IP Infrastructure Geolocation

Guan-Yan Cai, **Michael McCarrin**, Robert Beverly
Naval Postgraduate School

CAIDA AIMS-5
April 1, 2015

cMAND

Introduction

- IP Geolocation:
  - Given IP address, determine physical location
- IP Geolocation (commercially) used for:
  - Targeted advertising, recommendation systems
  - Reputation, security
- Hence, majority of existing work focuses on edge devices
- Less attention on infrastructure. e.g.,:
  - Routers
  - Servers
- Motivation:
  - Understand physical Internet topology better

cMAND

Prior Work

- Prior work on router geolocation:
  - DNS (undns, DRoP)
  - Latency (Yoshida)
  - Topology (Feldman)
- State-of-the-art technique: *DNS-based Router Positioning (DRoP)* by Huffaker et al.
  - Relies on geolocation clues within DNS PTR record of router's IP Use geolocation hints to generate rule sets
- Our focus:
  - Does not work for routers with no DNS PTRs (40.4% or 12.8M)

Intuition

- Our simple intuition:
  - Routers are frequently co-located with other routers
  - E.g., carrier neutral colo, hosting facility, etc
- Hence, if we can determine that a router with known location is co-located or near to a router with unknown location:
  - Provides a means to estimate (with a measurable upper bound) the location of unknown router IPs

cMAND

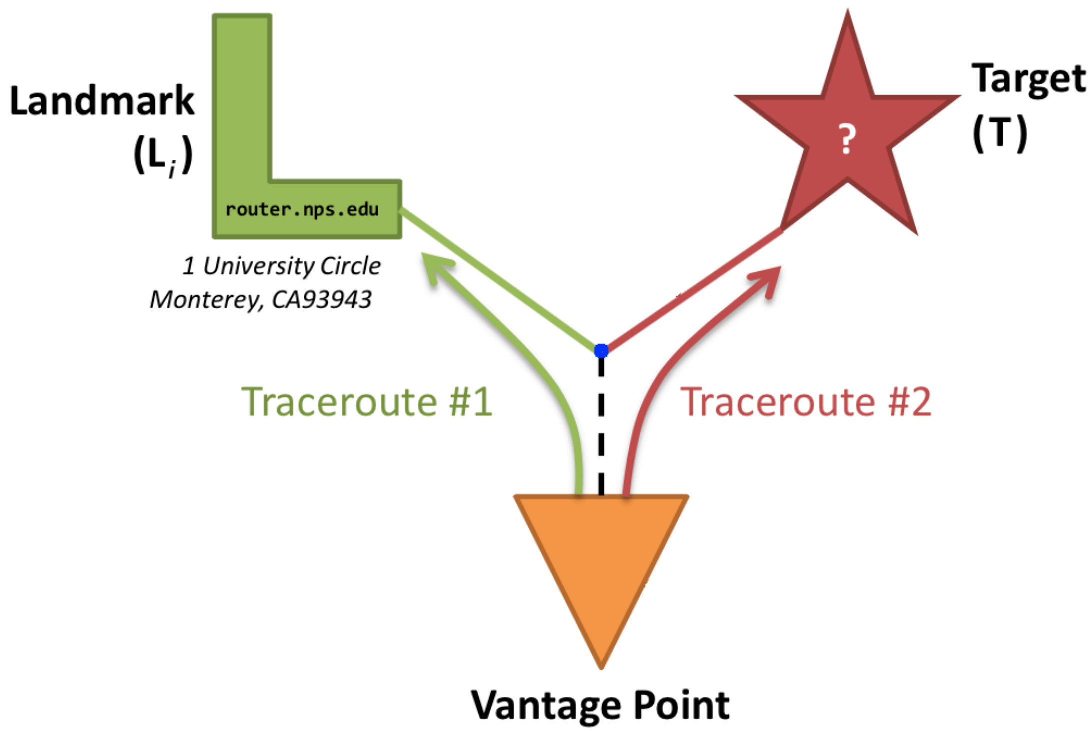
Methodology

- Leverage "Street-Level geolocation" technique (Wang et al. 2006):
  - Uses trace route to estimate latency between passive landmarks and target
  - This gets you more vantage points (via passive landmarks)
  - Accuracy is proportional to number of vantage points and nearest vantage point
- Apply Wang's technique to *router interfaces:*
  - Router interfaces (instead of web servers) as landmarks

cMAND

Methodology

- Geolocating* target, *T*, with landmarks, L_i:
  - Perform traceroutes to *T* and to L_i

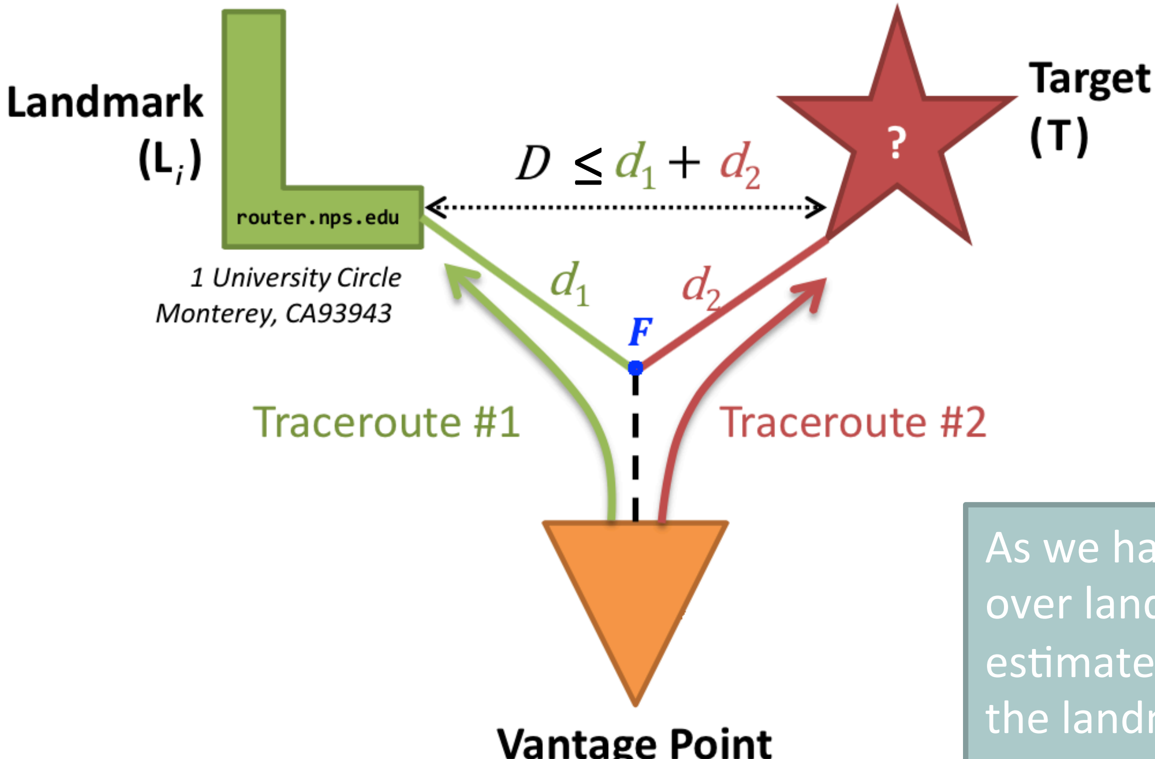


cMAND

* Technique adapted from Street-Level geolocation by Wang et al.

Methodology

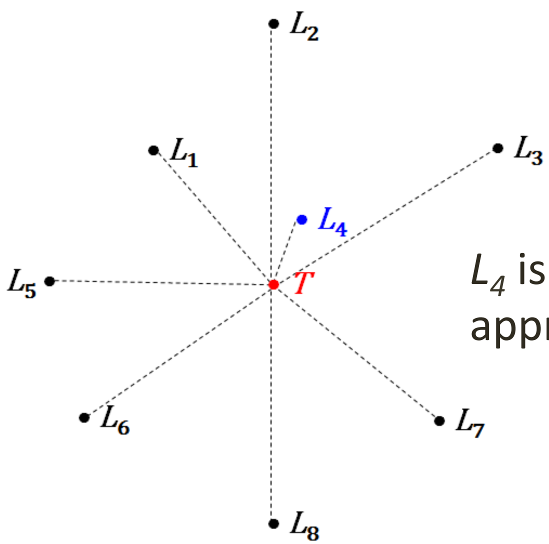
- Geolocating target, *T*, with landmarks, L_i:
  - Perform trace routes to *T* and to L_i
  - Determine point at which traceroutes diverge (F)
  - Estimate landmark to target delay, D, for all <L_i , *T*>



As we have no control over landmarks, we estimate delay between the landmark and target with $d_1 + d_2$

Methodology

- Geolocating target, *T*, with landmarks, L_i:
  - Perform trace routes to *T* and to L_i
  - Estimate delay (milliseconds), D, for all <L_i , *T*>
  - Find L_{min} that produces the least estimated delay for all <L_i , *T*> over all vantage points
  - Note, estimated delay is an upper bound (worst case)
  - Location of T = Location of L_{min}



L_4 is nearest to *T*, so *T*'s location is approximated as L_4's location.

cMAND

Experiment

- Use DRoP results as ground truth
- From DRoP's ~6M interfaces and ~8K unique locations:
  - Find locations with two interfaces that respond to trace route without anonymous hops (about half)
  - Half of them as landmarks (~4K)
  - Half of them as targets (~4K)
- Applied our methodology to geolocate all 4K targets
- Calculated *Error Distance* (km) i.e., geolocated position versus DRoP's location (Haversine distance)

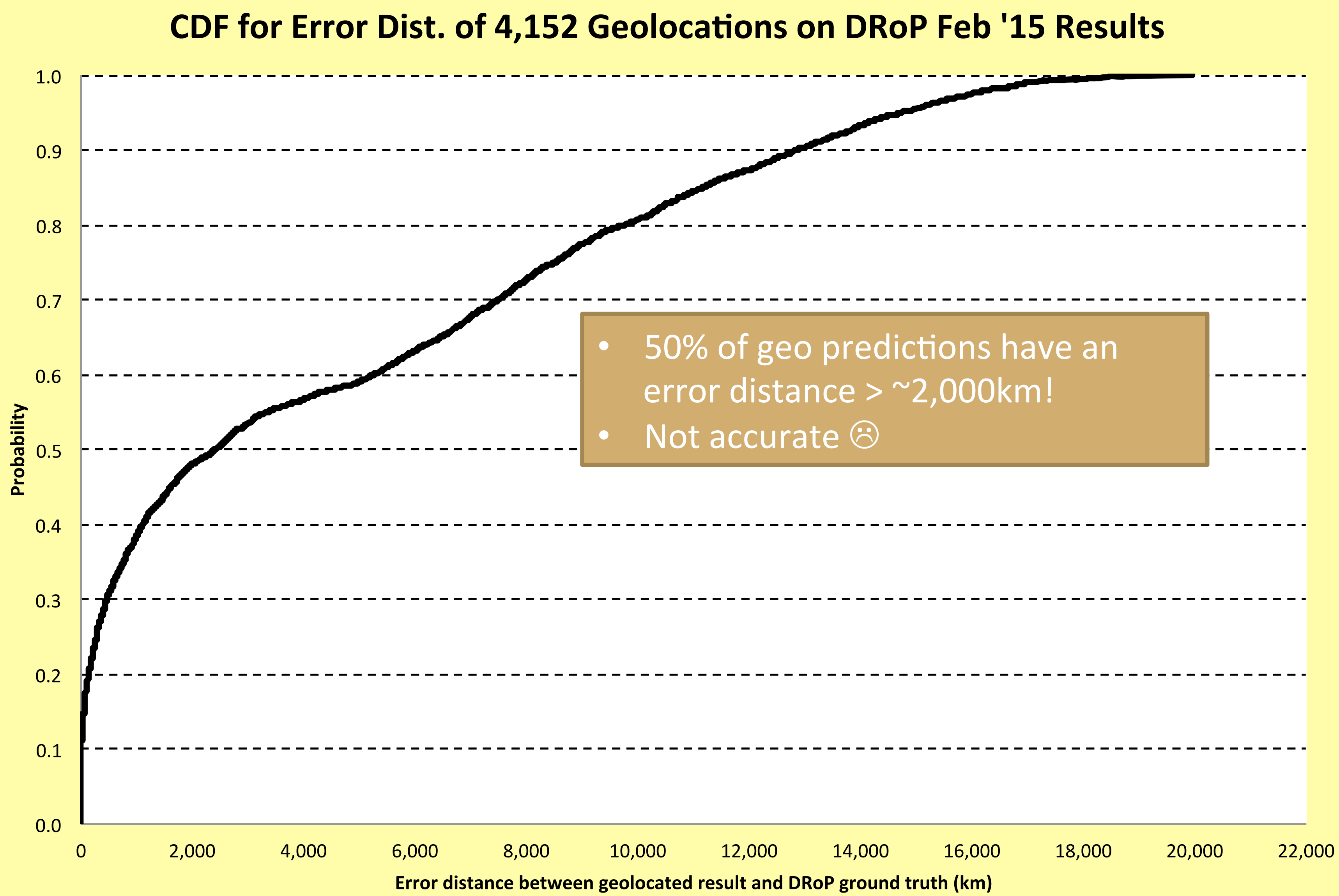
Target's geolocation
- Nearest landmark
- Location obtained from DRoP results

← *Error Distance (km)* →

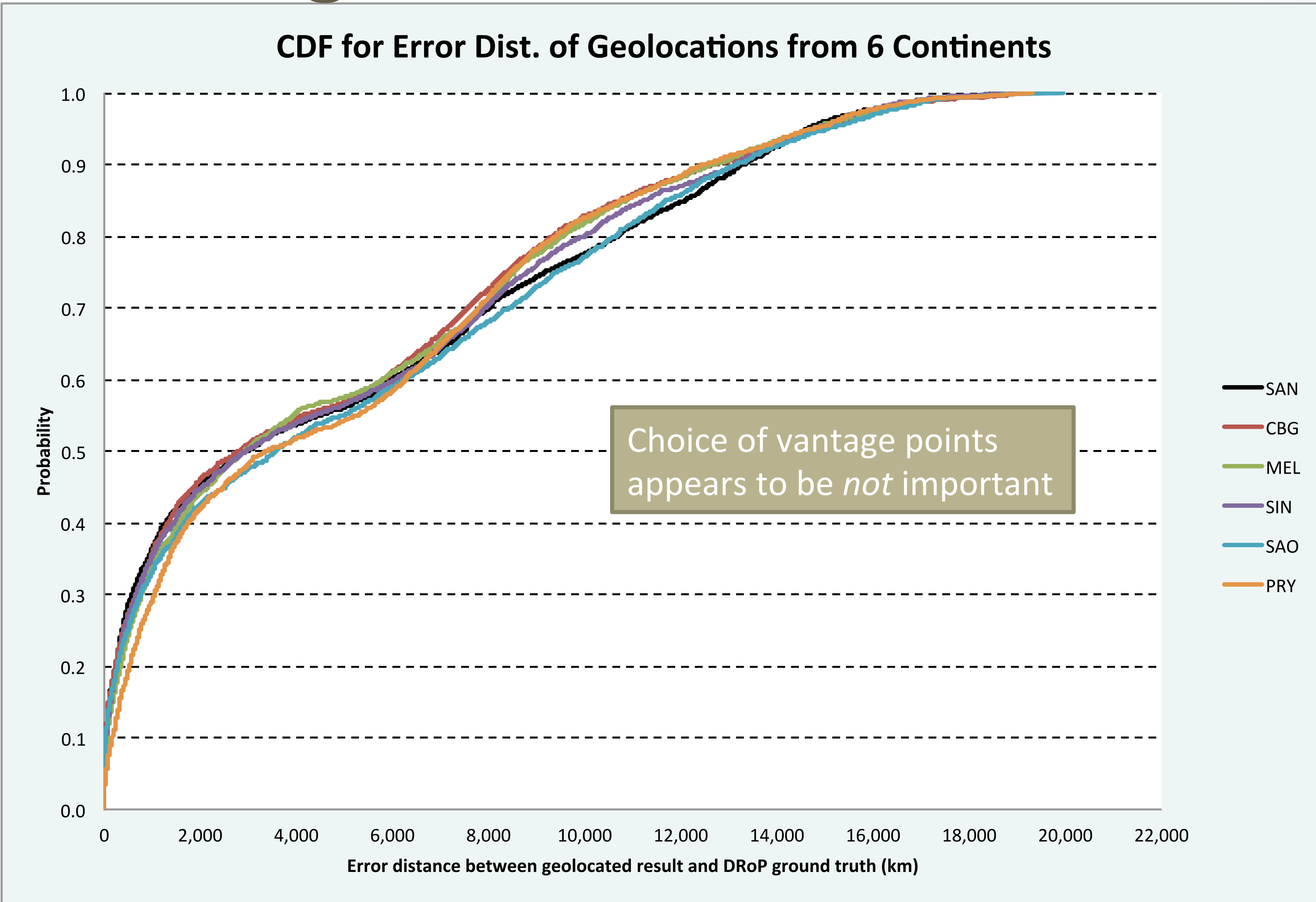
Target's ground truth location
- Location obtained from DRoP results

cMAND

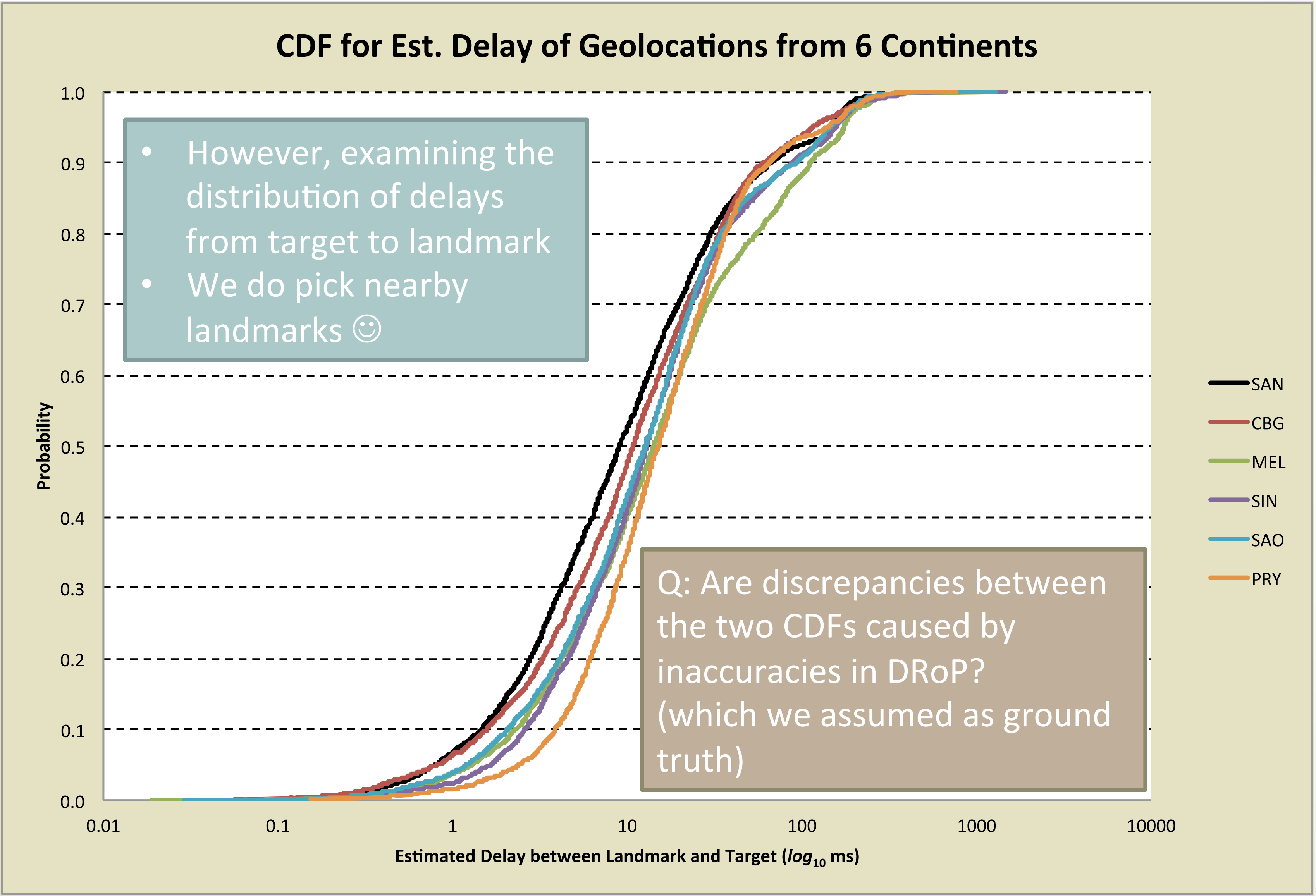
Results – Global Err. Dist.



Results – Err. Dist. from different Vantage Points



Results – Est. Delay from nearest landmark (multiple VPs)



- However, examining the distribution of delays from target to landmark
- We do pick nearby landmarks ☺

Q: Are discrepancies between the two CDFs caused by inaccuracies in DRoP? (which we assumed as ground truth)

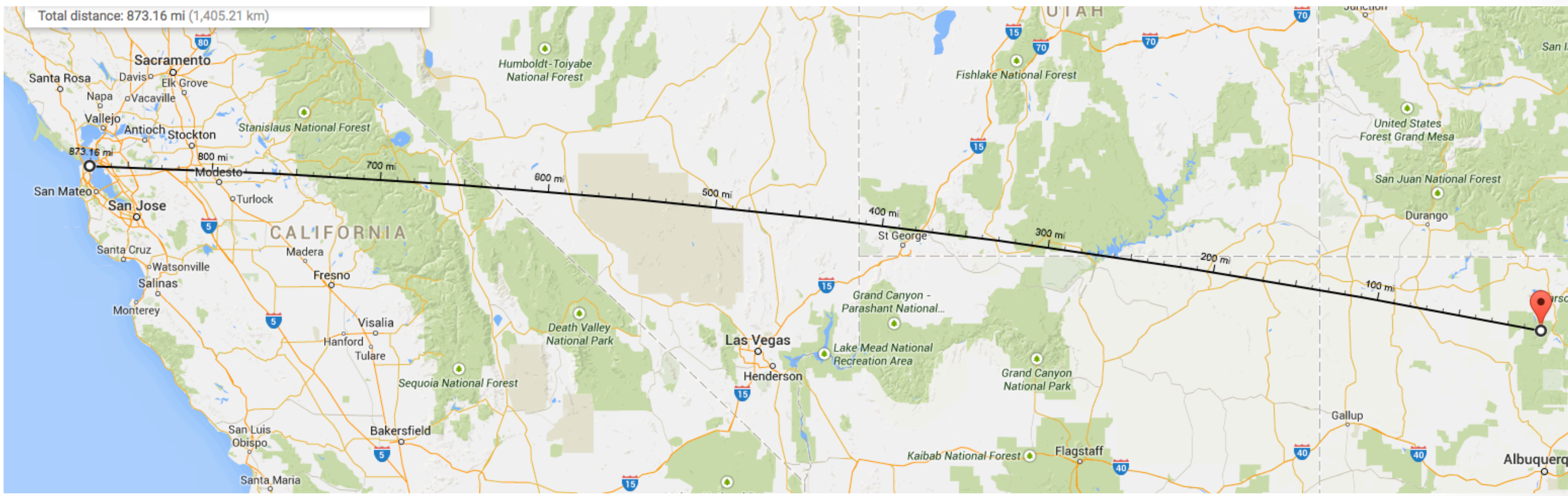
Evaluating DRoP

- Given our findings, we sought to better understand DRoP data:
  - Examine location inconsistencies
  - Use CBG to determine if locations are feasible
  - Use CBG to determine self-consistency of IPs believed to be at a particular location

cMAND

Errors in DRoP Locations

- How can there be errors in locations?
- E.g.
  - `251|us|ca|san francisco|36.3480163544573|-106.644463571429`
  - Where is that lat/long?



cMAND

Example: There are over 1900 San Franciscos

- Columbia alone has more than 100
- Some entries represent different places with the same name
- Others represent the same place with slightly different coordinates.
- Others are the same place with different spellings / nicknames / translations
- ***These problems emerge prior to DNS PTR record analysis.***

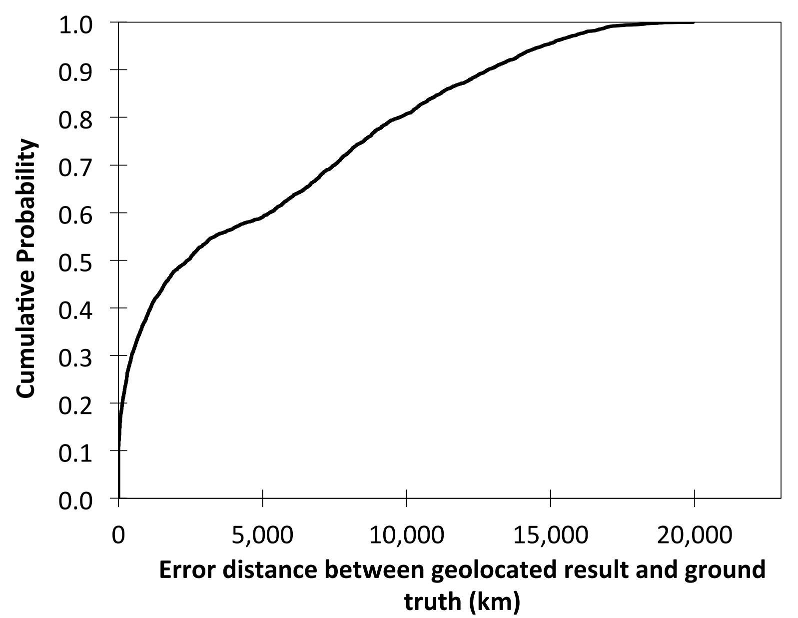
Name	Latitude	Longitude	Country
San Francisco	-20.71667	-64.7	Bolivia
San Francisco	-19.98922	-63.13761	Bolivia
San Francisco	-17.35	-61.15	Bolivia
San Francisco	-17.31667	-61.11667	Bolivia
San Francisco	-16.78333	-68.76667	Bolivia
San Francisco	-16.78333	-62.85	Bolivia
San Francisco	-16.66667	-65.18333	Bolivia
San Francisco	-15.26667	-65.51667	Bolivia
San Francisco	-15.2	-64.45	Bolivia
San Francisco	-15.08333	-65.16667	Bolivia
San Francisco	-14.18048	-62.80217	Bolivia
San Francisco	-13.91667	-63.7	Bolivia
San Francisco	-13.03333	-64.75	Bolivia
San Francisco	-11.83333	-66.81667	Bolivia
San Francisco	-11.59252	-69.08892	Bolivia
San Francisco	-11.20491	-69.06671	Bolivia
San Francisco	12.51667	-81.7	Columbia
San Francisco	10.92704	-72.81018	Columbia
San Francisco	8.72267	-75.5885	Columbia
San Francisco	8.71667	-74.63333	Columbia
San Francisco	8.69894	-75.43727	Columbia
San Francisco	8.45	-73.11667	Columbia
San Francisco	8.12039	-75.75981	Columbia
San Francisco	7.78811	-74.80846	Columbia
San Francisco	7.23535	-73.07099	Columbia
San Francisco	7.08333	-73.83333	Columbia
San Francisco	6.23333	-73.46667	Columbia
San Francisco	6.11667	-75.98333	Columbia
San Francisco	4.68333	-76.03333	Columbia

Excerpt from GeoNames allCountries.txt

cMAND

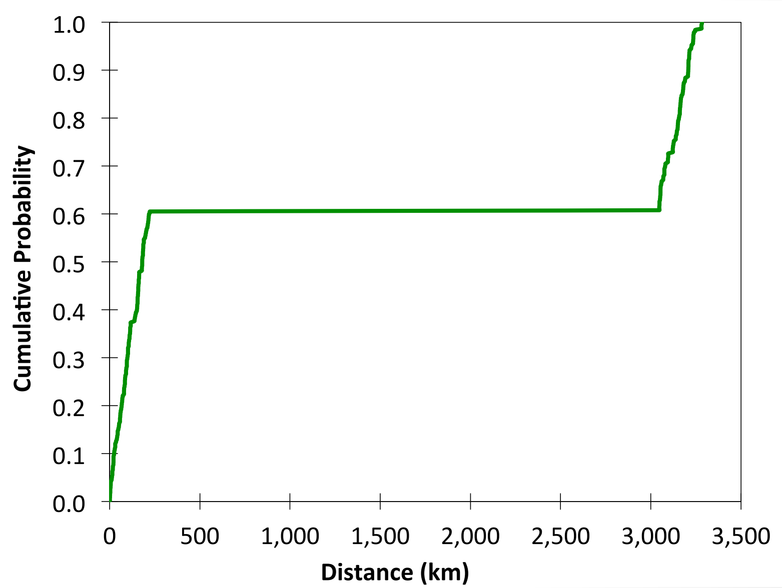
Finding Errors in DRoP IP to Location Mappings

- For each location, pick one responsive router interface
- Obtain 4,638 distinct locations with responsive interfaces
- Obtain RTTs from 22 Ark monitors to 4,638 interfaces (~100K RTTs)
- Use CBG on RTTs to determine possible region of interface
- Results:
  - 46% of these 4,638 interfaces *outside* of feasible boundaries imposed by CBG
  - CDF of distances from CBG centroid to DRoP location shows relatively large error distances



Focus on a DRoP Location (I)

- How self-consistent are IPs within a DRoP location:
  - Use Ark vantage points to gather RTTs
  - Use CBG to find centroids of feasible regions
  - For a given location, examine the pairwise N(N-1) distances between centroids
- Examined 20 router IPs from Chicago, IL:
- Results:
  - CDF of pairwise distances shows two modes!
  - Two distinct locations!
  - 60% in Chicago, IL
  - 40% in ocean 12mi west of Santa Barbara



cMAND

Focus on a DRoP Location (II)

- Two distinct locations:
  - 60% Chicago, IL
  - 40% 12 miles west of Santa Barbara
- What happened here?
- Examining a secondary IP geolocation database indicates that the 8 interfaces are in Chico, CA
- DNS PTR record contains non-standard geographic hint:
  - `cr1.chi2ca.sbcglobal.net`
  - "chi" == Chico
  - "chi" != Chicago
- Road Runner geo hint is consistent:
  - bu-ether25.chctilwc00w-bcr00.tbone.rr.com

cMAND

DRoP ambiguities/errors are pervasive

IP	PTR	DRoP Location	True Location
137.164.42.242	dc-pom-csu-lax-dc2-10ge.cenic.net	Port Moresby, Papua New Guinea	Los Angeles
128.83.10.110	tnh-gi5-5-nocb10.gw.utexas.edu	Erdaojiang, Jilin, China	Austin, TX
146.6.137.125	ccp-test.its.utexas.edu	Concepcion, Chile	Austin, TX
115.111.183.237	inpudiidnsprprd01.tatacommunications.com	Cumberland, RI	Nadu, India

Future Work

- Currently in active collaboration with CAIDA
- We can do some obvious things to improve name-to-coordinate mapping.
  - Some problems have already been fixed.
- How do we scale up error detection?
  - Get more out of fewer trace routes by intelligently selecting landmarks.
  - Start with CBG to get course granular.
  - Use landmarks within feasible region.
- Use existing CAIDA traceroutes?
  - Trade up control for speed / convenience
  - Might be good enough…

cMAND

Thank You

Questions?

cMAND