Investigating the Imprecision of IP Block-Based Geolocation

Bamba Gueye[1] and Steve Uhlig[2] and Serge Fdida[1]

[1] Université Pierre et Marie Curie (Paris 6)
Laboratoire LiP6/CNRS - UMR 7606
{Bamba.Gueye,Serge.Fdida}@lip6.fr
[2] Delft University of Technology
Network Architectures and Services
S.P.W.G.Uhlig@ewi.tudelft.nl

Abstract. The lack of adoption of a DNS-based geographic location service as proposed in RFC 1876 has lead to the deployment of alternative ways to locate Internet hosts. The two main alternatives rely either on active probing of individual hosts or on doing exhaustive tabulation of IP address ranges and their corresponding locations. Using active measurements, we show that the geographic span of blocks of IP addresses make their location difficult to choose. Using the single location for a block of IP addresses as an estimation of the location of its IP addresses leads to significant localization errors, whatever the choice made for the location of the block. Even using as the location of a block the one that minimizes the global localization error for all its IP addresses leads to large errors. The notion of the geographic span of a block of IP addresses is fuzzy, and depends in practice very much on the uncertainty associated to the location estimates of its IP addresses.
Keywords: geolocation, active measurements, exhaustive tabulation

1 Introduction

Location-aware applications have recently become more and more widespread [1–3]. One approach to locate Internet hosts is to push their location inside DNS records, as proposed in RFC 1876 [4]. Unfortunately, the adoption of this approach has been limited since it requires changes in the DNS records. There are also some geolocation services based on an exhaustive tabulation between IPs ranges and their corresponding locations. Examples of such services are *GeoURL* [1], the *Net World Map* project [2], and several commercial tools. Exhaustive tabulation is difficult to manage and to keep updated, and the accuracy of the locations is uncertain.

Padmanabhan et al. [5] developed three different techniques to map IPs to geographic locations and investigated the challenges in geolocation of Internet hosts. One of those techniques iteratively clusters IP addresses to map them to a single location. The authors of [5] observed that the accuracy of this method was related to the geographic spread of the hosts within these blocks of IP addresses.

In this paper, we quantify the extent to which locating all IP addresses within a block leads to an inaccurate geolocation of Internet hosts. We compare the location of blocks

of IP addresses as given from two datasets [6, 7] with IP address location estimates based on active measurements. We show that the geographic span of the blocks of IP addresses, together with the intrinsic uncertainty of the exact location of individual IP addresses makes the choice of the location of a block difficult. The notion of the geographic span of a block of IP addresses is itself fuzzy, and depends in practice very much on the uncertainty associated to the location estimates of the IP addresses that belong to it. Even the *optimal location* (location that minimizes the global localization error of all IP addresses within a given block) of IP addresses blocks leads to significant differences between the estimated location of IP addresses and the one attributed to the entire block. Note that throughout this paper we refer to "block of IP addresses" as block and "IP addresses" as IPs.

The paper is organized as follows. Section 2 presents the datasets used to infer the location of target hosts based on active measurements. In Section 3, we investigate the inherent imprecision of estimating the location of individual IP addresses using a single location for their block. Finally, we conclude in Section 4.

2 Datasets

We consider two datasets containing IP or block of IP addresses to location entries in this paper. The first dataset contains $292,362$ potential IPs of Web Clients that exchanged content over CoralCDN [6] and the second dataset is the database used by *GeoIP* [3]. For each IP address that composes the CoralCDN dataset, we seek its geographic location in the *GeoIP* [3] database. The GeoIP database provides also the block of IP addresses that this IP belongs to. Afterwards, we cross-check the location estimate obtained for each IP with its location estimate given by [7]. We found that $80,449$ hosts provide different location estimates whereas $211,913$ hosts have the same location estimate (at the city-level). Considering the IPs for which we found the same location estimate in the two databases, we apply the CBG technique [8] to find their geographic location estimate. For our measurements, we relied on 74 PlanetLab nodes spread all over the world as landmarks. During 3 weeks from 31 March until 19 April 2006, we conducted measurements to locate $25,775$ IPs among the $211,913$ whose location at the city-level agreed between the two databases. Among the set of IPs used, $7,016$ have not been located by CBG. These hosts may be private, behind firewalls, or simply do not respond to *ping* probes. Thus we use for our study the remaining $18,759$ IPs that were successfully located. The $18,759$ successfully localized correspond to 876 blocks. The number of IP addresses probed within these 876 blocks varies between 3 and 197.

3 Limitations of block-level geolocation

3.1 Geographic span of IP addresses blocks

Estimating the actual geographic area spanned by a block of IPs is tricky. Geolocation of IP addresses based on active measurements and exhaustive tabulation both contain some uncertainty. In the case of active measurements like CBG, the geolocation is given in the form of an area where the host lies, the *Confidence Region* (CR) [8]. Since all

IPs of a block are attributed to a single city-level location in the two used databases, it is impossible to estimate the span of blocks based on this information. Hence we have to rely on the estimates provided by CBG for each IP address. Note that we use as the location of an IP address the centroid of the CR computed by CBG [8].

For each block p, we compute the maximal distance between any two of its IPs $d_{max}(p)$ for which CBG gave us a location. We call $d_{max}(p)$ the *maximal span* of block p. Since $d_{max}(p)$ might be far larger than the typical distance between the locations of any two IP addresses within p, we also compute the median of the distance between any pair of IP addresses within p. We call this median distance the *median span* of block p.

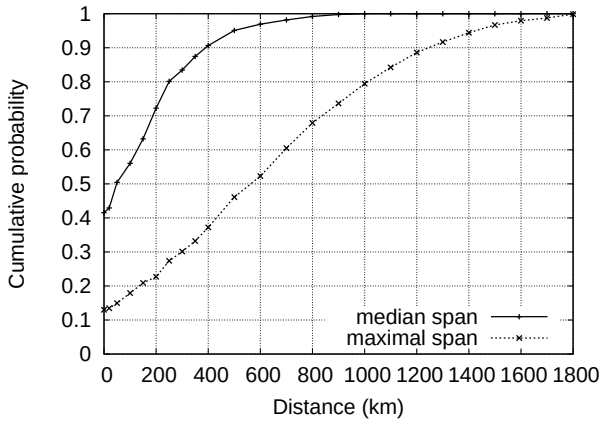

Fig. 1. Geographic span of blocks.

Figure 1 provides the CDF of both the maximal and the median spans over the 876 blocks of IPs. More than 10% of the blocks have a maximal span of 0, i.e. all their IPs have exactly the same location. More than 40% of the blocks have a median span of 0 Km. Half of the IPs of these blocks are located at the same spot. Note that having a very small span for a block requires that the uncertainty of the geolocation of its IPs be very small, which typically happens when the localized host is close to one or several landmarks. About 50% of the blocks have a maximal span larger than 500 Km. Only 5% of the blocks have a median span larger than 500 Km.

3.2 Optimal location of blocks of IP addresses

Assume that we want to have a location of a block that lies as close as possible to the locations of all IPs within this block. If we locate an IP at the centroid of the CR given by our active measurements, how large is the minimal geolocation error that we can expect when using as an estimate of the location of the IPs the location of the whole block? To answer this question, we compute the optimal location for each block of IP addresses, i.e. the location that minimizes the sum of the distances between the location of the block and the centroid of the CR of each of its IPs. If we were to do that, we would obtain approximation errors as the one shown on Fig. 2.

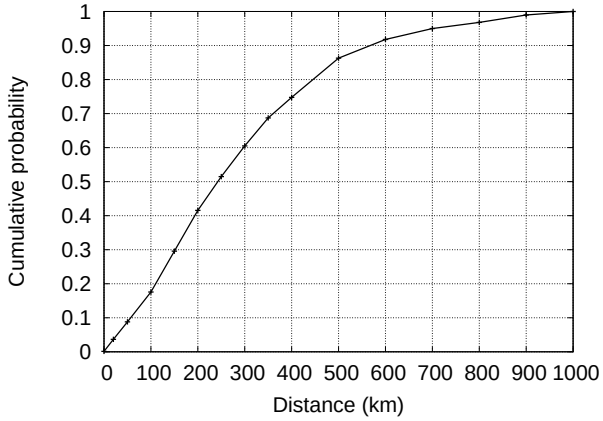

Fig. 2. Distribution of distances between optimally-located blocks and their IPs.

If we attribute to each block the optimal location, only a little bit more than 40% of the IPs would be located at most 200 km away from their estimated location based on active measurements. More than 10% of the IPs would still suffer from a wrong localization more than 500 km away from their estimated location. Even though this is a little better than the localization we have in the database, this would still be far from satisfactory compared to the localization active measurements are able to provide.

4 Conclusion

We investigated the imprecision of relying on the location of blocks of IP addresses to locate Internet hosts. We showed that the geographic area spanned by blocks can be large, far larger than the typical distance between any two IPs within a block. We showed that even using the optimal location of a block leads to large geolocation errors. Our work indicates that it is necessary to assess the quality of geolocation information coming from exhaustive tabulation, because it contains an implicit imprecision.

References

1. *GeoURL*, `http://www.geourl.org/`.
2. *Net World Map*, `http://www.networldmap.com/`.
3. MaxMind LLC, *GeoIP*, `http://www.maxmind.com/geoip/`.
4. C. Davis, P. Vixie, T. Goodwin, and I. Dickinson, "A means for expressing location information in the domain name system," *Internet RFC 1876*, Jan. 1996.
5. V. N. Padmanabhan and L. Subramanian, "An investigation of geographic mapping techniques for Internet hosts," in *SIGCOMM*, San Diego, CA, USA, Aug. 2001.
6. M. J. Freedman, E. Freudenthal, and D. Mazires, "Democratizing content publication with coral," in *Proc. of USENIX NSDI*, San Francisco, California, March 2004.
7. M. Freedman, M. Vutukuru, N. Feamster, and H. Balakrishnan, "Geographic locality of IP prefixes," in *Proc. ACM/SIGCOMM IMC*, Berkeley, CA, USA, Oct. 2005.
8. B. Gueye, A. Ziviani, M. Crovella, and S. Fdida, "Constraint-based geolocation of internet hosts," *IEEE/ACM Transactions on Networking*, to appear.