IP Geolocation Database Stability and Implications for Network Research

Matthieu Gouel, Kevin Vermeulen, Olivier Fourmaux, Timur Friedman, Robert Beverly

IP Geolocation Database Stability and Implications for Network Research

Matthieu Gouel
Sorbonne Université
matthieu.gouel@sorbonne-universite.fr

Kevin Vermeulen
Columbia University
kevin.vermeulen@columbia.edu

Olivier Fourmaux
Sorbonne Université
olivier.fourmaux@sorbonne-universite.fr

Timur Friedman
Sorbonne Université
timur.friedman@sorbonne-universite.fr

Robert Beverly
Naval Postgraduate School
rbeverly@nps.edu

***Abstract*—IP geolocation has myriad applications. While a body of prior research has investigated the accuracy of geolocation databases, we take a first look at their *stability*. Using a large collection of snapshots from a popular geolocation database, we examine the longitudinal evolution of its location mappings and address coverage. Across different classes of IP addresses, we find that significant differences can exist even between two successive weekly snapshots – a previously underappreciated source of potential error. To assess the sensitivity of research results to the geo database instance, we examine a prior study [1] that used geolocation. Using their data and methodology, we generate results for each database instance available during their measurement period, i.e., the hypothetical results had the authors used a different snapshot. We show that the median distance of addresses considered shifted over 100km from ground truth and the coverage differed by 30% – potentially impacting the conclusions of this prior study. Based on our findings, we recommend best practices when using geolocation databases for network research to encourage reproducibility and soundness.**

I. Introduction

Determining the physical location of Internet hosts is important for a range of applications including, but not limited to, advertising, content and language customization, security and forensics, and policy enforcement [2], [3], [4]. However, the Internet architecture includes no explicit notion of physical location and hosts may be unable or unwilling to share their location. As a result, the process of third-party IP geolocation – mapping an IP address to a physical location – emerged as a research topic [5] more than two decades ago and has since matured into commercial service offerings, e.g., [6], [7], [8].

IP addresses represent network attachment points, thus IP geolocation is often inferential. Commercial geolocation providers compete, so the methodologies for creating their databases are proprietary. State-of-the-art techniques include combining latency constraints [9], topology [2], registries [5], public data [10], and privileged feeds [11].

This work takes a fresh look at IP geolocation data from a *temporal* perspective. Specifically, we examine the longitudinal stability of locations in an IP geolocation database, the characteristics of location changes when they do occur, and the extent to which a particular instance of a geolocation database impacts conclusions that depend on locations. To wit, network and systems researchers frequently utilize available IP geolocation database snapshots. However, the date of the snapshot may only loosely align with the time of the lookup operation, or the lookups may span multiple snapshots, e.g., a long-running measurement campaign. We show that snapshots of the same geolocation database separated even closely in time can have a non-trivial effect on research results and findings.

For example, across database snapshots in three month window, we find up to 22% of IP addresses move more than 40km, while coverage (the simple presence or absence of an address in the database) varies by as much as 18%. Despite this temporal sensitivity, the *date* of the geolocation database snapshot is rarely reported in the academic literature – an omission that we show confounds scientific reproducibility.

We use 10 years of data from the most popular, publicly available, and frequently used database: MaxMind [7]. We use this large collection of snapshots to examine the longitudinal evolution of its location mappings and address coverage, as well as to conduct a reproducibility case study. Our contributions include:

- A survey of how recent systems and networking literature utilizes and depends on IP geolocation data.
- The first longitudinal study of a widely used IP geolocation database where we find significant short-term dynamics.
- A case study of prior research that depended on geolocation, showing that the results fundamentally differ based on the instance of the geolocation database used.
- Recommendations for the sound use of IP geolocation data in research.

Our findings provide several tangible lessons for the broader network research community:

- IP locations in geolocation databases can be highly dynamic, with non-negligible coverage and movement differences even over short ($<$ 3 month) time scales.
- Despite this variation, published network research frequently omits specific details of the geolocation snapshot. Not only does this hinder reproducibility, research results that depend on IP geolocation can significantly differ depending on the instance of the snapshot used.

- Researchers should ensure they publish details of the IP geolocation database, align lookups with measurements, and investigate the sensitivity of their results to different instances of the database.

II. Motivation

To better understand IP geolocation as used in the network and systems research community, we surveyed the academic literature. We performed full-text queries, over all time, on four popular digital libraries for three common geolocation databases, MaxMind [7], NetAcuity [12], and IP2Loc [13]. Table I shows the number of papers in each library. MaxMind is clearly the most popular by an order of magnitude. Therefore the analysis in the reminder of this work focuses on MaxMind.

TABLE I: References to geolocation databases

	ACM	IEEE	arXiv.org	Springer
MaxMind	171	373	96	162
NetAcuity	10	10	8	7
IP2Loc	3	3	0	0

A. MaxMind

Founded in 2002, MaxMind is a commercial entity specializing in IP geolocation and related services. MaxMind offers two IP geolocation databases, one that is free (GeoLite) and one that requires a license (GeoIP). The academic literature uses both GeoLite and GeoIP. GeoLite is available as a complete database "snapshot'.' Snapshots are currently updated weekly and available for public download. GeoLite snapshots contain variable length IP prefixes, each with an associated geolocation. The geolocation may include country, city, latitude/longitude, and accuracy (in km); however many prefixes only provide a geolocation at the country granularity. This work studies the IPv4 GeoLite databases. Henceforth, we refer to GeoLite (and its successor, GeoLite2) informally as "MaxMind" for simplicity.

B. Survey Methodology

We characterized the use of MaxMind across nine systems, security and networking conferences during the five year period from 2016-2020. To find papers in the literature using MaxMind, as well understand how it is used, we adopt a semi-automatic method: first, for a given conference venue, we obtain the complete proceedings and perform a case-insensitive search for the string "maxmind." We manually inspect each paper found to contain "maxmind" to determine whether the work utilizes the database or is simply referencing MaxMind. For example, in [14], "maxmind" appears only as a citation to the sentence "Current IP-based geolocation services do not provide city-level accuracy…" Only those papers that used MaxMind's database for their research are included.

Keeping in mind the variety of research questions and geolocation requirements inherent in the various papers, we sought to distinguish what was being geolocated and at what granularity. We manually extract from each paper the granularity required (country, city, or AS) and the type of IP addresses geolocated (all, end users, end host infrastructure, and router). The "end user" category contains IP addresses belonging to residential users (e.g., [15]), or, more broadly, end users issuing web traffic (e.g., [16]). The "end host infrastructure" category includes addresses belonging to Internet infrastructure, typically web [17], proxies [18], or DNS [19] servers. "Routers" include the IP addresses of network router interfaces. Finally, the "all" category contains papers that geolocate all types of addresses such as [20], [21]. Note that these sets are mutually exclusive, but a paper can use MaxMind on several types of IP addresses. For instance, [22] studies the Mirai botnet where the infected IP addresses can belong to both end users and end host infrastructure.

C. MaxMind in the Literature

Table II summarizes our findings. We follow the rhetorical structure of Scheitle et al. [23] to classify the impact of MaxMind on the paper's results.

- Affected "Y" are papers that use MaxMind in their methodology to obtain a result. For example, Papadopoulos et al. [16] use MaxMind to build a classifier to infer how much advertisers pay to reach users.
- Affected "V" are papers that do not use MaxMind to obtain results, but rather to *compare* their results. For example, Weinberg et al. [18] compare their inferred proxy locations to MaxMind's locations.

The "Date" column indicates whether or not the paper explicitly provides the MaxMind snapshot date. The last column indicates which MaxMind version is used: either free, paid or if the info was not available.

From a macro perspective, MaxMind is both used at country (53%) and city (37%) granularity. Second, it is mostly used to geolocate end users (38%) and end host infrastructure (49%) rather than routers (9%). Then, the majority of papers (86%) use MaxMind to obtain results, and few (11%) provide the snapshot date. Finally, we see that free and paid version of MaxMind are equally used by the community. Note that the totals do not sum to the number of papers as, e.g., a paper may use MaxMind for both AS and country information [20], or use both the free and paid version [1].

Lesson: MaxMind is the most popular geolocation database to support other research. Further, the results of many papers may be sensitive to geolocation variation, especially given the lack of snapshot dates, large windows of measurement or data, and no explicit alignment between data collected and the geolocation snapshot.

III. Metrics

This section defines metrics used to characterize the impact of selecting one geolocation database snapshot rather than a different snapshot in time. We assume that snapshots contain IP prefixes and their associated locations. For all IP prefixes, we expand them to their constituent set of individual addresses.

We define metrics using two concepts for comparing two geodatabase snapshots: *coverage difference* and *distance distribution*. For these definitions:

TABLE II: Literature survey of MaxMind use in academic venues (2016-2020). "Affected" column specifies if MaxMind was used in methodology (Y) or for validation (V).

Conference	Area	Papers	MaxMind							Affected		Snapshot date specified		Free (F) Paid(P) (N/A)		
			Granularity			IP type				Y	V	Y	N	F	P	N/A
			AS	Country	City	All	End user	End host infrastructure	Router							
IMC	Meas.	16	1	13	3	2	5	8	1	12	4	1	15	8	3	6
PAM	Meas.	6	0	2	4	1	3	2	0	5	1	0	6	3	1	2
TMA	Meas.	4	1	0	3	3	0	1	0	3	1	2	2	1	2	1
USENIX Sec	Security	10	0	7	3	0	4	7	0	10	0	2	8	1	4	5
CCS	Security	6	2	1	3	0	1	2	3	6	0	0	6	2	3	1
SIGCOMM	Systems	3	0	1	2	0	3	1	0	2	1	0	3	0	3	0
NSDI	Systems	1	0	0	1	0	0	0	1	1	0	0	1	0	0	1
CoNEXT	Systems	2	1	1	0	1	0	1	0	2	0	0	2	1	0	1
WWW	Web	10	0	7	3	0	4	7	0	10	0	2	8	2	2	6
Total	All	58	7	30	22	8	22	28	6	51	7	6	52	18	20	23

- Let A be a set of IPv4 addresses (either all address, or a population e.g., addresses known to be router interfaces).
- Let L be the set of locations present in a geodatabase (either latitude-longitude pairs, cities, or countries).
- Let M be a snapshot of this database, defined as a set of pairs (a, l) that map addresses to locations.
- Let $A_M \subset A$ be the set of addresses that appear in snapshot M that belong to population A.

A. Coverage difference

Intuitively, coverage difference means the portion of IP addresses that appear in one snapshot or another, but not both. Two identical snapshots have a coverage difference of zero while the coverage difference is one for two snapshots with no IP addresses in common.

Formally, coverage difference is an extension of the concept of 'coverage', which for snapshot M with respect to a set of IPv4 addresses A is:

$$\text{coverage}(M) = \frac{|A_M|}{|A|} \tag{1}$$

The coverage difference between two snapshots M_i and M_j on A is the Jaccard distance between A_{M_i} and A_{M_j}:

$$\text{covdiff}(M_i, M_j) = \frac{|(A_{M_i} \bigcup A_{M_j}) - (A_{M_i} \bigcap A_{M_j})|}{|A_{M_i} \bigcup A_{M_j}|} \tag{2}$$

B. Distance distribution

An address a can appear in a location in one geodatabase snapshot and different location in a second snapshot. Let $\text{dist}(a, M_i, M_j)$ be the Haversine distance [24] between the locations of a in M_i and M_j, using latitude-longitude values for each location. The distance distribution between the two snapshots is the set of distances, one for each address that appears in both snapshots:

$$\text{D}_{\text{dist}}(a, M_i, M_j) = \{\text{dist}(a, M_i, M_j) \mid a \in A_{M_i} \bigcap A_{M_j}\} \tag{3}$$

The definitions above serve to define the differences between two snapshots. Because we are not interested in only comparing two snapshots, but also knowing the average difference between two snapshots or the worst case, we need to compare those differences. For coverage, we can sort the pairs of snapshots by their Jaccard distance. To compare two distance distributions, we define the following metric:

$$\begin{aligned}\text{distdiff}(\text{M}_\text{i}, \text{M}_\text{j}) = (\text{mean}(\{\log_{10}(\text{dist}(a, M_i, M_j)) \\ \mid a \in A_{M_i} \bigcap A_{M_j}\})\end{aligned} \tag{4}$$

By taking the mean of the log of the distances, we prevent outliers (e.g., distances potentially up to 20,000 km) from disproportionately outweighing lower, but nonetheless meaningful, distances, e.g., on the order of 100 km.

Note that while the median is a more robust statistic, typically more than 50% of the address have zero distance, i.e., did not move between snapshots (§ V-C). Thus, the mean provides a meaningful non-zero measure. Other metrics that we define on the distance distributions are the quantiles of a distribution, along with the maximum value.

For both coverage and distance, the higher the metric, the larger the difference is between the two snapshots.

C. Distance

Our survey looks at the distribution of distance values per address. We define the maximum distance of an address a as being the maximum distance between two of its locations. Formally, the distribution of distances of a is:

$$D(a) = \{\text{dist}(l_i, l_j) : l_i, l_j \in \mathcal{L}_a\}\} \tag{5}$$

where $\mathcal{L}_a$ is the list of locations of a in a considered set of snapshots. The maximum distance of a is then:

$$\max(D(a)) \tag{6}$$

IV. Data

A. MaxMind snapshots

We collect 214 MaxMind snapshots spanning the ten year period from January 2010 to December 2019. There are two primary challenges in the raw data: (1) the snapshots we obtain are not uniformly distributed in time; and (2) IP

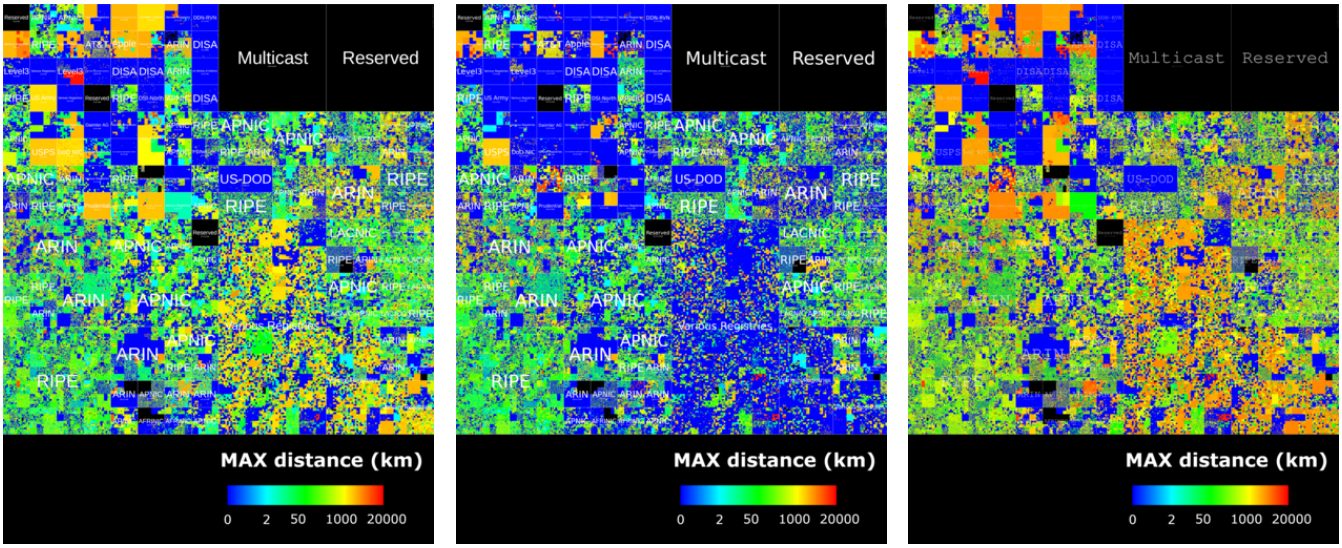


Fig. 1: Hilbert heatmap of maximum distance (Eq. 6) in 2018 (left) 2019 (center) and absolute difference (right) for the entire IPv4 address space (log scale; each pixel represents a /24). MaxMind exhibits a high degree of global geolocation dynamics and year-to-year variation.

addresses appear within prefixes of different networks and lengths over time. To utilize this data within the framework of our methodology and metrics, we pre-processed it.

1) Sampling the snapshots for time uniformity: Sec. III assumes a uniform distribution of snapshots in time. Our evaluation examines a ten year span from 2010-2019. Within this ten year period, we have at least one snapshot per month, but sometimes as many as one snapshot per week. Therefore to ensure uniformity, we simply down-sample so that the ten year period includes one snapshot per month. Our evaluation is conducted on this subset of snapshots such that they are uniformly distributed in time.

2) Prefixes of different lengths: A MaxMind snapshot contains a mapping of prefix blocks to geolocation. However, these prefixes may split, be aggregated, or even overlap in time. While our analysis is at the per-IP address granularity, rather than prefix, maintaining the geolocation for all IP addresses over time is inefficient. Our first step then is to find a data structure to efficiently store and query the snapshots. Over all prefixes in all snapshots, we construct the set of covering longest length prefixes and construct a Patricia trie [25]. We build one Patricia trie for each geolocation granularity: country, city, and coordinates. The Patricia trie contains, per prefix, all its locations over the period of time.

To handle prefix variation over time, we insert into the Patricia trie the longest prefixes seen in the snapshots. The resulting fine-grained prefixes will be inserted in a database. As an example, consider the prefix 1.0.0.0/23 located in London in the snapshot s1. On the other hand, in the snapshot s2, the prefix 1.0.0.0/24 is located at London and the prefix 1.0.1.0/24 is located at Paris. We place the two /24 prefixes in the trie for each of the snapshots, London being inserted for the two prefixes of the snapshot s1.

The prefixes considered in our database therefore do not necessarily correspond to BGP prefixes nor are the same as the initial prefixes on MaxMind snapshots. The resulting prefix lengths vary between /9 and /32, the most common being /29 with 19.2% of the total prefixes.

B. Different types of IP addresses

Sec. II has shown that researchers use MaxMind to locate three classes of IP addresses: end users, end hosts infrastructure and routers. We therefore collect and label three sets of IP addresses corresponding to these three types.

- **End users:** M-Lab [26] performs and records measurements to end users requesting performance tests (i.e., a "speedtest"). From the M-Lab public datasets we extract targets in the year 2019. We randomly sample these targets to obtain 6.7M IPv4 addresses in approximately 2M unique /24 prefixes.
- **End host infrastructure:** For end host infrastructure, we extract the daily top list made available by [23]. We perform an intersection of all 2019 lists in order to minimize the number of IP addresses that could be reassigned for other purpose. Because these top lists are volatile, our filtering for high-confidence end host infrastructure addresses produces 26,231 IP addresses in 16,942 /24 prefixes.
- **Routers:** We leverage both CAIDA ITDK dataset [27] and Diamond-Miner [28] public Internet topology datasets to collect IP addresses belonging to router interfaces. Both datasets are the result of Internet-wide traceroute style probing. We take the intersection of 2019-01, 2019-04 ITDK and 2019-08 Diamond-Miner datasets and obtain 730k IP addresses in more than 177k different /24 prefixes. By taking the intersection over time, the aim is again to

ensure the likelihood that the addresses indeed belong to routers.

C. Ethical Considerations

Our work does not involve human subjects, questionnaires, or personally identifiable information, and, hence, does not meet the standards for IRB review. The MaxMind data we analyze is covered by the Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) license which permits adaption of the database: "remix, transform, and build upon the material for any purpose, even commercially." Beginning in January 1, 2020, MaxMind adopted a more restrictive policy in order to comply with GDPR requirements [29]; our research does not analyze any data after 2019.

V. Evaluation

This section presents an evaluation of the MaxMind data snapshots from 2018 to 2019 using the metrics defined in the methodology. We examine the extent and impact of both location movement and coverage across several dimensions. Primary results are presented here, while a more exhaustive evaluation, including the full 10 years of longitudinal data, is available in an accompanying technical report [30].

A. Limitations

Our primary contribution in this work is to define metrics that characterize the dynamics of IP geolocation databases, and understand how these dynamics can impact network research that depends on geolocation. We keenly recognize that location changes within the database may be genuine or may be artifacts of the geolocation system's methodology. We do not investigate the root causes of the dynamics we observe. Indeed this limitation is fundamental – MaxMind's algorithm is proprietary so that we cannot provide true causal analysis. Rather, we focus on providing lesson for how researchers should view and utilize geo-databases.

B. Visualizing Internet-wide geo movement

Fig. 1 shows an exhaustive representation of the maximum distance change for each /24 of the entire IPv4 space for 2018 and 2019, as well as the absolute difference between the two years. Each pixel represents a /24, and the color represents the maximum distance between two locations; black pixels indicate that the IP address is not present in the database. If the /24 contains more specific entries in the MaxMind database, we take the maximum of the maximum distance of the IP addresses within the /24.

We see that the visualization of 2019 differs from 2018: Many of the various registries in the bottom center right and top left part of the plots have a maximum distance of more than 1000km in 2018, whereas they did not move in 2019. There is also a red square in the prefixes belonging to Level3, that had a maximum distance of 20,000km in 2018 but did not move in 2019. Surprisingly, there are also some IP addresses that were covered in 2018 (i.e., in this case, having lat/long coordinates) which are not covered in 2019. This is the case of some blocks of IP addresses in the bottom center left of the graph belonging to APNIC and AFRINIC.

All these differences between the two years are highlighted by the map on the right: we clearly see the center and the bottom left mainly colored in orange and red as well as some big prefixes on the top right. It reveals a significant dynamic change not only along the prefixes but also through time.

Overall, by looking at the Hilbert representations of each year over the 10 years dataset, it is difficult to perceive a trend that could lead us to say that prefixes are experiencing bigger or smaller distance change over years.

Lesson: These visualizations confirm not only the **high degree of global geolocation dynamics**, but also the presence of **year-to-year variation in geolocation movement**. An IP address can experience a maximum distance change of 0km in 2018 and more than 1000km in 2019, and vice versa.

C. Impact

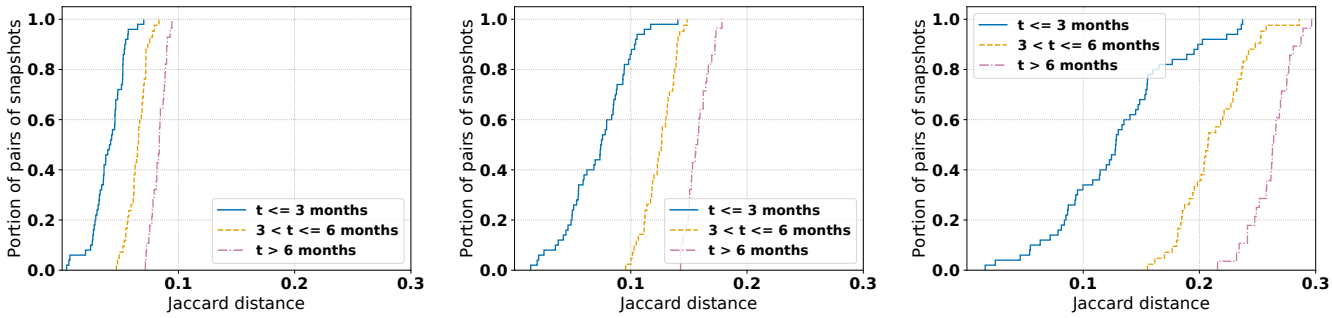
While the preceding analysis demonstrates how our metrics can shed light on the underlying dynamics of a geolocation database, we conclude this section with an analysis of the potential *impact* of selecting a particular snapshot of MaxMind versus a different snapshot, for instance as a researcher seeking to geolocate a population of IP addresses under study. To bound our results, we compare pairs of snapshots from 2019 within three time windows: when the snapshots differ by less than 3 months, between 3 to 6 months, and between 6 to 12 months. We evaluate the impact across the three IP classes: end users, end host infrastructure, and routers.

1) Coverage (Fig. 2): For coverage we show results at the city level as we find no country level coverage differences between snapshots; almost all IP addresses, across all classes of addresses, have a country geolocation present in the database.

Not unexpectedly, for all types of IP addresses, we observe that the coverage difference increases with time between the two snapshots. As shown in our tech report (Fig.5 of [30]), the overall coverage is globally constant, therefore this cannot be imputed to an increase of the total coverage.

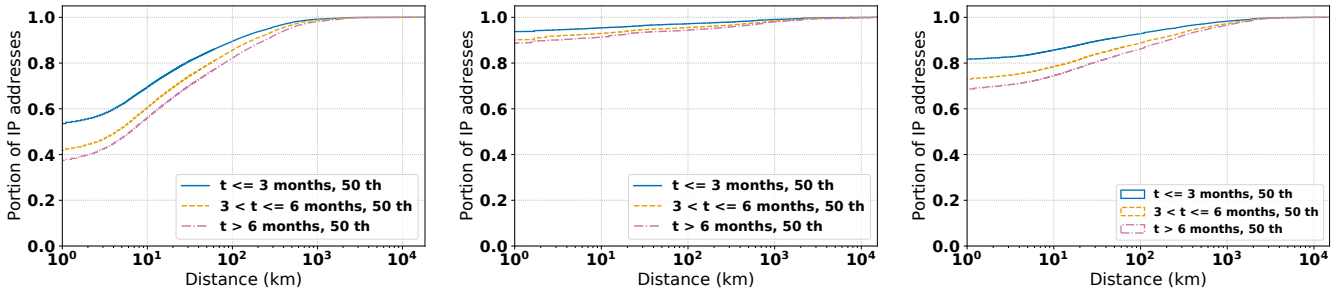
We see that even for two snapshots created within less than three months of each other, there is a significant coverage difference, up to 6%, 11% and 20% for end users, end host infrastructure and routers respectively. As seen in Fig. 2c, there is a 50% probability of more than 12% coverage difference between two snapshots created less than three months apart. Between two snapshots of more than six months and less than a year, the difference can be even worse, up to 9%, 17% and 30%.

2) Distance (Fig. 3): We first sort the pairs of snapshots by the metric defined in Eq. 4, the mean of the logarithmic distances (MLD). Recall, the higher the MLD, the more the snapshots differ. We compute distance across pairs of snapshots within the same time ranges as for coverage: less than three months, between three and six months, and between six and twelve months. From the MLD distribution, we then show the pair of snapshots corresponding to the median. For example, in Fig. 3a, one should read: On the end users dataset, 15% of



(a) End users (b) End host infrastructure (c) Routers

Fig. 2: Comparing pairs of database snapshots by city coverage difference (Eq. 2). Across all classes of IP addresses, there are significant coverage differences, even on among snapshots closely separated in time.



(a) End users (b) End host infrastructure (c) Routers

Fig. 3: Comparing pairs of database snapshots by IP address distance difference (Eq. 3). Up to 22% of addresses move more than 40km among snapshots in a three month window.

the IP addresses moved by at least 40km. This corresponds to the median result for a pair of snapshots that are less than three months apart.

Fig. 3 shows two trends. First, as one might expect, for all types of IP addresses, the more time between two snapshots, the more IP addresses move. Then, the percentage of IP addresses moving depend on the type of IP addresses. We observe that end users tend to move more than routers and end host infrastructure. In details, for a pair of snapshots that are more than six months apart, we have 28%, 8% and 18% of IP addresses that move more than 40km for respectively end users, end host infrastructure, and routers.

If we consider that 40km corresponds to most metropolitan areas [1], this implies that a non-trivial portion of IP addresses experience a location change out of the metro area – a significant change. However, distances greater than 1000km are rare, accounting for less than 5% of IP addresses across all addresses classes.

Lesson: There are non negligible differences in both coverage and distance of movement even for database snapshots created closely in time (< 3 months). Therefore: **we recommend, insofar as possible, aligning geolocation database snapshots with the measurements that produced them**, for instance by programmatically using an API to lookup IP addresses on-demand as they are gathered. Further, one should look at several snapshots closely spaced in time over the measurement period and more deeply **investigate IP addresses that experienced significant changes**.

D. A longitudinal study

Due to space constraints, we leave the 10 year longitudinal study for our accompanying tech report [30] and associated research [31]. In this paper, we define new metrics and extend the ones defined in Section III to enable the comparison of an arbitrary number of snapshots and the analysis of the dynamics of the geolocation of an IP address over time. One of the main results is that we find that a majority of IP addresses are mapped to at least two locations far from 40 km (a metropolitan area) within a year, with a high variance depending on the country and the type of IP address. This reinforces our call to be very cautious about the usage of MaxMind when data are collected during a period longer than few weeks.

VI. Use Case

Previous sections have shown two things. MaxMind is a widely used database (Sec. II), and selecting a particular snapshot in a time period can have a significant impact on the results (Sec. V). In this section, we concretely demonstrate

the potential impact on research that depends on MaxMind by reproducing the results from Gharaibeh et al.'s IMC 2017 work [1] with different MaxMind snapshots. Gharaibeh et al. study the accuracy of different databases for router geolocation, including MaxMind. Using the author's publicly available ground truth, we reproduce their accuracy results (see Sec. 5.2, Fig. 2 of [1]).

Surprisingly, the MaxMind snapshot that produces the largest impact on the results was created within only two months of the snapshot used by the authors. This snapshot shifts the median of the distance distribution to ground truth from more than 100 km to 40 km, which is close to the results of the paid version. Given this variability, the claim that the free version of MaxMind is worse than the paid version seems to depend on the specific instance studied.

A. Dataset

The Gharaibeh dataset consists of 16,586 router interface IP addresses with corresponding ground truth locations inferred either with RTT-based measurements or DNS-based techniques. The authors do not mention which specific snapshot of MaxMind they used, however: "The databases are accessed again on early July 2016, to geolocate the ground truth." We inquired with the authors for the exact snapshot date, but unfortunately they could not be more specific. We therefore select the closest snapshot as our reference, from July 8, 2016. Sec. VI-B confirms that the results of this snapshot are very close to those presented in the original paper.

The measurement period for the ground-truth collection and validation, however, spanned a larger time period. As stated in the paper: "Overall, between May 2016 and September 2017, 8,197 (69.1%) [...] have different hostnames, and 6.9% no longer have rDNS records." We therefore restrict our comparison between snapshots belonging to this period of time, on which the authors consider that the ground truth is valid.

B. Results

1) Distance to ground truth: We compute the distance to ground truth distribution of all the snapshots from May 2016 to September 2017. We then compare each of these distributions to the distribution of the reference snapshot, using the Kolmogorov-Smirnov (KS) test [32]. The KS test quantifies the dissimilarity between two distributions, with higher values indicating less similarity. Fig. 4a shows the distance to ground truth distribution of the snapshots corresponding to the 5th, 25th, 50th, 75th, and 100th percentiles of the KS distribution. We also show the reference snapshot.

First, we compare Fig.2 of Gharaibeh et al. with our reference snapshot. We infer that on Fig.2 of Gharaibeh et al. $\sim$8%, 47%, 50%, 55%, and 96%, are located at less than respectively 1, 40, 100, 1,000, and 10,000km from the ground truth, whereas it is 8%, 46%, 49%, 56% and 97% in our reference snapshot. Overall, the qualitative shape of the distribution is identical to the original figure, giving us confidence in our ability to reproduce the author's results.

However, we observe significant differences between results derived from the other snapshots versus the reference. The median shifts from 167km in the reference snapshot to respectively 57, 51, 41, 40 and 40km for the 5th, 25th, 50th, 75th and 100th percentile.

When we consider the snapshots dates with these percentiles, it is surprising to observe that the 100th percentile was created only two months after the reference snapshot, whereas the 5th percentile is a snapshot taken one month later and the 25th percentile corresponds to a snapshot taken nine months later. This implies that MaxMind did not improve over time for these addresses, but also that there are significant differences in the results within a relatively short time.

Finally, we look at the comparison between the free and paid version of MaxMind. On Fig.2 of Gharaibeh et al. we infer that the paid version has a median between 30 and 40km, so that the difference between this distribution and the different snapshots of Fig. 4a is less pronounced than the difference between the free and paid version of their graph. Therefore the conclusion that the paid version performs better than the free one should be taken with caution.

2) Coverage: Finally, we examine coverage variability. In Gharaibeh et al., the authors only compute the distance to ground truth if the IP address is covered by MaxMind at the city level.

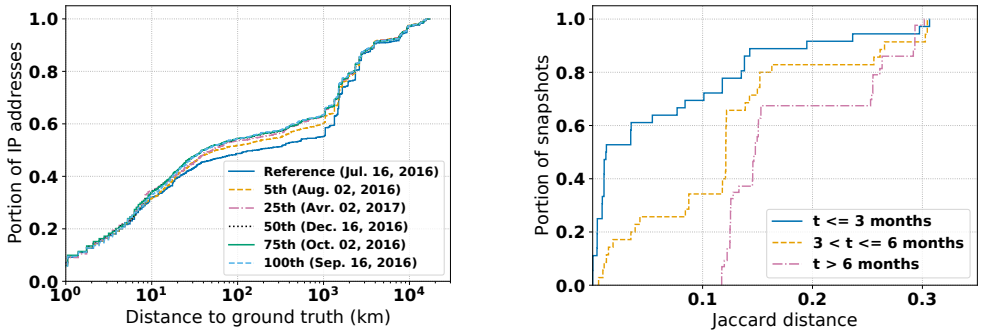
Fig. 4b shows the distribution of the coverage difference (Eq. 2) as we did in § V-C1, but only comparing snapshots with the reference snapshot. We observe that even with snapshots taken three months apart from the reference snapshot, 30% of the snapshots have more than 10% of coverage difference. It is even worse for snapshots between 3 and 6 months and snapshots with more than 6 months of difference, with a coverage difference up to 30%.

Lesson: Work is being published in the network research community without specifying the MaxMind snapshot dates. Had the authors used a different snapshot, the set of IP addresses over which they would have computed their accuracy measures – and, hence, their results – would have significantly changed.

VII. Related work

Mapping IP addresses to the physical world is an important topic that has seen two decades of research. Early efforts used landmarks, hosts with known position, to assign locations to unknown targets at coarse granularity [5]. Landmark-based geolocation was subsequently enhanced to use latency constraints [9], network topology [2], and population densities [10] to improve accuracy. Because the accuracy of latency-based techniques is often proportional to the distance between the target and its nearest landmark, Wang et al. developed techniques to find and utilize additional landmarks [4].

IP geolocation has since matured, with several competing commercial offerings including [6], [7], [8]. While the exact methodology of these commercial services is proprietary, they likely use a combination of databases (e.g., whois and DNS), topology, latency, and privileged data feeds from providers [11].



(a) Distance to ground truth CDF on different snapshots

(b) Coverage difference on city with the reference snapshot

Fig. 4: Reproducing result of [1] with different snapshots demonstrates the sensitivity of the results to the instance of the geo database. The median distance of addresses used shifted over 100km and coverage differed by up to 30%

Even so, the inference-based nature of IP geolocation imparts errors and inaccuracies even in commercial databases [3], [33]. demonstrated by several prior analyses. For instance, Poese et al. found 50-90% of ground-truth locations to be geolocated with greater than 50km of error [34]; most recently Komosnỳ et al. studied eight commercial geolocation databases and found mean errors ranging from 50-657km [35]. Geolocation of network infrastructure, including routers, is known to be particularly problematic [36], [1]. However, as shown in Sec. II, MaxMind is still widely used, for the simple reason that there exists no other alternative than geolocation database to get an Internet scale IP geolocation mapping.

Our work looks at IP geolocation through a novel lens by analyzing the longitudinal characteristics of a popular geolocation database. By showing the stability of locations at different granularities and timescales, we offer a first look at the error bounds for particular classes of applications that utilize geolocations, as well as offer practical lessons for consumers of IP geolocation data.

VIII. Conclusion

Physical mapping of Internet hosts and resources is critical in this day and age. Techniques to perform IP geolocation have matured into commercial offerings. While the accuracy of these geolocation databases has been extensively studied, little attention has been paid to understand the way they have evolved over time. Our work demonstrates that a commonly used geolocation database, MaxMind, exhibits significant changes in address coverage and locations, especially when considering particular subsets of addresses.

These changes can occur even on short timescales, including on the order of a typical measurement study duration. In this way we highlight the importance of geolocation lookups that are contemporaneous with the time an IP address is measured, observed, or gathered. Via a case study, we demonstrate the potential for a large discrepancy in results depending on the particular date of a geolocation snapshot. Similar large variances in auxiliary data sources at short time scales have been demonstrated in the past, e.g., for DNS and Internet top lists [23]. Thus, a take-away of our work is to encourage alignment of geolocation lookups with measurements, publishing the exact date of a geolocation snapshot or lookup methodology, and rigorously investigating addresses that change geolocation significantly over the course of a measurement study. In the spirit of similar measurement best practices [37], we hope to encourage more sound and reproducible measurement research. Because MaxMind does not provide access to historical data, we provide historical snapshots on demand to the community.

In future work, we plan to more deeply investigate the root causes of the geolocation movement we observe, characterize IPv6 geolocation, and work toward integrating our findings into more robust geolocation services.

Acknowledgments

We thank the anonymous reviewers and our shepherd, Roman Kolcun. Justin Rohrer and Jon Culbert provided invaluable feedback while CAIDA and Peter Boothe contributed archived data essential to our study. Matthieu Gouel, Olivier Fourmaux, and Timur Friedman are associated with Sorbonne Université, CNRS, Laboratoire d'informatique de Paris 6, LIP6, F-75005 Paris, France. Matthieu Gouel and Timur Friedman are associated with the Laboratory of Information, Networking and Communication Sciences, LINCS, F-75013 Paris, France. Matthieu Gouel, Olivier Fourmaux, and Timur Friedman were supported in part by a university research grant from the French Ministry of Defense. Robert Beverly was supported in part by National Science Foundation (NSF) grant CNS-1855614. Views and conclusions are those of the authors and should not be interpreted as representing the official policies or position of the U.S. government or the NSF.

References

[1] M. Gharaibeh, A. Shah, B. Huffaker, H. Zhang, R. Ensafi, and C. Papadopoulos, "A look at router geolocation in public and commercial databases," in *Proc. ACM IMC*, 2017, pp. 463–469.

[2] E. Katz-Bassett, J. P. John, A. Krishnamurthy, D. Wetherall, T. Anderson, and Y. Chawathe, "Towards IP geolocation using delay and topology measurements," in *Proc. ACM SIGCOMM*, 2006, pp. 71–84.

[3] B. Huffaker, M. Fomenkov, and K. Claffy, "Geocompare: A Comparison of Public and Commercial Geolocation Databases," 2011, pp. 1–12.

[4] Y. Wang, D. Burgener, M. Flores, A. Kuzmanovic, and C. Huang, "Towards Street-Level Client-Independent IP Geolocation," in *USENIX NSDI*, vol. 11, 2011, pp. 27–27.

[5] V. N. Padmanabhan and L. Subramanian, "An Investigation of Geographic Mapping Techniques for Internet Hosts," in *Proc. ACM SIGCOMM*, 2001, p. 173–185. [Online]. Available: https://doi.org/10.1145/383059.383073

[6] Akamai, "Edgescape," 2020, https://developer.akamai.com/edgescape.

[7] "Maxmind," 2020, https://www.maxmind.com/en/home.

[8] Hexasoft, "Ip2location," 2020, https://www.ip2location.com.

[9] B. Gueye, A. Ziviani, M. Crovella, and S. Fdida, "Constraint-based geolocation of internet hosts," *IEEE/ACM Transactions on Networking*, vol. 14, no. 6, pp. 1219–1232, 2006.

[10] B. Eriksson, P. Barford, J. Sommers, and R. Nowak, "A learning-based approach for IP geolocation," in *PAM*. Springer, 2010, pp. 171–180.

[11] E. Kline, K. Duleba, Z. Szamonek, S. Moser, and W. Kumari, "A Format for Self-Published IP Geolocation Feeds," RFC 8805 (Informational), RFC Editor, Fremont, CA, USA, Aug. 2020. [Online]. Available: https://www.rfc-editor.org/rfc/rfc8805.txt

[12] "Netacuity database," 2020, https://www.digitalelement.com/solutions/.

[13] "Ip2loc database," 2020, https://www.ip2location.com.

[14] V. Kotronis, G. Nomikos, L. Manassakis, D. Mavrommatis, and X. Dimitropoulos, "Shortcuts through colocation facilities," in *Proc. ACM IMC*, 2017, pp. 470–476.

[15] R. Padmanabhan, A. Schulman, D. Levin, and N. Spring, "Residential links under the weather," in *Proc. SIGCOMM*, 2019, pp. 145–158.

[16] P. Papadopoulos, N. Kourtellis, P. R. Rodriguez, and N. Laoutaris, "If you are not paying for it, you are the product: How much do advertisers pay to reach you?" in *Proc. ACM IMC*, 2017, pp. 142–156.

[17] J. Deng, G. Tyson, F. Cuadrado, and S. Uhlig, "Internet scale user-generated live video streaming: The twitch case," in *PAM*. Springer, 2017, pp. 60–71.

[18] 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 *Proc. IMC*, 2018, pp. 203–217.

[19] P. Pearce, B. Jones, F. Li, R. Ensafi, N. Feamster, N. Weaver, and V. Paxson, "Global Measurement of DNS Manipulation," in *USENIX Security)*, 2017, pp. 307–323.

[20] Y. Lee and N. Spring, "Identifying and aggregating homogeneous IPv4 /24 blocks with hobbit," in *Proc. ACM IMC*, 2016, pp. 151–165.

[21] P. Winter, R. Padmanabhan, A. King, and A. Dainotti, "Geo-locating BGP prefixes," in *IEEE TMA*, 2019, pp. 9–16.

[22] M. Antonakakis, T. April, M. Bailey, M. Bernhard, E. Bursztein, J. Cochran, Z. Durumeric, J. A. Halderman, L. Invernizzi, M. Kallitsis *et al.*, "Understanding the mirai botnet," in *USENIX Security*, 2017, pp. 1093–1110.

[23] Q. Scheitle, O. Hohlfeld, J. Gamba, J. Jelten, T. Zimmermann, S. D. Strowes, and N. Vallina-Rodriguez, "A Long Way to the Top: Significance, Structure, and Stability of Internet Top Lists," in *Proc. ACM IMC*, 2018, p. 478–493. [Online]. Available: https://doi.org/10.1145/3278532.3278574

[24] F. Cajori, *A history of mathematical notations*. Courier Corporation, 1993, vol. 1.

[25] K. Sklower, "A tree-based packet routing table for berkeley unix." in *USENIX Winter*, vol. 1991, 1991, pp. 93–99.

[26] "M-lab public datasets," 2020, https://www.measurementlab.net/data/.

[27] "Caida itdk," http://www.caida.org/data/internet-topology-data-kit/index.xml, 2020.

[28] K. Vermeulen, J. P. Rohrer, R. Beverly, O. Fourmaux, and T. Friedman, "Diamond-Miner: Comprehensive Discovery of the Internet's Topology Diamonds," in *USENIX NSDI*, 2020, pp. 479–493.

[29] "Significant changes to accessing and using geolite2 databases," Dec. 2019, https://blog.maxmind.com/2019/12/18/significant-changes-to-accessing-and-using-geolite2-databases/.

[30] M. Gouel, K. Vermeulen, O. Fourmaux, T. Friedman, and R. Beverly, "Longitudinal Study of an IP Geolocation Database," Tech. Rep., 2021, https://arxiv.org/abs/2107.03988.

[31] J. A. Culbert, "Toward Understanding the Longitudinal Stability of an IP Geolocation Database," Master's thesis, Naval Postgraduate School, Mar. 2020, http://hdl.handle.net/10945/64848.

[32] H. W. Lilliefors, "On the Kolmogorov-Smirnov test for normality with mean and variance unknown," *Journal of the American statistical Association*, vol. 62, no. 318, pp. 399–402, 1967.

[33] Y. Shavitt and N. Zilberman, "A geolocation databases study," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 10, pp. 2044–2056, 2011.

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

[35] D. Komosnỳ, M. Vozňák, and S. U. Rehman, "Location accuracy of commercial ip address geolocation databases," 2017.

[36] B. Huffaker, M. Fomenkov *et al.*, "DRoP: DNS-based Router Positioning," *ACM SIGCOMM Computer Communication Review*, vol. 44, no. 3, pp. 5–13, 2014.

[37] V. Paxson, "Strategies for Sound Internet Measurement," in *Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement*, 2004, p. 263–271. [Online]. Available: https://doi.org/10.1145/1028788.1028824