Assessing the geographic resolution of exhaustive tabulation for geolocating Internet hosts

S.S. Siwpersad[1], Bamba Gueye[2], and Steve Uhlig[1]

[1] Delft University of Technology, The Netherlands
mail@swiep.com,S.P.W.G.Uhlig@ewi.tudelft.nl
[2] Université de Liège, Belgium
cabgueye@ulg.ac.be

Abstract. Geolocation of Internet hosts relies mainly on exhaustive tabulation techniques. Those techniques consist in building a database, that keeps the mapping between IP blocks and a geographic location. Relying on a single location for a whole IP block requires using a coarse enough geographic resolution. As this geographic resolution is not made explicit in databases, we try in this paper to better understand it by comparing the location estimates of databases with a well-established active measurements-based geolocation technique.
We show that the geographic resolution of geolocation databases is far coarser than the resolution provided by active measurements for individual IP addresses. Given the lack of information in databases about the expected location error within each IP block, one cannot have much confidence in the accuracy of their location estimates. Geolocation databases should either provide information about the expected accuracy of the location estimates within each block, or reveal information about how their location estimates have been built, unless databases have to be trusted blindly.
Keywords: geolocation, exhaustive tabulation, active measurements

1 Introduction

Location-aware applications have recently become more and more widespread. Examples of such applications comprise targeted advertising on web pages, displaying local events and regional weather, automatic selection of a language to first display content, restricted content delivery following regional policies, and authorization of transactions only when performed from pre-established locations. Each application may have a different requirement on the resolution of the location estimation. Nevertheless, as IP addresses are in general allocated in an arbitrary fashion, there is no strict relationship between an IP address and the physical location of the corresponding physical interface.

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. When coupled with a script embedded in a website and upon a client access to the website being detected, a request can be sent instantly to the database. This request can be to check if the IP address has an exact or longest prefix match (LPM) with a corresponding geographic location and coordinate. Since there is no actual measurement involved but

merely a simple lookup, the request can be served in a matter of milliseconds. The expected time for which a website should be fully loaded, without causing any nuisance, is in general within one second. Most commercial database providers offer highly optimized scripts as well as abundantly documented application programming interfaces, which meet this short expected response time. The database-driven geolocation thus seems to be a useful approach.

Examples of geolocation databases are *GeoURL* [1], the *Net World Map* project [2], and free [3] or commercial tools [4–9]. Exhaustive tabulation is difficult to manage and to keep updated, and the accuracy of the locations is unclear. In practice however, most location-aware applications seem to get a sufficiently good geographic resolution for their purposes.

In this paper, we try to better understand the resolution of geolocation databases, by comparing their location estimates with a well-known active measurements-based geolocation technique, CBG [10]. We show that, as expected, the geographic resolution of databases is far coarser than the resolution provided by active measurements, typically several times coarser than the confidence given by active measurements. As most geolocation databases do not give confidence in the accuracy of their location records, they are likely not to be trustworthy sources of geolocation information if precise IP address-level locations are required. Applications that require as much accuracy as possible would thus typically have to rely on active measurements, not databases. To improve the quality of current geolocation databases, we believe that the database records should contain information about the expected confidence in the location estimates.

The remainder of the paper is structured as follows. Section 2 introduces the datasets used. Section 3 studies the geographic resolution of databases. Section 4 describes our active measurements for geolocating Internet hosts. In Section 5, we compare the resolution of active measurements with location estimates from databases. Finally, we conclude in Section 6.

2 Datasets

During the past few years, a growing number of companies have spent a lot of effort in creating databases for geolocation purposes. Most of these companies, like Maxmind [11], Hexasoft [8] and Quova [9], provide commercially available databases with periodic updates. There are also freely available databases such as Host IP [3].

One of the problems of geolocation databases is that typically one does not know much about the methodology used by the database provider to gather their geographic information. One has to blindly rely on the claimed geographic resolution they provide. There are four basic geographic resolution levels that occur in most databases: zipcode, city, country and continent. Note that some databases may use more resolutions than those four, like regions that may relate to countries, continents, or some intermediate resolution. In most instances, we expect that the zipcode and the city granularity will be very similar. The country resolution is widely recognized to be the typical one that is reliable from databases. Many databases do not give any information about the expected geographic resolution of the database records, and when they do, not all records do contain this information. The price of commercial databases increases with improved

geographic resolution, or with additional information about attributes of IP blocks like ISP, connection type of hosts, and in a single instance confidence about the location estimates. Note that we know one example of geolocation database that provides a notion of confidence related to the uncertainty about where the end-user actually lies compared to the location estimate [9]. This notion of confidence is however not quantitative, i.e. it does not express how far an IP address belonging to the IP block is expected to be from the location estimate provided, rather the type of host or connection that the host is using.

In the sequel of this paper, we restrict our attention to two databases. These commercial databases, GeoIP by Maxmind [11] and IP2Location by Hexasoft [8], are used because of their popularity (see [8, 11] for a listing of some of their customers) and their expected reliability. The number of IP blocks and the coverage in IP addresses of

$Database$	$Public\ blocks$	$Special\ blocks$	$Total\ blocks$	$Public\ addresses$	$Total\ addresses$
Maxmind	3,278,391	2	3,278,393	2,322,257,277	2,355,811,965
Hexasoft	5,111,309	44	5,111,353	3,991,797,760	4,294,967,296

Table 1. Overview of the 2 selected databases.

the two databases is shown in Table 1. Maxmind contains more than 3 million blocks, and Hexasoft more than 5 million blocks. Note that a few blocks, called special blocks according to RFC3330 [12], should not be considered.

3 Geographic resolution of databases

Based on the information provided in the geolocation databases, it is hard to say anything about the actual geographic resolution of the location estimates. We merely know that most records contain either a city or a country name. 73.1% of the databases records in Maxmind contain a city name (66.6% for Hexasoft), then if no city name can be found, 3.4% of the records contain a country name (33.2% for Hexasoft). When neither a city name nor a country name is present in the record, a continent name or a federation of countries will typically be found. Note that sometimes records contain geographic coordinates only. While the area of countries and continents are well-defined, the area of a city depends much on what is meant by the boundaries of the considered city. For example, taking the largest 250 cities in the world[3] shows well how much the area of a city can vary, especially depending on whether the suburbs or the "metro" area are considered to be part of the city or not.

When we analyze the number of unique cities in both Maxmind and Hexasoft, we obtain $110,349$ unique cities in Maxmind and $15,133$ in Hexasoft. $100,087$ cities in Maxmind occur each in a single IP block ($12,918$ for Hexasoft), and $10,262$ cities occur each in multiple IP blocks ($2,215$ in Hexasoft). When several IP blocks have

[3] `http://www.citymayors.com/statistics/largest-cities-area-250.html`

the same city information, they will have the same location estimate in the database. Note that a city is defined by a city name, but also a country and a continent when this information is available in the databases. Some city names occur in several countries and/or continents. When we compare the occurrence of unique city names (string-wise), we observe that among a total of $7,844$ unique city names present in the databases, $7,618$ are present in one database only, and 226 are in both.

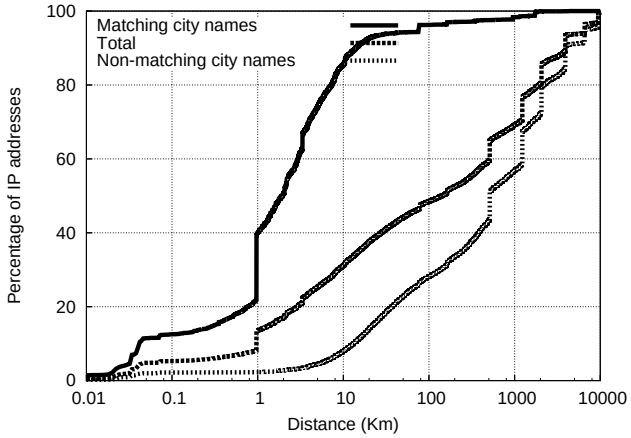


Fig. 1. Difference in location between Maxmind and Hexasoft.

In geolocation databases, a unique location is associated to a given city. It is thus impossible to infer directly the geographic resolution used by the databases by comparing the location estimates of different IP blocks for a given city. However, we can compare the location estimates from Maxmind and Hexasoft, hoping that the difference between their location estimates will give us an indication of their geographic resolution. We rely on a free database, Host IP [3], that contains $1,356,506$ IP blocks, to perform lookups in the two other databases. For each IP block of Host IP, we take an IP address and use it to lookup the two databases. We then compute the difference between the two location estimates returned by the databases. Figure 1 displays the cumulative distribution of the distance between the locations given by the two databases when performing a lookup on IP addresses from the Host IP database. We provide three different curves, one for the distribution of the distance when the city strings match between Maxmind and Hexasoft, when they do not match, and irrespective of the city-level match. Among the $1,264,892$ IP addresses looked up, $377,736$ have the same city-level name in the databases, while $887,156$ do not have matching city names. We see on the curve that corresponds to matching cities that the difference in location between the databases tends to be far smaller than when the city names do not match. Depending on whether the city names match between the two databases entries, the typical distance between their location estimates differs much. When the IP blocks from the two databases have the same city name information, their locations are very close, typically less than 10Km.

When the city names do not match on the other hand, the locations differ more than usual. Globally, about 50% of the IP lookups give a difference smaller than 100Km. If the differences observed between the databases were to reflect in some way differences in geographic resolutions used by them, then we would deduce that those resolutions go from 1Km up to thousands of Km.

4 Measurements-based geolocation

Given that we cannot obtain the actual geographic location of many IP addresses in the Internet, we need to rely on location estimates. To obtain location estimates for a large enough number of IP hosts, we need accurate location estimates. For this, we rely on active measurements. Active measurements have the advantage of providing an explicit estimate of their accuracy.

Previous works on measurement-based geolocation of Internet hosts [13,14] use the positions of reference hosts, called landmarks, with a well-known geographic location as the possible location estimates for the target host. This leads to a discrete space of answers; the number of answers is equal to the number of reference hosts, which can limit the accuracy of the resulting location estimation. This is because the closest reference host may still be far from the target. To overcome this limitation, the authors of [10] propose the Constraint-Based Geolocation (*CBG*) approach, which infers the geographic location of Internet hosts using *multilateration*. Multilateration refers to the process of estimating a position using a sufficient number of distances to some fixed points. As a result, multilateration establishes a continuous space of answers instead of a discrete one. This multilateration with distance constraints provides an overestimation of the distance from each landmark to the target host to be located, thus determining a region, *i.e.* confidence region, that hopefully encloses the location of the target hosts [10]. For instance, the confidence region allows a location-aware application to assess whether the estimate is sufficiently accurate for its needs.

Although showing relatively accurate results in most cases, these measurement-based approaches may have their accuracy disturbed by many sources of distortion that affect delay measurements. For example, delay distortion may be introduced by the circuitous Internet paths that tend to unnecessarily inflate the end-to-end delay [15–17] and by the potential existence of bottleneck links along the paths. To deal with these sources of distortion, *GeoBuD*, *Octant*, and *TBG* were proposed by [18–20]. The GeoBuD technique shows that estimating buffering delays, by *traceroute* measurements, at intermediate hops along the traceroute path between a landmarks and a target host enables to improve the accuracy of geolocation of Internet hosts. In the same way, Topology-Based Geolocation (TBG) and Octant which are an extension of multilateration techniques with topology information were proposed. TBG additionally uses inter-router latencies on the landmark to target network paths to find a physical placement of the routers and target that minimizes inconsistencies with the network latencies. TBG relies on a global optimization that minimizes average position error for the routers and target. Octant differs from TBG by providing a geometric solution technique rather than one based on global optimization. Although it considers intermediate routers as additional landmarks, Octant also uses geographic and demographic information. Geographic and

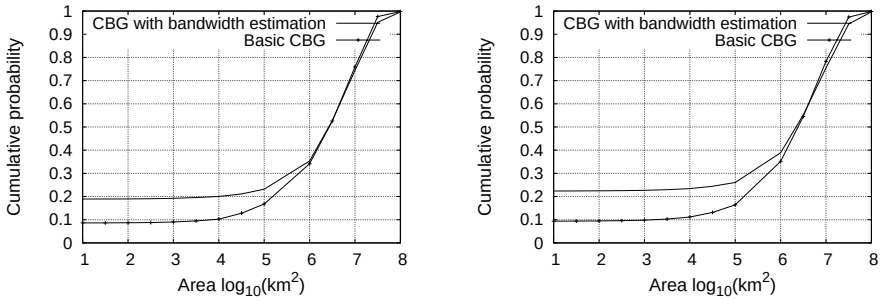
demographic constraints are used in Octant to reduce the region size where the target may be located. Only landmasses and areas with non-zero population are considered as possible target locations [19]. Furthermore, it takes into account queuing delays by using height as an extra dimension. It requires significantly computational time and resources. All these techniques generate a huge amount of overhead in the network for a small gain in accuracy.

To illustrate the marginal improvement of complex measurement-based geolocation techniques, we do not only consider CBG, but also add to it estimation of the bottleneck bandwidth on the path. The bottleneck bandwidth can be defined as the maximum throughput that is ideally obtained across the slowest link over a network path. CBG with bandwidth estimation allows the improvement of the geolocation estimation given by CBG. Additional delay distortions caused by the bottleneck along the path are removed from the overestimations of distance constraints that define the region enclosing the target host in CBG, allowing tighter overestimations that result in a smaller region. Smaller regions that still enclose the target host provide a more accurate location estimation.

4.1 CBG with bandwidth estimation

To estimate the bottleneck bandwidth over a network path between each landmark and a given target host, we use *SProbe* [21]. SProbe estimates bottleneck bandwidth in uncooperative environments, *i.e.* a measurement software is only deployed locally on the measurement host. SProbe relies on the exploitation of the *TCP* protocol. It sends two *SYN* packets to an inactive port on the remote host to which it appends 1460 bytes of data. Since the port is inactive, the remote host answers to these packets with two *RST* packets of 40 bytes each. For the native traceroute used by Octant, TBG, and GeoBuD, three packets are sent to each intermediate hops between a source and a destination causing an important overhead. SProbe produces accurate and fast estimates using little amount of probing data, so that it can scale to a large number of estimates.

For our evaluation, we rely on 39 *PlanetLab* nodes [22] as landmarks and we use a subset of the two commercial databases (Maxmind and Hexasoft) as input for hosts to be localized. Each landmark estimates the bottleneck bandwidth towards a given target host by sending 7 SYN packets. We found in Section 3 that there are 226 city names that are unique and can be found in both databases. Using these city names we find $41,797$ IP blocks from Maxmind matching those city names. Since we need "*pingable*" addresses within each IP block to be used in measurements, we use the single ping approach to find at least one IP address per block. The single ping approach consists in brute-force probing all IPs within a prefix, and stopping the probing within the prefix as soon as a single IP address has answered. We find $18,805$ IP blocks which have at least one pingable IP address for Maxmind. For the Hexasoft database, we have $41,758$ IP blocks among which $15,823$ contain at least one pingable IP address. Using the set of pingable addresses, Figure 2 presents the cumulative distribution of the confidence region in km^2 for location estimates in both the Maxmind and Hexasoft databases. Figure 2(a) shows that CBG with bandwidth estimation assigns a confidence region with a total less than 10^4 km^2 for about 20% of the location estimates, whereas the basic CBG has only 10% for the same confidence region. For IP addresses that



(a) CBG using the Maxmind dataset (b) CBG using the Hexasoft dataset

Fig. 2. Confidence region.

are given a confidence region between 10^4 km^2 and 10^6 km^2, bandwidth estimation is less and and less useful. Finally, when the confidence region is larger than 10^6 km^2, bandwidth estimation is useless, or even makes the confidence region larger than the classical CBG technique.

Measurement-based geolocation techniques assume that the target host is able to answer measurements. Active measurements will be impractical when we rely on *ICMP echo* probes for instance, which can be filtered by a firewall. We observe that for most IP blocks, we get only a few IP addresses that answer our probes, typically only one.

5 Comparison between databases and active measurements

Having discussed the geographic resolution of geolocation databases in Section 3 and presented the confidence area obtained with active measurements in Section 4, we use the active measurements introduced in Section 4 to check the resolution of geolocation databases. When comparing geolocation based on active measurements and databases, several situations may occur. One possibility is when databases and active measurements give the same location for an IP address, i.e. databases give a location that lies within the confidence region given by active measurements. This situation is not typical, given the coarse geographic resolution of database records. When location estimates from the databases do not belong to the confidence region provided by active measurements, we would tend to doubt the accuracy of databases rather than expecting that the confidence region suffers from measurements biases, as the confidence region is made from higher bounds on the distance constraints.

Let us now measure the distance between the border of the confidence region given by CBG and the location estimates of the databases. If CBG is correct in its estimation of the location, then this distance should provide a lower bound on the actual geolocation error made by the database. Figure 3 shows the cumulative distribution of the minimal distance between the location estimates of the Maxmind dataset (results for Hexasoft are similar) and the border of the confidence region given by CBG, with and without using bandwidth estimation. This minimal distance first tells whether the location estimates from databases are within the confidence region or not. If the distance

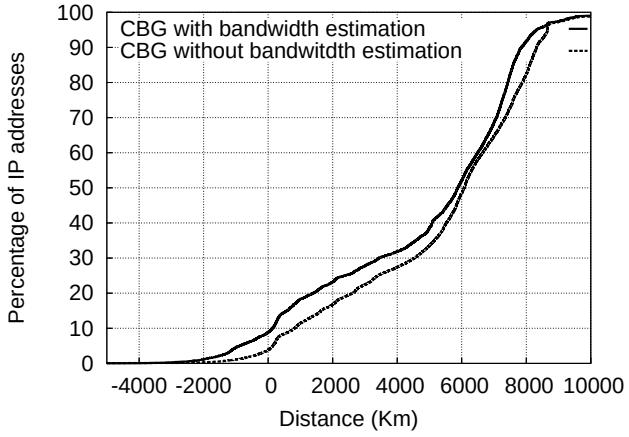


Fig. 3. Distance between the database results and the border of the CBG confidence region (Maxmind dataset).

is negative on Figure 3, it means that databases are within the confidence region. If the confidence region is small and the location estimate of the database lies within the confidence region, then we expect that it is likely that the database estimate is correct. We observe on Figure 3 that more than 90% of the probed IP addresses have a database location estimate that lies outside the confidence region, and quite far away from it. Note that in a few cases the distance on Figure 3 is negative and large, meaning that the confidence region is pretty large.

The large distances shown in Figure 3 suggest that the geographic resolution of databases is poor, compared to the confidence region given by CBG. To quantify the relative resolution of databases compared to the confidence region given by CBG, we plot in Figure 4 the ratio of the difference between the CBG estimate and the locations given by the Maxmind dataset (results for Hexasoft are similar), divided by the uncertainty in the CBG estimate (radius of the confidence region). Let us denote the location given by CBG by $loc_{cbg}(IP)$, the radius of CBG's confidence region by $radius_{cbg}(IP)$, and the location given by a database by $loc_{database}(IP)$, then the ratio we compute is $\mid \frac{loc_{database}(IP) - loc_{cbg}(IP)}{radius_{cbg}(IP)} \mid$. A ratio smaller than 1 means that the location estimate given by the database is within the confidence region. In this case, we would tend to trust the location estimate given by the database. A ratio larger than 1 means that the location estimate given by the database lies outside the confidence region. In that case, it is likely that the geographic resolution of the database is too coarse to give an accurate location estimate for the considered IP address. We observe on Figure 4 that the ratio is typically far larger than 1, meaning that the geographic resolution of the databases compared to the confidence in the active measurements estimates is poor, relative to the confidence region of CBG. For only less than 10% of the probed IP addresses, the databases have a good enough geographic resolution to make them comparable to the accuracy of active measurements. Note that those results do not suggest that location

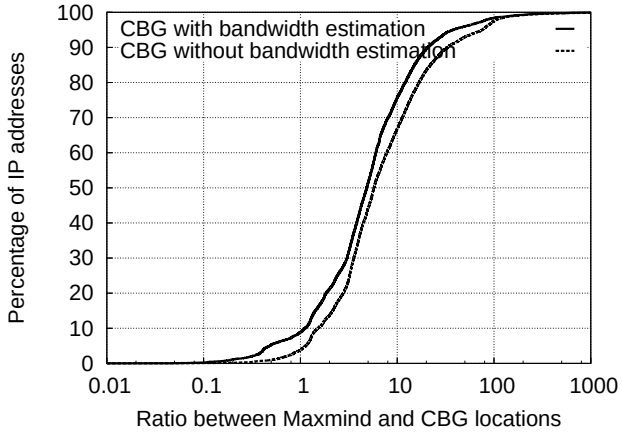


Fig. 4. Ratio of the distance of the databases to the distance of CBG with respect to the CBG location estimate (Maxmind dataset).

estimates provided by databases are incorrect, but rather that the geographic resolution at which databases give mappings from IP blocks to locations are too coarse to provide accuracy at the level of individual IP addresses.

6 Conclusion

In this paper, we assessed the geographic resolution of geolocation databases. We described the typical content of such databases, showing that they do not contain information to give confidence in the expected accuracy of their location estimates. We illustrated the relative coarse resolution databases provide, by showing how large the span of cities is, and how much the location estimates differ between the considered databases.

We carried out active measurements in order to compare the geographic resolution of databases to a more accurate standard. We quantified the accuracy of active measurements, and tried to improve them by adding bandwidth measurements to reduce the bias from bottleneck links.

Our comparison of the active measurements and the location estimates from the databases demonstrated the coarse geographic resolution of databases location estimates. We showed that not only the distance between the location estimate of the databases and the location given by active measurements is very large, but that also difference between the database location estimates from the active measurements estimates, divided by the accuracy expected from the active measurements, is very large.

Our work shows that the geographic resolution of geolocation databases is coarse compared to the one of active measurements. That does not mean that the location estimates given by databases are not good enough. Information about the geographic resolution of the databases can be embedded in them, for example by giving an estimate of the city-level span for each record. In general, we do not expect that active measure-

ments will be so helpful to improve the geographic resolution of geolocation databases, simply because databases work at the level of IP blocks. However, in particular cases where better accuracy is required for specific IP addresses, active measurements have great potential to provide better location estimates than databases.

Acknowledgments

Bamba Gueye is supported by the IST ANA project.

References

1. *GeoURL*, http://www.geourl.org/.
2. *Net World Map*, http://www.networldmap.com/.
3. Host ip, http://www.hostip.info.
4. Digital Island Inc, http://www.digitalisland.com/.
5. Akamai Inc, http://www.akalai.com/.
6. *GeoNetMap*, http://www.geobytes.com/GeoNetMap.htm.
7. *WhereIsIP*, http://www.jufsoft.com/whereisip/.
8. H. D; S; Bhd, *Ip2location LLC*, http://www.ip2location.com.
9. *GeoPoint*, http://www.quova.com/.
10. B. Gueye, A. Ziviani, M. Crovella, and S. Fdida, "Constraint-based geolocation of internet hosts," *IEEE/ACM Trans. Netw.*, vol. 14, no. 6, pp. 1219–1232, 2006.
11. MaxMind LLC, *MaxMind*, http://www.maxmind.com.
12. IANA, "Special-use IPv4 addresses," Tech. Rep., Internet RFC 3330, Sept. 2002, http://www.rfc-editor.org/rfc/rfc3330.txt.
13. A. Ziviani, S. Fdida, J. F. de Rezende, and O. C. M. B. Duarte, "Improving the accuracy of measurement-based geographic location of Internet hosts," *Computer Networks, Elsevier Science*, vol. 47, no. 4, pp. 503–523, Mar. 2005.
14. V. N. Padmanabhan and L. Subramanian, "An investigation of geographic mapping techniques for Internet hosts," in *Proc. of ACM SIGCOMM*, San Diego, CA, USA, Aug. 2001.
15. H. Tangmunarunkit, R. Govindan, S. Shenker, and D. Estrin, "The impact of routing policy on internet paths," in *Proc. of IEEE INFOCOM*, Anchorage, AK, USA, Apr. 2001.
16. L. Subramanian, V. Padmanabhan, and R. Katz, "Geographic properties of Internet routing," in *Proc. USENIX*, Monterey, CA, USA, June 2002.
17. H. Zheng, E. K. Lua, M. Pias, and T. Griffin, "Internet Routing Policies and Round-Trip-Times," in *in Proc. of PAM Workshop*, Boston, MA, USA, Apr. 2005.
18. B. Gueye, S. Uhlig, A. Ziviani, and S. Fdida, "Leveraging buffering delay estimation for geolocation of Internet host," in *Proc. IFIP Networking Conference*, Coimbra, Portugal, May 2006, Lecture Notes in Computer Science (LNCS) 3976, pp. 319–330.
19. B. Wong, I. Stoyanov, and E. Gün Sirer, "Geolocalization on the internet through constraint satisfaction," in *WORLDS'06: Proceedings of the 3rd conference on USENIX Workshop on Real, Large Distributed Systems*, Berkeley, CA, USA, 2006, pp. 1–1, USENIX Association.
20. 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 IMC measurement*, Rio de Janeriro, Brazil, Oct. 2006.
21. S. Saroiu, P. K. Gummadi, and S. D. Gribble, "Sprobe: A fast technique for measuring bottleneck bandwidth in uncooperative environments," in *Proc. of IEEE INFOCOM*, New York, NY, USA, June 2002.
22. *PlanetLab: An open platform for developing, deploying, and accessing planetary-scale services*, 2002, http://www.planet-lab.org.