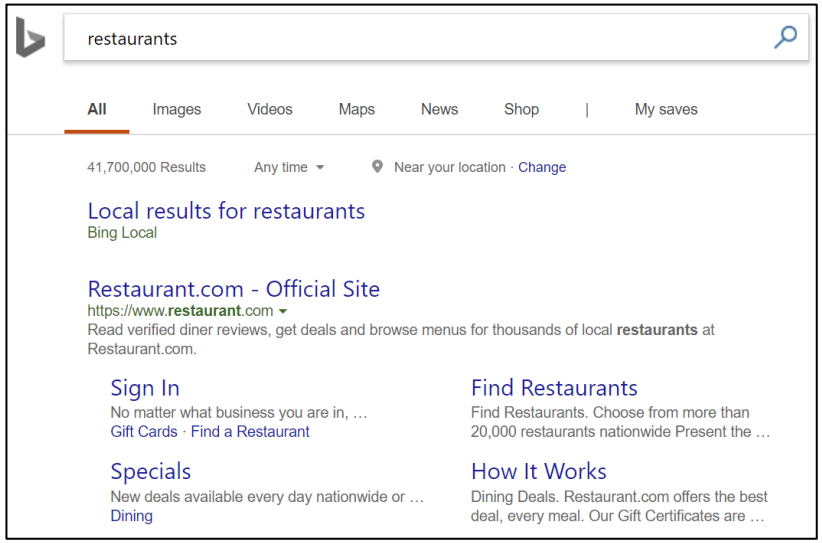
3.

MOTIVATION

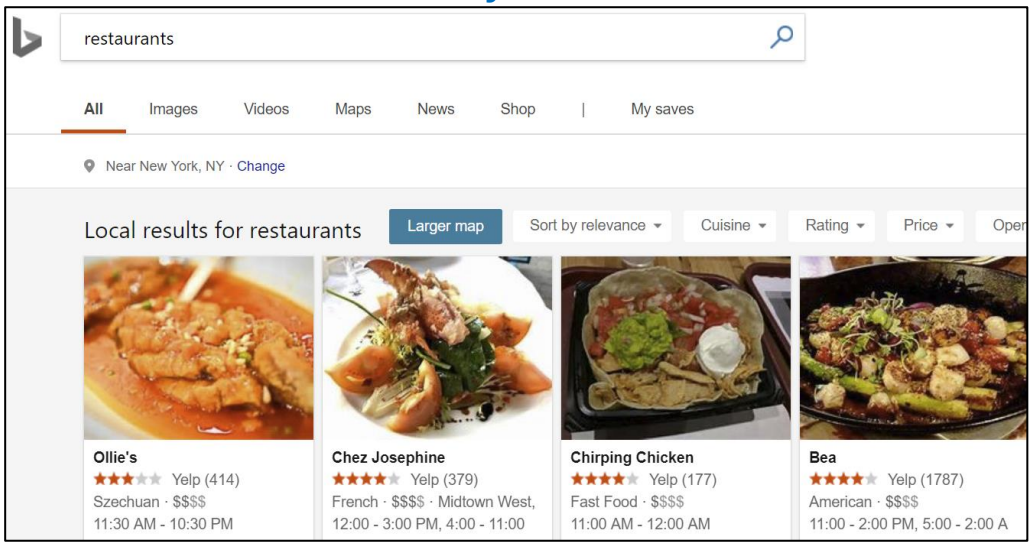
Impact of inaccurate IP geolocation in the context of search engines

Example of Missing IP Geolocation

User Location: **Unknown**

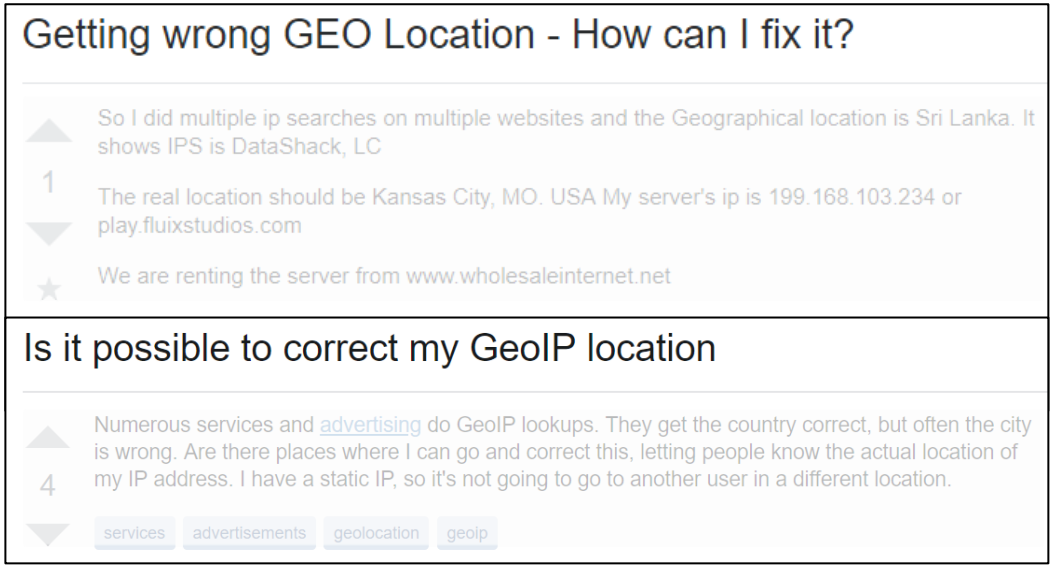


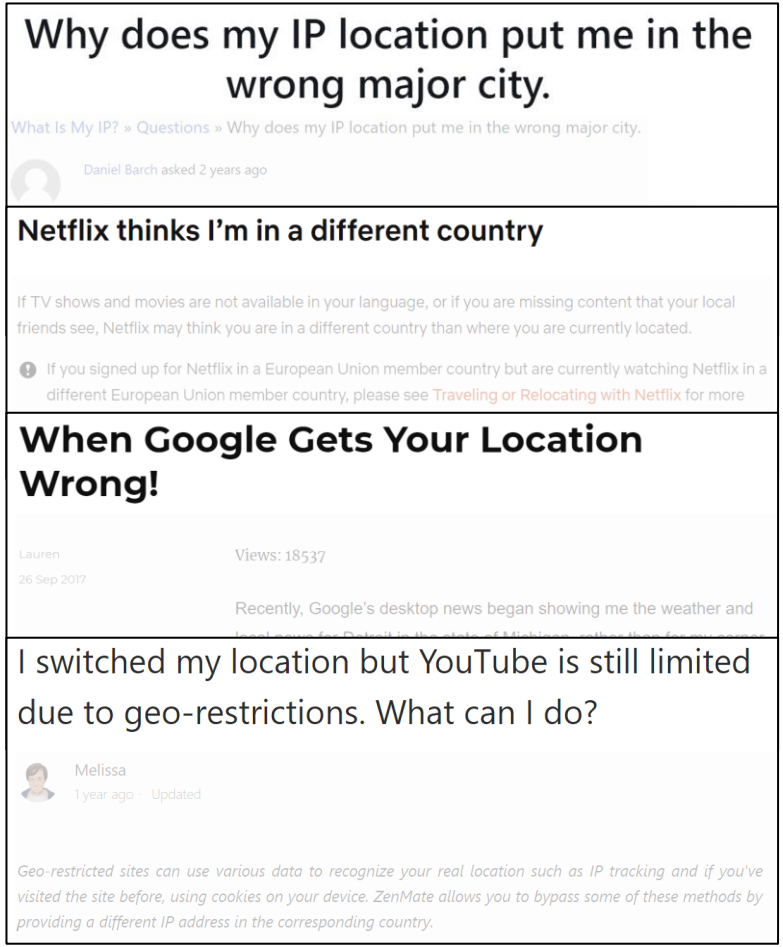
User Location: **New York City**



Motivation

- IP Geolocation is not a solved problem
- Commercial geolocation services are proprietary black boxes
- Previous research lacks accuracy and scale





Impact of Incorrect IP Geolocation in Search Engines

- **Correct location vs incorrect location**
  - Mined 7 days of search engine logs in US market from Oct 2014
  - Intersect with separate unused ground truth
  - Compare ground truth location with Bing IP geolocation
    - Impressions with error > 15km vs error ≤ 15km

Metrics	Local Intent
Overall Click-Through rate (any link)	-4.3%
Algorithmic Click-Through rate (on search results)	-1.1%
Algorithmic Click Success	-6.1%
Ads Click-Through Rate	-17.9%
Ads Click Success	-15.2%
Ad Revenue	-40.3%

IP Geolocation Approaches

- **Overall goal**: Create an IP geolocation database from scratch
- Part of a larger research project
- Approaches
  - Reverse DNS
  - Geographic Clicks
    - GPS
    - Web Index
  - Query Logs
  - Interpolation
  - Traceroutes
  - WHOIS
- Some approaches used in production today in Bing

4.

PRIVACY AND DATASETS

Privacy-Conscious IP Geolocation

Your Apps Know Where You Were Last Night, and They're Not Keeping It Secret

Dozens of companies use smartphone locations to help advertisers and even hedge funds. They say it's anonymous, but the data shows how personal it is.

Apps sending location, secretly.

Posted by Joel Reardon on May 14, 2018

REVEAL MOBILE AUDIENCE INSIGHTS

SAFEGRAPH

FYSICAL

kiip

- **Protecting User Privacy**
  - Ground Truth Sets
  - Training data

Privacy (continued)

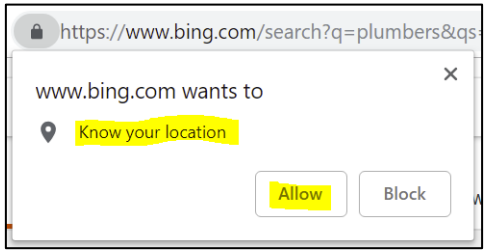
- IP Geolocation
  - Does not track individual's user location
  - Uses IP addresses and not accounts
  - Is coarse, it maps IP ranges to cities
  - Arguably, more privacy conscious than user-level GPS tracking

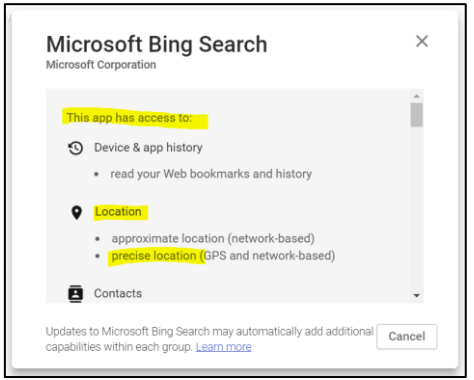
Public Traceroute Dataset

- **9 billion traceroutes**
- Collected between January and November 2017
- From Archipelago (Ark) Measurement Infrastructure of CAIDA
- 208 servers located in 63 countries
- Every 48 hours a random IP address is chosen in each /24 prefix, then the chosen IP addresses are individually probed by random Ark servers.

Proprietary Ground Truth Set

- **8.9 million IP addresses with known location**
- From Bing query logs: 28-day period ending on December 1st, 2017
- Opt-in and anonymized data from browsers and apps
- IPs from wired, wireless, and cell networks
- From all countries in the world
- Only used in training and testing, individual data points are not part of the final geolocation database

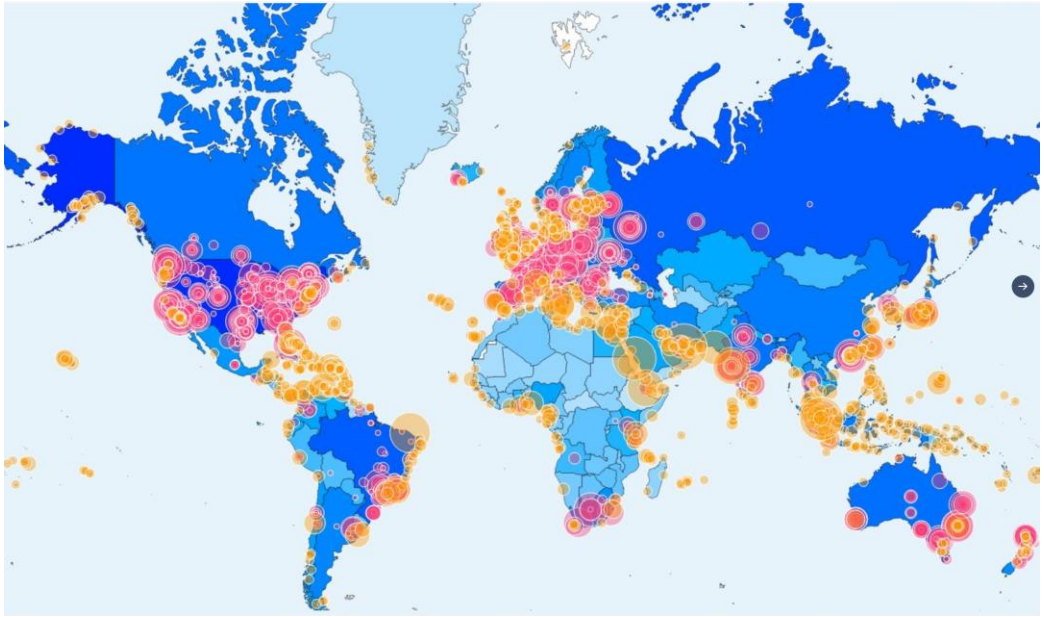




PeeringDB

- Smaller alternative to our proprietary ground truth set
- 400 IP ranges spanning 128,000 IP addresses, along with geographic coordinates
- Disadvantage: heavily infrastructure focused vs end users

PeeringDB

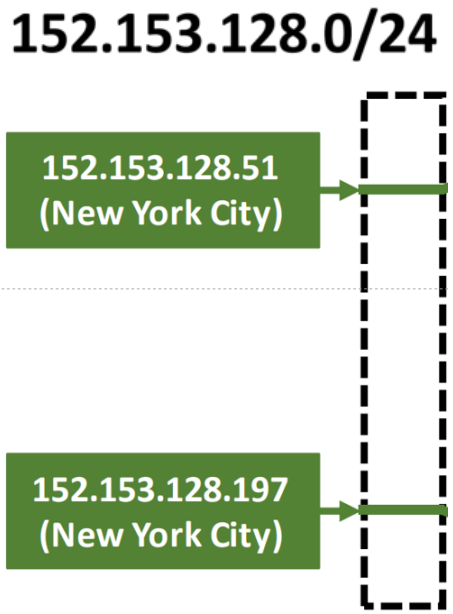


5.

IP LOCATION INTERPOLATION

Extrapolating location of entire IP range from a few individual IPs

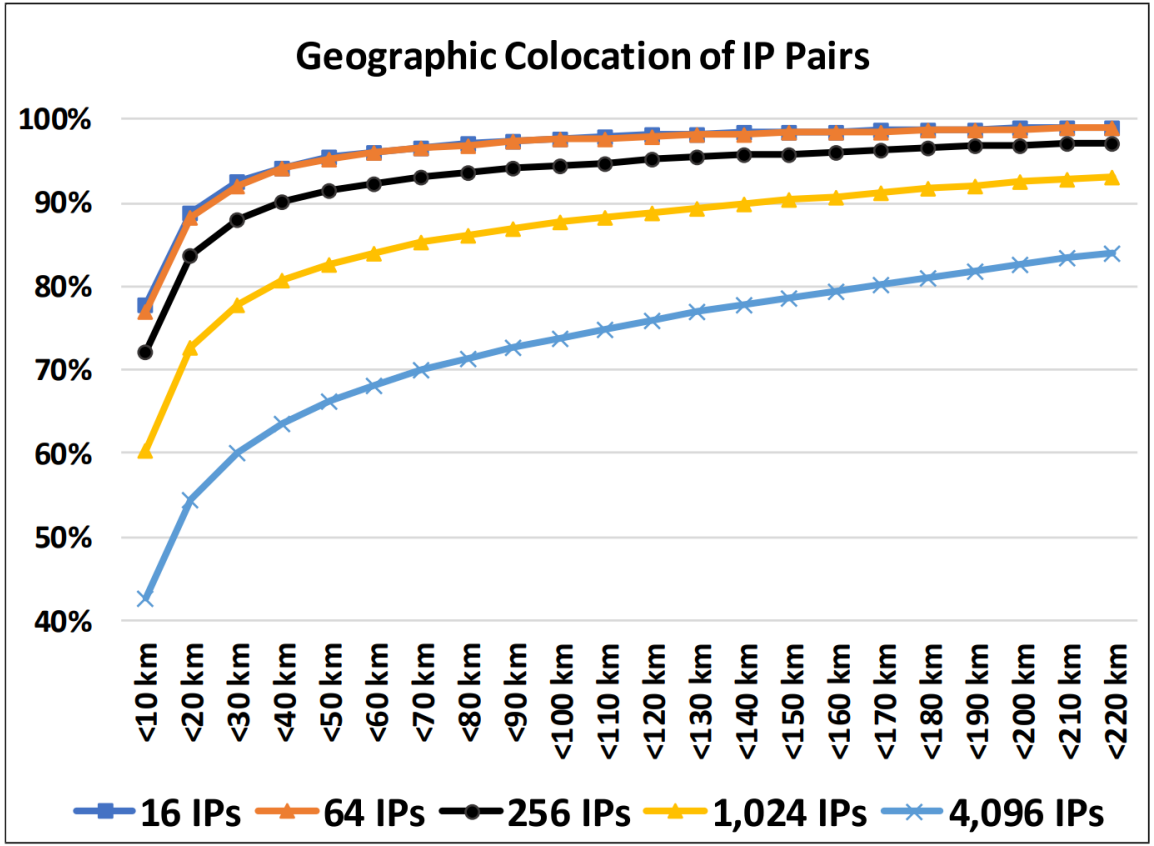
IP Colocation



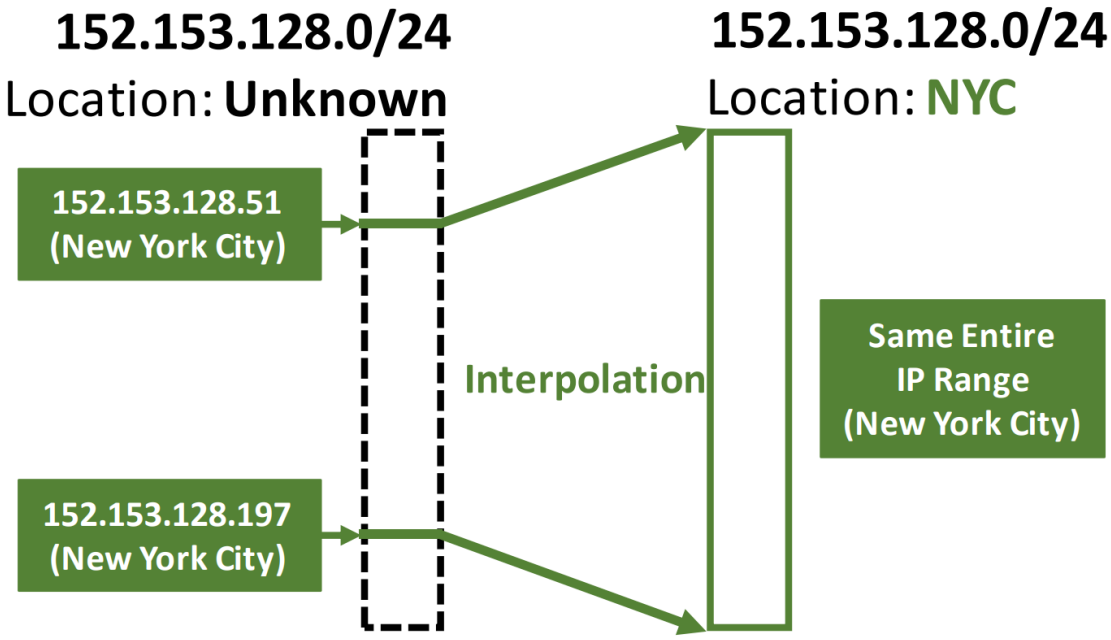
	Hosts	Netmask
/30	4	255.255.255.252
/29	8	255.255.255.248
/28	16	255.255.255.240
/27	32	255.255.255.224
/26	64	255.255.255.192
/25	128	255.255.255.128
/24	256	255.255.255.0
/23	512	255.255.254.0
/22	1024	255.255.252.0
/21	2048	255.255.248.0
/20	4096	255.255.240.0
/19	8192	255.255.224.0
/18	16384	255.255.192.0
/17	32768	255.255.128.0
/16	65536	255.255.0.0

- Two IP addresses are collocated if they are in the same IP range.
- IP ranges can have different sizes
- In example above the 2 IPs are also geographically colocated

Colocation of IP Range Addresses

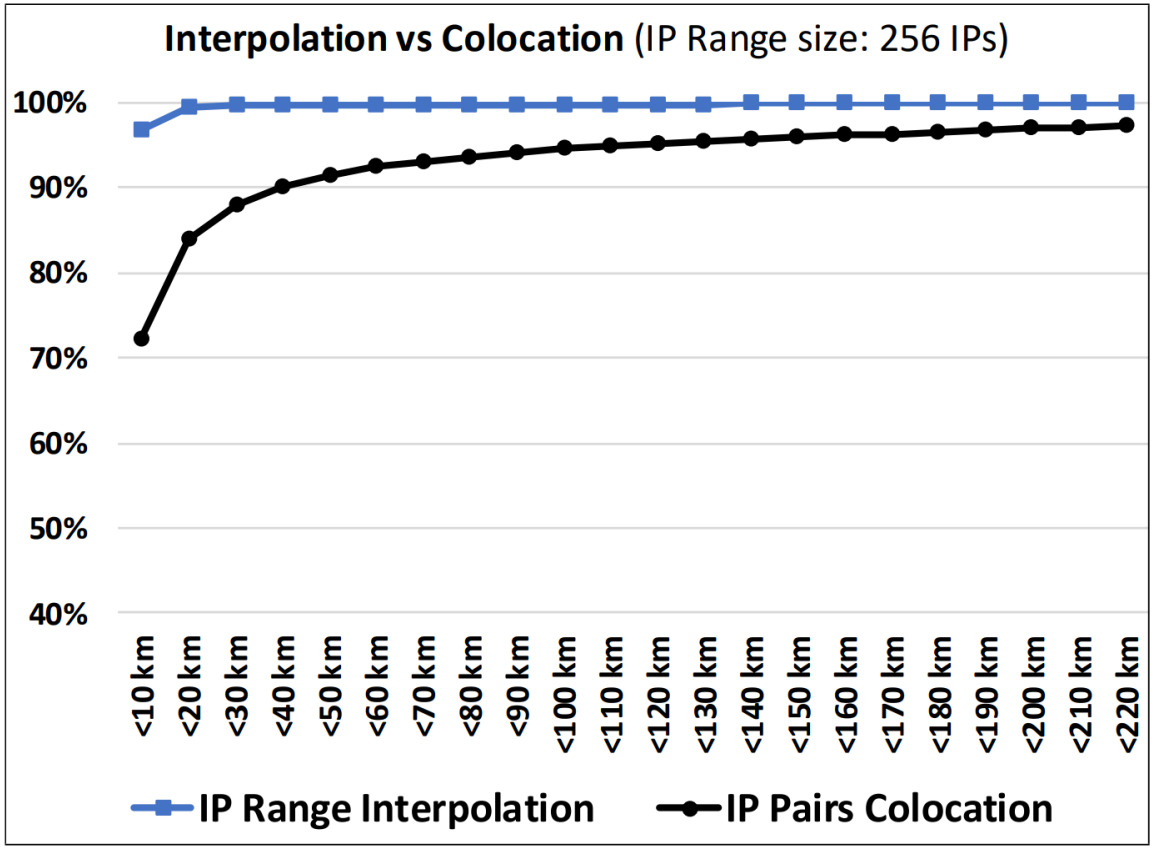


Location Interpolation



Example of IP range interpolation. Since IP range 152.153.128.0/24 contains two IP addresses with the same known location (coordinates in New York City), we propagate that location to the entire IP range.

Colocation vs Interpolation



6.

TRACEROUTE LOCATION PROPAGATION

Background, Motivation, Research Questions, Contributions

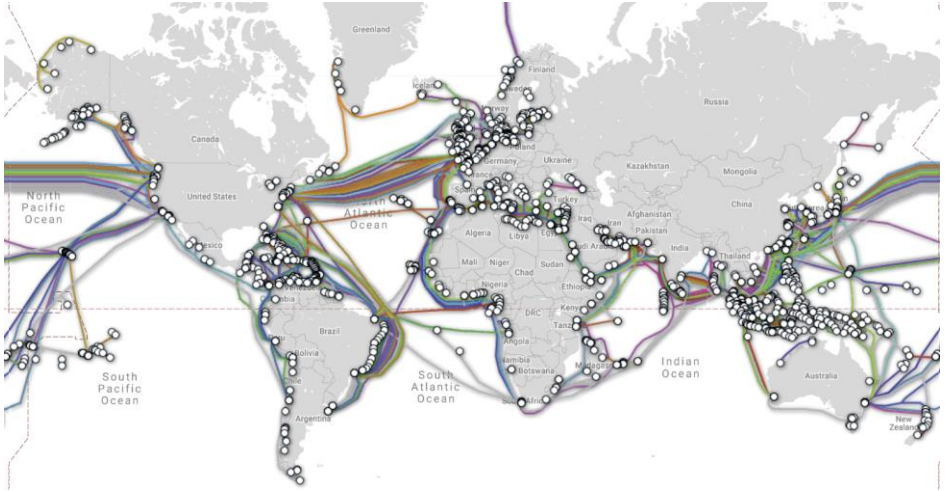
Ping and Traceroute

Command Prompt

```
C:\Users\zmart>ping www.lehigh.edu

Pinging www.lehigh.edu [128.180.1.16] with 32 bytes of data:
Reply from 128.180.1.16: bytes=32 time=79ms TTL=50
Reply from 128.180.1.16: bytes=32 time=79ms TTL=50
Reply from 128.180.1.16: bytes=32 time=79ms TTL=50
Reply from 128.180.1.16: bytes=32 time=79ms TTL=50
```

Command Prompt

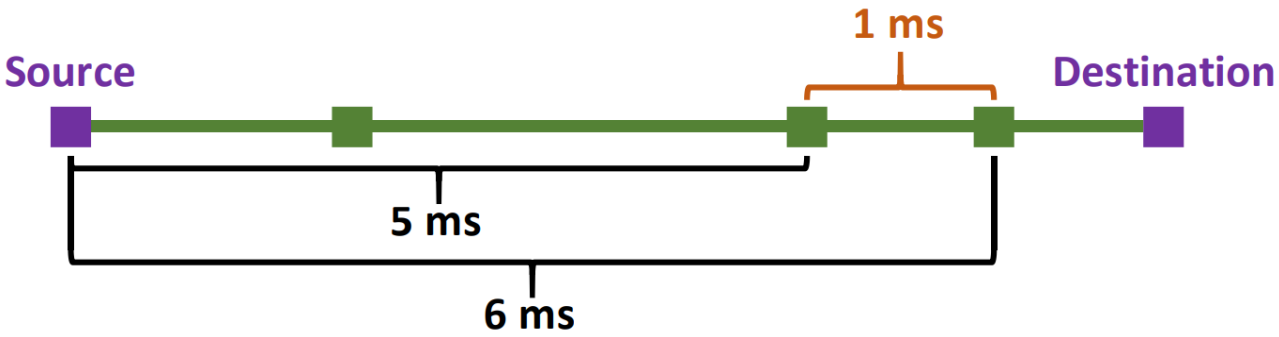
```
C:\Users\zmart>tracert www.lehigh.edu

Tracing route to www.lehigh.edu [128.180.1.16]
over a maximum of 30 hops:

  1     1 ms     1 ms     1 ms  Soul.ziggystardust.selfhost.corp.microsoft.com [10.0.10.2]
  2    <1 ms    <1 ms    <1 ms  10.0.10.1
  3     2 ms     1 ms     1 ms  50-46-216-1.evrt.wa.frontiernet.net [50.46.216.1]
  4     2 ms     2 ms     2 ms  172.76.21.177
  5     3 ms     3 ms     3 ms  ae3---0.cor02.sttl.wa.frontiernet.net [74.40.1.101]
  6     3 ms     3 ms     3 ms  ae1---0.cbr01.sttl.wa.frontiernet.net [74.40.5.126]
  7     3 ms     3 ms     3 ms  74.43.94.7
  8    77 ms    77 ms    78 ms  216.156.16.80.ptr.us.xo.net [216.156.16.80]
  9    75 ms    76 ms    82 ms  207.88.12.228.ptr.us.xo.net [207.88.12.228]
 10    76 ms    76 ms    75 ms  207.88.12.144.ptr.us.xo.net [207.88.12.144]
 11    78 ms    76 ms    77 ms  207.88.12.190.ptr.us.xo.net [207.88.12.190]
 12    77 ms    77 ms    78 ms  207.88.12.189.ptr.us.xo.net [207.88.12.189]
 13    78 ms    78 ms    77 ms  207.88.12.164.ptr.us.xo.net [207.88.12.164]
 14    77 ms    76 ms    76 ms  207.88.12.213.ptr.us.xo.net [207.88.12.213]
 15    80 ms    78 ms    78 ms  207.88.12.214.ptr.us.xo.net [207.88.12.214]
 16    78 ms    78 ms    77 ms  207.88.12.207.ptr.us.xo.net [207.88.12.207]
 17    79 ms    77 ms    78 ms  216.156.16.45.ptr.us.xo.net [216.156.16.45]
 18    79 ms    79 ms    79 ms  mail.tfains.net [209.118.23.114]
 19    80 ms    80 ms    79 ms  128.180.18.202
 20    79 ms    79 ms    79 ms  ws1iv.cc.lehigh.edu [128.180.1.16]
```



Speed of light in fiber

- **2/3 speed of light**: 199 km/ms (124 miles/ms)
- **4/9 speed of light**: 133 km/ms (83 miles/ms)

Latency Neighbors



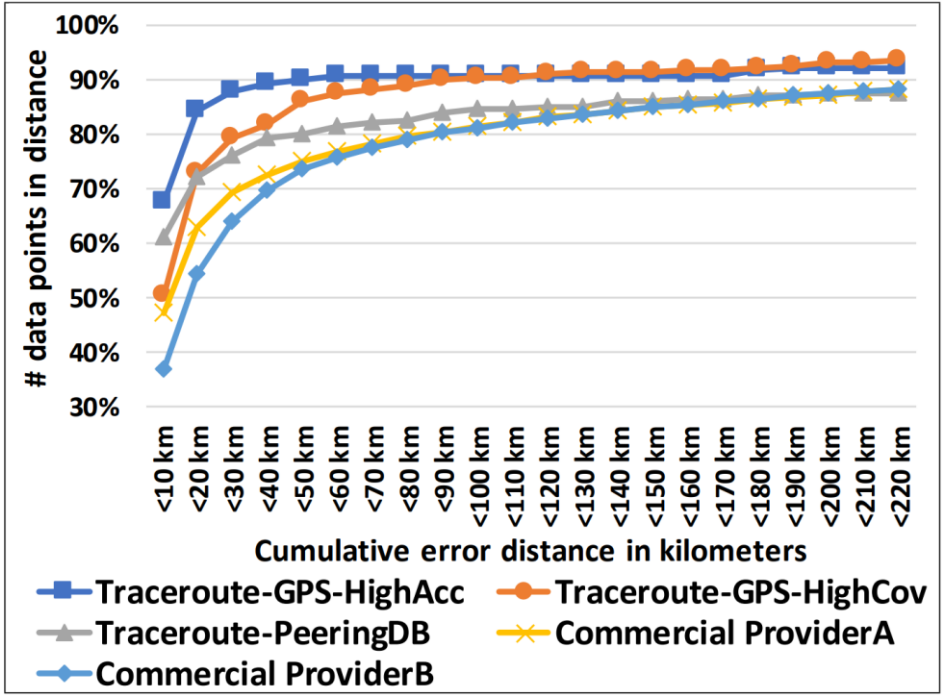
Approach

Combine traceroute with IP interpolation

- Extract latency neighbors
- Propagate locations through neighbors
- Apply interpolation to increase coverage
- Evaluate against ground truth

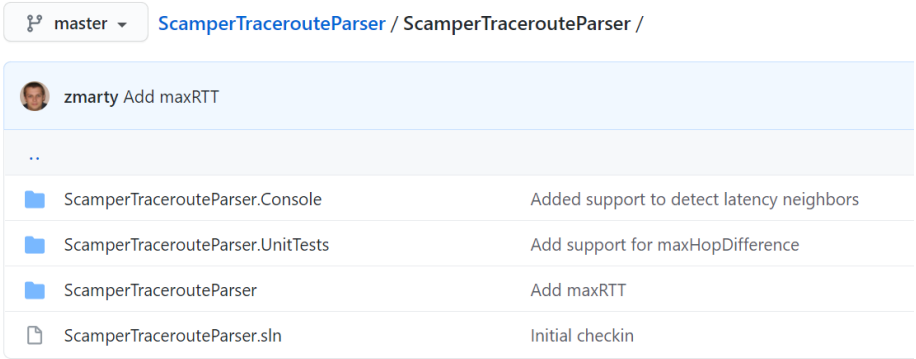
Commercial Evaluation

	Median error	% Err <10km	RMSE in km
Traceroute-GPS-HighAcc	**4.3 km**	**67.7%**	**329.3**
Traceroute-GPS-HighCov	10.1 km	50.5%	423.6
Traceroute-PeeringDB	8.4 km	61.1%	2124.9
Commercial Provider A	11.1 km	47.2%	545.9
Commercial Provider B	16.7 km	36.7%	545.3



Reproducing Results: Open Source Components

- Traceroute Dataset Parsing Library
  - https://github.com/zmarty/ScamperTracerouteParser
  - C# library that can parse the text traceroute dumps extracted from binary warts files using sc_analysis_dump (part of the CAIDA Scamper project).
- PeeringDB parsing and generation library
  - https://github.com/zmarty/PeeringDBToTSV
  - Tiny C# library which reads the PeeringDB SQLite databases and can output the data in TSV format more suitable for IP geolocation research



THANK YOU

Ovidiu Dan, Vaibhav Parikh, Brian. D. Davison

7.

RELATED WORK

Brief summary of two decades of related work in IP geolocation

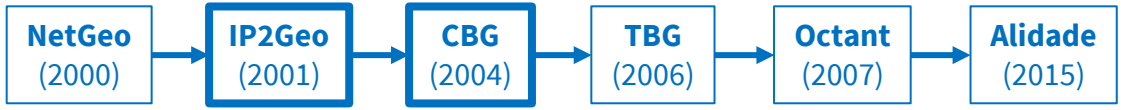
Approaches in Related Work

- **Network delay**
- **Network Topology**
  - Traceroute
  - BGP / ASN
- **Reverse DNS**
- **IP Interpolation**
- **WHOIS databases**
- **Social networks**
- **Web Mining**
- **Crowdsourcing**

		Ping	Trace-route	Reverse DNS	Inter-polation	WHOIS	BGP AS	Crowd Social	Other
2000	NetGeo [112]			✓		✓			✓
2001	GeoTrack [69, 70]		✓	✓					
2001	GeoPing [69, 70]	✓							
2001	GeoCluster [69, 70]				✓		✓		
2002	undns [84]			✓					
2004	Ziviani [71]	✓							
2004	CBG [2, 40]	✓							
2004	PBE [82]	✓	✓						
2004	RBE [82]	✓	✓						
2006	TBG [83]		✓	✓					
2006	GeoBuD [89]	✓	✓						
2007	Octant [31]	✓	✓	✓		✓			
2009	SG [72]	✓							
2009	Structon [32]		✓		✓		✓		✓
2010	SocialGraph [101]							✓	
2010	WBG [113]		✓			✓			
2010	Geo-RhOL [88]	✓	✓						
2010	NB-LHP [90]	✓	✓						✓
2011	Spotter [30]	✓							
2011	WebCBG [100]	✓	✓						✓
2011	HawkEyes [118]	✓	✓						
2012	SDP [75]	✓							
2012	Posit [91, 92]	✓	✓						
2013	GeoGet [114]								✓
2013	Checkin-Geo [63]				✓			✓	✓
2013	PCFL [95]	✓	✓						✓
2013	RUEL [95]	✓	✓	✓					✓
2013	PathAudit [93]			✓					
2013	AdvancedGeo [122]		✓	✓		✓			✓
2014	DRoP [97]		✓	✓					
2015	Dragoon [119–121]	✓	✓						
2015	Alidade [33]	✓	✓	✓	✓	✓	✓		✓
2016	GeoSpeed [60]				✓			✓	
2016	AIG [116]	✓							
2016	Neural-RBF [76]	✓							
2017	HLOC [99]	✓		✓					
2018	CRLB [77]	✓							

Network Delay and Network Topology

		Ping	Trace-route	Reverse DNS	Inter-polation	WHOIS	BGP AS	Crowd Social	Other
2000	NetGeo [112]			✓		✓			✓
2001	GeoTrack [69, 70]		✓	✓					
2001	GeoPing [69, 70]	✓							
2001	GeoCluster [69, 70]				✓		✓		
2002	undns [84]			✓					
2004	Ziviani [71]	✓							
2004	CBG [2, 40]	✓							
2004	PBE [82]	✓	✓						
2004	RBE [82]	✓	✓						
2006	TBG [83]		✓	✓					
2006	GeoBuD [89]	✓	✓						
2007	Octant [31]	✓	✓	✓		✓			
2009	SG [72]	✓							
2009	Structon [32]		✓		✓		✓		✓
2010	SocialGraph [101]							✓	
2010	WBG [113]		✓			✓			
2010	Geo-RhOL [88]	✓	✓						
2010	NB-LHP [90]	✓	✓						✓
2011	Spotter [30]	✓							
2011	WebCBG [100]	✓	✓						✓
2011	HawkEyes [118]	✓	✓						
2012	SDP [75]	✓							
2012	Posit [91, 92]	✓	✓						
2013	GeoGet [114]								✓
2013	Checkin-Geo [63]				✓			✓	✓
2013	PCFL [95]	✓	✓						✓
2013	RUEL [95]	✓	✓	✓					✓
2013	PathAudit [93]			✓					
2013	AdvancedGeo [122]		✓	✓		✓			✓
2014	DRoP [97]		✓	✓					
2015	Dragoon [119–121]	✓	✓						
2015	Alidade [33]	✓	✓	✓	✓	✓	✓		✓
2016	GeoSpeed [60]				✓			✓	
2016	AIG [116]	✓							
2016	Neural-RBF [76]	✓							
2017	HLOC [99]	✓		✓					
2018	CRLB [77]	✓							



Network Delay and Network Topology

- Ping, Traceroute, BGP, ASN
- Active probes
- Router table dumps
- Most studied

GeoPing (2001)

- Measure ping from each probe servers to all the others
- For a new target IP, measure ping from all servers to it
- Pick probe server vector with smallest latency
- Assign probe server location to target IP

Active Probe Servers
Location: **Known**

	Known Location IP 1	Known Location IP 2	Known Location IP 3	Known Location IP N
Known Location IP 1	**0ms**	13ms	43ms	12ms
Known Location IP 2	11ms	**0ms**	9ms	207ms
Known Location IP 3	42ms	17ms	**0ms**	4ms
Known Location IP N	132ms	43ms	0.5ms	**0ms**

Pick vector w/ smallest distance

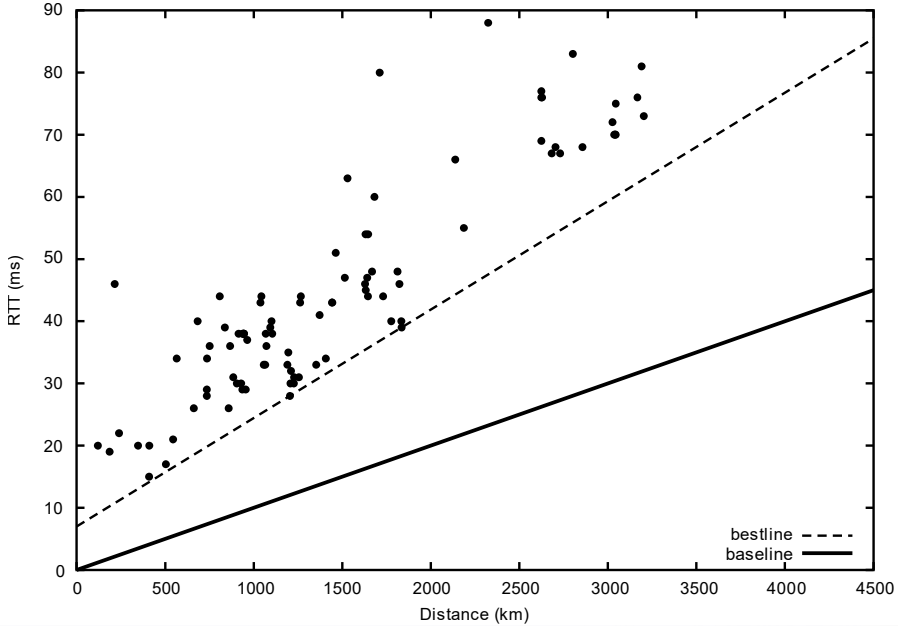
New target IP
Location: **Unknown**

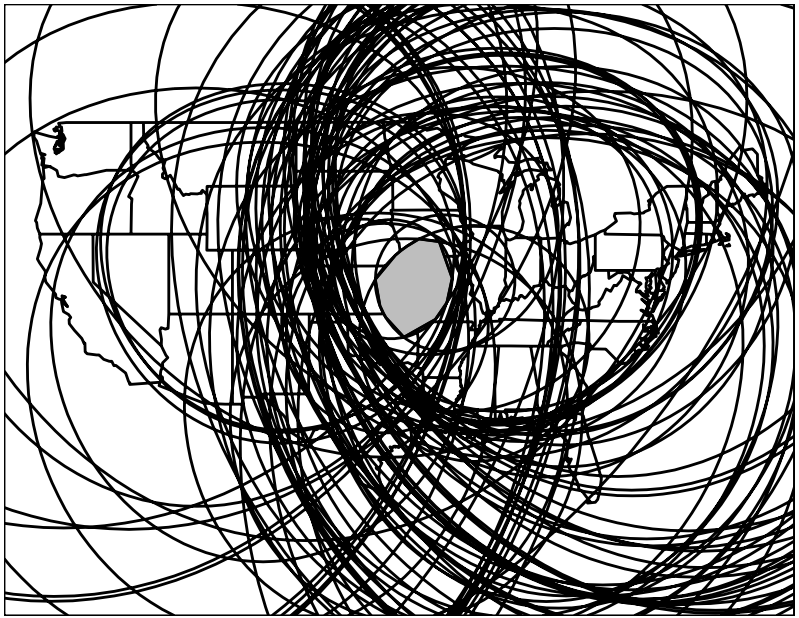
	IP 1	IP 2	IP 3	IP N
~~Unknown Location~~	10ms	0.1ms	9ms	204ms

Network Delay

CBG (2004)

- Determine *bestline* for each probe server ➔ delay-to-distance graph
- Draw circles around each active probe server
- Intersect circles and find center of intersection ➔ *multilateration*
- Assign center location to target

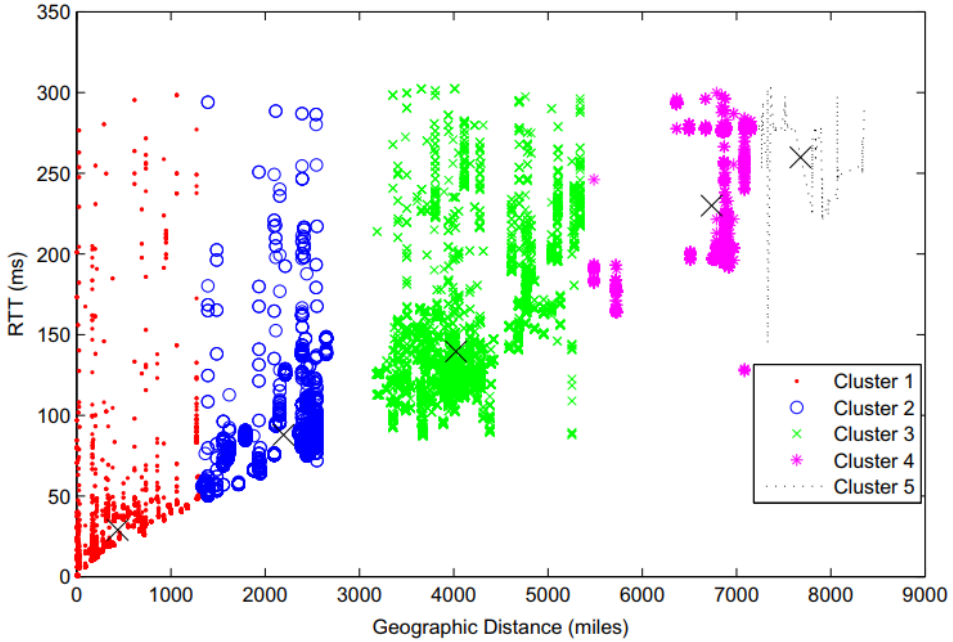




Bestline Estimation Research

NetGeo (2000) → **IP2Geo** (2001) → **CBG** (2004) → **TBG** (2006) → **Octant** (2007) → **Alidade** (2015)

- **SG** (2009) — Kernel density estimation
- **Spotter** (2011) — Fit single common baseline
- **SDP** (2012) — 2-step K-Means Clustering
- **AIG** (2016) — 2-step Coarse Granular
- **Neural-RBF** (2016) — 2-step + neural network



SDP (2012)

- Cluster bestline data points
- Only use probes in closest cluster

TBG (2006)

- Use intermediate routers on the traceroute path to constrain distances from latency measurements
- Treat geolocation as an optimization problem on a graph
  - Vertices – active probe servers and targets
  - Edges – traceroute paths with network delay

Reverse DNS

		Ping	Trace-route	Reverse DNS	Inter-polation	WHOIS	BGP AS	Crowd Social	Other
2000	NetGeo [112]			✓		✓			✓
2001	GeoTrack [69, 70]		✓	✓					
2001	GeoPing [69, 70]	✓							
2001	GeoCluster [69, 70]				✓		✓		
2002	undns [84]			✓					
2004	Ziviani [71]	✓							
2004	CBG [2, 40]	✓							
2004	PBE [82]	✓	✓						
2004	RBE [82]	✓	✓						
2006	TBG [83]		✓	✓					
2006	GeoBuD [89]	✓	✓						
2007	Octant [31]	✓	✓	✓		✓			
2009	SG [72]	✓							
2009	Structon [32]		✓		✓		✓		✓
2010	SocialGraph [101]							✓	
2010	WBG [113]		✓			✓			
2010	Geo-RhOL [88]	✓	✓						
2010	NB-LHP [90]	✓	✓						✓
2011	Spotter [30]	✓							
2011	WebCBG [100]	✓	✓						✓
2011	HawkEyes [118]	✓	✓						
2012	SDP [75]	✓							
2012	Posit [91, 92]	✓	✓						
2013	GeoGet [114]								✓
2013	Checkin-Geo [63]				✓			✓	✓
2013	PCFL [95]	✓	✓						✓
2013	RUEL [95]	✓	✓	✓					✓
2013	PathAudit [93]			✓					
2013	AdvancedGeo [122]		✓	✓		✓			✓
2014	DRoP [97]		✓	✓					
2015	Dragoon [119–121]	✓	✓						
2015	Alidade [33]	✓	✓	✓	✓	✓	✓		✓
2016	GeoSpeed [60]				✓			✓	
2016	AIG [116]	✓							
2016	Neural-RBF [76]	✓							
2017	HLOC [99]	✓		✓					
2018	CRLB [77]	✓							

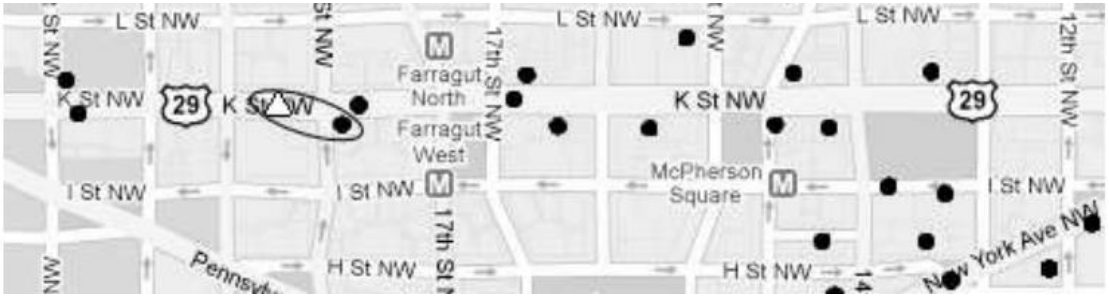
hostname	ccr21.	par01.	atlas.	cogentco.com
position	2	1	0	
hint	ccr	par	atlas	
location	|	|	Salas Atlas, ES	
	|	Paris, FR		
	Concord, CA			

- **undns (2002)**
  - Manual rules
  - US centric
- **DRoP (2014)**
  - Automatic parsing rules
  - Rules validated with network delay
- **DDec (2015)**
  - Combines undns with DRoP
  - Gives precedence to undns
- **HLOC (2017)**
  - Use location hints directly

Crowdsourcing and Social Networks

		Ping	Trace-route	Reverse DNS	Inter-polation	WHOIS	BGP AS	Crowd Social	Other
2000	NetGeo [112]			✓		✓			✓
2001	GeoTrack [69, 70]		✓	✓					
2001	GeoPing [69, 70]	✓							
2001	GeoCluster [69, 70]				✓		✓		
2002	undns [84]			✓					
2004	Ziviani [71]	✓							
2004	CBG [2, 40]	✓							
2004	PBE [82]	✓	✓						
2004	RBE [82]	✓	✓						
2006	TBG [83]		✓	✓					
2006	GeoBuD [89]	✓	✓						
2007	Octant [31]	✓	✓	✓		✓			
2009	SG [72]	✓							
2009	Structon [32]		✓		✓		✓		✓
2010	SocialGraph [101]							✓	
2010	WBG [113]		✓			✓			
2010	Geo-RhOL [88]	✓	✓						
2010	NB-LHP [90]	✓	✓						✓
2011	Spotter [30]	✓							
2011	WebCBG [100]	✓	✓						✓
2011	HawkEyes [118]	✓	✓						
2012	SDP [75]	✓							
2012	Posit [91, 92]	✓	✓						
2013	GeoGet [114]								✓
2013	Checkin-Geo [63]				✓			✓	✓
2013	PCFL [95]	✓	✓						✓
2013	RUEL [95]	✓	✓	✓					✓
2013	PathAudit [93]			✓					
2013	AdvancedGeo [122]		✓	✓		✓			✓
2014	DRoP [97]		✓	✓					
2015	Dragoon [119–121]	✓	✓						
2015	Alidade [33]	✓	✓	✓	✓	✓	✓		✓
2016	GeoSpeed [60]				✓			✓	
2016	AIG [116]	✓							
2016	Neural-RBF [76]	✓							
2017	HLOC [99]	✓		✓					
2018	CRLB [77]	✓							

- **Mine websites**
  - Chinese websites - Structon (2009)
  - Yellow Pages + CBG - WebCBG (2011)
- **Mine social graphs**
  - Extrapolate location from friends - SocialGraph (2010)
  - Locations from social checkins - Checkin-Geo (2014)
- **WHOIS databases**
  - Combine WHOIS with traceroutes - WBG (2010)
- **HTTP Network Delay**
  - Use GET as "ping" for multilateration - GeoGet (2013)
- **Crowdsourcing**
  - Crowdsourced Internet speed test - Speed (2016)



Limitations of previous approaches

- **General**
  - Small ground truth set
  - Insufficient geographic diversity
  - Lack of networking environment diversity
  - Poor city accuracy
  - Non-existent or vague evaluation
  - Usage of commercial IP geolocation databases for training or testing
- **Network Delay and Network Topology**
  - Require access to geographically diverse infrastructure
  - Active probes, ICMP is often filtered out
  - Not scalable, as locating even a single IP takes a long time (billions of IPs)
  - Cramér–Rao lower bound theoretical limit – best-case error of 20 kilometers