AFRL-RI-RS-TR-2014-115

MAPPING THE WHOLE INTERNET

DUKE UNIVERSITY

MAY 2014

FINAL TECHNICAL REPORT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

STINFO COPY

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE

■ **AIR FORCE MATERIEL COMMAND** ■ **UNITED STATES AIR FORCE** ■ **ROME, NY 13441**

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09. This report is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RI-RS-TR-2014-115 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE DIRECTOR:

/ S /
Robert L. Kaminski
Work Unit Manager

/ S /
WARREN H. DEBANY, JR.
Technical Advisor
Information Exploitation
and Operations Division
Information Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
MAY 2014	FINAL TECHNICAL REPORT	Aug 2011 – Dec 2013

4. TITLE AND SUBTITLE

MAPPING THE WHOLE INTERNET

5a. CONTRACT NUMBER
FA8750-11-1-0262

5b. GRANT NUMBER
N/A

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Bruce Maggs

5d. PROJECT NUMBER
BYU1

5e. TASK NUMBER
DU

5f. WORK UNIT NUMBER
K2

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Duke University
Office of Research Administration
2200 W. Main St STE 710
Durham, NC 27705-4677

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory/RIG
525 Brooks Road
Rome NY 13441-4505

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/RI

11. SPONSOR/MONITOR'S REPORT NUMBER
AFRL-RI-RS-TR-2014-115

12. DISTRIBUTION AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited. This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This effort developed techniques for building better IP geolocation systems. Geolocation has many applications, such as presenting advertisements for local business establishments on web pages to debugging network performance issues to attributing attack traffic to country of origin. The effort developed a prototype geolocation database called Alidade. Like other geolocation databases, Alidade precomputes location estimates for all of IP space. Alidade, is fundamentally different from previous systems described in the academic literature, however, because it computes predictions for the entire IP address space and does not issue any measurement probes of its own either before or after it is presented with queries.

15. SUBJECT TERMS

IP address space, IP geolocation

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			**ROBERT L. KAMINSKI**
U	U	U	UU	24	19b. TELEPHONE NUMBER (*Include area code*) **N/A**

Table of Contents

List of Figures

1.0 Summary

This effort developed techniques for building better IP geolocation systems. Geolocation has many applications, such as presenting advertisements for local business establishments on web pages to debugging network performance issues to attributing attack traffic to country of origin. The effort developed a prototype geolocation database called Alidade. Like other geolocation databases, Alidade precomputes location estimates for all of IP space. Alidade, is fundamentally different from previous systems described in the academic literature, however, because it computes predictions for the entire IP address space and does not issue any measurement probes of its own either before or after it is presented with queries.

2.0 Introduction

During the period of this contract, the PI worked to develop techniques for building better IP geolocation systems. These systems accept queries of the form, "Where is 128.2.205.42?" and then provide predictions, such as, "128.2.205.42 is in Pittsburgh, Pennsylvania." Geolocation has many applications, such as presenting advertisements for local business establishments on web pages to debugging network performance issues to attributing attack traffic to country of origin. Geolocation systems generally fall into two categories. Commercial systems provide precomputed address-to-location mappings for all IP addresses. We refer to such systems as geolocation databases. Upon presenting a geolocation database with a target IP address, a location estimate is provided immediately. Almost all systems reported in the academic literature, on the other hand, employ active measurements, issuing probes to a target after it has been specified, but before estimating the location of the target. These systems use constraints derived from the measurements to improve the accuracy of their predictions. Both approaches have their advantages. The active measurement approach may be more accurate, while the geolocation database approach is not intrusive and can answer queries quickly, even when off-line.

For the past several years, the PI has worked to develop a prototype geolocation database called Alidade. Like other geolocation databases, Alidade precomputes location estimates for all of IP space. Indeed, using the available constraints,

Alidade computes a joint solution for all addresses. Alidade, is fundamentally different from previous systems described in the academic literature, however, because it computes predictions for the entire IP address space and does not issue any measurement probes of its own, either before or after it is presented with queries. Instead, Alidade fuses available data sets of various types, attempting to resolve conflicts in the data and to find mutually compatible solutions for all addresses.

Commercial geolocation databases also provide precomputed answers for all IP addresses. Like Alidade, the commercial products do not issue any probes when presented with geolocation queries. Alidade competes head-to-head with these databases, and, as Figure 1 shows, outperforms even the best of them on a large ground-truth data set provided by a Tier-1 ISP. We compare and contrast Alidade's geolocation accuracy with that of six other geolocation database systems: *EdgeScape (ES)*, *MaxMind GeoCity (MM)*, *MaxMind GeoCity2 Lite (MML)*, *DB-IP (DBIP)*, *IP2Location (IP2L)*, and *IPligence (IPLG)*. The systems were presented with 100,000 targets sampled uniformly at random from the ground-truth data set. Figure 1 shows the error distance (in km) on a log-scale along the x-axis and the Empirical Cumulative Distribution Function (ECDF) of these errors along the y-axis; we define error distance as the distance between the point-based prediction made for a target address and its ground-truth location. Alidade outperforms the other six systems with 79% of its targets geolocated to within a 10 km error. Because the exact methods used to compile the commercial databases are proprietary, we do not know for certain why Alidade is more accurate.

No single source of input data suffices on its own to make good predictions. The data sets ingested by Alidade include latency and path measurements collected for other purposes, e.g., traceroute data from iPlane [14] and CAIDA's Archipelago (Ark) measurement infrastructure [3], and client-server round-trip times measured by a Content Delivery Network (CDN). Alidade also relies on a tool called *HostParser* that translates domain names into geographical locations, much as the Undns tool [19] does. To provide coverage over the entire IP address space, Alidade leverages data from the Internet registries too. The extent to which the registry entry for an address is trusted is mitigated by the position of the corresponding Autonomous System (AS) in the AS hierarchy produced by CAIDA [4].

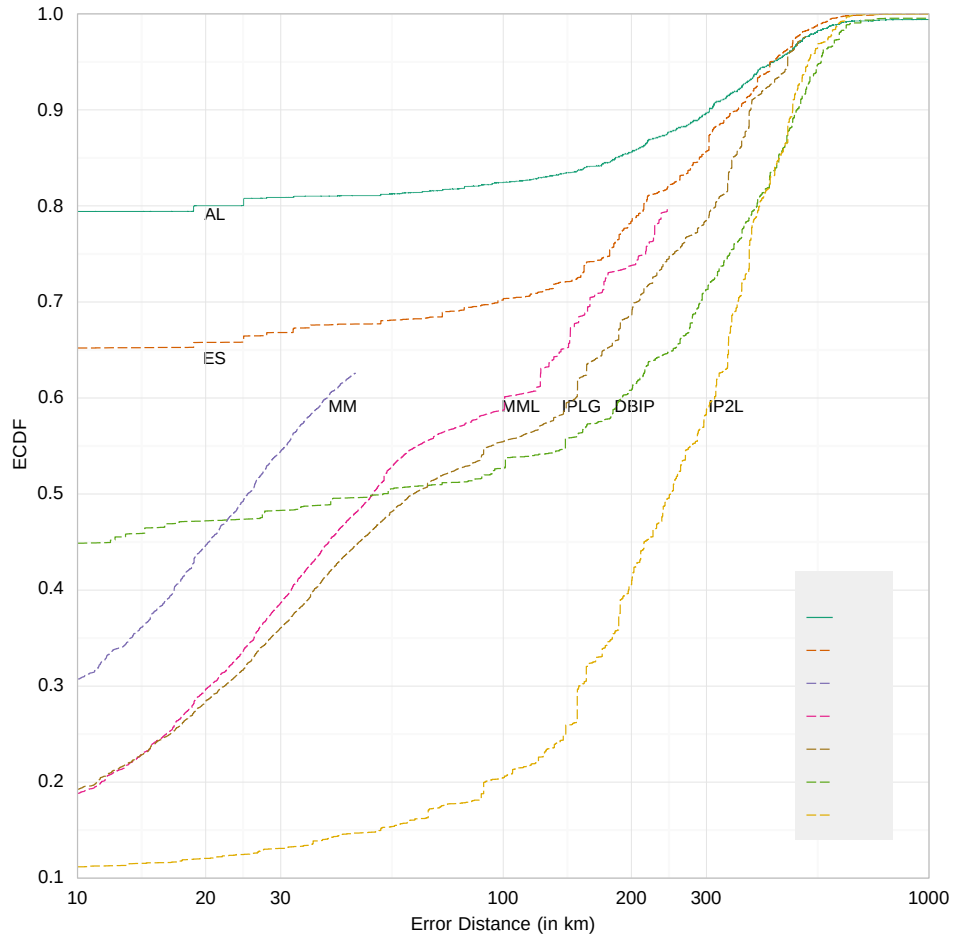

Figure 1: Comparison of Alidade's geolocation accuracy with six commercial geolocation databases

At its core, Alidade is a constraint-based *passive* geolocation system, inspired by Octant [21], but able to incorporate a wider variety of non-measurement data sources. Alidade uses latency measurements only when they are issued from hosts with known geographical locations, e.g., PlanetLab nodes. We call these hosts and/or their IP addresses *landmarks*. Alidade's estimate of the location of an address with an unknown location, which we call a *target*, is represented as a *polygonal* region on the surface of the Earth that should (if the prediction is correct) contain the address. The predictions made by commercial geolocation systems, in contrast, generally consist of a single latitude-longitude point or the name of a city or country. To facilitate a comparison with these systems, Alidade selects a single point to represent the polygon region.

Figure 2 shows an example of an answer region computed by Alidade. The region bounded by the dark green line represents the area resulting from intersecting constraints derived from latency measurements. In this example the intersection happens to be a circular region. The polygon in blue is a country-level hint (Germany) inferred from one of the Internet registries. Since the registry data does not conflict with the constraints derived from the measurements, Alidade uses it to further refine its prediction. In this example, Alidade has also identified a city-level hint (Kaiserslautern, a district in the Rhineland-Palatinate state of Germany) by examining the names of the routers on a traceroute path to the target. The city-level hint is indicated in the figure by the tiny red polygon inside the larger blue one. Ultimately Alidade pins the target in this demonstration to Kaiserslautern, which is consistent with the ground truth location of the target.

Figure 2: Example of a prediction made by Alidade for a target.

To process large volumes of data, Alidade is structured as a map-reduce (Hadoop) application. (Indeed, we started by porting Octant to Hadoop.) We conducted our experiments using a cluster of 40 8-core servers, each with 32GB of RAM. Each component of Alidade exhibits "embarrassing parallelism" and is implemented as a map-reduce job. In a later section we provide a breakdown of where the Alidade application spends most of its time, e.g., in "preprocessing" measurement data.

3.0 Related Work

Past work on IP geolocation can be loosely categorized into *active* approaches that perform on-demand network measurements to derive constraints on a target's geographic location, and *passive* approaches that rely only on previously collected information to geolocate a target. Both approaches have advantages and disadvantages. Active approaches may be more accurate, but predictions may not be available until new measurements have been taken. Passive approaches can precompute predictions and hence answer queries immediately, without even requiring network access at query time. Importantly, passive approaches are also unobtrusive, and do not risk alerting or annoying the target of a prediction. But passive approaches may not have the target-specific measurement data that would enable better accuracy.

Alidade takes a passive geolocation approach, but Alidade does not rely exclusively on coarse-grained and potentially error-prone data, such as the WHOIS database and hostname-to-location hints. Instead, Alidade filters the hints provided by these data sets by applying constraints derived from large volumes of passively collected network measurements.

In the following sections we examine both active and passive approaches, noting where Alidade borrows techniques.

3.1 Active Approaches

Much of the prior work in geolocating IP addresses relies on on-demand network measurements. *IP2Geo* [17] is an early IP geolocation system that introduces two active IP geolocation techniques. The first technique is *GeoPing*, which requires a deployment of landmarks of known geographic locations that can perform all-pairs latency measurements. To predict the location a target, all landmarks probe the target. *GeoPing* then selects the landmark that has the most similar latency profile (the set of latency measurements from other landmarks) to the user-specified target. It then uses the landmark's location as the prediction for the target. Although this technique is simple and easy to deploy, the location of a target cannot be accurately predicted unless there is a landmark nearby and that landmark has a similar latency profile. At present, Alidade doesn't compile latency profiles or compare the latency profiles of targets and landmarks. The second technique is *GeoTrack*, which performs traceroutes from landmarks to the target to discover routers on the traceroute paths whose DNS names can be interpreted geographically. From this set of routers, GeoTrack locates the target at the closest router's location, where distance is determined in terms of estimated network latency. Alidade's "extrapolator" applies a variation of this technique. By relying only on this relatively incomplete data source, however, GeoTrack's geolocation accuracy is inconsistent.

In contrast to locating the target at the closest landmark or router, Constraint-Based Geolocation (CBG) [9] determines the location of a target by creating circles on the surface of the earth around each landmark, where each circle represents a constraint that bounds the possible location of the target. The size of each circle is a function of the latency between the landmark and target. CBG combines constraints by

intersecting the circles, and selects the middle of the intersection as its best estimate of the target's location. One risk in taking this approach is that a single corrupt measurement can lead to an empty intersection. At its core, Alidade is a CBG approach.

Octant [21] builds on CBG by providing a general framework that can combine both positive and negative constraints, that is, information on where the target is likely and unlikely to be, respectively. To handle uncertain or error-prone data sources, Octant combines constraints using a weight-based mechanism that can limit the impact of erroneous measurements. Alidade builds on the Octant framework. In order to process large volumes of measurement data and to geolocate all of the IP address space, Alidade restructures the framework into a parallel Hadoop application so that more memory and compute cycles can be applied.

Topology-Based Geolocation (TBG) [12] uses traceroutes from the landmarks to the target to discover the routers along the network paths and determine inter-router latencies. With this data, TBG performs a global optimization to find a physical placement of the routers and the target that minimizes inconsistencies with the network latencies. By attempting to globally optimize the placement of both the routers and the target, TBG is more sensitive to measurement errors, such as inflated latencies, than constraint-based solutions, where errors tend to be more localized. To some extent Alidade applies this approach too. In particular, Alidade uses all available estimated latencies between pairs of addresses (landmarks, routers, and end hosts) to jointly predict the locations of the routers and end hosts.

Several systems [7, 22, 2, 13] have applied statistical approaches to construct landmark-specific functions that map measured latencies to geographical distances. These systems generally have significant computational requirements, and are currently unable to make use of non-latency-based constraints. *Posit* [6] presents a more recent statistical approach that, while still requiring active measurements, is able to significantly reduce the required number of on-demand probes by precomputing a statistical embedding. At present, Alidade does not construct a sophisticated model of the relationship between latency and distance. Instead, Alidade uniformly assumes that datagrams travel at two-thirds the speed of light, which is very close to the speed of light in optical fiber. Hence, in

converting latency to distance, Alidade does not model circuitous fiber paths, nor does it model queuing delays or any other sorts of delays. The resulting constraints tend to be loose, but they are also hard. In particular, provided that no measurements are corrupt and no faster-than-fiber technologies, such as microwave transmission, are employed, the intersection of a set of constraints derived by Alidade from direct latency measurements must contain the actual location of the target. Other work has suggested that if latency is to be converted to distance by a simple multiplicative factor, four-ninths the speed of light might be used. The smaller constant leads to smaller intersection areas, but these areas might be empty or might not contain the target.

Guo et al. [10] propose mining physical addresses displayed on publicly accessible Web sites that are hosted by Web servers with IP addresses in the same prefix as the target address, and using these physical addresses as hints to improve geolocation accuracy and as sources of ground truth to support evaluations. Caruso [5] (as part of the Alidade project) and Wang et al. [20] extend this approach by combining the mined information with latency measurements to offer finer-grained geolocation results. Although these systems produce accurate results in certain experiments, it is difficult to ascertain their actual effectiveness in general. First, it is tricky to determine when an organization is hosting its own Web site. Furthermore, even when an organization does host its own site, for the technique to work the site must list a physical address that is close to that of the hosting location. In previous experiments the best results were obtained when the set of geolocation targets were biased towards belonging to organizations that typically host their own Web servers and publish physical address information on their web pages, e.g., in one experiment reported in [20], university Web servers hosting Web pages listing campus addresses were used as landmarks and PlanetLab nodes were used as targets. Nevertheless, scraped address information from locally-hosted Web sites is a rich source of geographic data, and Alidade includes this information as one of its many data sources.

Gill et al. [8] propose two broad classes of attacks on active measurement-based geolocation approaches. The first misleads geolocation systems by injecting delays to latency probes from specific landmarks at the target, thereby altering the geolocation result by moving the centroid of the constraint intersection in a

CBG-based approach. The second targets topology-aware geolocation approaches by altering inter-router latencies in traceroutes, which enables powerful adversaries to place geolocation targets at arbitrary locations. Alidade does not attempt to detect possible adversaries. Unlike active approaches, however, where latency probes can often be easily identified, Alidade also uses a large body of passively collected measurements that piggyback real user TCP connection requests and replies. Adversaries must therefore delay legitimate TCP traffic rather than just latency probes in order to distort much of Alidade's input data.

3.2 Passive Approaches

Although active geolocation approaches can be highly accurate, their dependence on performing on-demand network measurements make them unsuitable for many location-aware applications. Most commercial geolocation systems, such as *MaxMind GeoCity* [15], *EdgeScape* [1], *IPInfoDB* [11], and *HostIP.Info* [16] have instead adopted passive approaches, where they offer their users a pre-computed IP-to-location database that can identify a target's location without additional network access. Unfortunately, the exact methodology for creating these databases are generally proprietary; only the expected accuracy of these databases are typically published. However, the common understanding is that these databases rely on a combination of domain registry information, ISP provided data, host name hints, latency measurements, and other heuristics. Alidade relies on many of the same sources, except that the ISP-supplied ground-truth geolocation data (from one Tier-1 ISP) is used only for evaluation purposes and not as an input to Alidade.

Poese et al. [18] performs an analysis of the accuracy of commercial geolocation databases. They report that while geolocation databases are extremely accurate at the country level, they perform poorly at the city level. Note that Poese et al. did not analyze EdgeScape (or Alidade).

In addition to GeoPing and GeoTrack, IP2Geo [17] also introduces *GeoCluster*, a passive approach that partitions the IP address space into geographically co-located clusters. GeoCluster then assigns each cluster to a geographic location based on the geographic information extracted from user

registration and usage databases. The effectiveness of this approach is largely limited by the availability of such databases, the geographic coverage of the users in the databases, and the accuracy and freshness of the self-reported user location information. At present, no such data is available to us, but if it were, it could be used as an input to Alidade.

4.0 Evaluation

During the period of the contract, we were able to evaluate the performance of the Alidade prototype by comparing its answers with that of six commercial geolocation databases – *EdgeScape (ES)*, *MaxMind GeoCity (MM)*, *MaxMind GeoCity2 Lite (MML)*, *DB-IP (DBIP)*, *IP2Location (IP2L)* and *IPligence (IPLG)*. We used the latest versions, updated in September 2013, of all the databases except for MaxMind GeoCity, for which the last update available to us was made in early June 2013. This is one of the reasons that we have included two databases from the same provider in our study. MaxMind GeoLite2 City, the free version of MaxMind, has also been widely used in academic research for evaluation of geolocation systems.

The database of IP-address-to-location mappings generated by Alidade was generated from a set of input data sets that included both measurement and non-measurement data. The non-measurement data consisted of HostParser hints for approximately 700 million addresses, of which roughly 207 million contain city-level predictions, location hints compiled from various Internet registries, AS hierarchy data from CAIDA, ground-truth locations of landmarks, and shape files for cities and countries along with accompanying metadata. Much of the measurement data for the experiment was provided by a Content Delivery Network (CDN) and consisted of traceroutes between CDN servers and hundreds of thousands of resolving DNS servers collected over a period of three months (recorded by the CDN for network mapping purposes), traceroutes from CDN servers to a small fraction of end user addresses collected over a period of six months, one week of ping measurements from CDN servers to routers (recorded by the CDN to estimate network performance), and one month of round-trip latency values recorded between CDN servers and end-user machines for a small fraction of TCP connections. The database of results created using these measurement and non-measurement inputs was used as input to the querying

engine to geolocate the targets in the evaluation data set. The selection of targets for evaluation was performed after Alidade's database was finalized; Alidade's results had no influence on selection of targets for the performance comparison.

The ground-truth data used for the evaluation is a list of city locations for approximately 24 million IP addresses provided by a European Tier-1 network provider. We refer to this dataset as *EuroGT*. One peculiarity of this data set is that it contains only 73 distinct city locations, although presumably this provider has infrastructure in more than 73 cities.

We define error distance as the geographic distance between a system's point-based prediction for a target and the target's ground-truth location. Although, Alidade outputs polygonal regions as answers, it also computes a point-based estimate, which is always contained in the polygonal region. This enables a head-to-head performance comparison of Alidade with the other geolocation databases, all of which provide point-based predictions. Alidade uses various heuristics to output a point-based answer. Picking the center of a city enclosed by the polygonal answer, is an example of such a heuristic.

We begin by analyzing the effectiveness of relying solely on hints derived from the registry or from the names of the target addresses. These are the primary sources of non-measurement data used by Alidade. Figure 3 shows the ECDFs of errors for the complete 24-million-address EuroGT dataset (Plots use log-scale for the x-axis, unless mentioned otherwise)

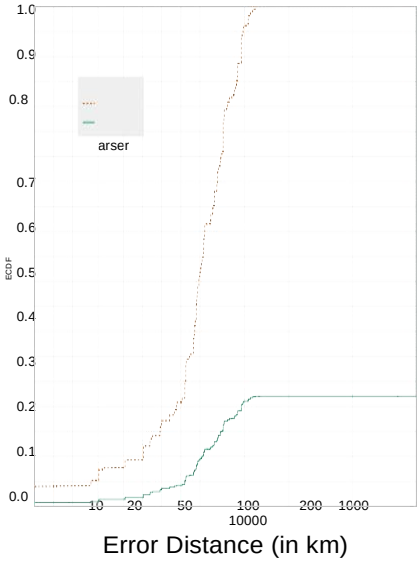

Figure 3: EuroGT: Using just registry or HostParser

using only HostParser or registry. HostParser provides answers to just a little over 20% of the targets; for targets with no answers (approximately, 18 million) we assumed an error distance of 10,000km. Registry, by comparison, performs better, with a median error distance of 214km. The results indicate that these two data sources alone are not sufficient to make accurate predictions.

For the comparison of Alidade with commercial geolocation databases we selected a set of 100,000 targets uniformly at random from the EuroGT dataset and for each target computed the error distance for Alidade and for each of the others. Figure 1 presents the ECDFs of error distance for each database. Since the ground truth for the EuroGT dataset is only at the city level, we begin the ECDF plots at an error distance of 10km. Alidade (AL) outperforms the other geolocation databases with 79% of targets located with an error of 10km or less. Akamai's EdgeScape (ES) provides the best results from among the commercial databases.

To gauge the importance of measurement data, we compare ECDFs for those targets for which any kind of measurement data is available (e.g., the target appeared on a traceroute path) with

those for which no measurement data is available, as shown in Figure 4. In this figure, the curve for the targets with measurements is labeled *meas*, while that for targets without measurements is labeled *nomeas*. Of the 100,000 targets, measurement data was available for 80,764 targets, while no measurement data was available for the remaining 19,236 targets The plot confirms the hypothesis that it helps to have measurements in addition to data from the registries or HostParser. Alidade records what information was actually used to compute the polygon region representing Alidade's prediction for the location of any target. Taking advantage of this feature, we categorized targets based on which datasets and techniques used to predict their locations. Figure 5 shows the ECDFs of a few such categories. The curve *meas+hp+reg+ext* with 5225 (5%) of the targets represents the set of targets that benefited from the use of HostParser, registry, and extrapolator hints, in addition to having measurements. The *meas+hp+reg* and *meas+reg+ext* ECDFs represent similar ECDFs with extrapolator hints unavailable in the former and HostParser unavailable in the latter; they account for 3.5% and 19% of the sample, respectively.

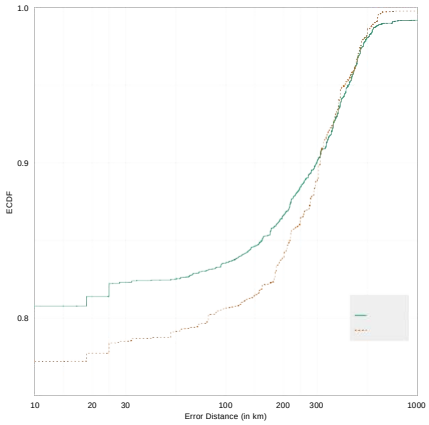

Figure 4: Impact of measurements on accuracy

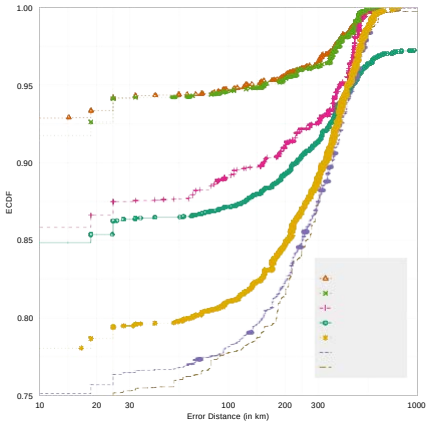

Figure 5: Impact of non-measurement data on accuracy

Curiously, the category that dominates the rest on an absolute scale, shown as *hp*-ECDF, is the set of targets for which only a HostParser hint was used. This category contains 5931 (6%) of the targets. This finding reinforces our intuition that we can treat HostParser hints with high confidence. Note that for the set of 100,000 targets, there were only 167 targets for which the only information used was a HostParser hint and a measurement. All of the other targets with both a HostParser hint and a measurement also had other hints.

The remaining curves are explained as follows. Targets geolocated using just extrapolator, the *ext*-ECDF, account for another 19% of the 100,000 IPs chosen. The curve *meas.misc* with 18% of the sample refers to targets that had measurements and maybe hints from other sources. The remaining targets (29% of the sample) that had no measurements for geolocation are represented by the *others*-CDF.

5.0 Results

During the period the PI worked to evaluate Alidade against ground-truth data sets other than the European Tier-1 ISP data set. These data sets include the locations of networked GPS receivers distributed throughout the world as part of an

experiment on continental drift, the locations of network time protocol servers (NTP servers) that report their coordinates, the locations of servers belonging to the PlanetLab system, and the locations of servers belonging to the MLab system.

As shown in Figures 6, 7, 8, and 9, Alidade generally outperforms the commercial geolocation systems on these ground-truth data sets, with a few exceptions.

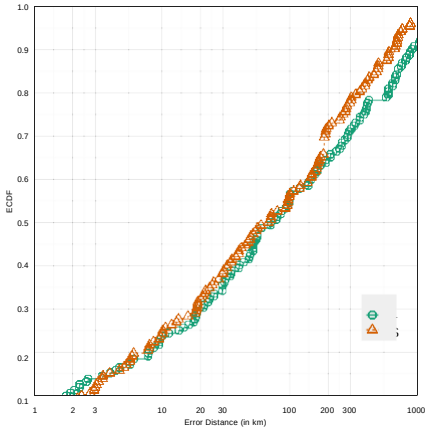

Figure 6: Performance on GPS ground-truth data set

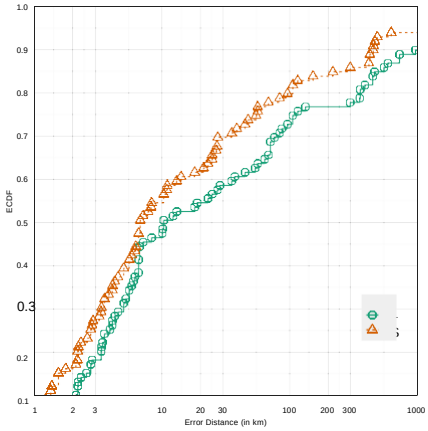

Figure 7: Performance on NTP server ground-truth data set

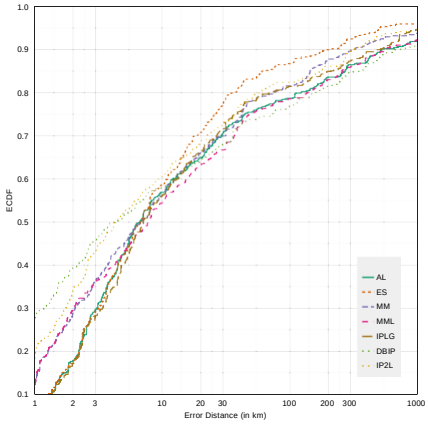

Figure 8: Performance on Planet-Lab ground-truth data set

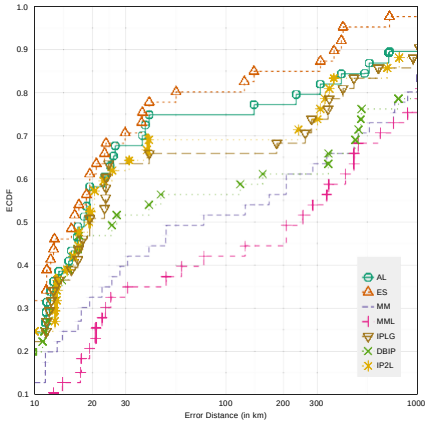

Figure 9: Performance on MLab ground-truth data set

In studying the instances in which Alidade made incorrect predictions, two deficiencies were discovered in Alidade's algorithms. First, the extrapolator was too aggressive in applying hints derived from router names on the paths to end hosts. We plan to experiment with allowing a hint from a router name only if the router lies in the same autonomous system (AS) as the target. Second, it was discovered that landmarks (locations for which we know the ground truth) were not being used by the aggregator. i.e., when evaluating a prefix that contains a target for which no measurements are available, the aggregator does not consider landmarks that are also contained in the prefix. We plan to experiment with a change to aggregator to include landmarks in its prediction algorithm.

6.0 Conclusions

This effort developed techniques for building better IP geolocation systems. Geolocation has many applications, such as presenting advertisements for local business establishments on web pages to debugging network performance issues to attributing attack traffic to country of origin. The developed system, Alidade, is fundamentally different from previous systems described in the academic literature. It computes predictions for the entire IP address space and does not issue any measurement probes of its own, either before or after it is presented with queries. Active measurement approaches may be more accurate however, the geolocation database approach developed is not intrusive and can answer queries quickly, even when off-line.

7.0 References

[1] Akamai Technologies, Inc. EdgePlatform. http://www4.akamai.com/ html/technology/products/edgescape.html, 2013

[2] M. J. Arif, S. Karunasekera, and S. Kulkarni. GeoWeight: Internet Host Geolocation Based on a Probability Model for Latency Measurements. In *Proceedings of the Thirty-Third Australasian Conferenc on Computer Science - Volume 102*, ACSC '10, pages 89–98, Darlinghurst, Australia, Australia, January 2010. Australian Computer Society, Inc.

[3] CAIDA. Archipelago measurement infrastructure. http://www.caida. org/projects/ark/, 2013.

[4] CAIDA. AS Rank: AS Ranking. http://as-rank.caida.org/, 02/19/2013 2013.

[5] Nicole Caruso. A Distributed System For Large-Scale Geolocalization Of Internet Hosts. diploma thesis, Cornell University, Ithaca, NY, 2011.

[6] Brian Eriksson, Paul Barford, Bruce Maggs, and Robert Nowak. Posit: a lightweight approach for IP geolocation. *SIGMETRICS Perform. Eval. Rev.*, 40(2):2–11, October 2012.

[7] Brian Eriksson, Paul Barford, Joel Sommers, and Robert Nowak. A Learning-Based Approach for IP Geolocation. In *Proc. of the 11th International Conf. on Passive and Active Measurement*, PAM'10, pages 171–180, Berlin, Heidelberg, April 2010. Springer-Verlag.

[8] Phillipa Gill, Yashar Ganjali, Bernard Wong, and David Lie. Dude, where's that IP? Circumventing measurement-based IP geolocation. In *Proceedings of the 19th USENIX conference on Security*, USENIX Security'10, pages 16–16, Berkeley, CA, USA, 2010. USENIX Association.

[9] Bamba Gueye, Artur Ziviani, Mark Crovella, and Serge Fdida. Constraint-Based Geolocation of Internet Hosts. In *ACM Internet Measurement Conference*, Taormina, Sicily, Italy, October 2004.

[10] Chuanxiong Guo, Yunxin Liu, Wenchao Shen, H.J. Wang, Qing Yu, and Yongguang Zhang. Mining the web and the internet for accurate ip address geolocations. In *INFOCOM 2009, IEEE*, pages 2841–2845, 2009.

[11] IP2Location.com. IPInfoDB. http://www.ip2location.com/, 2013. [12] Ethan Katz-Bassett, John P. John, Arvind Krishnamurthy, David Wetherall, Thomas Anderson, and Yatin Chawathe. Towards IP Geolo-cation Using Delay and Topology Measurements. In *Proceedings of the6th ACM SIGCOMM conference on Internet measurement*, IMC '06, pages 71–84, New York, NY, USA, October 2006. ACM.

[13] S. Laki, P. Mátray, P. Hága, T. Sebok, I. Csabai, and G. Vattay. Spot-

ter: A Model Based Active Geolocation Service. In *IEEE INFOCOM*, April 2011.

[14] Harsha V. Madhyastha, Tomas Isdal, Michael Piatek, Colin Dixon, Thomas Anderson, Arvind Krishnamurthy, and Arun Venkataramani. iPlane: An Information Plane for Distributed Services. In *OSDI '06: Proceedings of the 7th symposium on Operating systems design and implementation*, pages 367–380, Berkeley, CA, USA, 2006. USENIX Association.

[15] MaxMind, Inc. GeoIP City. http://www.maxmind.com/en/geolocation_landing, 2013.

[16] Net Industries, LLC. hostip.info. http://www4.akamai.com/html/technology/products/edgescape.html, 2013.

[17] Venkata N. Padmanabhan and Lakshminarayanan Subramanian. An Investigation of Geographic Mapping Techniques for Internet Hosts. In *Proceedings of ACM SIGCOMM conference*, San Diego, CA, USA, August 2001.

[18] Ingmar Poese, Steve Uhlig, Mohamed Ali Kaafar, Benoit Donnet, and Bamba Gueye. IP Geolocation Databases: Unreliable? *SIGCOMM Comput. Commun. Rev.*, 41(2):53–56, April 2011.

[19] Neil Spring, Ratul Mahajan, David Wetherall, and Thomas Anderson. Measuring ISP Topologies With Rocketfuel. *IEEE/ACM Trans. Netw.*, 12:2–16, February 2004.

[20] Yong Wang, Daniel Burgener, Marcel Flores, Aleksandar Kuzmanovic, and Cheng Huang. Towards Street-Level Client-Independent IP Geolocation. In *Proceedings of the 8th USENIX conference on Networked systems design and implementation*, NSDI'11, pages 27–27, Berkeley, CA, USA, April 2011. USENIX Association.

[21] Bernard Wong, Ivan Stoyanov, and Emin Gün Sirer. Octant: A Comprehensive Framework for the Geolocalization of Internet Hosts. In *NSDI*, April 2007.

[22] Inja Youn, Brian L. Mark, and Dana Richards. Statistical Geolocation of Internet Hosts. In *ICCCN*, pages 1–6, 2009.

List of Acronyms

AS	Autonomous System
CBG	Constraint-Based Geolocation
CND	Content Delivery Network
DNS	Domain Name Server
ECDF	Empirical Cumulative Distribution Function
GB	Gigabyte
IP	Internet Protocol
NTP	Network Time Protocol
PI	Principal Investigator
RAM	Random Access Memory
TBG	Topology-Based Geolocation
TCP	Transmission Control Protocol