Technische Universität Berlin

Forschungsberichte der Fakultät IV – Elektrotechnik und Informatik

IP Geolocation Databases: Unreliable?

Ingmar Poese, Mohamed Ali Kaafar, Benoit Donnet,
Bamba Gueye, Steve Uhlig

Bericht-Nr. 2011 – 03
ISSN 1436-9915

IP Geolocation Databases: Unreliable?

Technical Report 2011-03
ISSN:1436-9915

Ingmar Poese*, Mohamed Ali Kaafar‡, Benoit Donnet†, Bamba Gueye**, Steve Uhlig*
∗ Deutsche Telekom Lab./TU Berlin, Germany
‡ INRIA Rhône-Alpes, France
† Université catholique de Louvain, Belgium
∗∗ Université Cheikh Anta Diop de Dakar, Senegal

***Abstract*—The most widely used technique for IP geolocation consists in building a database to keep the mapping between IP blocks and a geographic location. Several databases are available and are frequently used by many services and web sites in the Internet. Contrary to widespread belief, geolocation databases are far from being as reliable as they claim. In this paper, we conduct a comparison of several current geolocation databases -both commercial and free- to understand the limitations in their usability.**

First, the vast majority of entries in the databases refer only to a few popular countries (e.g., U.S.). This creates an imbalance in the representation of countries across the IP blocks of the databases. Second, these entries do not reflect the original allocation of IP blocks, nor BGP announcements. In addition, we quantify the accuracy of geolocation databases on a large European ISP based on ground truth information, as well as on two tier-1 ISPs based on DNS names containing geographic clues. This is the first study using a ground truth showing that the overly fine granularity of database entries makes their accuracy worse, not better. Their blocks, often as fine as /29 prefixes, have geolocations inaccurate by hundreds of kilometers in a significant fraction of the cases. All in all, geolocation databases can claim country-level accuracy, but certainly not city-level.

I. Introduction

With the emergence of Internet services requiring location information, *IP geolocation techniques* (i.e., mapping an IP address to the geographic location of the corresponding host) becomes a key enabler for many of these services. Examples of such services comprise targeted advertising on web pages, displaying local events and regional weather, automatic selection of languages to first display content and restricted content delivery following regional policies.

Two main paradigms exist to geolocate IP addresses: active and passive. Active IP geolocation techniques, typically based on delay measurements [1], [2], [3], [4], may achieve desirable properties such as accuracy (i.e., active measurements provide better results compared to geolocation database in many cases). However, these properties come at the expense of lack of scalability, high measurement overhead, and very high response time ranging from tens of seconds to several minutes to localize a single IP address. This is several orders of magnitude slower than what is achievable with the passive approach, i.e., database-driven geolocation.

Database-driven geolocation usually consists of a database-engine (e.g., SQL/MySQL) containing records for a range of IP addresses, which are called *blocks* or *prefixes*. Geolocation prefixes may span non-CIDR subsets of the address space, and may span only a couple of IP addresses. Examples of geolocation databases are *GeoURL* [5], the *Net World Map* project [6], and are provided as free [7], [8], [9] or commercial tools [10], [11], [12], [13], [14].

The other side of the coin with geolocation databases is that, besides the difficulty to manage and update them, their accuracy is more than questionable [15], [16], especially due to lack of information about the methodology used to build them. The crux of the problem is that prefixes within databases are not clearly related to IP prefixes as advertised in the routing system, nor to how those routing prefixes are used by their owners (e.g., ISPs, enterprises, etc). Indeed, even if many commercial geolocation databases claim to provide a sufficient geographic resolution, e.g., at the country-level, their bias towards specific countries make us doubt their ability to geolocate arbitrary end-hosts in the Internet.

Few works focus on geolocation databases and their accuracy. Freedman et al. studied the geographic locality of IP prefixes based on active measurements [17]. Siwpersad et al. assessed the geographic resolution of geolocation databases [16]. Based on active measurements, the authors of [16], [17] showed the inaccuracies of geolocation databases by pinpointing the natural geographic span of IP addresses blocks.

In this paper, we go further by questioning the reliability of the information contained in geolocation databases. As the databases are expected to be able to correctly geolocate IP addresses, we find a surprising low number of unique geographic locations, tens of thousands, compared to the large number of blocks (up to several millions) in many databases. In addition, we observe that a few countries are over-represented in these databases, making the geographic sampling of the databases not fairly spread across the world.

One of our salient findings is that these entries do not reflect the address space of IP blocks as originally allocated to their owners or as announced by BGP. Locations discrepancies between the databases, coupled with the fine granularity of their blocks, often /29, shed serious doubt on the accuracy of their geolocation.

Finally, to confirm our doubts about the inability of databases to provide city-level accuracy, we confront the geolocations of three databases on the prefixes advertised by several large ISPs, based on ground truth information. We find that most of the blocks of the databases incorrectly geolocate

prefixes, with errors being systematically in the order of a few hundreds of kilometers.

The remainder of this paper is organized as follows: Sec. II provides a description of the five databases considered throughout this paper; Sec. III investigates whether a geolocation database is constructed following one or several objective criteria; Sec. IV determines whether geolocation databases are reliable; Sec. V-A confronts three commercial databases with the network of a large European ISP for which ground truth is available, while Sec. V-B confronts the same three databases with the network of two tier-1 ISPs, based on DNS names containing geographic clues. Finally, Sec. VI concludes this paper by summarizing its main achievements.

II. Dataset

Database	Blocks	(lat; long)	Countries	Cities
HostIP	8,892,291	33,680	238	23,700
IP2Location	6,709,973	17,183	240	13,690
InfoDB	3,539,029	169,209	237	98,143
Maxmind	3,562,204	203,255	244	175,035
Software77	99,134	227	225	0

TABLE I
General characteristics of the studied geolocation databases

In this paper, we consider five IP geolocation databases. Two are commercial (*Maxmind* [14] and *IP2Location* [12]) and three are freely available (*InfoDB* [8], *HostIP* [7], and *Software77* [9]). Although these databases share some information about their construction processes, comments about how they are built are vague and technically evasive. As reported in [8], InfoDB is, for instance, built upon the free Maxmind database version, and incremented by the IANA (Internet Assigned Numbers Authority) locality information. The HostIP database is based on users' contributions. Finally, Software77 is managed by Webnet77, an enterprise offering Web hosting solutions.

Typically, a geolocation database entry is composed of a pair of values, corresponding to the integer representation of the minimum and maximum address of a block. Each block is then associated with several information helpful for localization: country code, city, latitude and longitude, and Zip code.

Table I shows the number of entries (i.e., the number of IP blocks) recorded in each database (column labeled "Blocks"). Most databases contain several millions of IP blocks. Only Software77 has much less entries: $99,134$. HostIP has the highest number of entries because it is composed exclusively of /24 prefixes. Compared to the more than $300,000$ prefixes advertised in BGP routing, one might be led to believe that the geographic resolution of the geolocation databases is much finer than the natural one from BGP routing [17].

Table I provides also the number of countries and cities retrieved from the databases locations. From the number of countries, we can infer that most of the world countries are covered. However, containing blocks for most countries does not imply that countries are properly sampled, neither from an address space perspective nor from a geographic location one. Fig. 1 shows the cumulative fraction of blocks from

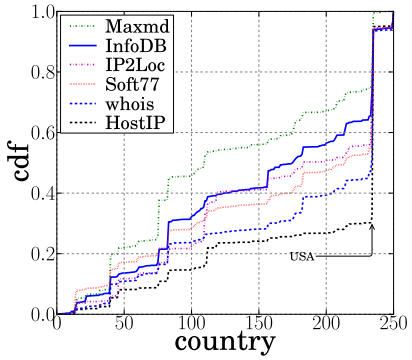


Fig. 1. Common countries distribution

the databases across countries. Note that countries on Fig. 1 (horizontal axis) have been alphabetically ordered based on their ISO country codes.

Again, we stress the number of countries represented in all databases that gives the impression that they cover fairly all countries in the world. This is misleading as more than 45% of the entries in these databases are concentrated in a single country: the United States (see Fig. 1). The five databases display a similar shape of their cumulative number of blocks across countries. The big jump around country 230 corresponds to the over-representation of the U.S in terms of database blocks compared to other countries. It is worth to notice that countries distribution observed in whois database (see Fig. 1) presents the same behavior than geolocation databases.

From Table I, we also notice the strong difference between the number of IP blocks and the number of unique (latitude, longitude) pairs. The perfect example of this is HostIP. While it contains roughly 8 millions of IP blocks, those blocks only refer to $33,000$ (latitude, longitude) pairs. This observation casts some doubts upon the true geographic resolution of the databases.

III. Database Construction

In this section, we investigate whether the construction of geolocation databases follows some global patterns. We focus on two aspects. First, we check how similar the blocks of the databases are from the official address allocations and prefixes advertised in BGP routing (Sec. III-A). Second, we evaluate whether the construction of a database follows any demographic property, such as the amount of connected users in a given country (Sec. III-B).

A. Prefixes

Comparing the subnet size of database entries with those from the official allocations by the Internet routing registries and BGP routing tables is enlightening (see Fig. 2). HostIP is not plotted as it is exclusively made of /24 prefixes. We show results as for the period of February 2010, but it is worth noticing that we observed similar results for other periods in 2009.

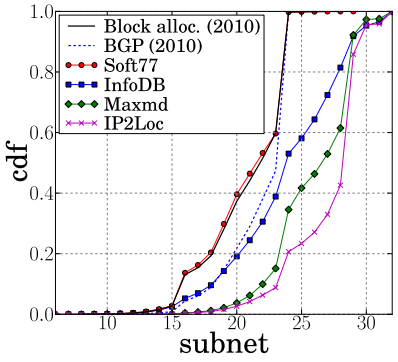


Fig. 2. Prefix distribution

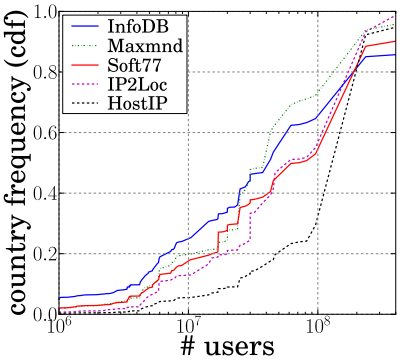


Fig. 3. Fraction of database prefixes as a function of Internet users

Most allocated blocks and BGP prefixes are between /16 and /24. Very few allocations and BGP prefixes are subnets smaller than 256 IP addresses (/24). BGP prefixes are slightly more de-aggregated than the original allocations. The Software77 database is made of entries that have the same subnet size distribution as the original address space allocation. 95.97% of the entries in Software77 correspond to IP blocks as allocated in February 2010. As expected from their sheer size, the other databases have a significant fraction of their blocks smaller than /24 subnets. These databases split official address space allocations and BGP prefixes into finer blocks.

Prefixes advertised by BGP and allocated blocks could, however, constitute a first approximation to the databases entries. Nevertheless, most of the IP blocks from Maxmind and IP2Location correspond to subnets smaller than /25. In essence, Maxmind and IP2location entries substantially differ from BGP and official allocations by more than 50% from a blocks size perspective. With such fine IP blocks, we should expect a very high geographic accuracy. Again, because the way these databases are built is kept secret, we can only infer some of their characteristics. In particular, from these first observations, all the studied databases, except Software77, are clearly not related to official allocations and BGP routing tables. Even if the entries would closely match allocated or advertised prefixes, we would not expect that the locations attributed to them in the databases would be reliable. We believe this because the locations contained in the databases do not have to be related to how address space is actually allocated and used by its owners. We will demonstrate this point in Sec. IV and V.

B. Internet Demographics

The Internet is a worldwide communication infrastructure. Its deployment and usage however differ across different regions of the world. In the same way as address space allocation is biased towards certain regions of the world, geolocation databases should also reflect their usage. It is worth to notice that, throughout this section, we only focus on the countries that are common to all databases.

1) Internet Users: A factor that is likely to explain the number of databases blocks kept per country is the amount of *Internet users* per country, i.e., the number of people in a given country being connected to the Internet. The more popular the Internet in a given country, the more we expect to see entries in the databases for this country. Internet users statistics are from 31st December 2009 [18].

We consider each country seen in the databases[1], and rank them according to the amount of people connected to the Internet (horizontal axis of Fig. 3 in logarithmic scale). We then compute the fraction of blocks recorded in the different databases for each value of the number of Internet users and plot it in a cumulative way (vertical axis of Fig. 3).

Fig. 3 shows that there is a strong relationship between the number of Internet users in a country and the importance of that country in the databases in terms of IP blocks. Countries with less than 1 million users are more or less non-existent.

There is an exception to the general tendency drawn in Fig. 3: a few countries with a large amount of population connected to the Internet are under-represented. The perfect example of this is China with roughly 400 million Internet users but a low database representation (between 1% and 5%, depending on the database). Others examples are India, Japan, or Germany. The most represented country, U.S., is also one of the countries with the largest community of Internet users (roughly 230 million people).

In addition, we cross-checked the amount of Internet users per country with the whole population of a country, leading so to an *Internet penetration rate* (results are not shown here due to space constraints). In essence, more than 75% of the countries recorded in all the databases have a penetration rate higher than 0.6. The more popular the Internet among the population, the more frequent the country within the databases entries.

Geolocation databases are therefore clearly biased towards Internet usage. Again we note that HostIP is much more impacted than the other databases by the over-representation of the U.S. in its entries. This is expected since HostIP is based on users contributions, that are most likely to be U.S. Internet users.

2) Per-capita GDP: We expected that geolocation databases not only target countries with many Internet users. Furthermore, given that databases are used for electronic

[1]Thus assuming that the country given in the database is correct for any given block, which is reasonable.

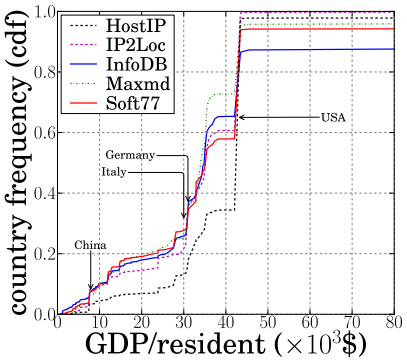


Fig. 4. Fraction of database prefixes as a function of per-capita GDP

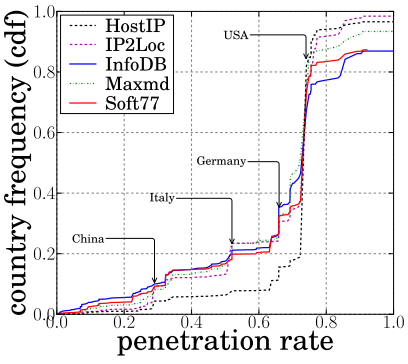


Fig. 5. Fraction of database prefixes as a function of Internet penetration rate

commerce, we also expect that they target Internet users that are more likely to spend money on electronic transactions. We therefore expect that the economic importance is reflected in the geolocation databases. We capture economic importance through the per capita *Gross Domestic Product* (GDP). We choose this measure because most economists use it when looking at per-capita welfare and comparing living conditions or use of resources across countries. Internet users statistics are from 31st December 2009 [18].

Fig. 4 is similar to Fig. 3, but instead of Internet users, the horizontal axis shows the per capita GDP (in US dollar). In addition, we point several countries (China, Italy, Germany, and United States) on Fig. 4

We observe a strong correlation between the number of prefixes in the databases and the per capita GDP. Indeed, countries with higher incomes have more opportunity to benefit from Internet services (Internet access, electronic commerce, online games, etc.) than those with low incomes. As a consequence, it is not necessary for geolocation databases to keep a lot of entries for countries having a low per capita GDP.

It is worth noticing that income, education, age are the principal factors determining the profile of Internet users. So, this exception can be explained by the fact that most of chinese is located to rural areas, and thus can have either a lack of computer skills or "no need" of getting online. Nevertheless, nowadays with the growth in the number of Internet users in China one can expect a rise of the number of entries hosted by China in new geolocation databases.

Fig. 3 also illustrates the rapid growth of country frequency according to the number of users upper than 1,000,000. Indeed, more a country owns an important number of users, more the country's appearances increases in databases. It should be noted that, these broad patterns are noticed for all countries considered as rich (*e.g* USA, Germany).

The first factor that is likely to explain how many prefixes databases keep per country is the *Internet penetration* rate, i.e., the percentage of the population being connected to Internet. The more popular Internet is in a given country, the more we expect to see entries in geolocation databases for this country.

The big jump observed in Fig. 5 around 0.7 is due to the number of prefixes owned by the United States in all databases. Besides, we observe a high correlation between the Internet penetration rate and the number of entries in the database. In essence, more than 75% of the countries recorded in all the databases have a penetration rate higher than 0.6. Put simply, the more popular Internet among the population, the more frequent the country within the databases entries. It should be noted that one country, the Falkland Islands, has an Internet penetration rate of 1, i.e. all users are connected to the Internet. Geolocation databases are therefore clearly biased towards Internet usage. Again we note though that, HostIP is much more impacted than the other databases by the over representation of US in its entries. This is expected since HostIP is based on users contributions, that are most likely US Internet users that do provide US locations.

IV. Databases Reliability

In this section, we are interested to know whether geolocation databases are *reliable*. By reliable we mean that considering mutual comparison for a given IP address, the geolocation provided by the databases is the same (or very close). To this end, we perform an experiment based on a large set of randomly generated IP addresses. We evaluate to which extent databases' answers would match when geolocating arbitrary IP addresses

To this end, we perform two kind of experiments. First, we compute the intersection between each pair of databases, and verify whether the geolocation provided for the intersection is the same in the databases pair (Sec. IV-A). Second, based on a large set of randomly generated IP addresses, we evaluate to which extent databases' answers would match when geolocating arbitrary IP addresses (Sec. IV-B).

A. Databases Overlap

In this section, we consider the overlap that might exist between the five studied databases. First, we observe the common entries that the databases may share. The intersection has been computed by considering that two blocks match if they have the same starting IP address. As the distribution of block sizes strongly differ from one database to another (see Fig. 2), requiring an exact match on the first and last IP address of a block would have led to a very small intersection size. Table II shows the size of the intersection between the databases. Other non shown intersections are empty. The

Database 1	Database 2	Size
InfoDB	HostIP	19,481
	IP2Loc	5,213
	Maxmd	4,725
	Soft77	124
Maxmd	IP2Loc	2,701,034
	Soft77	84,469
Soft77	IP2Loc	85,577

TABLE II
DATABASES INTERSECTION SIZE.

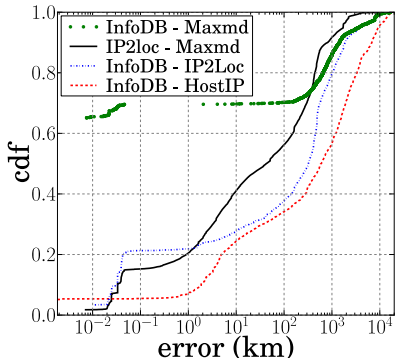


Fig. 6. Database differences for intersections

largest intersection occurs between Maxmind and IP2location that share $2,701,034$ IP blocks. This is more than 75% of the number of blocks of Maxmind. The other pairs of databases share very few blocks. Based on this observation, one might be tempted to think that IP2Location and Maxmind share similar methodologies to construct their databases' entries.

We evaluate how these common blocks are localized by the databases. We want to understand whether common blocks also share common locations. Fig. 6 depicts the CDF of the distance differences (x-axis, logarithmic scale) as returned by the pairs of studied databases for the common blocks. Software77 has not been included in the plot as it only returns countries, but no (latitude, longitude) pairs. The majority of blocks in common between Maxmind and InfoDB (65%) share the same localizations. This is expected since InfoDB is built originally from the free Maxmind database and augmented with other sources such as IANA assignments. However, the proportion of shared locations for other databases pairs is very low. For instance, although they share a significant proportion of prefixes, IP2Location and Maxmind do localize only a tiny proportion of these common prefixes in the same locations (1.7%). We can conclude that even though their blocks selection methodology is quite similar, the process of assigning locations to the entries differs substantially. This suggests that the databases rely on very different location input and methodologies. In turn, widely differing methodologies cast doubts on the ability of any database to accurately geolocate Internet hosts. between IP2Location and Maxmind, for instance).

B. Location Discrepancy

Now, we consider the differences in geolocation across databases when randomly sampling IP addresses across the available blocks.

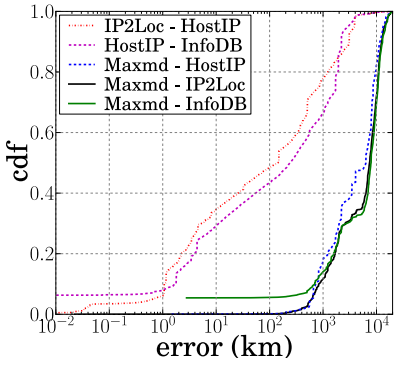


Fig. 7. Database discrepancies for randomly generated IP addresses

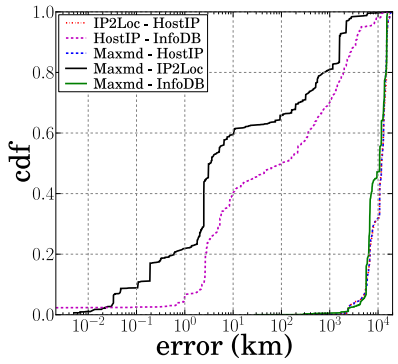


Fig. 8. Database discrepancies for CDNs

We randomly generate 10^6 IP addresses, each byte of an address being randomly selected between 0 and 255. We then geographically localize each of those IP addresses using four databases: Maxmind, IP2Location, HostIP, and InfoDB. Then, we evaluate the difference between the locations returned by each database pair (in km), assuming that these locations are correct. Note that Software77 is not considered here as the number of recorded blocks is too small.

Fig. 7 plots the cumulative distribution of distance difference (in km - x-axis in logarithmic scale) for the four considered databases. We notice first that a low proportion of IP addresses are identically geolocated by a given pair of databases. For instance, in only 5% of the cases, InfoDB and Maxmind provide the same answer. This is roughly the same proportion for HostIP and InfoDB. Fig. 7 confirms that these databases disagree on a vast majority of IP addresses locations. In particular in all our comparisons more than 50% of the provided locations are at least 100km away from each other. Interestingly enough, locations as returned by Maxmind exhibits the largest distance differences compared to other databases, with more than half of the sampled error distances larger than $7,000$ km.

Finally, it is worth noticing that we obtained very similar results to the random 10^6 IP addresses when using a set of $30,000$ IP addresses collected from various content delivery networks (CDNs), as demonstrated by Fig. 8.

	Exact	Smaller	Larger	Partial
IP2Location	32,429	70,963	3,531	373
Maxmind	27,917	79,735	4,092	128
InfoDB	9,954	51,399	1,763	104

TABLE III
MATCHING PREFIXES FROM AN EUROPEAN ISP AGAINST IP2LOCATION, MAXMIND AND INFODB

V. ASSESSING DATABASES ACCURACY ON ISP ADDRESS SPACE

In this section, we first confront three databases with the network of a large European ISP for which we have ground truth about its allocated prefixes and their geographic locations (Sec. V-A). Because of the limited prefix range over which the large European ISP provides ground truth, we augment our use case with information as close as possible to ground truth – DNS names that provide location hints – from two very large transit ISPs (Sec. V-B). We limit ourselves to IP2Location, Maxmind, and InfoDB because Software77 provides only a per-country localization and HostIP is limited to /24 blocks.

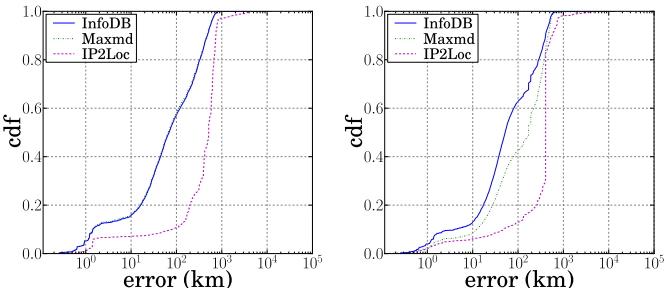
A. ISP Ground Truth

We extracted the complete routing table from a backbone router of a large European ISP. This dump contained a total of about $380,000$ prefixes (both internal and external). From these prefixes, those originated by the ISP were extracted. This list was further trimmed down by dropping all entries not advertised by the ISP to external networks. This leaves us with 357 BGP prefixes advertised by the ISP and reachable from the global Internet that can be matched against the databases. We call this set of prefixes the *ground_truth_set*, since we have POP-level locations for them.

Fig. III shows how the blocks of the three geolocation databases match the prefixes of the ISP (*ground_truth_set*). Four outcomes are possible for the match: *Exact* (the block is present and the same), *Smaller* (the block is present but smaller in the database), *Larger* (the block is present but larger in the database), and *Partial* (the block from the database overlaps with one prefix from the *ground_truth_set*).

The number of geolocation blocks that are smaller than prefixes from the ISP is almost as large as the full set of prefixes from *ground_truth_set*. Surprisingly, the databases also have prefixes that match exactly those from *ground_truth_set* in about 40% (IP2Location), 34% (Maxmind), and 12% (InfoDB) of the cases. Databases therefore rely on the official allocations and advertisements from the ISP, but also try to split the blocks into more specific subsets for geolocation purposes. Few blocks from the databases are bigger than those advertised by the ISP or partially match one from the ISP.

The next step is to extract the city-level position of the routers advertising the subnets inside the ISP, giving us ground truth about the actual location where the prefix is being used by the ISP. To determine the exact location of the prefix, we relied on a passive trace of all IGP messages of one of the backbone routers of the ISP. Thanks to the internal naming scheme of the ISP, we obtained GPS coordinates of the PoP in which each backbone router lies, and associated each prefix



(a) exact match (b) subset match

Fig. 9. Geolocation error of databases for large ISP network with ground truth information

advertised on that router to the location of the router. These coordinates for each prefix are our ground truth used to assess the accuracy of the databases.

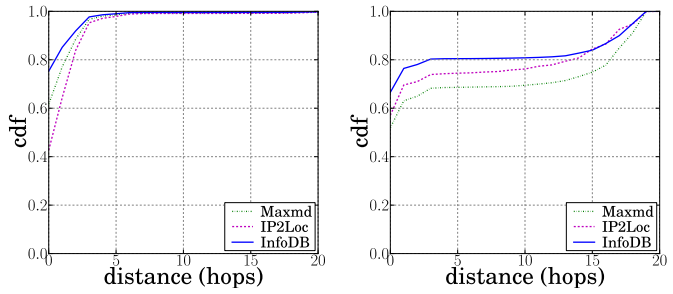
Fig. 9 shows the distribution of the distances between the position reported by IGP and the one reported by the databases, when looking at blocks of the databases that do exactly match (Fig. 9(a)) or are smaller than prefixes advertised by the ISP (Fig. 9(b)). The x-axis (in log-scale) gives a distance (in Km) that we consider as an error from the part of the databases, given the ground truth from the ISP. A value of 10 on the x-axis, for instance, shows the fraction of database prefixes that are less than 10Km away from the ground truth.

From exact matches (Fig. 9(a)), we observe that Maxmind and InfoDB have the same distance distribution to the ground truth (both curves overlap). This is due to the fact that InfoDB is based on the free version of the Maxmind database. Less than 20% of the exact matches for Maxmind and InfoDB are within a few tens of Km from the ground truth. The rest of the blocks have errors distributed between 10Km and 800Km. Note that 800Km is the maximal distance in the country of the considered ISP. IP2Location has much larger errors than Maxmind and InfoDB for the exactly matching blocks, with errors ranging between 200Km and 800Km.

For databases blocks smaller than the ISP prefixes (Fig. 9(b)), we observe two interesting behaviors. First, InfoDB and Maxmind have different error distributions, with Maxmind being actually worse than InfoDB. This is unexpected given that InfoDB is based on the free version of Maxmind. The explanation has to do with the commercial version of the Maxmind database that splits the prefixes from the ISP into very small blocks, many containing only eight IP addresses. Splitting is intended to improve the accuracy of the geolocation, but turns out to make geolocation worse given that many small blocks have incorrect locations.

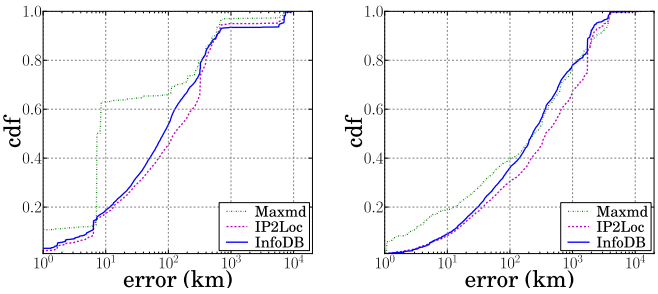
The second observation we make from Fig. 9(b) is the big jump for IP2Location around an error of 400Km for about 50% of the blocks smaller than the ISP prefixes. By checking those blocks, we notice that these belong to a few prefixes from the ISP that are advertised but partly unused. These large prefixes are currently advertised from a single location in the ISP network. A large number of database blocks consistently mislocate subsets of these prefixes.

We report the high success rates in providing the correct



(a) Orange (b) Level3

Fig. 10. Distance to the destination distribution for DNS hints



(a) Orange (b) Level3

Fig. 11. Geolocation error of databases for tier-1 ISP networks based on DNS information

country of the considered IP blocks (between 96% and 98% depending on the database). We conclude that some databases actually do a decent job at geolocating some of the address space of the ISP. In most of the cases however, the location given by the databases is off by several hundreds, even thousands of kilometers. Furthermore, by trying to split the address space into too small blocks, the databases do make mistakes that are hard to detect unless one relies on ground truth information from the ISP that owns the address space. To conclude this section, we cannot trust the databases for the ISP at the granularity of cities, especially given large relative errors they make compared to the span of the considered country (800Km). Their country-level information however seems globally accurate.

B. DNS-based Assessment of Geolocation

Obtaining ground truth information about how allocated address space is being used by ISPs, as we did in Sec. V-A, is difficult since it requires access to confidential information, e.g., IGP routing messages and router configuration. Without such an information, assessing the accuracy of geolocation database records can be done by carrying traceroutes towards a prefix and trying to locate the prefix using location hints from DNS names. Indeed, ISPs sometimes rely on naming conventions for their routers [19].

We select the address space advertised by two tier-1 ISPs, Level3 (AS 3356) and Orange (AS 3215), from BGP dump of June, 30$^{\text{th}}$ 2010. All database records that belong to these BGP prefixes are searched in geolocation databases, leading to 347,736 IP blocks. For each IP block, we perform traceroutes towards the first IP address inside that block. Next we run a DNS lookup for each IP address on the traceroute, starting at the closest to the traceroute destination and working backwards through the traceroute until a DNS query succeeds in resolving the IP address of the router on the path. As shown in Fig. 10, in the vast majority of the cases, the hop with the DNS name we use to estimate the IP block's location is very close to the traceroute destination. In addition, in 66% of the cases, we succeed in resolving a DNS name. The DNS name returned is then searched for location hints. A location hint stands for a string that potentially indicates a city name. This is done by parsing the DNS name looking for simple strings as done by the UNDNS tool [19], and then querying the Google maps service to find the likely location referred to by the hint. If Google maps returns coordinates and a name matching the hint for the location, we deem the IP block to be located close to these coordinates. If more than one suggestion is provided, or if no location hint is found in the DNS name, we simply discard the IP block. We have then been able to find 158 locations (double-checked manually), leading to a DNS-based estimation of the location for more than 165,000 IP blocks, i.e. 48% of the original blocks.

In summary, for each IP block, we selected as its location the geographic coordinates of the router that was closest in hop count to an IP address from the IP block as seen from the traceroutes and that returned a usable DNS name. We stress that these considered locations are only estimations, and as such would likely add an uncertainty of tens of kilometers to the actual locations. However, these estimates can be good indicators of whether geolocation databases' returned locations are sufficiently close to an hypothetical ground truth location.

On Fig. 11, we compare the distance inferred thanks to the DNS hints, with the location provided by geolocation databases. In a similar way to the ISP with ground truth information (Fig. 9), the two tier-1 ISPs confirm the limited accuracy of geolocation databases for most of the IP blocks. Maxmind performs well in the case of Orange (see Fig. 11(a)), thanks to the high concentration of block on a few cities, e.g., Paris. Most of the blocks from Orange are located within 100Km of the location inferred thanks to DNS. For Level3 (see Fig. 11(b)), more than 60% of the IP blocks are mislocated by the databases by more than 100Km. Similarly to the European ISP discussed in Sec. V-A, most of the blocks of Orange have location errors bounded by the diameter of the country in which the ISP has most of its address space, which is less than $1,000$Km both for Orange and the European ISP. In the case of Level3, location errors are larger than $1,000$Km for more than 20% of the studied blocks.

Based on location hints provided by DNS names, we measured similar location errors of geolocation databases to those measured based on ground truth information. By no means do our measurements allow us to make general claims about the accuracy of geolocation databases over the whole address space. Much more extensive measurements are necessary for this. However, given that the studied ISPs are mostly present in Europe and the United States, we believe that the different ISPs we studied cannot be unfortunate cases where

geolocation databases happen to provide poor accuracy at the city-level, and satisfactory accuracy only at the country-level. Our findings here confirm our ground truth-based conclusions.

VI. Conclusion

This paper questioned the reliability of several popular geolocation databases. Given that these databases are frequently used by many services and web sites in the Internet and they do not provide much information about their information sources, the quality of their geolocation information should be checked.

Our findings indicate that geolocation databases often successfully geolocate IP addresses at the country-level. However, their bias towards a few popular countries, mostly those having a large number of Internet users, makes them unusable as general-purpose geolocation services. We observe significant differences among the locations they return for a given IP address, often in the order of hundreds of kilometers. Our results based on a ground truth information from a large Europen ISP, coupled with a study of two other major ISPs where DNS names contain geographic clues, show that the databases perform poorly on the address space of those ISPs. One of the reasons we could identify for their poor geolocation abilities is the way databases try to split prefixes advertised by the studied ISPs into very small blocks. Instead of improving the geolocation accuracy, significant errors are introduced for a large number of blocks, especially at the city-level.

Acknowledgements

Mr. Donnet's work is supported by the FNRS (Fonds National de la Recherche Scientifique, rue d'Egmont 5 – 1000 Bruxelles, Belgium.).

References

[1] 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, December 2006.
[2] V. N. Padmanabhan and L. Subramanian, "An investigation of geographic mapping techniques for Internet hosts," in *Proc. ACM SIGCOMM*, August 2001.
[3] E. Katz-Bassett, J. John, A. Krishnamurthy, D. Wetherall, T. Anderson, and Y. Chawathe, "Towards IP geolocation using delay and topology measurements," in *Proc. ACM SIGCOMM IMC Conference*, October 2006.
[4] B. Wong, I. Stoyanov, and E. G. Sirer, "Golocalization on the Internet through constraint satisfaction," in *Proc. USENIX WORLDS Workshop*, November 2005.
[5] GeoURL, "The GeoURL ICBM address server," http://www.geourl.org.
[6] Net World Map, "The net world map project," http://www.networldmap.com.
[7] Host IP, "My IP address lookup and geotargeting community geotarget IP project," http://www.hostip.info.
[8] IPInfoDB, "Free IP address geolocation tools," http://ipinfodb.com/.
[9] Software 77, "Free IP to country database," http://software77.net/geo-ip/.
[10] Akamai Inc., "Akamai," http://www.akamai.com.
[11] GeoBytes Inc., "GeoNetMap - geobytes' IP address to geographic location database," http://www.geobytes.com/GeoNetMap.htm.
[12] Hexasoft Development Sdn. Bhd, "IP address geolocation to identify website visitor's geographical location," http://www.ip2location.com.
[13] Quova Inc., "GeoPoint - IP geolocation experts," http://www.quova.com.
[14] MaxMind, "Geolocation and online fraud prevention from MaxMind," http://www.maxmind.com/.
[15] B. Gueye, S. Uhlig, and S. Fdida, "Investigating the imprecision of IP block-based geolocation," in *Proc. PAM Conference*, April 2007.
[16] S. Siwpersad, B. Gueye, and S. Uhlig, "Assessing the geographic resolution of exhaustive tabulation for geolocating Internet hosts," in *Proc. PAM Conference*, April 2008.
[17] M. Freedman, M. Vutukurum, N. Feamster, and H. Balakrishnan, "Geographic locality of IP prefixes," in *Proc. ACM SIGCOMM IMC Conference*, October 2005.
[18] Miniwats Marketing Group, "Internet world stat: Usage and population statistics," 2009, http://www.internetworldstats.com.
[19] N. Spring, D. Wetherall, and T. Anderson, "Scriptroute: A public Internet measurement facility," in *Proc. USENIX USITS Symposium*, March 2003.