Selection of Landmarks for Efficient Active Geolocation

Shinyoung Cho*, Zachary Weinberg†, Arani Bhattacharya‡, Sophia Dai*, Ramsha Rauf*
*Smith College, {scho,sdai33,rrauf}@smith.edu
†Independent researcher, zack@owlfolio.org
‡IIIT-Delhi, arani@iiitd.ac.in

***Abstract*—A reliable way of estimating the location of an Internet host is to infer it from packet round-trip times between that host (the *target*) and several hosts in known locations (the *landmarks*). This technique is known as *active geolocation*. A major drawback of active geolocation is that it can be very slow, especially when many targets need to be located and when the landmarks are far away from the targets.**

In this work, we improve the efficiency of an existing active geolocation procedure by minimizing the number of landmarks it requires to locate a set of targets. We evaluate several algorithms for selecting an optimal set of landmarks from a larger pool: purely random selection, clustering based on geography and network topology, and incremental addition of landmarks far from those already used, according to two different distance metrics. We find that the most effective method is initial random selection of 100 landmarks, followed by incremental addition of landmarks while maximizing the Autonomous System (AS) and geographic diversity of the pool. Using this method, we can verify the location of a target using only 32% of a pool of 780 landmarks, with the same accuracy as if the entire pool had been used.

Our code is publicly available for use and improvement. The code can be accessed at https://github.com/grace71/tma24-vp-ls under an open-source license.

I. Introduction

Many forms of research on the global Internet require performing measurements from "vantage points" in known locations, distributed all over the world. To give just a few examples, Dang et al. [4] and Zhang et al. [38] used global vantage points to study server latency as experienced by people all over the world; and Filasto and Appelbaum [10], Niaki et al. [25] and Sundara Raman et al. [31] use them on an ongoing basis to detect communication blockades imposed by governments. If some vantage points are not physically located where the researchers think they are, observations from those vantage points are invalid [20, 25].

The easiest and fastest way to determine the location of an Internet host is to look up its IP address in an "IP-to-location" database. Several organizations maintain such databases as either free or paid services. Unfortunately, these databases are not updated frequently enough for research purposes [2]. Ramesh et al. [27] and Du et al. [7] found that some types of commonly used vantage points are moved so frequently that one ought to re-check their locations on a daily basis. IP-to-location databases are also notoriously full of errors [13, 26, 29, 30].

978-3-903176-64-5

A more trustworthy way to locate vantage points is *active geolocation*, which works by measuring packet round-trip times (RTTs) between the host whose location is uncertain (the *target*) and a set of hosts in known locations (the *landmarks*). Since packets travel through the network at finite speeds, each RTT measurement gives an upper bound on the distance between the target and one landmark. Combining these upper bounds produces an estimate of the target's location. Active geolocation has been known to be feasible since at least 2004, and is now widely used by the network research community [e.g. 3, 7, 35].

Active geolocation can be very slow, especially if many targets need to be located. For example, data collection for "How to Catch when Proxies Lie" [35] required a full day for each of the larger VPN providers they tested [34]. Daily location re-checks of large sets of vantage points are impossible if the check itself takes a day. Worse, as Hu et al. [16] point out, incoming "ping" packets from all available landmarks may be enough network traffic that the target experiences it as a distributed denial-of-service attack.

An obvious way to speed up active geolocation of many targets is to reduce the number of landmarks used. As we explain in Section II, previous research has shown that only a few landmarks are needed for an accurate location estimate, as long as they are close to the target. The only problem is, how do you know which landmarks are near each target?

In this paper, we experiment with several different ways to

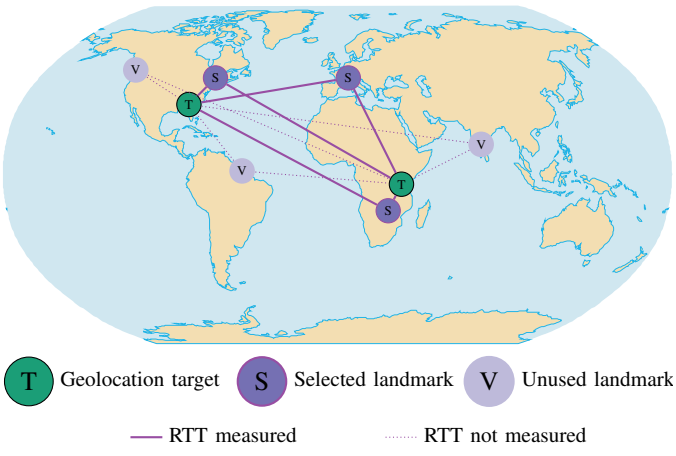


Fig. 1. Selection of landmarks for active geolocation.

minimize the size of a pool of landmarks that will be used to locate some set of targets. The basic idea is illustrated in Fig. 1: to locate both of the targets Ⓣ, one could use the landmarks labeled Ⓢ and skip those labeled Ⓥ. The challenge is to choose a subset of the landmarks for optimal efficiency, while maintaining accuracy. All our selection techniques can be used with any active geolocation algorithm, since they only affect which landmarks are used to take RTT measurements and the order in which they are chosen. The geolocation algorithm we used for testing is described in Section III-C.

Our objective is to speed up the process of geolocating all of the targets, while maintaining an existing level of accuracy. Accuracy improvements may also be possible, but they are reserved for future work. Thus, all of our results are described in terms of *agreement* between geolocation results obtained from a subset of the available landmarks, and those obtained from the whole pool. An ideal outcome is 100% agreement, with as small a subset as possible. We find that random selection of landmarks works fairly well for small subsets of the pool (< 100 landmarks) but cannot match the full pool for accuracy, even when large subsets are used. Conversely, strategic choice of landmarks, maximizing the "diversity" of the subset in some sense, *can* produce a subset that matches the full pool, but only by using 70% or more of the landmarks, which does not meet our efficiency goals. A hybrid approach, combining random selection and diversity maximization, is much more successful. It can reproduce all the same location estimates as the full pool using only 32% of the landmarks.

Organization of the paper: Section II summarizes previous work on active geolocation. Section III describes the active geolocation algorithm we tested our selection techniques with, and the landmarks and targets we used. Section IV presents each selection technique we tested and analyzes the results. Section V closes with some discussion.

II. Previous Work

Research into algorithms for active geolocation has been ongoing for more than two decades. The main line of research focuses on increasingly sophisticated statistical models of the relationship between packet travel time and distance [e.g. 8, 14, 20, 21, 36]. Some proposed algorithms incorporate information from route traces and/or prior knowledge of where hosts are likely to be [e.g. 19, 23, 32, 33, 37].

While sophisticated models can be useful at the scale of cities or provinces, Katz-Bassett et al. [19] reported in 2006 that *all* the models they tried were unreliable at larger scales, where queuing delay and network topology play a strong role. Weinberg et al. [35] found that this was still true a decade later. Candela et al. [3] quantified the problem: one-way travel times greater than 30 milliseconds are likely to be dominated by effects other than great-circle distance. This is especially true when the route takes a long detour, or when the "last hop" link is slow, both of which are common for routes outside Europe and North America [4]. Xie et al. [37] measured patterns of variation in end-to-end travel time over long distances and found that no simple mathematical model could account for them, even within Europe and North America. Thus, no matter how sophisticated the model is, landmarks far away from a target are much less useful than those near by. When landmarks near a target are available, Darwich et al. [5] find that only a few are necessary to produce a location estimate good to within tens of kilometers.

The conclusion is clear: active geolocation should use only landmarks near the target. In addition to the accuracy benefits, this should reduce the number of landmarks needed enough to make the process be fast and non-disruptive. But if the target could be anywhere in the world, how do you know which landmarks are near it? Only a few researchers have tried to address this question. Hu et al. [16] tackles it via amortization: assuming that all the hosts in each /24 subnet are physically near each other, it is only necessary to use all available landmarks on a few "representatives" in each subnet to identify appropriate landmarks to use for the rest of them. Unfortunately, Jiang et al. [18] and Darwich et al. [5] have both found that this assumption does not hold anymore. Even if it were, it only helps when one is trying to geolocate many hosts in every subnet of interest.

Du et al. [7] instead use prior information about the geographical service area of the target's AS to make an informed choice of what landmarks are likely to be nearby. Jiang et al. [18] use an incremental approach: choose a landmark at random to ping the target, estimate where the target might be, then choose additional landmarks that are best able to reduce the uncertainty in that estimate. These approaches are suitable for geolocating one target at a time.

Our work answers a different question: given a *group* of targets, believed to be distributed over the entire world, and a pool of available landmarks, also distributed over the entire world, what is the smallest subset of the pool that can accurately geolocate *all* of the targets? While this is still an exercise in optimizing the use of landmarks, it has more in common with Metis [1] and Holterbach et al. [15], both of which seek to choose small numbers of "probes" that can, together, collect a representative sample of network traffic all over the world. Metis in particular works by optimizing distance diversity, much as we do (see Section IV-B).

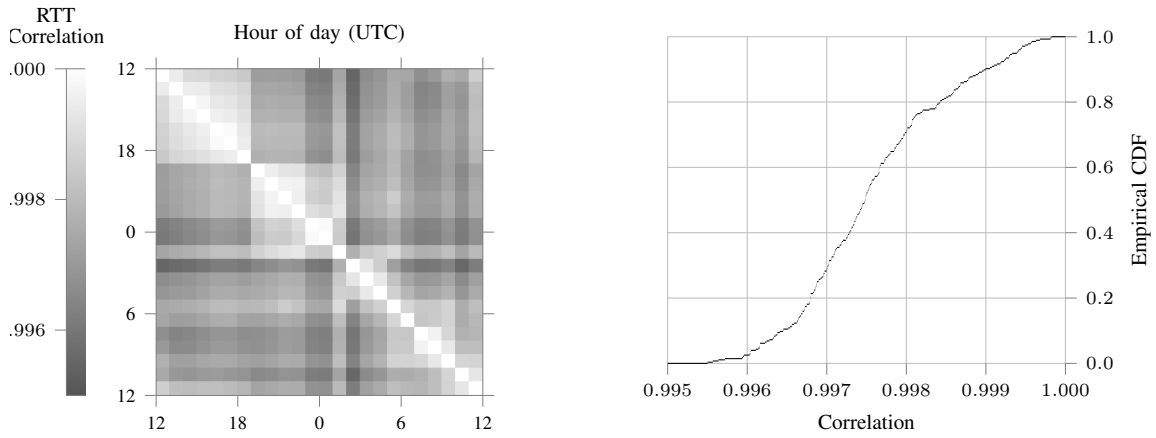
III. Geolocation Landmarks, Targets, and Algorithm

In order to test our selection techniques, we needed hosts in known locations to use as landmarks, hosts in *unknown* locations to use as targets, and an active geolocation algorithm.

A. Landmarks

Active geolocation requires a large set of landmark hosts, in known physical locations, distributed all over the world. They must be in reliable, continuous operation, and available on demand to use as the source of ping packets. Public measurement "constellations," such as RIPE Atlas [28], provide just such a set of landmarks.

RIPE Atlas's nodes are divided into two classes. "Probes" are numerous and geographically diverse, but they run on cheap



(a) Correlation matrix. Darker shades indicate *less* correlation.

(b) Cumulative distribution of correlations.

Fig. 2. Correlation across one-hour intervals within a day, of the minimum RTT between each ordered pair of RIPE anchors.

hardware and have limited bandwidth. "Anchors" are fewer, and not as broadly distributed, but they use server-grade hardware and have fast, efficient connectivity. Probes can only generate a few ping packets per second [5], and their network connections may be very slow, systematically inflating the distance estimates they produce [4, 12]. Also, probes are operated by volunteers and may be deactivated or moved without notice [2]. Largely for this last reason, we elected to use only anchors for this study. However, including probes as geolocation landmarks does improve geolocation precision especially in Africa and South America [3]. Algorithms that incrementally geolocate a single target [e.g. 7, 18] can make effective use of probes despite their limitations. Finding a way to apply this kind of algorithm to the geolocation of many targets at once is a priority for our future work. At the time of the study (December, 2022), there were 780 anchors; their geographic distribution is summarized in Table I.

The anchors are programmed to measure round-trip times between themselves and all the other anchors, constantly, and upload the results to a public database [3, 6]. We can calibrate active geolocation algorithms for the RIPE anchors using only the data in this database.

Routes through the global Internet are constantly changing. The typical latency of a long-distance route is stable over a period of hours to single-digit days, but not longer than that [6]. This could cause the calibration of an active geolocation algorithm to degrade rapidly. To get a sense of *how* rapidly, we retrieved one full day's worth of anchor-to-anchor RTT measurements, and computed the correlation of minimum RTTs from hour to hour. The correlation matrix is shown in Fig. 2, along with the statistical distribution of correlations (empirical cumulative distribution function). Changes from hour to hour are visible in the matrix—notably, RTTs within the UTC 12h00–19h00 period, typical working hours in the Americas, are better correlated with each other than with RTTs observed at any other time of day. However, the minimum RTT at any given hour is always at least 99.5% correlated with the minimum RTT at any other hour. Thus, we feel safe assuming that calibrations will not degrade significantly over the course of a single day, and that the time of day when we measure RTTs to target nodes is unimportant.

B. Targets

We used 559 commercial VPN endpoints as the targets for geolocation; their geographic distribution is summarized in Table II. Commercial VPNs are convenient vantage points for global network measurements [25, 35], but at the same time, the VPN services have strong incentives to make exaggerated claims of global coverage while actually concentrating their servers in operationally convenient locations [4, 27]. Location claims by VPN services should, therefore, be verified before using VPN servers in research.

TABLE I
WORLDWIDE DISTRIBUTION OF RIPE ATLAS ANCHORS, AND CITIES AND COUNTRIES THAT CONTAIN THEM, AT THE TIME OF THE STUDY.

Continent	# of countries	cities	landmarks
Asia	31	71	122
Europe	36	270	438
South America	8	20	28
Oceania	3	11	25
Africa	9	14	18
North America	9	95	149

TABLE II
WORLDWIDE DISTRIBUTION OF TARGETS USED IN THIS STUDY, AND THE COUNTRIES HOSTING THEM, ACCORDING TO THE TARGETS' OPERATORS.

Continent	# of countries	targets
Asia	51	110
Europe	47	120
South America	13	27
Oceania	20	41
Africa	50	103
North America	34	176

C. Active Geolocation Algorithm

We test our landmark selection techniques using a simple active geolocation algorithm developed for internal use by ICLab [25, 35]. Given a set of landmarks, a target, and a "claimed location"—a country where the target supposedly is—ICLab's algorithm produces a yes-or-no judgment of whether the claimed location is accurate. It works by assuming each measurement packet traveled the shortest possible great-circle distance from its source landmark to the nearest border of the claimed country, and no further. Each measured RTT is converted to the minimum *speed* that the measurement packet would have had to travel to cover that distance. If any of these minimum speed estimates is greater than a calibrated speed limit, the claimed location is rejected, otherwise it is accepted.

The speed limit is calibrated using a simplified version of the calibration procedure for CBG [14]: For all pairs of landmarks, divide the distance between that pair by the minimum RTT measured between that pair, producing a travel speed estimate. Take the fastest of all such speed estimates as the speed limit. For this study, we used a calibrated speed limit of 153 km/ms ($0.51\,c$). This is a little higher than Katz-Bassett et al. [19]'s estimated "speed of Internet" (133 km/ms, $0.44\,c$), but well below the theoretical limit of 200 km/ms ($0.67\,c$), the speed of light in long-distance optical fiber.[1] We presume the difference from Katz-Bassett et al.'s estimate reflects improvements to global network latency since 2006.

In most cases, measurement packets will have to travel much farther than just to the nearest border of the claimed country. Dividing an RTT measurement by a shorter distance than the packet actually would have had to travel produces a speed estimate lower than the packet's true speed. Therefore, ICLab's algorithm errs systematically on the side of acceptance. Reducing the number of landmarks used for a measurement can only increase this systematic error, because removing data points from a measurement that rejected a claim can convert it into a measurement that accepts a claim, but not vice versa. In the rest of the paper, whenever we discuss "agreement" between results from the full landmark pool and results from a reduced set of landmarks, keep in mind that each disagreement means the data from the full landmark pool rejected a claim and the data from the reduced set did not.

D. Ethical issues

Our experiments generated a relatively small amount of network traffic: roughly two million ICMP Echo Request and Echo Reply packets, generated using the RIPE Atlas measurement API, over the course of five days. All were transmitted between RIPE anchors and commercial VPN servers. Each VPN server received 2400 packets over the course of a few minutes: not enough to disrupt service. We are customers of the VPN services whose servers we located, and we honored the terms of service of both RIPE Atlas and the VPN services.

[1]The refractive index of fused silica is 1.45–1.55 depending on wavelength [24]. "Hollow core" fibers with an effective refractive index very close to 1 are not yet usable for long-distance communication [11].

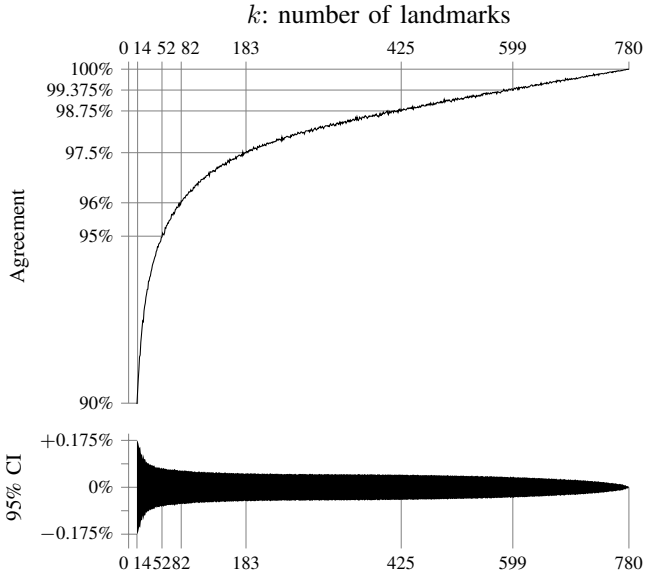


Fig. 3. Mean agreement with the full landmark pool, for randomly selected landmark sets of all possible sizes. Relative 95% confidence interval presented on a different scale for legibility.

We recorded only the round-trip time of each ping exchange plus public metadata about the hosts involved, such as IP addresses, hostnames, and AS numbers. None of the hosts involved were personal computers, therefore this metadata is not personally identifying information.

IV. Landmark Selection

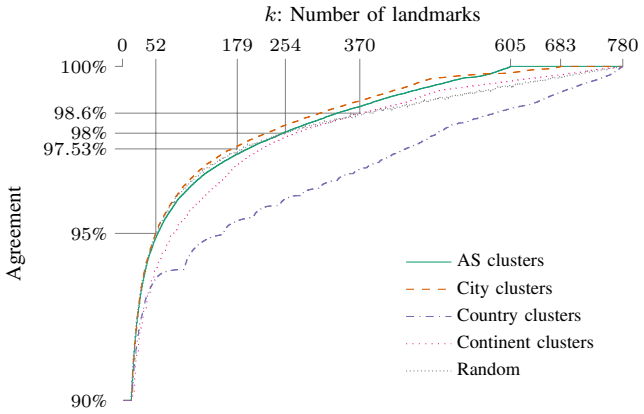
In this section, we describe each of the landmark selection algorithms we tested and characterize their performance relative to the full pool of available landmarks.

A. Random Selection

Mathematical analysis of sampling from large scale-free graphs indicates that a small, purely random sample is likely to represent the complete graph as well or better than any more structured sample [22].

Therefore, we expect that a small random sample will give us a reasonable first estimate of how many landmarks are actually *needed* for accurate geolocation of a target. We generated subsets of the landmark pool uniformly at random (without replacement). For each possible subset size (1 to 780), we selected 1,000 random subsets. Fig. 3 shows the mean agreement between the geolocation results obtained from random subsets and the geolocation results obtained from the complete landmark pool, as a function of subset size, with a 95% confidence interval.

We need only fourteen randomly selected landmarks to match 90% of the geolocation results from the complete landmark pool. However, the marginal benefit of adding more randomly selected landmarks falls off rapidly thereafter. 95% agreement requires 50 landmarks, 97.5% agreement requires 180 landmarks, and full agreement is not achieved until all landmarks are in use. This matches the prediction of sampling



(a) Agreement with the full landmark pool for cluster-based subsets of all sizes.

Type	# clusters	Mean agreement vs. random
ASes	534	99.28% > 99.20%
Cities	481	99.62 > 98.99
Countries	96	93.88 ≪ 96.32
Continents	6	83.08 < 84.05

(b) Cluster types, counts, and agreement for subset size = number of clusters.

Fig. 4. Effectiveness of clustering selection.

theory, and confirms that it is *possible* to geolocate a target using only a small number of landmarks, but doing so reliably will require more strategic selection of the landmarks.

B. Diversity Metrics

Landmarks that are geographically and topologically close to each other can be expected to produce similar round-trip time measurements for the same target. Therefore, we suspect we can find a small set of landmarks that can still accurately geolocate all targets by maximizing diversity of the landmarks' locations. Mathematically, let $\mathbb{V} = \{v_1, \ldots, v_n\}$ denote the full pool of landmarks. For any subset of the landmarks, $\mathbb{S} \subseteq \mathbb{V}$, define the diversity $D(\mathbb{S})$ with respect to a distance metric $d(a, b)$ as:

$$D(\mathbb{S}) = \sum_{s,\, t \in \mathbb{S}} d(s, t) \tag{1}$$

The hypothesis is that small landmark subsets that maximize $D(\mathbb{S})$, with respect to a well-chosen distance metric d, can yield geolocation results that perfectly agree with the results from the full landmark pool.

C. Clustering Selection

Our first few candidates for d will group the landmarks into *clusters* based on the Autonomous System (AS) that operates them, or the city, country, or continent where they are physically located. Intuitively, AS clusters should reflect Internet topology, whereas location clusters reflect geography. The cluster distance metric is binary:

$$d_C(a, b) = \begin{cases} 0 & \text{if } C(a) = C(b) \\ 1 & \text{if } C(a) \neq C(b) \end{cases} \tag{2}$$

where $C(v)$ identifies the cluster that landmark v belongs to.

For each of the four types of cluster, we generated 1,000 landmark subsets using one landmark selected at random from each cluster. (Thus, the subset size is equal to the number of clusters.) The mean agreement for all the subsets of one type is shown in Fig. 4b, along with the size of those subsets, and a comparison with subsets of the same size selected purely at random. Country- and continent-based cluster selection does not outperform random selection; city- and AS-based selection based does outperform random selection, but only slightly.

We then extended this procedure to subsets of all sizes by choosing landmarks at random with the constraint that, as far as is possible, every cluster must contribute the same number of landmarks. For example, a subset of size 13 using continent clusters would have two landmarks from five of the six continents and three from the sixth. Which continent was the sixth would be randomized. Fig. 4a shows mean agreement with the full landmark pool for cluster-based subsets of all sizes. We generated 1,000 landmark subsets for each cluster. City- and AS-based selection outperforms random selection for most subset sizes, and both can achieve perfect (100%) agreement without using the entire pool. Continent-based selection can also outperform random, but only when 370 landmarks or more are used, and only by a little; country-based selection is always worse than random. This is probably because country-based clustering does not choose enough landmarks within geographically large countries. Most city and AS clusters contain only one or two landmarks, and none contain more than 26. Country clusters are much more uneven: 31 country clusters also have only one landmark, but another 31 have more than five, and two have more than 100.

D. Greatest-Distance Selection

Another pair of candidates for d use actual geographic distances,

$$d_g(a, b) = \begin{array}{l}\text{distance between } a \text{ and } b \text{ over the} \\ \text{surface of the Earth}\end{array} \tag{3}$$

and measured round-trip times as a proxy for distances through the network,

$$d_r(a, b) = \text{minimum RTT measured from } a \text{ to } b. \tag{4}$$

Note that the metric d_r might not be symmetric (i.e. it's possible that $d_r(a, b) \neq d_r(b, a)$), making it not formally a distance metric, but this does not cause problems as $D(\mathbb{S})$ always counts both $d(a, b)$ and $d(b, a)$.

To select landmark subsets that maximize D for either d_g or d_r, we use a greedy algorithm inspired by Prim's algorithm for spanning trees [17]. For any landmark $v \in \mathbb{V}$, define $m(v)$ as the largest distance from v to any other landmark, according to the d in use:

$$m(v) = \max_{w \in \mathbb{V}} d(v, w) \tag{5}$$

One variant of the algorithm begins by choosing a landmark within the target's claimed location, whenever possible, and failing that, a landmark that maximizes m. The other variant

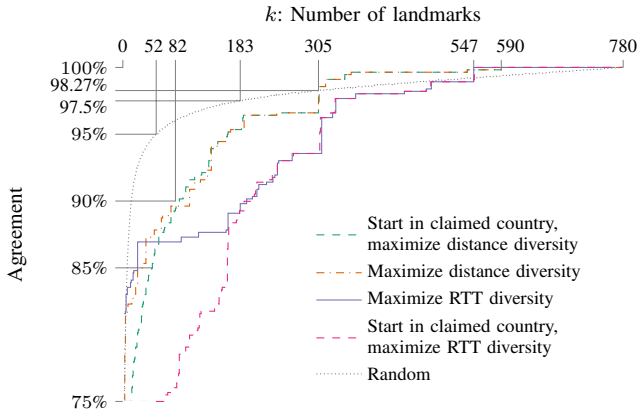


Fig. 5. Distance-diversity maximization: agreement with the full landmark pool.

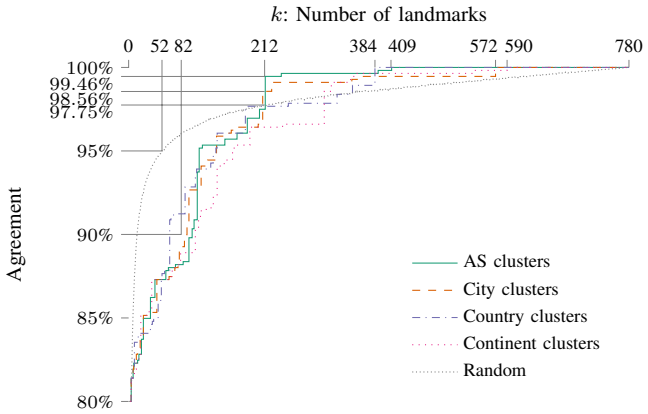


Fig. 6. Hybrid 1 (clustering + geographic distance maximization): agreement with the full landmark pool.

always begins with a landmark that maximizes m. In both cases, we continue by choosing landmarks that maximize m from the landmarks that have not already been selected, until we reach the desired subset size k. Whenever there is more than one landmark that meets the criterion, we pick one at random. (This algorithm is greedy because adding to $\mathbb{S}$ a landmark from $\mathbb{V}\setminus\mathbb{S}$ that maximizes m increases D as much as possible at that stage of the process. Since adding a landmark to $\mathbb{S}$ can never decrease $D(\mathbb{S})$, nor can it prevent any other landmark from being added to $\mathbb{S}$ in the future, a greedy algorithm produces an optimal solution.)

Because it is unusual for any two pairs of landmarks to be at *exactly* the same distance from each other, greatest-distance selection makes far fewer random choices than clustering selection. Therefore, we only generated one landmark subset of each possible size using each possible variant of the algorithm. Fig. 5 shows agreement with the full landmark pool for all four variants as a function of subset size. For small sizes, none of the variants can outperform random selection. However, geographic distance maximization becomes better than random selection when 305 or more landmarks are in use (39% of the pool). It achieves perfect agreement with the full landmark pool when 590 or more landmarks are in use (75.6% of the pool), which random selection never does.

RTT-distance maximization is less effective than geographic distance maximization for small to medium-sized subsets, but when 547 or more landmarks are in use (70% of the pool) it passes both random selection and geographic distance maximization and jumps all the way to perfect agreement. These phenomena may be related to the nonlinear relationship between great-circle routes and actual network paths over longer distances [4, 9, 37]. Since the difference between 547 and 590 is small, and geographic distance maximization is more effective than RTT-distance maximization for most subset sizes, we did not experiment any further with RTT-distance maximization.

The different rules for choosing the first landmark have a notable effect on agreement for small subset sizes in both RTT-distance and geographic distance maximization. But, this effect ceases to make a difference when 182 or more landmarks for RTT-distance and 132 or more landmarks for geographic distance are in use. For this reason, we did not experiment any further with using a landmark in the claimed country as the first landmark.

E. Hybrid Selection

Random and clustering selection outperforms geographic distance maximization at small subset sizes; the opposite is true for large subset sizes. This suggests that hybrid approaches might perform better than any one alone.

1) Hybrid 1: Clustering and Greatest Distance: To combine clustering with greatest-distance selection, we define a hybrid metric,

$$d_h(a,b) = W_C \cdot d_C(a,b) + W_d \cdot d_d(a,b) \tag{6}$$

where d_C is any of the clustering metrics, d_d is the geographic distance metric, and W_C and W_d are weights. Landmark subsets that maximize diversity according to this metric can be generated using the same greedy algorithm as we used to maximize diversity according to distance alone.

We only tested $W_d = 1$ and W_C greater than the maximum d_d, i.e. cluster diversity is given overwhelmingly more weight than distance diversity. These weights permit a minor optimization: if the landmark subset does not include a representative from each cluster, select the next landmark from clusters that are not yet represented.

Fig. 6 shows agreement with the full landmark pool for landmark subsets that maximize the hybrid metric, using all four types of clusters, as a function of subset size. Unlike clustering by itself, the hybrid is worse than random selection at small subset sizes. However, if either AS or city-based clustering is used, the hybrid becomes better than random selection when 210 landmarks or more are in use (27% of the pool). This is 100 fewer landmarks than were required to beat random selection using geographic distance maximization alone. Country-based clustering is also better than random selection at 210 landmarks or so, but only by a tiny fraction, and it falls

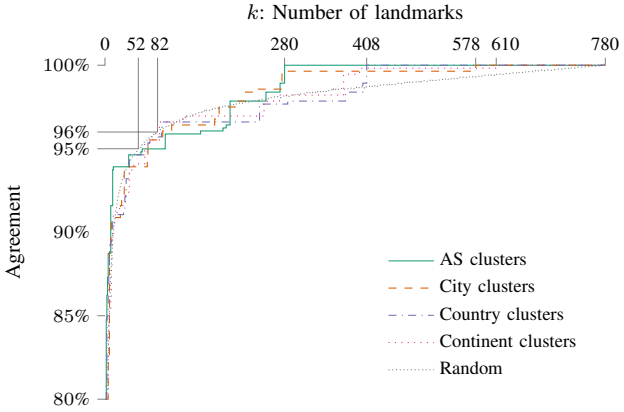


Fig. 7. Hybrid 2 (first 100 random, then clustering + geographic distance maximization): agreement with the full landmark pool.

TABLE III
NUMBER OF LANDMARKS REQUIRED TO ACHIEVE 100% AGREEMENT WITH THE FULL POOL, FOR EACH SELECTION PROCEDURE THAT ACHIEVED 100% AGREEMENT WITHOUT USING THE FULL POOL.

				# landmarks to...	
Shorthand	Metric	Cluster by	First 100 random?	beat random	perfect agreement
CLUSTER-CITY		Cities		179	683
H2-CONTINENT	Geodesic	Continents	Yes	85	610
CLUSTER-AS		ASes		254	605
DIST-GEO	Geodesic			305	590
H1-CONTINENT	Geodesic	Continents		305	590
H2-CITY	Geodesic	Cities	Yes	179	578
DIST-RTT	Travel time			547	547
H1-AS	Geodesic	ASes		213	410
H2-COUNTRY	Geodesic	Countries	Yes	88	408
H1-COUNTRY	Geodesic	Countries		182	384
H2-AS	Geodesic	ASes	Yes	195	280

off again with more. However, the hybrid with country-based clustering performs much better than country-based clustering alone, confirming our suspicion that country-based clustering performs poorly by itself because of inadequate geographic diversity within large countries. Strikingly, this variant achieves perfect agreement with the full pool before any of the other variants, at 384 landmarks (49% of the pool). AS clustering is not far behind, reaching perfect agreement at 409 landmarks; city and continent clustering don't get there until 572 and 590 landmarks, respectively. (Continent clustering is almost indistinguishable from geographic distance maximization alone. This is because there are only six continents.)

Overall, AS-based clustering is the best choice of these four variants, outperforming both random selection and the other three variants for most subsets of 212 or more landmarks.

2) Hybrid 2: Random, Then Hybrid 1: None of the selection algorithms tested so far is substantially better than purely random selection for small subset sizes. Random selection reaches 96% agreement with the full pool using only 82 landmarks; distance maximization and hybrid 1 can only reach 90% agreement with 82 landmarks.

The last variation we investigated was to begin the process by selecting up to 100 landmarks at random and then expand those subsets using hybrid 1. Fig. 7 shows agreement with the full landmark pool for the landmark subsets generated this way, using all four types of clusters, as a function of subset size. (The hybrid curves deviate from the comparison even when $k \leq 100$ because the "random" comparison curve is an average of 1000, whereas each of the four hybrid curves in Fig. 7 is based on just one landmark set of each size.)

Comparing with Fig. 6, the principal effect of this modification is to bring performance up to parity with random selection, or nearly so, throughout the range—not just for small subsets. The modification also accelerates AS-based clustering's convergence to full agreement with the landmark pool, which it now reaches at only 280 landmarks, 130 fewer than were required for hybrid 1. However, the other three cluster types require slightly *more* landmarks to reach full agreement.

F. Summary and Analysis

Table III summarizes the performance of each selection procedure that was able to achieve 100% agreement with the full landmark pool without using all available landmarks. (Thus, pure random selection and two variants of cluster selection are omitted. Geodesic maximization starting from the claimed country is also omitted because its performance is identical to pure geodesic maximization.) They are sorted in descending order of the number of landmarks needed for 100% agreement.

Although many of the procedures we tested can reduce the number of landmarks needed for 100% agreement somewhat, there is a clear winner: algorithm H2-AS (hybrid 2 with AS-based clustering) requires only 280 landmarks, 36% of the pool, for perfect agreement. Assuming that each landmark sends three ICMP Echo Request packets to each target for geolocation, the total number of request packets required to geolocate 559 targets would be reduced from 1,308,060 to 469,560, and we could expect the time required for the process to drop proportionally.

Use of (partial) random selection introduces the possibility of geolocation results being unstable over time. Perfect agreement using 280 landmarks today might not be perfect agreement anymore next week, because a different group of landmarks might be selected and an important one might get left out. If this is a concern, algorithm H1-COUNTRY is the runner-up, achieving perfect agreement with 384 landmarks (49% of the pool; 643,968 request packets required), and it will always pick the same subset for a given k.

If one is willing to accept some deviation from the results produced by the full landmark pool, then algorithm H1-AS (hybrid 1 with AS-based clustering) should also be considered, as it achieves 99.46% agreement using only 213 landmarks, 27% of the pool. 355,527 ICMP packets would need to be sent. Algorithms H2-AS, H2-CONTINENT, and H1-COUNTRY do not *significantly* outperform random selection up till the point where they jump to 100% accuracy, despite their low numbers in the "beat random" column of Table III.

V. Conclusion

In this paper, we demonstrated that it is possible to reduce by two-thirds the number of landmarks required to actively geolocate an entire set of target hosts, with no change in the overall results. If small changes are acceptable, the number can be reduced further.

We wish to highlight how much more effective cluster-based selection was when clusters were defined based on cities or ASes than when they were defined based on countries or continents. This once again demonstrates the importance of fine-grained diversity in one's measurement constellation, consistent with previous observations by Candela et al. [3] and Appel et al. [1].

The fact that geographic distance diversity is a more effective selection criterion than RTT-based (topological) distance diversity may seem to go against expectations from previous work such as Dang et al. [4] and Xie et al. [37]. If long-distance RTTs are dominated by factors other than geographic distance, shouldn't it be more effective to maximize diversity according to a metric that honors those factors? This logic is incorrect because those other factors are *confounding* factors. Adding a landmark that's a long way away from the current subset as measured by round-trip time may mean adding a landmark that has large RTTs to *everything*, and therefore does not contribute effectively to *any* location estimate. Adding landmarks far away in geographic distance, on the other hand, is an effective way to select landmarks that are near *targets* far away from the current subset, and therefore poorly located.

A priority for future work is to combine our selection rules with an incremental active geolocation algorithm, such as those described by Du et al. [7] and Jiang et al. [18]. We expect that the combination will perform better than either alone, both by reducing the number of landmarks needed even more, and by rapidly locating the target within a region small enough to use sophisticated delay-distance models. Also, unlike ICLab's simple geolocation algorithm, incremental algorithms can make effective use of RIPE Atlas probes. They were excluded from this study because their locations might be outdated [2], and because they might be attached to slow networks [12]. Incremental geolocation is naturally robust to landmarks that produce high distance estimates for either of these reasons, and including probes will give us much better geographic diversity, especially outside of Europe and North America.

Acknowledgments

We would like to thank our shepherd, Brian Trammell, and the anonymous reviewers for their feedback. We acknowledge the contributions of Evelyn Gao, Maggie Hollis, and Way Zheng, who were part of this project for a limited period. This work was partially funded by Science and Engineering Research Board (SERB), Department of Science and Technology of Government of India, under the CORE Research Grant (CRG/2022/005096).

References

[1] M. Appel, E. Aben, and R. Fontugne, "Metis: Better Atlas Vantage Point Selection for Everyone," in *Network Traffic Measurement and Analysis*, 2022. https://tma.ifip.org/2022/wp-content/uploads/sites/11/2022/06/tma2022-paper18.pdf

[2] V. Bajpai, S. J. Eravuchira, J. Schönwälder, R. Kisteleki, and E. Aben, "Vantage Point Selection for IPv6 Measurements: Benefits and Limitations of RIPE Atlas Tags," in *Integrated Network and Service Management*, 2017, pp. 37–44. DOI:10.23919/INM.2017.7987262

[3] M. Candela, E. Gregori, V. Luconi, and A. Vecchio, "Using RIPE Atlas for Geolocating IP Infrastructure," *IEEE Access*, vol. 7, pp. 48 816–48 829, 2019. DOI:10.1109/ACCESS.2019.2909691

[4] T. K. Dang, N. Mohan, L. Corneo, A. Zavodovski, J. Ott, and J. Kangasharju, "Cloudy with a Chance of Short RTTs: Analyzing Cloud Connectivity in the Internet," in *Internet Measurement Conference*, 2021, pp. 62–79. DOI:10.1145/3487552.3487854

[5] O. Darwich, H. Rimlinger, M. Dreyfus, M. Gouel, and K. Vermeulen, "Replication: Towards a publicly available internet scale ip geolocation dataset," in *Internet Measurement Conference*, 2023, pp. 1–15. DOI:10.1145/3618257.3624801

[6] L. Davisson, J. Jakovleski, N. Ngo, C. Pham, and J. Sommers, "Reassessing the Constancy of End-to-End Internet Latency," in *Network Traffic Measurement and Analysis*, 2021. https://par.nsf.gov/biblio/10386604

[7] B. Du, M. Candela, B. Huffaker, A. C. Snoeren, and k. claffy, "RIPE IPmap Active Geolocation: Mechanism and Performance Evaluation," *SIGCOMM Computer Communication Review*, vol. 50, no. 2, pp. 3–10, 2020. DOI:10.1145/3402413.3402415

[8] B. Eriksson, P. Barford, J. Sommers, and R. Nowak, "A Learning-based Approach for IP Geolocation," in *Passive and Active Measurement*, 2010, pp. 171–180. DOI:10.1007/978-3-642-12334-4_18

[9] R. Fanou, P. Francois, and E. Aben, "On the Diversity of Interdomain Routing in Africa," in *Passive and Active Measurement*, 2015, pp. 41–54. DOI:10.1007/978-3-319-15509-8_4

[10] A. Filasto and J. Appelbaum, "OONI: Open Observatory of Network Interference," in *Free and Open Communications on the Internet*, 2012. https://www.usenix.org/system/files/conference/foci12/foci12-final12.pdf

[11] E. N. Fokoua, S. A. Mousavi, G. T. Jasion, D. J. Richardson, and F. Poletti, "Loss in hollow-core optical fibers: mechanisms, scaling rules, and limits," *Advances in Optics and Photonics*, vol. 15, no. 1, pp. 1–85, 2023. DOI:10.1364/AOP.470592

[12] R. Fontugne, A. Shah, and K. Cho, "Persistent Last-mile Congestion: Not so Uncommon," in *Internet Measurement Conference*, 2020, pp. 420–427. DOI:10.1145/3419394.3423648

[13] M. Gharaibeh, A. Shah, B. Huffaker, H. Zhang, R. Ensafi, and C. Papadopoulos, "A Look at Router Geolocation in Public and Commercial Databases," in *Internet Measurement Conference*, 2017, pp. 463–469. DOI:10.1145/3131365.3131380

[14] B. Gueye, A. Ziviani, M. Crovella, and S. Fdida, "Constraint-based geolocation of internet hosts," in *Internet Measurement Conference*, 2004, pp. 288–293. DOI:10.1145/1028788.1028828

[15] T. Holterbach, E. Aben, C. Pelsser, R. Bush, and L. Vanbever, "Measurement Vantage Point Selection using a Similarity Metric," in *Applied Networking Research Workshop*, 2017, pp. 1–3. DOI:10.1145/3106328.3106334

[16] Z. Hu, J. Heidemann, and Y. Pradkin, "Towards geolocation of millions of IP addresses," in *Proceedings of the 2012 Internet Measurement Conference*, 2012, p. 123–130. DOI:10.1145/2398776.2398790

[17] V. Jarník, "O jistém problému minimálním," *Práce moravské přírodovědecké společnosti*, vol. 6, no. 4, pp. 57–63, 1930. DOI:10338.dmlcz/500726

[18] N. Jiang, J. H. Wang, J. Wang, and P. Wang, "TinyG: Accurate IP Geolocation Using a Tiny Number of Probers," in *Network and Service Management*, 2023, pp. 1–9. DOI:10.23919/CNSM59352.2023.10327884

[19] E. Katz-Bassett, J. P. John, A. Krishnamurthy, D. Wetherall, T. Anderson, and Y. Chawathe, "Towards IP Geolocation using Delay and Topology Measurements," in *Internet Measurement Conference*, 2006, pp. 71–84. DOI:10.1145/1177080.1177090

[20] K. Kohls and C. Diaz, "VerLoc: Verifiable Localization in Decentralized Systems," in *USENIX Security Symposium*, 2022, pp. 2637–2654. https://www.usenix.org/conference/usenixsecurity22/presentation/kohls

[21] S. Laki, P. Mátray, P. Hága, T. Sebők, I. Csabai, and G. Vattay, "Spotter: A Model Based Active Geolocation Service," in *International Conference on Computer Communications*, 2011, pp. 3173–3181. DOI:10.1109/INFCOM.2011.5935165

[22] J. Leskovec and C. Faloutsos, "Sampling from Large Graphs," in *Knowledge Discovery and Data Mining*, 2006, pp. 631–636. DOI:10.1145/1150402.1150479

[23] C. Liu, X. Luo, F. Yuan, and F. Liu, "RNBG: A Ranking Nodes Based IP Geolocation Method," in *International Conference on Computer Communications Workshops*, 2020, pp. 80–84. DOI:10.1109/INFOCOMWKSHPS50562.2020.9162976

[24] I. H. Malitson, "Interspecimen Comparison of the Refractive Index of Fused Silica," *Journal of the Optical Society of America*, vol. 55, no. 10, pp. 1205–1209, 1965. DOI:10.1364/JOSA.55.001205

[25] A. A. Niaki, S. Cho, Z. Weinberg, N. P. Hoang, A. Razaghpanah, N. Christin, and P. Gill, "ICLab: A Global, Longitudinal Internet Censorship Measurement Platform," in *Security and Privacy*, 2020, pp. 135–151. DOI:10.1109/SP40000.2020.00014

[26] I. Poese, S. Uhlig, M. A. Kaafar, B. Donnet, and B. Gueye, "IP Geolocation Databases: Unreliable?" *SIGCOMM Computer Communication Review*, vol. 41, no. 2, pp. 53–56, 2011. DOI:10.1145/1971162.1971171

[27] R. Ramesh, L. Evdokimov, D. Xue, and R. Ensafi, "VPNalyzer: Systematic Investigation of the VPN Ecosystem," in *Network and Distributed System Security*, 2022, pp. 24–28. https://www.ndss-symposium.org/wp-content/uploads/2022-285-paper.pdf

[28] RIPE NCC Staff, "RIPE Atlas: A Global Internet Measurement Network," *Internet Protocol Journal*, vol. 18, no. 3, pp. 2–26, 2015. https://www-static.ripe.net/static/rnd-ui/atlas/static/page/InternetProtocolJournal-18-3.pdf

[29] Q. Scheitle, O. Gasser, P. Sattler, and G. Carle, "HLOC: Hints-based Geolocation Leveraging Multiple Measurement Frameworks," in *Network Traffic Measurement and Analysis*, 2017, pp. 1–9. DOI:10.23919/TMA.2017.8002903

[30] Y. Shavitt and N. Zilberman, "A Geolocation Databases Study," *Selected Areas in Communications*, vol. 29, no. 10, pp. 2044–2056, 2011. DOI:10.1109/JSAC.2011.111214

[31] R. Sundara Raman, P. Shenoy, K. Kohls, and R. Ensafi, "Censored Planet: An Internet-wide, Longitudinal Censorship Observatory," in *Computer and Communications Security*, 2020, pp. 49–66. DOI:10.1145/3372297.3417883

[32] W. Tai, B. Chen, T. Zhong, Y. Wang, K. Chen, and F. Zhou, "RIPGeo: Robust Street-Level IP Geolocation," in *Mobile Data Management*, 2023, pp. 138–147. DOI:10.1109/MDM58254.2023.00031

[33] Y. Wang, D. Burgener, M. Flores, A. Kuzmanovic, and C. Huang, "Towards Street-Level Client-Independent IP Geolocation," in *Networked Systems Design and Implementation*, vol. 8, 2011. https://www.usenix.org/event/nsdi11/tech/full_papers/Wang_Yong.pdf

[34] Z. Weinberg, personal communication, 2024.

[35] Z. Weinberg, S. Cho, N. Christin, V. Sekar, and P. Gill, "How to Catch When Proxies Lie: Verifying the Physical Locations of Network Proxies with Active Geolocation," in *Internet Measurement Conference*, 2018, pp. 203–217. DOI:10.1145/3278532.3278551

[36] B. Wong, I. Stoyanov, and E. G. Sirer, "Octant: A Comprehensive Framework for the Geolocalization of Internet Hosts," in *Networked Systems Design and Implementation*, vol. 4, 2007. https://www.usenix.org/legacy/events/nsdi07/tech/full_papers/wong/wong.pdf

[37] Y. Xie, Z. Zhang, Y. Liu, E. Chen, and N. Li, "Evaluation Method of IP Geolocation Database Based on City Delay Characteristics," *Electronics*, vol. 13, no. 1, 2024. DOI:10.3390/electronics13010015

[38] F. Zhang, C. Lu, B. Liu, H. Duan, and Y. Liu, "Measuring the Practical Effect of DNS Root Server Instances: A China-Wide Case Study," in *Passive and Active Measurement*, 2022, pp. 247–263. DOI:10.1007/978-3-030-98785-5_11