

Posit: An Adaptive Framework for Lightweight IP Geolocation

Brian Eriksson

Department of Computer Science
Boston University
eriksson@cs.bu.edu

Paul Barford

Department of Computer Science
University of Wisconsin and Qualys
pb@cs.wisc.edu

Bruce Maggs

Department of Computer Science
Duke University and Akamai
bmm@cs.duke.edu

Robert Nowak

Department of Electrical and Computer Engineering
University of Wisconsin
nowak@ece.wisc.edu

July 11, 2011

BU/CS Technical Report 2011

Abstract

Location-specific Internet services are predicated on the ability to identify the geographic position of IP hosts accurately. Fundamental to prior geolocation techniques is their reliance on landmarks with known coordinates whose distance from target hosts is intrinsically tied to the ability to make accurate location estimates. In this paper, we introduce a new lightweight framework for IP geolocation that we call *Posit*. The *Posit* framework geolocates by automatically adapting to the geographic distribution of the measurement infrastructure relative to each target host. This lightweight framework requires only a small number of Ping measurements conducted to end host targets in conjunction with a computationally efficient geographic embedding methodology. We demonstrate that *Posit* performs significantly better than all existing geolocation tools across a wide spectrum of measurement infrastructures with varying geographic densities. *Posit* is shown to geolocate hosts with median error improvements of over 50% with respect to all current measurement-based IP geolocation methodologies.

1 Introduction

There are many ways in which the structure and topology of the Internet can be characterized. The router-level topology (*e.g.*, [1]), autonomous system-level topology (*e.g.*, [2]), end-to-end path characteristics (*e.g.*, [3]), and latencies between hosts (*e.g.*, [4]) have all been examined extensively in prior work. One characteristic of Internet structure that has significant implications for advertisers, application developers, network operators, and network security analysts is to identify the geographic location, or *geolocation*, of networked devices, such as routers or end hosts.

The ultimate goal of IP geolocation is to find the precise latitude/longitude coordinates of a target Internet device. There are considerable challenges in finding the geographic location of a given end host in the Internet. First, the size and complexity of the Internet today, coupled with its highly diffuse ownership means that there is no single authority with this information. Second, no standard protocol provides the geographic position of an Internet device on the globe (although domain names can include a location record). Third, non-mobile Internet devices are not typically equipped with location identification capability (*e.g.*, GPS), although this may change in the future. However, even GPS-equipped devices may choose not to report location information due to privacy concerns. Finally, measurement-based geolocation can be confused by NAT'ed devices or by users who are trying to anonymize their communications [5].

IP geolocation methods that are currently used largely fall into two categories. The first is database-specific approaches in which a geolocation database is established by examining network address space allocations and user-entered geographical data. While this can be effective for providers who offer service in a restricted geographic region (*e.g.*, a university or a small town), it will fail for providers with a large geographic footprint unless coupled with additional information. The second method is to use active probe-based measurements to place the target host within some specified geographic region. The accuracy of these probe-based techniques is intrinsically dependent on the

geographic proximity of target hosts and the measurement infrastructure. The result is geolocation estimates with relatively high median error and high error variability when measurement resources are geographically distant to a target host.

This paper proposes a novel approach to IP geolocation that we call *Posit*. The Posit framework considers three categories of devices in the network.

- *Monitors* - The set of network resources with known geographic location and the ability to send Ping measurements to both landmarks and targets.
- *Landmarks* - The set of hosts in the network with known and very accurate geolocation information. These nodes respond to probes, but we do not require the ability to send probes from the landmarks.
- *Targets* - The set of hosts with unknown geographic location that we aim to geolocate. Ideally, these nodes respond to probes, but we do not require this ability.

The first goal of our work is to develop a measurement-based IP geolocation framework that provides highly accurate estimates and significantly reduces estimation error over prior methods. We achieve this through a framework that automatically adapts to the geographic distance of monitors and passive landmarks relative to the target being geolocated. This is in contrast to prior studies that focus on a single dataset with unspecified levels of monitor/landmark distance to target end hosts. The second goal of our work is to develop a geolocation framework that is lightweight in the sense that it only relies on a small number of measurements in order to establish location estimates.

We begin by demonstrating a component of the Posit framework that relies solely on observed hop count distances. This hop information can be found from either lightweight Ping-like measurements or from passively observed network traffic, which require no additional probes injected into the network. Hop count-based geolocation is critical for target end hosts that block probes and are therefore incapable of being geolocated using prior techniques. Hop-based Posit adapts to the geographic density of the measurement infrastructure by only returning the subset of target hosts with highest geolocation confidence.

For end hosts that respond to probes, we demonstrate latency-based Posit, where we rely on measured latencies from monitors to our targets. To estimate geographic location, Posit uses a new statistical embedding process that encourages the targets to cluster geographically without the need for explicitly defined population or geographic data. Critical to our embedding algorithm is the ability to estimate the probability of geographic distance between targets and our set of landmarks without any direct latency measurements required between the two sets of Internet hosts. This significantly reduces the network load of the Posit methodology in contrast to other geolocation techniques that exploit passive landmarks (*e.g.*, [6]).

We examine and validate our Posit geolocation framework on two separate data sets. First using 431 commercial end hosts, and second using 283 domain names with LOC records that have been validated. It is important to note that the exact coordinates of *all* hosts used in this study were known, which gave us a strong foundation for evaluation. By thoroughly exploring the design space, our results demonstrate significant improvements using the Posit framework over all competing geolocation techniques. These results also demonstrate the necessity of comparing the various prior geolocation methodologies on the same measurement infrastructure, since measurement infrastructure density plays a central role in the variation of geolocation accuracy for all techniques. Specifically, our results show that simply relying on prior published error metrics will not accurately characterize the true performance of geolocation methodologies.

Our analysis of the hop-based Posit technique shows that over 20% of targets can be geolocated with median error of less than 200 miles. Using incomplete passive measurements, we find that only 25 passively observed hop counts are required to resolve 20% of targets with median error of 275 miles. It is important to note that these results show how nodes that are *impossible* to map with any other measurement-based method can be geolocated with Posit.

For the latency-based Posit methodology, across a wide spectrum of geographic densities on a commercial node dataset (in terms of the geographically closest monitor/landmark), Posit returns geolocation estimates with median error of only 26 miles. These datasets include a significant subset of targets that are hundreds of miles away from the nearest measurement infrastructure resource, a critical regime when considering non-US geolocation. In comparison with the best competing prior geolocation techniques on the same measurement datasets, we see improvements of over 50%. Of particular importance is that, to the best of our knowledge, this is the first study that examines varying the measurement infrastructures in terms of distance to our geolocation targets. Across all considered infrastructure regimes, we find Posit outperforming the prior geolocation methodologies with median geolocation error reductions

ranging from 33% to 75% over the best competing methodology. Our results show the critical importance of considering these varying infrastructures in validation of geolocation performance.

The paper is organized as follows. In Section 2, we review prior studies that inform our work. The datasets used for the experiments are described in Section 3. Section 4 details the two components of Posit. The first component, geolocation to landmark locations based on hop count proximity is described in Section 5 with results on our geolocation dataset. Next, the latency-based Posit geolocation algorithm is developed in Section 6. Finally, the experimental performance of the latency-based Posit methodology is explored in Section 7, with the conclusions of the paper in Section 8.

2 Related Work

Considerable prior work has been performed on the subject of IP geolocation [6, 7, 8, 9, 10]. We describe the details of the prior methods that we compare against in this paper in Section 7. While we are informed by this work and our motivation for highly accurate estimates is the same, the methodology described in this paper makes new contributions that reduce measurement load on the network and improve estimate accuracy over prior geolocation algorithms. Unlike the state-of-the-art Street-Level [6], Octant [7], and Topology-based [10] methodologies, no `traceroute` probes are necessary in the Posit methodology. In addition to significantly decreasing the network load, this avoids the well known problems of interface disambiguation [1, 11] and dependency on unreliable unDNS naming conventions [12].

Standard to all previous IP geolocation algorithms is the use of latency as a proxy for distance measurements. While some algorithms use latency measurements solely as an upper bound constraint on possible geographic locations [9], others have tried to directly estimate distance from the latency values (*e.g.*, the spline-based method of [7]). More recent work [13, 14, 15] has used nonparametric estimation of distance probability given observed latency. All of these prior methods are predicated on the observation of latency between the end host target with unknown geolocation and a monitor with known geolocation. In addition, the Posit methodology will transform these measurements to estimate the probability of distance between targets and landmarks without having any direct measurements between the two sets of Internet resources.

Similar to prior geolocation techniques (*i.e.*, [6, 8]), the Posit framework relies on the procurement of a large set of *passive landmarks* – Internet hosts with known latitude/longitude coordinates that respond to measurement probes. In contrast to prior work, Posit exploits the large collection of existing Internet infrastructure with publicly available geolocation information. Landmarks used by Posit in this paper are domain names with location information (via careful analysis of DNS LOC records [16]). To the best of our knowledge, this is the first time that DNS has been used for IP geolocation. Posit does not rely exclusively on this resource, and in future empirical studies we expect to add other nodes to our landmark database (such as a set of stratum 0/1 Network Time Protocol (NTP) servers [17]). Indeed, the novel method for identifying additional landmarks described in [6] could be used in parallel with Posit to increase the measurement node density.

3 Datasets

To evaluate Posit, we use a set of measurements collected from 431 commercial hosts with known latitude/longitude coordinates that belong to Akamai Technologies in North America. During the weekend of January 16-17, 2010, pairwise bidirectional hop count and latency measurements were conducted on those nodes. The measurements were collected using standard ICMP ECHO requests via the MTR tool [18]. The servers are used in Akamai’s production CDN so during the measurement period, they may have been performing other tasks (such as serving HTTP content), which could have had an effect on latency measurements. In addition, we only consider a single latency measurement between each commercial host in order to minimize network load, which may introduce additional inaccuracies due to queuing delay.

We also consider a set of 283 domain names with valid DNS LOC records in the continental United States. The locations obtained by the DNS LOC records were verified using both commercial IP geolocation databases and verification that no resource violated latency speed-of-light constraints. Standard ICMP ECHO requests via the MTR tool were performed on January 22, 2011 from the set of 431 commercial hosts to all 283 domain names.¹ The geographic distribution of these resources is seen in Figure 1. The figure highlights the non-uniform geographic distribution of

¹The authors would like to thank Rick Weber and KC Ng from Akamai Technologies for supplying us this data.

landmarks, which we argue will also be the case for other types of landmarks derived from existing Internet infrastructure.

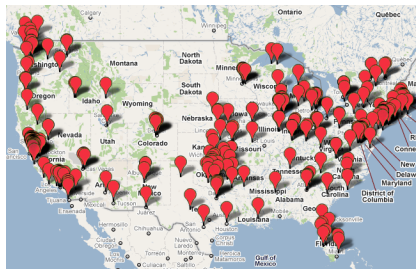


Figure 1: Geographic placement of our set of domain names with known geographic location.

While the size of our validation set may appear small compared with prior work [6, 13], in contrast to these larger prior studies we have ground truth knowledge for the location of all of our hosts under consideration. Also, we refer to ICMP probes throughout the paper as a means for gathering latency and hop count data, but recognize that ICMP is sometimes blocked or that hosts will not respond to those probes. ICMP probing in and of itself is not intrinsic to our ability to geolocate resources using Posit.

4 Posit Framework Summary

The Posit IP geolocation framework is divided into two distinct components.

1. *Hop-Based Posit Geolocation* - Using acquired hop count vectors and the known geolocation for a set of landmarks, we intelligently select a subset of targets (with sufficient confidence) that are geolocated by exploiting network topology structure.
2. *Latency-Based Posit Geolocation* - Using latency measurements from a set of monitors to both target end hosts and landmarks, we use our adaptive statistical embedding methodology to efficiently geolocate each target end host.

We describe both of these components in detail in the following sections. As stated previously, we developed both techniques, since latency-based methods will be infeasible for end host targets that block probes. It is important to note that the hop-based Posit methodology is also applicable for adversarial targets, which is a previously unexplored geolocation regime, but one with important implications for network security.

5 Hop-Based Posit Geolocation

Consider T landmarks in the network with known geographic location. Given a set of geographically diverse landmarks, it is intuitive that simply mapping each target to the geographically closest landmark may result in a highly accurate estimate. While this knowledge of geographically closest landmark is not known a priori, we claim that accurate estimates can be made by hop count measurements from the set of monitors to our set of landmarks and targets.

To generate hop counts, ICMP probes were sent from a set of M diverse monitors in the Internet (such as Planetlab nodes [19]). From these probes, we construct a set of *hop count vectors*.

For target $i = \{1, 2, \dots, N\}$,

$$\mathbf{h}_i^{target} = \begin{bmatrix} h_{i,1}^{target} & h_{i,2}^{target} & \dots & h_{i,M}^{target} \end{bmatrix}$$

Where $h_{i,k}^{target}$ is the observed hop count between target i and monitor k .

And for landmark $j = \{1, 2, \dots, T\}$,

$$\mathbf{h}_j^{land} = \begin{bmatrix} h_{j,1}^{land} & h_{j,2}^{land} & \dots & h_{j,M}^{land} \end{bmatrix}$$

Where $h_{j,k}^{land}$ is the observed hop count between landmark j and monitor k .

From previous work [20], it was shown that hop count vectors contain enough information to cluster targets in a topologically significant manner. This methodology relies on the existence of *border routers* from [21], where multiple targets can share the same egress router to the core of the network. If two target end hosts share the same border router, this implies that all paths to the network core from both targets will be shared past the egress point. Given this shared path property, we can determine that two targets (i, j) are topologically close in the network if they have the hop count property $h_{i,k} = h_{j,k} + C$ for every monitor k in the network located past the shared border router (for a constant integer C). A visual example of this can be seen in Figure 2.

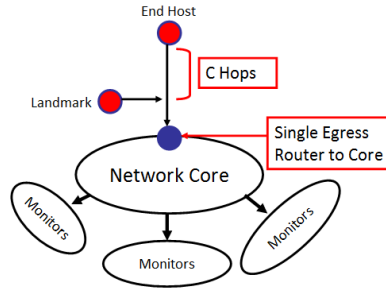


Figure 2: Example of network where a target end host is C hops away from a landmark, with both sharing the same border router.

We hypothesize that topologically close targets in this regime are also geographically close. Using the methodology from [20], we can geolocate each target (i) to the estimated topologically closest landmark (c_i) by finding the landmark such that the hop count difference vector has the smallest variance. Therefore, the *hop-based Posit methodology* consists of finding a landmark c_i such that,

$$c_i = \underset{j}{\operatorname{argmin}} (\hat{\sigma}^2 (\mathbf{h}_i^{target} - \mathbf{h}_j^{land})) \quad (1)$$

Where the sample variance,

$$\hat{\sigma}^2 (\mathbf{x}) = \frac{1}{N} \sum_{i=1}^N (x_i - \frac{1}{N} \sum_{j=1}^N x_j)^2$$

For two hop vectors separated by a constant integer offset (indicating a shared border node), then

$$\hat{\sigma}^2 (\mathbf{h}_i - \mathbf{h}_j) = \hat{\sigma}^2 (\mathbf{h}_i - (\mathbf{h}_i + C\mathbf{1}^T)) = \hat{\sigma}^2 (C\mathbf{1}^T) = 0$$

When no landmarks lie in a topologically close region with respect to a target, our methodology could potentially result in wildly inaccurate geolocation estimation. Fortunately, the hop difference variance between the target and the mapped landmark ($\hat{\sigma}^2 (\mathbf{h}_i^{target} - \mathbf{h}_{c_i}^{land})$) returns a powerful indicator of the geolocation accuracy of this hop-based method. To avoid erroneous estimates, we will exploit this estimated variance to resolve a subset of target hosts we are most confident in geolocation performance (a fraction of the targets, λ). The full hop-based Posit framework can be found in Algorithm 1.

To compare against our Posit hop-based geolocation methodology, we use two natural analogues from prior latency-based geolocation methodologies. The first will be a hop vector nearest neighbor approach, where the landmark with the hop count vector with the smallest Euclidean distance with respect to the target will be considered (*i.e.*, $\operatorname{argmin}_k \|\mathbf{h}_i^{target} - \mathbf{h}_k^{land}\|$). The second methodology will simply map the target to the monitor with the shortest hop distance (*i.e.*, $\operatorname{argmin}_j (h_{i,j}^{target})$). These two methodologies can be consider hop-based versions of the Shortest Ping and GeoPing algorithms [8], respectively.

Algorithm 1 - Hop-based Posit Framework

Given:

- Hop count vectors from the M monitors to our set of N targets, \mathbf{h}_i^{target} for $i = \{1, 2, \dots, N\}$.
- Hop count vectors from the M monitors to our set of T landmarks, \mathbf{h}_j^{land} for $j = \{1, 2, \dots, T\}$.
- Confidence fraction, $\lambda \in [0, 1]$.

Methodology:**For each target**, $i = \{1, 2, \dots, N\}$

- Find the closest inferred landmark, $c_i = \underset{j}{\operatorname{argmin}} (\hat{\sigma}^2 (\mathbf{h}_i^{target} - \mathbf{h}_j^{land}))$.
- Assign geolocation of target i as the location of landmark c_i .

Return the geolocation of the fraction of targets, λ , with the smallest hop variance.

Table 1: Average geolocation error for hop-based Posit using the Commercial Dataset with number of monitors, $M = 50$ and the fraction of targets λ at 20% and 40%.

Methodology	Mean Error (in miles)	Median Error (in miles)
Hop-Based Posit ($\lambda = 0.2$)	277.80	195.16
Hop-Based Posit ($\lambda = 0.4$)	360.13	223.47
Nearest Neighbor	640.00	486.21
Shortest Hop	1226.03	1165.07

5.1 Commercial Dataset Hop-Based Posit Results

Consider partitioning our commercial node host set from Section 3. For each target, we randomly selected 50 monitors and 150 landmarks from the remaining set of commercial nodes. Using only hop counts from the collection of monitor nodes, we estimate the geolocation of the targets using the hop-based Posit methodology. To compare against this hop-based method, we use hop-based nearest-neighbor geolocation and shortest hop geolocation.

To thoroughly evaluate the hop-based geolocation methods, we select monitors and landmarks for the set of targets such that each target belongs to one of three different monitor geographic density regimes, where the closest monitor lies 10 to 75 miles, 75 to 150 miles, or 150 to 250 miles from the target. In terms of the landmarks, each target belongs to one of three different landmark geographic density regimes, where the closest landmark lies 0.1 to 5 miles, 5 to 15 miles, or 15 to 30 miles from the target. The set of test targets were equally distributed across all pairs of monitor-landmark geographic density.

Across all possible measurement infrastructure density regimes, we aggregate the results in Table 1. It is shown that using just 50 observed hop counts, we can geolocate 20% of our set of targets with median error of 195.16 miles using our collection of landmarks. The cumulative distribution of the geolocation errors can be seen in Figure 3. By only returning estimates of the most confident targets, the hop-based Posit methodology avoids the considerable error caused by targets that cannot possibly be accurately recovered using the measurement infrastructure.

5.2 Passive Hop-based Geolocation Using a Subset of Monitors

Our objective is to create a geolocation method that is accurate and robust to the realities of Internet configurations and traffic dynamics. To that end, we must assume that it could be the case that ICMP or other forms of active probes will be blocked by providers. This would seem to present an impossible challenge for geolocation. However, we assume

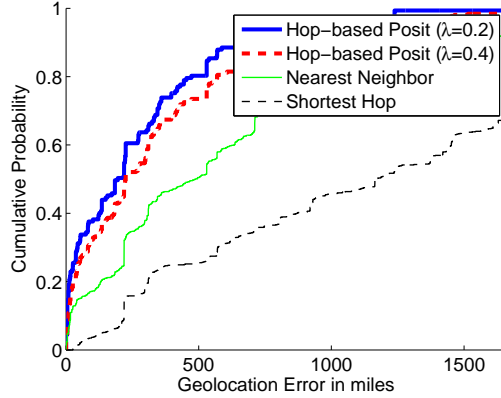


Figure 3: Cumulative distribution of geolocation error for Hop-based Posit using the Commercial Dataset with number of monitors $M = 50$ and number of landmarks $T = 150$ for targets across all measurement infrastructure densities.

Table 2: Geolocation estimation errors for hop-based Posit Geolocation errors using synthesized passive hop counts from the Commercial Data set with number of monitors, $M = 50$ and the fraction of targets λ at 20% and 40%.

λ	Observed Hops	Mean Error (in miles)	Median Error (in miles)
$\lambda = 0.2$	5	789.83	694.18
	25	321.46	272.85
	45	280.07	184.26
$\lambda = 0.4$	5	836.08	769.52
	25	417.18	350.76
	45	367.93	272.85

that we have the ability to monitor packets and thereby collect TTL counts at monitors, then we can use the techniques described in [22] to infer the number of hops between monitors and our set of targets.²

Due to the passive nature of these measurements, each target would be unlikely to obtain hop measurements to all M of our monitors. Instead, it should be assumed that we only have hop measurements to a random subset of our monitors. The objective of this analysis is to assess the impact on accuracy of the hop-based method in what we consider to be a more realistic measurement scenario. In this regime, we find the landmark hop vector with the minimum variance with respect to the passively observed hop count indices, \mathcal{I}_i^p ,

$$c_i^{passive} = \underset{j}{\operatorname{argmin}} (\sigma^2 (\mathbf{h}_i^{end}(\mathcal{I}_i^p) - \mathbf{h}_j^{land}(\mathcal{I}_i^p))) \quad (2)$$

Where $\mathbf{h}(\mathcal{I}) = [h_{\mathcal{I}_1} \ h_{\mathcal{I}_2} \ \dots \ h_{\mathcal{I}_{|\mathcal{I}|}}]$ is the subvector of hop counts with respect to indices \mathcal{I} .

Using complete hop count vectors to $M = 50$ monitors, the observation of passive measurements can be synthesized by withholding a specified number of hop elements in randomly chosen locations. In Table 2, error results for hop-based Posit geolocation using simulated incomplete passive measurements are shown for both average and median error rates for the commercial node dataset. These results highlight the fact that there is only a modest decline in accuracy when at least half of passive measurements from monitors are available. This suggests a profile for the accuracy of Posit’s geolocation estimates for nodes that are currently impossible to geolocate by other measurement-based methods.

²Assuming that our monitors would be co-located in popular network locations capable of observing a large number of passive measurements.

6 Latency-based Posit Geolocation

We are motivated by the notion that geolocation accuracy is tied to measurement infrastructure density and hypothesize that accuracy can be improved by adapting the algorithm to the proximity of landmarks and monitors. Consider observed latency measurements between our target end host and the monitors. From latency probes, we construct a set of *latency vectors*. For target $i = \{1, 2, \dots, N\}$,

$$\mathbf{I}_i^{target} = \begin{bmatrix} l_{i,1}^{target} & l_{i,2}^{target} & \dots & l_{i,M}^{target} \end{bmatrix}$$

Where $l_{i,k}^{target}$ is the observed latency between target i and monitor k .

Additionally, consider latency measurements to each landmark $j = \{1, 2, \dots, T\}$,

$$\mathbf{I}_j^{land} = \begin{bmatrix} l_{j,1}^{land} & l_{j,2}^{land} & \dots & l_{j,M}^{land} \end{bmatrix}$$

Where $l_{j,k}^{land}$ is the observed latency between landmark j and monitor k .

The latency-based Posit algorithm estimates the geographic location of the target end host using only these observed latency measurements vectors from our set of monitors. We learn distance probability likelihood distributions between the target and the set of landmarks using the methodology from Section 6.1 (which requires no direct measurements between the targets and landmarks), and then use these estimated probabilities to estimate geolocation using the statistical embedding algorithm from Section 6.2.

6.1 Landmark-to-Target Distance Likelihood Estimation

Consider the latency vectors for a target end host and a single landmark, \mathbf{I}_i^{target} and \mathbf{I}_j^{land} . Prior work in [23] has shown that by weighting short latency values using the Canberra distance, $d_{i,j}^{canberra} = \sum_{k=1}^M \frac{|l_{j,k}^{land} - l_{i,k}^{target}|}{l_{j,k}^{land} + l_{i,k}^{target}}$, we can obtain a more accurate estimation of distance from the two latency signatures, (compared to taking the Euclidean norm, $\|\mathbf{I}_j^{land} - \mathbf{I}_i^{target}\|_2$). Further extending this idea, we introduce the concept of thresholding the latency values to only consider the set of short latency indices, $\mathcal{I}_{i,j}$, the indices of the two vectors where at least one of the values is below some specified delay threshold, λ_{lat} ,

$$\mathcal{I}_{i,j} = \{k : l_{i,k}^{target} \leq \lambda_{lat} \text{ or } l_{j,k}^{land} \leq \lambda_{lat}\} \quad (3)$$

The problem becomes how to choose the distance transformation for the short latency elements ($\mathcal{I}_{i,j}$), such that we achieve the closest relationship between observed latency and geographic distance. We propose three possible transformations using only the short latency elements:

The L1-norm for the short latency indices (*i.e.*, threshold L1 distance) :

$$v_{i,j}^{L1} = \frac{1}{|\mathcal{I}_{i,j}|} \|\mathbf{I}_i^{target}(\mathcal{I}_{i,j}) - \mathbf{I}_j^{land}(\mathcal{I}_{i,j})\|_1 \quad (4)$$

Where $\mathbf{1}(\mathcal{I}) = [l_{\mathcal{I}_1} \quad l_{\mathcal{I}_2} \quad \dots \quad l_{|\mathcal{I}|}]$ is the subvector of latency with respect to indices \mathcal{I} .

The L2-norm for the short latency indices (*i.e.*, threshold L2 distance):

$$v_{i,j}^{L2} = \frac{1}{|\mathcal{I}_{i,j}|} \|\mathbf{I}_i^{target}(\mathcal{I}_{i,j}) - \mathbf{I}_j^{land}(\mathcal{I}_{i,j})\|_2 \quad (5)$$

The Canberra distance for the short latency indices (*i.e.*, threshold Canberra distance):

$$v_{i,j}^{can.} = \frac{1}{|\mathcal{I}_{i,j}|} \sum_{k \in \mathcal{I}_{i,j}} \frac{|l_{j,k}^{land} - l_{i,k}^{target}|}{l_{j,k}^{land} + l_{i,k}^{target}} \quad (6)$$

To evaluate performance of the various latency distance metrics, we use the R^2 coefficient of determination metric [24], which measures the quality of the linear relationship between the geographic distance and the distance metric value. By definition, $R^2 = 1$ if there is a perfect linear trend between the geographic distance values and distance

Table 3: Coefficient of Determination (R^2) - Measure of linear fit quality of latency distance metric to true geographic distance.

Distance Metric	Commercial Dataset
L1 Norm	0.672
L2 Norm	0.109
Canberra	0.604
Threshold L1 Norm	0.761
Threshold L2 Norm	0.343
Threshold Canberra	0.466

metric, and $R^2 = 0$ if the two sets of values are uncorrelated (with no linear trend). In Table 3, we show the coefficient of determination, R^2 , for the six latency-based distance metrics with respect to the true geographic distance between the targets and landmarks. This experiment was performed for the Commercial node dataset with the set latency threshold $\lambda_{lat} = 10$.

As seen in the table, the threshold L1 norm distance metric from Equation 4 results in a better linear fit to the true geographic distance given the observed latency values compared with all other distance metrics. This matches our intuition, as the L1-norm will weight smaller latency deviations more than either the L2-norm or Canberra distances. The strong linear trend between the threshold L1 norm values and the true geographic distance indicates the potential to obtain accurate geographic distance estimates between targets and landmarks without the need for direct measurements.

Although these results indicate correlation between our threshold L1 norm metric and geographic distance, simply obtaining a single distance estimate for each target-landmark pair may obscure characteristics of the network. To avoid erroneous estimates, we will instead consider learning likelihood distributions, $\hat{p}_{land}(d | v_{i,j})$, the probability of a target being d distance away from a landmark given threshold L1-norm value $v_{i,j}$. In order to learn these distribution functions, we exploit a set of training targets with known measurements and geographic locations, and off-the-shelf kernel density estimator techniques [25].

6.2 Statistical Embedding Algorithm

Given the estimated likelihood probability of distance for each target to the set of landmarks, $\hat{p}_{land}(d | v_{i,j})$, our goal is to estimate each target’s latitude/longitude coordinates, \mathbf{x}_i . In addition to our estimated likelihood distributions between landmarks and targets, there is additional information we can exploit when geolocating our targets. For example, the latency observations between the monitors and targets. Similar to recent statistical geolocation methodologies (e.g., [13, 14, 15]), we can construct likelihood probabilities from observed latencies to the monitors, $\hat{p}_{monitor}(d | l_{i,j})$ (the probability of being d miles away from monitor j given observed latency $l_{i,j}$), using a training set of targets with known location. Additionally, using Constraint-based Geolocation [9] we can obtain the constraint region, \mathbf{C}_i , the set of feasible latitude/longitude coordinates given our observed latency values from the monitors.

We assume that the resulting embedding coordinates of the set of targets should be *sparse*. This is the case where we confine geolocation to a small subset of geographic locations (e.g., cities) where we expect the target to be geographically located. To enforce this restriction, we only consider the known locations of landmarks and monitors in the infrastructure that are contained in the geographic constraint region (\mathbf{C}_i). Given prior work on geographic clustering of Internet resources [26], consider this to be constraining target geolocation to areas of high population density in the geography (where our monitors/landmarks are likely to be located). In contrast to previous work [7, 13] which requires explicitly defined population and/or geographic data as input into the algorithm, Posit does not require a priori knowledge of the population density or geographic properties of the region of interest.

Given the set of feasible latitude/longitude coordinates in the constraint region \mathbf{C}_i found by Constraint-based geolocation, we can define the set of constrained resource coordinates, \mathbf{C}_{R_i} , as the coordinates of the landmarks ($\mathbf{T} = \{\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_T\}$) and monitors ($\mathbf{M} = \{\mathbf{m}_1, \mathbf{m}_2, \dots, \mathbf{m}_M\}$) found in the constraint region,

$$\mathbf{C}_{R_i} = \begin{cases} \mathbf{C}_i \cap \{\mathbf{M} \cup \mathbf{T}\} & \text{if } \mathbf{C}_i \cap \{\mathbf{M} \cup \mathbf{T}\} \neq \emptyset \\ \mathbf{C}_i & \text{if } \mathbf{C}_i \cap \{\mathbf{M} \cup \mathbf{T}\} = \emptyset \end{cases} \quad (7)$$

Where, if none of the monitors or landmarks are found in the region determined by Constraint-based geolocation, we geolocate with respect to the entire constraint region (\mathbf{C}_i).

We aim to find the most probable constrained resource location ($\mathbf{c}_r \in \mathbf{C}_{R_i}$) given our set of trained likelihood distributions and observed measurements to the monitors. Assuming independence between the measurements, we can find the geographic coordinate that maximizes the log-likelihood given our observed latency values ($l_{i,k}$) and landmark distance metrics ($v_{i,j}$). But, using the standard log-likelihood, all observed latency and landmark distance values would be weighted equally.

We argue that due to non-line-of-sight routing for a vast majority of paths through the Internet with medium/long path length, many latency measurements will be a very poor indication of distance in the Internet. Informed by the geolocation improvement by using measurement weights in prior geolocation methodologies [7, 13], the latency of the path should also imply some degree of importance. Shorter latency values should hold more weight than longer latency values, as they are likely to be the result of short paths through the network with possible direct-line-of-sight routing. Therefore, we will weight each measurement value using an exponential, where we construct the weights such that the latency measurement of $l_{i,j}$ milliseconds (between host target i and landmark j) is weighted by $w_m(l_{i,j}) = \exp(-\phi_{monitor}l_{i,j})$. Where the tuning parameter, $\phi_{monitor} > 0$.

Similarly, the likelihood derived from threshold L1 norm metric, $v_{i,k}$ (between host target i and landmark k), is weighted by, $w_l(v_{i,k}) = \exp(-\phi_{land}v_{i,k})$. Where the tuning parameter, $\phi_{land} > 0$.

Finally, the Statistical Embedding geolocation algorithm estimates the geographic location coordinates \mathbf{x}_i using the resource location that maximizes the weighted log-likelihood,

$$\hat{\mathbf{x}}_i = \underset{\mathbf{x} \in \mathbf{C}_{R_i}}{\operatorname{argmax}} \left(\sum_{j=1}^T w_l(v_{i,j}) \log(\hat{p}_{land}(d(\mathbf{x}, \mathbf{t}_j) | v_{i,j})) + \sum_{k=1}^M w_m(l_{i,k}) \log(\hat{p}_{monitor}(d(\mathbf{x}, \mathbf{m}_k) | l_{i,k})) \right) \quad (8)$$

Where $d(\mathbf{x}, \mathbf{y})$ is the geographic distance between latitude/longitude coordinates \mathbf{x} and \mathbf{y} , and \mathbf{C}_{R_i} is the set of feasible resource coordinates from Equation 7. We see that this computationally lightweight methodology only requires $O(|\mathbf{C}_R|(T + M))$ operations to geolocate each target.

6.3 Latency-Based Posit IP Geolocation Algorithm Summary

We now summarize the full latency-based Posit geolocation algorithm. We exploit the distance likelihood distributions (described in Section 6.1), and then use the statistical embedding algorithm (described in Section 6.2) to estimate geographic location for our set of test targets. All tuning parameters ($\lambda_{lat}, \phi_{land}, \phi_{monitor}$) are found by an efficient bisection search through the training set. After careful inspection of our data, the likelihood distributions are constructed from the training sets for landmark-based distance likelihoods ($\hat{p}_{land}(d | v)$) using the threshold L1 distance ranges, $\{(0, 5], (5, 10], \dots, (75, 80]\}$, and the monitor-based distance likelihoods ($\hat{p}_{monitor}(d | l)$) are constructed for observed latency ranges $\{(0, 10], (10, 20], \dots, (140, 150]\}$ (in milliseconds). The complete latency-based Posit IP Geolocation methodology is presented in Algorithm 2.

7 Latency-Based Experiments

Using both the commercial node dataset and domain name dataset described in Section 3, we now evaluate the performance of the latency-based Posit algorithm (which for the rest of this section will simply be referred to as ‘‘Posit’’).

7.1 Comparison Methodologies

To evaluate relative performance of the Posit algorithm, we compare against all current state-of-the-art geolocation methodologies³.

³The MATLAB code used in this paper for both our Posit methodology and all comparison methods will be made publicly available before the conference.

Algorithm 2 - Latency-based Posit IP Geolocation Algorithm

Given:

- Latency vectors from the M monitors to our set of N targets, \mathbf{I}_i^{target} for $i = \{1, 2, \dots, N\}$.
- Latency vectors from the M monitors to our set of T landmarks, \mathbf{I}_j^{land} for $j = \{1, 2, \dots, T\}$.
- Training set of targets with known geolocation and latency measurements to the monitors.

Initialize:

- Learn the likelihood distributions, $\hat{p}_{land}(d | v)$ and $\hat{p}_{monitor}(d | l)$ using the known training set target locations.
- Use the training set to find the optimal values of tuning parameters $(\lambda_{lat}, \phi_{land}, \phi_{monitor})$ with respect to the training set geolocation error rate.

Methodology:**For each target**, $i = \{1, 2, \dots, N\}$

- Resolve the threshold L1 distances, $v_{i,k}$ for $k = \{1, 2, \dots, T\}$ using Equation 4.
 - Using the learned distributions $(\hat{p}_{land}(d | v)$ and $\hat{p}_{monitor}(d | l)$), use the Statistical Embedding methodology (Equation 8) to estimate the target geolocation.
-

7.1.1 GeoPing and Shortest Ping Algorithms

Some of the first IP geolocation methodologies developed were the Shortest Ping and GeoPing techniques from [8]. The *Shortest Ping* technique uses a series of latency measurements from a set of monitors to a target, and then maps that target’s geolocation to the monitor that has the shortest observed latency value. We expect Shortest Ping to work well in our evaluation for instances where the monitor placement is dense. However, in instances where monitors are not near targets, the Shortest Ping methodology’s accuracy will decline and the strength of our Posit methodology will be highlighted.

Meanwhile, the *GeoPing* algorithm was the first IP geolocation algorithm proposed that exploited existing Internet infrastructure with known location (*i.e.*, landmarks). Using latency measurements from a set of monitors, the target latency vector is compared with the latency vectors from the set of landmarks. The geolocation of the target is set to the location of the landmark with the smallest Euclidean distance between latency vectors. This methodology accuracy is strongly dependent on the location of the landmarks with respect to the target.

7.1.2 Constraint-Based Geolocation Algorithm

To generate *Constraint-Based Geolocation* (CBG) geolocation estimates, we implemented the algorithm described in [9]. CBG is the current state-of-the-art IP geolocation methodology using only Ping-based measurements. The basic intuition behind CBG is that each latency measurement to a set of monitors with known location can be considered a series of constraints, where given speed-of-light in fiber assumptions, and self-calibration using a set of training data, we can determine a feasible geographic region given each latency measurement. Given a series of latency measurements, the possible geographic placement is considered the intersection of many constraint regions, with the estimated location being the centroid of this intersection region. The size of this final constraint region will be correlated with the smallest individual constraint region size, which is dependent on the shortest observed latency (*i.e.*, likely the geographically closest monitor) to the target.

7.1.3 Octant-Based Geolocation Algorithm

Building on the Constraint-based Geolocation approach, the *Octant* algorithm [7] is the current state-of-the-art measurement-based geolocation methodology. In contrast to the Posit algorithm, Octant uses both ping-based measurements to the targets *and* given geographic information from unDNS [12] of routers along the path to the targets.

Our implementation, which we refer to as *Octant-Based Geolocation*⁴, includes the Octant methodology’s use of both “positive” and “negative” geographic constraints from latency measurements, the iterative refinement of the feasible constraint region, unDNS intermediate node information, point selection through Monte Carlo simulation, latency “heights”, and spline approximation of latency to distance. Missing from our implementation of Octant is the use of geographic/population information to aid in geolocation, as we feel the lack of process description in [7] could potentially bias our implementation of this component. In our experiments, unDNS-derived geographic information is derived from the last hop router encountered along the path before the target. For our commercial set of 431 nodes, it was found that only 71 nodes had available last hop unDNS information down to the city location. To give our Octant-based methodology every opportunity, this unDNS information will not be made available to the Posit framework (as Posit only requires latency measurements). As with the Posit experiments, the numerous tuning parameters in the Octant algorithm will be trained based on minimizing the geolocation error on a training set of targets with known geographic location. Similar to the CBG approach it is based on, we expect Octant to perform the best when monitors are close to the targets.

7.1.4 Statistical Geolocation

Considerable recent work in IP geolocation [13, 14, 15] has been framed as finding the geographic location that maximizes the likelihood probability, with respect to probability distributions learned from a training set of targets with known locations. While the construction of the probability distributions varies (nonparametric kernel density estimators in [13, 14], parametric log-normal model in [15]), all three methodologies assume conditional independence between measurements in order to efficiently calculate the geographic location with the maximum likelihood given observed measurements. Here we use the methodology from [13] using a training set of target end hosts with latency measurements and known geolocation to generate kernel density estimates [25] of distance given observed latency.

7.1.5 Street-Level Geolocation

The most recent contribution to the geolocation literature is the Street-Level geolocation methodology of [6]. Using a novel landmark identification methodology, the Street-Level approach maps to the closest estimated resource using estimated latency between targets and landmarks from `traceroute` probes. In contrast to the Street-Level approach, Posit does not require any `traceroute` probes and allows the use of simple latency information between landmarks and monitors. For an apples-to-apples comparison with other techniques, we test the Street-Level methodology on our declared set of passive landmarks that are available to every other geolocation methodology. We call this modified approach, *Street-Level Based* geolocation. Given the dependency on CBG, this methodology will return the most accurate geolocation results when both the monitors and the landmarks are geographically close to the targets.

While the Street-Level paper declares accuracy of sub-1 kilometer, it lacks performance comparison with competing geolocation techniques on the same dataset, merely stating the previously published results for each methodology. Our results will demonstrate that application of prior methodologies *on the same dataset* is critical for a fair and accurate comparison of geolocation accuracy performance. To give Street-Level every opportunity, the estimated pairwise latency between the targets and the landmarks (derived from `traceroute`), will not be made available to our Posit methodology.

7.1.6 Database-specific Geolocation Methodologies

We will also compare geolocation accuracy with both the *Maxmind* database⁵ [27] and the *IP2Location* database [28]. Both of these databases are commercially available IP lookup packages. Unfortunately, due to the commercial nature of both products, the methodology used for geolocation is not known.

7.2 Geolocation Probing Complexity

The number of network probes required for each geolocation methodology is seen in Table 4. From the table, it can be seen that latency-based Posit requires the same number of probes as the GeoPing methodology, where the only measurements required are latency probes from each monitor to the set of targets and set of landmarks. This

⁴We were unable to get access to specific Octant code used in [7] to compare it with Posit for this study.

⁵The Maxmind database tested here is the purchased web service MaxMind GeoIP City Database.

Table 4: Probing complexity for all measurement-based geolocation methodologies (given N targets, M monitors, and T landmarks).

Methodology	Ping-like Measurements	traceroute Measurements
Posit	$O(M(N+T))$	0
Shortest Ping	$O(MN)$	0
GeoPing	$O(M(N+T))$	0
Constraint-Based	$O(MN)$	0
Octant	$O(MN)$	$O(N)$
Street-Level	$O(MN)$	$O(M(N+T))$
Statistical	$O(MN)$	0

is in contrast to more recent geolocation methods, such as Street-Level geolocation, which for their pairwise latency estimation approach requires a `traceroute` probe from every monitor to the set of targets and landmarks. Another methodology that uses data derived from `traceroute` probes is Octant, which resolves router unDNS hints and latency observations to further constrain their target geolocation estimates.

7.3 Geolocation Experiments using Domain Name Landmarks

The first set of experiments use the 431 commercial nodes as targets to test the performance of the Posit geolocation algorithm. For each target, we randomly select 25 monitor nodes from the set of 430 remaining commercial nodes. In addition, for each target we select 75 domain names (out of 283 total) with known location as landmarks which will aid in our geolocation. To assess performance of the Posit, Constraint-Based, Octant, and Statistical geolocation algorithms (all of which require a training set of targets with known geolocation), we will perform hold-out Cross Validation [25] where randomly selected 50% of the targets are held out as training data with reported geolocation results with respect to the remaining random 50% of the targets used as test data.

To thoroughly evaluate all geolocation methods, we select monitors and landmarks for the set of targets such that each target belongs to one of three different monitor geographic density regimes, where the closest monitor lies 10 to 75 miles, 75 to 150 miles, or 150 to 250 miles from the target.⁶ Also, each target belongs to one of three different landmark geographic density regimes, where the closest landmark lies 0.1 to 5 miles, 5 to 15 miles, or 15 to 30 miles from the target. The landmark density is set considerably closer to the targets due to the significantly larger potential set of landmarks available in the Internet, as recently shown in [6]. The set of test targets are distributed equally across the possible pairs of monitor/landmark density regimes.

Aggregated across all monitor/landmark density regimes, the results in Table 5 show the improvements of the Posit methodology over all existing geolocation methodologies. On this aggregated dataset, we find that Posit returned median error performance of only 26.15 miles, 50% less than all other methodologies. The cumulative distribution of the errors can be seen in Figure 4. Clear improvements are seen using the Posit framework over all the competing techniques for over 75% of the targets. The intuition behind why Posit improves on the other methods is that our landmark distance likelihoods exploit information unseen by previous techniques and our statistical embedding methodology automatically adapts to the most confident subset of measurement infrastructure locations.

7.3.1 Dense Monitor Experiments

To further present the power of the Posit framework, we compare our performance against the density regimes that are most-advantageous for the competing geolocation methodologies. Using the densest monitor regime (where, for each target, at least one monitor is within 10 to 75 miles), we compare performance of Posit on targets in this regime with other latency-dependent methodologies (*i.e.*, Shortest Ping, Constraint-based, Octant) which we expect to take the most advantage of the geographically close monitors. The error metrics are seen in Table 6 and the cumulative distribution of errors can be seen in Figure 5-(Left). The results show that Posit outperforms by obtaining median error at least 34% less than all other methodologies. These improvements are seen even with the Octant methodology given every opportunity with unDNS information that is not supplied to our Posit framework.

⁶The threshold of 250 miles was chosen due to over 90% of the population of the United States residing within 250 miles of the most populous 100 United States cities ([29]).

Table 5: Geolocation error (in miles) for all geolocation methodologies using the commercial dataset with number of monitors $M = 25$ and number of landmarks $T = 75$ for targets across all measurement infrastructure densities.

Methodology	Mean Error	Median Error
Posit	82.18	26.15
Shortest Ping	160.93	182.30
GeoPing	167.93	126.11
Constraint-Based	140.94	108.20
Octant Based	94.82	53.91
Street-Level Based	169.03	101.17
Statistical	159.80	150.11
IP2Location	943.03	644.65
MaxMind	909.43	556.88

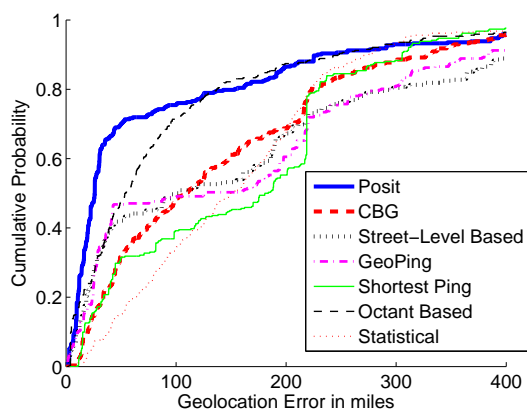


Figure 4: Cumulative distribution of geolocation error for Posit using the commercial dataset (with number of monitors $M = 25$ and number of landmarks $T = 75$) for targets across all measurement infrastructure densities.

7.3.2 Dense Landmark Experiments

Using the densest landmark regime (where, for each target, at least one landmark is within 0.1 to 5 miles), we compare performance of Posit on targets in this regime with other landmark-dependent methodologies (*i.e.*, GeoPing and Street-

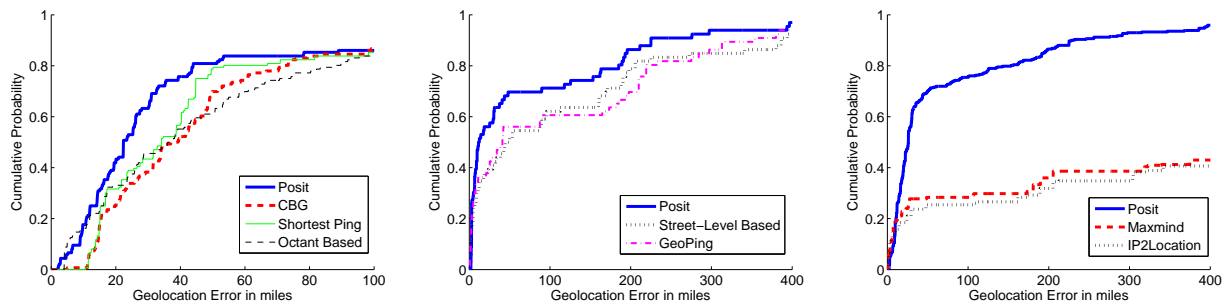


Figure 5: Cumulative distribution of geolocation error for Posit using the commercial dataset with number of monitors $M = 25$ and number of landmarks $T = 75$. (Left) - Compared with monitor-based techniques for targets with close monitor density. (Center) - Compared with landmark-based techniques for targets with close landmark density. (Right) - Aggregated results compared with commercial geolocation methods.

Table 6: The geolocation error (in miles) for monitor-based methodologies using the commercial dataset with number of monitors $M = 25$ and number of landmarks $T = 75$ for targets in a monitor dense regime.

Methodology	Mean Error	Median Error
Posit	53.09	22.38
Shortest Ping	65.86	34.13
Constraint-Based	67.85	38.20
Octant Based	70.74	36.43
Statistical	180.79	198.30

Table 7: The geolocation error (in miles) for landmark-based methodologies using the commercial dataset with number of monitors $M = 25$ and number of landmarks $T = 75$ for targets in a landmark dense regime.

Methodology	Mean Error	Median Error
Posit	81.31	11.98
GeoPing	129.52	41.08
Street-Level Based	133.30	47.22

Level). The error metrics are seen in Table 7 and the cumulative distribution of errors can be seen in Figure 5-(Center). Again we see that Posit drastically outperforms the other methods in this regime with median error 74% less than both competing methodologies. These improvements are seen even with the Street-Level based methodology given every opportunity with `traceroute`-based pairwise latency estimates between the targets and landmarks that are not supplied to our Posit framework.

In terms of the Street-Level geolocation methodology, one concern might be that our resolved accuracy deviates significantly from the previously published results. Inspection of the `traceroute` derived pairwise latency estimates reveal that due to the many disadvantages of `traceroute`-based probing (*e.g.* aliased router interfaces, routers that block ICMP, invisible MPLS routing) the resulting pairwise latency estimates can be wildly inaccurate. We validated our Street-Level implementation by performing their landmark constraint-based methodology using directly observed pairwise latency between the landmarks and the targets, as opposed to their methodology of estimating this data from `traceroute` probes. In these highly idealized tests, our Street-Level based implementation returned a mean error of 78.25 miles and median error of 30.11, which is still less accurate than our Posit methodology in terms of median error.

7.3.3 Sparse Density Experiments

While Posit performs well for regimes where either the monitors or the landmarks are geographically dense with respect to the targets, we now examine the performance when the distribution of the measurement infrastructure is *sparse*. We consider a sparse density regime where all monitors are greater than 150 miles away from our targets and the landmarks are greater than 15 miles away from our targets. For the results seen in Table 8, we again find that our Posit methodology outperforms all competing methodologies with median error 60% less than all other techniques. We additionally note the large degradation in geolocation performance for all methodologies in this regime. This further emphasizes the importance of characterizing geolocation methodologies across multiple measurement infrastructures with varying geographic densities.

7.3.4 Database Geolocation Experiments

Finally, in Figure 5-(Right) we see the performance of Posit with respect to the cumulative distribution of geolocation errors for the database-specific algorithms (Maxmind and IP2Location). Our Posit methodology has significantly better mean and median error than both the database-specific techniques. We see from the distributions that for roughly 20% the targets, the commercial techniques succeed with accurate geolocation, while the remaining 80% of targets deviate wildly from their true locations. While this could be biased by all our commercial node targets being owned by the same provider, this still demonstrates a fundamental failing of database-specific techniques.

Table 8: The geolocation error (in miles) for monitor-based methodologies using the commercial dataset with number of monitors $M = 25$ and number of landmarks $T = 75$ for targets in a monitor sparse and landmark sparse regime.

Methodology	Mean Error	Median Error
Posit	125.55	30.54
GeoPing	175.52	162.28
Shortest Ping	259.27	218.84
Constraint-Based	212.56	215.97
Street-Level Based	243.55	202.82
Octant Based	129.36	79.31
Statistical	133.52	87.58

Table 9: The geolocation error (in miles) for all geolocation methodologies (for number of monitors $M = 25$ and number of commercial node landmarks $T = 75$) for targets across all measurement infrastructure densities.

Methodology	Mean Error	Median Error
Posit	116.56	46.14
Shortest Ping	244.28	218.80
GeoPing	206.88	170.63
Constraint-Based	203.81	185.98
Octant Based	161.58	95.34
Street-Level Based	258.61	218.80
Statistical	168.72	121.13
IP2Location	768.53	405.84
MaxMind	693.62	374.66

7.4 Geolocation Experiments using Commercial Landmarks

For our second set of experiments, we use monitors, targets, and landmarks all chosen from the set of 431 commercial nodes to test the performance of the Posit geolocation algorithm. For each target, we randomly select 25 monitor nodes and 75 landmarks from the set of 430 remaining commercial nodes (holding out the target node). In contrast to the previous set of experiments, we evaluate geolocation using a measurements infrastructure that is farther away from the targets, where the closest monitor lies 50 to 100 miles, 100 to 200 miles, or 200 to 300 miles from the target. Also, each target belongs to one of three different landmark geographic density regimes, where the closest landmark lies 0.1 to 10 miles, 10 to 25 miles, or 25 to 50 miles from the target. Again, we will perform hold-out Cross Validation where 50% of the targets randomly selected are held out as training data with the remaining randomly selected 50% of the targets are used as test data.

Aggregated across all monitor/landmark density regimes, the results in Table 9 show the improvements of the Posit methodology over all existing geolocation methodologies. On this aggregated dataset with targets distant from the measurement infrastructure, we find that Posit returned median error performance of 46.14 miles, again 50% less than all other methodologies. The cumulative distribution of the errors can be seen in Figure 6-(Left). Clear improvements are seen through the use of the Posit framework over all the competing techniques for over 85% of the targets.

In Figure 6, we also show the performance of Posit against the competing methodologies in regimes where these prior algorithms should perform very well. For targets that are close to the monitors (with closest monitor within 50 to 100 miles of each target), in Figure 6-(Center), we show that Posit resolves the targets with median error of 42.43 miles, in contrast to the best competing geolocation technique (Octant Based) which obtains 68.54 mile median accuracy performance. In terms of landmark-based methodologies, in Figure 6-(Right) it is shown that Posit drastically outperforms all competing techniques, reducing the median error from 128.69 miles using the GeoPing approach, to only 31.00 miles using Posit.

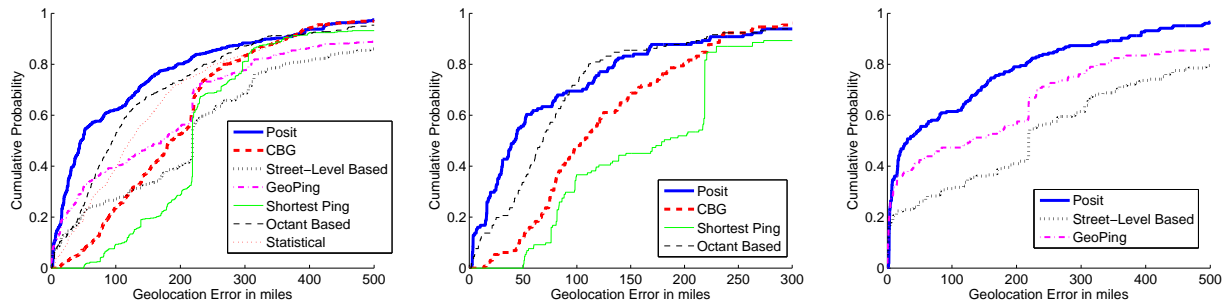


Figure 6: Cumulative distribution of geolocation error for Posit using the commercial dataset with number of monitors $M = 25$ and number of commercial node landmarks $T = 75$. (Left) - Compared across all measurement infrastructure densities. (Center) - Compared with monitor-based techniques for targets with close monitor density. (Right) - Compared with landmark-based techniques for targets with close landmark density.

8 Conclusions

The ability to determine the geographic coordinates of an IP host can be used in many different location-specific applications. Median and worst case errors in predictions made by prior geolocation methods render them ineffective for some classes of location-aware services. The goal of our work is to develop an IP geolocation methodology that is highly accurate and can compute estimates based on a relatively simple set of measurements.

In this paper, we described a new method for IP geolocation that we call Posit. Our approach is based on the insight that adapting to the geographic density of the measurement infrastructure is critical to geolocation performance. We first describe a clustering methodology based on hop count distance, which is easily established with lightweight ping-like measurements, or by passively observed hop counts. We then describe a latency-based methodology, where to estimate geographic location, Posit uses a distance likelihood estimation methodology combined with a new statistical embedding process, which mitigates the effects of noisy distance estimation from measurements.

We assess the capabilities of Posit using a data set of hop count and latency measurements collected from hundreds of hosts in the Internet with precisely known geographic coordinates. Our results show that using latency measurements, Posit is able to identify the geographic location of target hosts with a median error of only 26 miles. We compare this with an implementation of the current state-of-the-art measurement-based geolocation methodology, which produces geolocation estimates with median errors of 53 miles on the same dataset. We also compare Posit’s estimates against two commercial IP geolocation services, which produce mean error estimates that are nearly a factor of 9 higher than Posit. These results highlight the powerful capabilities of our approach.

The results of our study motivate future work in a number of areas. First, we plan to expand the scope of the Posit infrastructure to include a larger set of landmarks, which will further improve our estimation accuracy. Also, we plan to begin building an IP geolocation database using Posit that we plan to make available to the community.

References

- [1] R. Sherwood, A. Bender, and N. Spring, “DisCarte: A Disjunctive Internet Cartographer,” in *Proceedings of ACM SIGCOMM Conference*, Seattle, WA, August 2008.
- [2] D. Magoni and J. Pansiot, “Analysis of the Autonomous System Network Topology,” in *ACM SIGCOMM Computer Communications Review*, vol. 31, no. 3, July 2001.
- [3] V. Paxson, “End-to-End Routing Behavior in the Internet,” in *IEEE/ACM Transactions on Networking*, vol. 5, no. 5, October 1997.
- [4] T. S. E. Ng and H. Zhang, “Predicting Internet Network Distance with Coordinates-Based Approaches,” in *Proceedings of IEEE INFOCOM Conference*, New York, NY, June 2002.
- [5] J. Muir and P. Oorschot, “Internet geolocation: Evasion and Counterevasion,” in *ACM Computing Surveys*, vol. 42, December 2009.

- [6] Y. Wang, D. Burgener, M. Flores, A. Kuzmanovic, and C. Huang, "Towards Street-Level Client-Independent IP Geolocation," in *Proceedings of USENIX NSDI 2011*, vol. 5, no. 5, Boston, MA, March 2011.
- [7] B. Wong, I. Stoyanov, and E. Sirer, "Octant: A comprehensive framework for the geolocation of internet hosts," in *USENIX NSDI Conference*, April 2007.
- [8] V. N. Padmanabhan and L. Subramanian, "An Investigation of Geographic Mapping Techniques for Internet Hosts," in *Proceedings of ACM SIGCOMM Conference*, San Diego, CA, August 2001.
- [9] B. Gueye, A. Ziviani, M. Crovella, and S. Fdida, "Constraint-based geolocation of internet hosts," in *IEEE/ACM Transactions on Networking*, December 2006.
- [10] E. Katz-Bassett, J. John, A. Krishnamurthy, D. Wetherall, T. Anderson, and Y. Chawathe, "Towards IP Geolocation Using Delay and Topology Measurements," in *Proceedings of ACM Internet Measurements Conference*, October 2006.
- [11] N. Spring, R. Mahajan, and D. Wetherall, "Measuring ISP Topologies with Rocketfuel," in *Proceedings of ACM SIGCOMM Conference*, Pittsburgh, PA, August 2002.
- [12] M. Zhang, Y. Ruan, V. Pai, and J. Rexford, "How DNS Misnaming Distorts Internet Topology Mapping," in *USENIX Annual Technical Conference*, 2006.
- [13] B. Eriksson, P. Barford, J. Sommers, and R. Nowak, "A Learning-based Approach for IP Geolocation," in *Proceedings of Passive and Active Measurements Conference*, Zurich, Switzerland, April 2010.
- [14] I. Youn, B. Mark, and D. Richards, "Statistical Geolocation of Internet Hosts," in *Proceedings of IEEE International Conference on Computer Communications and Networks (ICCCN)*, San Francisco, CA, August 2009.
- [15] M. Arif, S. Karunasekera, S. Kulkarni, A. Gunatilaka, and B. Ristic, "Internet Host Geolocation Using Maximum Likelihood Estimation Technique," in *Proceedings of IEEE International Conference on Advanced Information Networking and Applications (AINA)*, Perth, Australia, April 2010.
- [16] "A Means for Expressing Location Information in the Domain Name System - RFC 1876."
- [17] "Network Time Protocol (Version 3) Specification - RFC 1305."
- [18] "The MTR Tool," <http://www.bitwizard.nl/mtr>.
- [19] B. Chun, D. Culler, T. Roscoe, A. Bavier, L. Peterson, M. Wawrzoniak, and M. Bowman, "PlanetLab: An Overlay Testbed for Broad-Coverage Services," *SIGCOMM CCR*, vol. 33, no. 3, pp. 3–12, 2003.
- [20] B. Eriksson, P. Barford, and R. Nowak, "Network Discovery from Passive Measurements," in *Proceedings of ACM SIGCOMM Conference*, Seattle, Washington, August 2008.
- [21] P. Barford, A. Bestavros, J. Byers, and M. Crovella, "On the Marginal Utility of Network Topology Measurements," in *Proceedings of ACM Internet Measurement Workshop*, October 2001.
- [22] C. Jin, H. Wang, and K. Shin, "Hop-Count Filtering: An Effective Defense Against Spoofed Traffic," in *Proceedings of IEEE INFOCOM Conference*, San Francisco, CA, April 2003.
- [23] A. Ziviani, S. Fdida, J. de Rezende, and O. Duarte, "Towards a Measurement-based Geographic Location Service," in *Proceedings of Passive and Active Measurements Conference*, Antibes Juan-les-Pins, France, April 2004.
- [24] L. Wasserman, "All of nonparametric statistics (springer texts in statistics)." Springer, May 2007.
- [25] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer, 2001.
- [26] A. Lakhina, J. Byers, M. Crovella, and I. Matta, "On the Geographic Location of Internet Resources," in *IEEE Journal on Selected Areas in Communications*, August 2003.

[27] “Maxmind IP Geolocation,” <http://www.maxmind.com/>.

[28] “IP2Location Geolocation,” <http://www.ip2location.com/>.

[29] “U.S. Census Bureau, <http://www.census.gov/>.”