IP Geolocation Databases: Unreliable?

Ingmar Poese, Steve Uhlig
Deutsche Telekom Laboratories/TU Berlin, Germany
{ingmar,steve}@net.t-labs.tu-berlin.de

Mohamed Ali Kaafar
INRIA Rhône-Alpes, France
mohamed-ali.kaafar@inrialpes.fr

Benoit Donnet
Université catholique de Louvain, Belgium
benoit.donnet@uclouvain.be

Bamba Gueye
Université Cheikh Anta Diop de Dakar, Senegal
bamba.gueye@ucad.edu.sn

This article is an editorial note submitted to CCR. It has NOT been peer reviewed.
The authors take full responsibility for this article's technical content. Comments can be posted through CCR Online.

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 have an insight of 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. This is the first study using a ground truth showing that the overly fine granularity of database entries makes their accuracy worse, not better. Geolocation databases can claim country-level accuracy, but certainly not city-level.

Categories and Subject Descriptors

C.2.1 [**Network Architecture and Design**]: Network Topology

General Terms

Geolocation, Measurement

Keywords

accuracy, reliability

1. 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 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.[1]

The remainder of this paper is organized as follows: Sec. 2 describes the geolocation databases we consider in this work; Sec. 3 confronts three commercial databases with the network of a large European ISP for which ground truth is available; finally, Sec. 4 concludes this work.

2. DATABASES

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 1: 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 1 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].

[1]An extended version of this work can be found in [18].

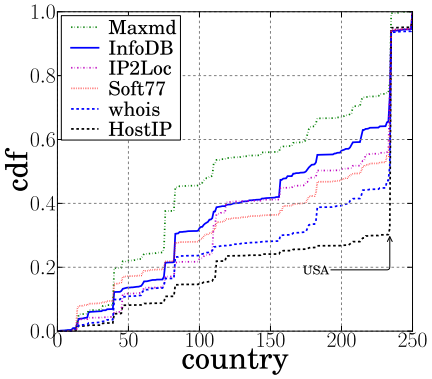


Figure 1: Countries distribution

Table 1 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 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 for countries distribution of various databases used). The five databases display a similar shape of their cumulative number of blocks across countries. The big jump at country 235 corresponds to the over-representation of the United States 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 1, 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.

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.

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

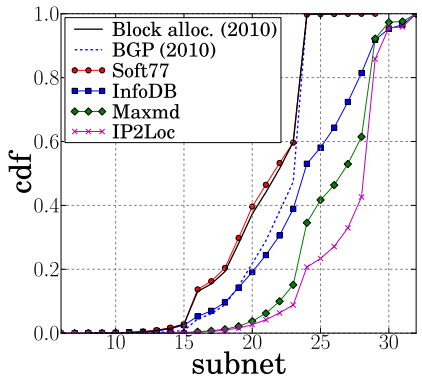
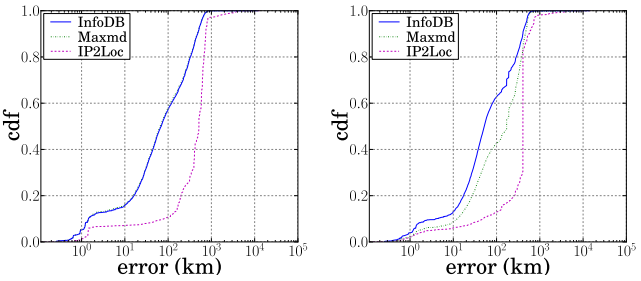


Figure 2: Prefix distribution



(a) exact match (b) subset match

Figure 3: Geolocation error of databases for large ISP network with ground truth information

	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 2: Matching prefixes from an European ISP against IP2Location, Maxmind and InfoDB

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.

3. ISP GROUNDTRUTH

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. 2 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 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. 3 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. 3(a)) or are smaller than prefixes advertised by the ISP (Fig. 3(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. 3(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. 3(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. 3(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 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.

4. 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 makes them unusable as general-purpose geolocation services. Our results based on a ground truth information from a large European ISP show that the databases perform poorly on the address space of this ISP. One of the reasons we could identify for their poor geolocation abilities is the way databases try to split prefixes advertised by the studied ISP 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.).

5. 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] I. Poese, M. A. Kaafar, B. Donnet, B. Gueye, and S. Uhlig, "IP geolocation databases: Unreliable?" Technische Universität Berlin, Fakultät Elektrotechnik und Informatik, Technical Report 2011-03, February 2011, see http://www.net.t-labs.tu-berlin.de/papers/PKDGU-IGDU-11.pdf.