Determining the Geographic Location of Internet Hosts[*]

Venkata N. Padmanabhan[†]
Microsoft Research
padmanab@microsoft.com

Lakshminarayanan Subramanian[‡]
University of California at Berkeley
lakme@cs.berkeley.edu

ABSTRACT

We study the problem of determining the geographic location of an Internet host knowing only its IP address. We have developed three distinct techniques, *GeoTrack*, *GeoPing*, and *GeoCluster*, to address this problem. These techniques exploit information derived from the DNS, network delay measurements, and inter-domain routing. We have evaluated our techniques using extensive and varied datasets.

1. INTRODUCTION

In this paper, we study the problem of determining the geographic location of an Internet host knowing only its IP address. While an interesting problem in its own right, *location mapping* is key to enabling a large and interesting class of location-aware applications for Internet hosts. Examples of such applications include targeted advertising and territorial rights management.

Many of the current location mapping systems are based on the *Whois*[2] database. For each block of IP addresses, Whois typically records the name, address, and other information pertaining to the organization which registered the address block. Whois-based tools use the address information to infer the location corresponding to an IP address. The problem, however, is that for a large ISP (e.g., AT&T) or a geographically-dispersed organization (e.g., IBM), the location registered with Whois may correspond to the head office of the organization and may offer little indication of the actual location of a host with a specific IP address.

An alternative approach used in a few systems is based on performing a *traceroute*[3] to determine the network path from a probe machine to the target host. The DNS names of routers on the path often indicate location (e.g., *corerouter1.SanFrancisco.cw.net* indicates the city of San Francisco). Tools that exploit such information include VisualRoute and GTrace.

We have developed and evaluated three distinct techniques to address the location mapping problem. The first, *GeoTrack*, is a refinement of existing traceroute-based techniques. The other techniques, *GeoPing* and *GeoCluster*, employ novel alternative approaches, as we describe next. We refer to our location mapping system that incorporates these three techniques as *IP2Geo*.

2. IP2GEO

We now discuss each of the three techniques that comprise IP2Geo in some detail.

2.1 GeoTrack

GeoTrack uses the traceroute-based approach described in Section 1. It determines the network path from a probe machine to the target host and then tries to infer location from the DNS names of router interfaces (*router labels*). The location of the last router (i.e., one closest to the target host) with a recognizable label is used as an estimate of the location of the target host. GeoTrack incorporates several refinements of this basic approach, including:

- *Delay-based verification:* We measure the round-trip time (RTT) to each intermediate router along the path. If the difference in RTT between two adjacent routers is smaller than a threshold (5 ms by default), then we check to see if the corresponding location estimates are less than a threshold distance apart (250 km by default). If not, we mark the location estimates for these routers as suspect and ignore them.
- *Multi-source tracing:* We initiate traceroutes from multiple, geographically-dispersed sources. The location estimates for the target host obtained from all of these runs are aggregated (e.g., using simple majority polling) to obtain a consensus estimate. The consensus estimate tends to be more robust than any individual estimate because of the diversity of network paths (and so ISPs) traversed by each run of traceroute.

2.2 GeoPing

GeoPing exploits the relationship between network delay and geographic distance to estimate the location of a target host. Conventional wisdom in the networking community

[*] A full-length version of this paper is available as [5]

[†] http://www.research.microsoft.com/~padmanab/

[‡] http://www.cs.berkeley.edu/~lakme/

has suggested that there is poor correlation between network delay and geographic distance. However, the extensive data that we have gathered and analyzed suggests otherwise, presumably because our data corresponds to the U.S., which is "well-connected" with relatively few circuitous network routes. While the relationship between delay and distance is not perfect and is hard to characterize mathematically, there is a marked knee in the cumulative distribution function of distance for any (narrow) delay range. Furthermore, the distance corresponding to the knee increases steadily as we move to higher delay ranges.

Motivated by RADAR [1], GeoPing employs an empirical approach, termed *nearest neighbor in delay space (NNDS)*, to exploiting the relationship between delay and distance. As the first (offline) step, we construct a *delay map*, where each entry contains: (a) the coordinates of a host at a known location, and (b) a delay vector, $DV = (d_1, \ldots, d_N)$, containing the measured (minimum) delay to the host from N probes at known locations. Given a target host, T, we first construct a delay vector, $DV' = (d'_1, \ldots, d'_N)$, for it using the probes. We then search through the delay map to find a delay vector, DV, that best matches DV', where the quality of the match is quantified using the Euclidean distance between DV and DV'. The location corresponding to the best match yields GeoPing's estimate of the location of the target host T.

2.3 GeoCluster

GeoCluster groups together IP addresses that correspond to hosts likely to be co-located, i.e., form a *geographic cluster*. It then uses partial (and possibly inaccurate) IP-to-location mapping information (derived from sources such as a user registration database) to infer the likely location of the geographic cluster.

To identify geographic clusters, we start with the algorithm proposed in [4] for determining *topological* clusters. We use the address prefixes (APs) contained in BGP routing tables to define the initial set of geographic clusters. We then prune this list using a *sub-clustering* algorithm that tests to see if there is sufficient consensus in the location estimates corresponding to a cluster (based on the partial IP-to-location mapping information). If not, the algorithm sub-divides the cluster into two halves (yielding two more specific APs in place of the original one) and repeats the process.

Given a target IP address, GeoCluster first determines the geographic cluster to which it belongs (using longest prefix match) and then estimates its location to be that of the geographic cluster (assuming this information is available). GeoCluster uses the spread in the geographic locations corresponding to a cluster (i.e., the *dispersion*) to provide an indication of the accuracy of the location estimate. This is particularly useful in the context of clients that connect via proxies. GeoCluster would (correctly) refrain from making a location estimate for such clients whereas other techniques (including GeoTrack and GeoPing) would mistake the proxy's location for the client's.

3. EXPERIMENTAL RESULTS

We evaluated IP2Geo through extensive experiments. Here we present a sample of our results. We deployed a set of 14 probe machines at geographically dispersed locations in the U.S. These machines were used both to initiate traceroutes for GeoTrack and to measure delay for GeoPing. For GeoCluster, we obtained partial IP-to-location mapping information from several sources including a Web-based email site, a Web hosting site, and an online TV program guide.

Figure 1 plots the CDF of the error in the location estimate for the three techniques. The target hosts were chosen from servers at university campuses across the U.S. We observe that GeoCluster performs the best, with a median error of 28 km compared to 108 km and 382 km for GeoTrack and GeoPing, respectively.

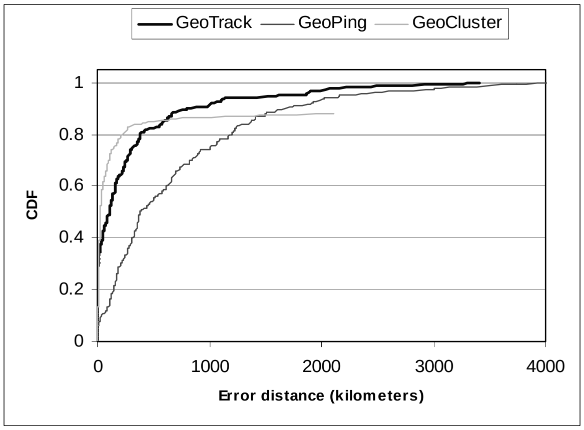

Figure 1: CDF of the error distance for GeoTrack, GeoPing, and GeoCluster.

4. SUMMARY AND CONTRIBUTIONS

We have developed a system called IP2Geo for determining the geographic location of an Internet host knowing only its IP address. IP2Geo incorporates three techniques, GeoTrack, GeoPing, and GeoCluster, that approach the problem from different directions. Our findings suggest that GeoCluster is the most promising one among these techniques and can often place a host to within a metropolitan area.

Acknowledgements

We are indebted to the many people who helped us obtain interesting datasets and access to probe machines. A more detailed acknowledgement is included in [5].

5. REFERENCES

[1] P. Bahl and V. N. Padmanabhan. RADAR: An In-Building RF-Based User Location and Tracking System. *IEEE INFOCOM*, March 2000.

[2] K. Harrenstien, M. Stahl, E. Feinler, NICKNAME/WHOIS, *RFC-954, IETF*, October 1985.

[3] V. Jacobson, Traceroute software, 1989, *ftp://ftp.ee.lbl.gov/traceroute.tar.Z*

[4] B. Krishnamurthy, J. Wang. On Network Aware Clustering of Web Clients. *ACM SIGCOMM 2000*, Stockholm, 2000.

[5] V. N. Padmanabhan and L. Subramanian. Determining the Geographic Location of Internet Hosts. *Microsoft Research Technical Report MSR-TR-2000-110*, November 2000.