IP Geolocation in Metropolitan Areas

Satinder Pal Singh
Bobby Bhattacharjee
Randolph Baden
Richard La
Choon Lee
Mark Shayman

University of Maryland,
College Park, 20742, USA

ABSTRACT

Current IP geoloation techniques can geolocate an IP address to a region approximately 700 square miles, roughly the size of a metropolitan area. We model geolocation as a pattern-recognition problem, and introduce techniques that geolocate addresses to within 5 miles inside a metropolitan area. We propose two complementary algorithms: The first algorithm, Pattern Based Geolocation (PBG), models the distribution of latencies to the target and compares it to those of the reference landmarks to resolve an address to within 5 miles in a metropolitan area. The second approach, Perturbation Augmented PBG (PAPBG), provides higher resolution by sending extra traffic in the network. While sending an aggregate of 600 Kbps extra traffic to 20 nodes for approximately 2 minutes, PAPBG geolocates addresses to within 3 miles.

Catetogies and Subject Descriptors C.2.3 [Computer-Communication Systems]: Network Operations — Public Networks

General Terms: Experimentation, Measurement

Keywords: Geolocation, Pattern Recognition, Probability Mass Function, Perturbation, Divergence

1. INTRODUCTION

IP Geolocation algorithms map IP addresses to geographic locations. Geolocation can be used for targeted advertising, efficient content distribution, location-specific content customization, and critical emergency services including E-911 for Voice-over-IP telephones [7, 3].

State-of-the-art IP geolocation techniques resolve addresses to approximately 30 miles [5, 8], roughly the diameter of a metropolitan area. In this paper, we present two new approaches for finer resolution IP Geolocation inside a metropolitan area. Our work departs from prior measurement-based geolocation approaches, all of which correlate latency with distance.

We model geolocation as a *pattern recognition problem*. Our algorithms identify and extract patterns from network statistics to geolocate an IP address. We propose a new Pattern Based Geolocation (PBG), which captures patterns in the distribution of latencies or Round Trip Times (RTTs) observed to a target. PBG models the signature of background traffic in the vicinity of the target and uses this 'signature' to geolocate the target to approximately 5 miles of its actual location. To further improve the resolution of PBG, we develop Perturbation Augumented PBG (PAPBG), which is inspired by Stochastic Resonance [1]. PAPBG sends a small amount of signal traffic in the network to enhance the signature of background traffic. At the cost sending an additional 600 Kbps aggregrate traffic to 20 nodes for approximately 2 minutes, PAPBG gives a higher resolution in the location estimate and geolocates the target to within 3 miles.

SIGMETRICS'11, June 7–11, 2011, San Jose, California, USA.
ACM 978-1-4503-0262-3/11/06.

2. OUR APPROACH

Our infrastructure consists of a collection of probe and landmark nodes. We administer three probe nodes in Maryland, USA: one in the city of College Park on Qwest network and one each in Silver Spring and Potomac on Verizon network. We have 20 landmarks distributed over 700 square miles large Washington D.C.-Baltimore metropolitan area on Comcast (12) and Verizon (8) networks. The mean pairwise distance between the landmarks is 8.4 miles for Comcast network and 11.8 miles for Verizon network.

Our techniques geolocate a target to the 'best matching' landmark in the testbed. Given a set of landmarks, the best possible estimate of a target's geographic location is the landmark which is geographically closest to it. Suppose s_{min} is the distance between the target and the geographically closest landmark. Let s^* be the distance between 'the best matching landmark' given by a geolocation algorithm and the target. Then the error of the location estimate of that geolocation algorithm is $\mathcal{E} = s^* - s_{min}$. Here $\mathcal{E} \geq 0$, with equality when the 'best matching landmark' given by the geolocation algorithm is in fact the geographically closest landmark.

Exisiting measurement based geolocation techniques assume correlation between distance and RTTs [5]. However, in a metropolitan area this correlation does not exist bacause propogation delay is a small component of RTTs, and the dominant component is queuing delay [2]. Our approach models IP geolocation as a pattern recognition problem and aims to geolocate a target by identifying, extracting and matching 'patterns from RTT sequences'.

2.1 Pattern Based Geolocation

Pattern Based Geolocation (PBG) uses the *distribution* of the RTT values as pattern for geolocation. First, we construct Probability Mass Functions (PMFs) from the col-

lected RTT sequences to model the distribution of RTTs using 'k Nearest Neighbor' density estimation method [4].

Next, we compare the PMFs of the landmarks to the PMF of the target to get the best match in shape. In our problem we encounter frequent cases where PMFs are similar in shape but shifted by a few milliseconds. To address this, we introduce a new distance metric called "Shifted Symmetrized Divergence" distance, (d_{SSD}), defined as:

$$d_{SSD}(p\|q) = a \times \min_s \big(d_{SD}(p\|q_s)\big) + (1-a) \times \phi(s_{min}) \quad (1)$$

Here

p, q	=	two PMFs
q_s	=	PMF q shifted by s
d_{SD}	=	Symmetric Kullback-Leibler Divergence [4]
s_{min}	=	$\arg\min_s \big(d_{SD}(p\|q_s)\big)$
ϕ	=	penalty function for shift
a	=	weight

Using d_{SSD} each probe node does PBG computations to obtain divergence values for each landmark. We finally output the landmark with the minimum mean divergence over all probe nodes as the target's location estimate. The two parameters involved in PBG computations, ϕ and a, are chosen empirically using a training dataset (See Section 3).

2.2 Perturbation Augmented PBG

PBG relies on the background traffic in the vicinity of a target. However, in some instances, the background traffic signature is not strong enough, and PBG fails to map the target to geographically close landmark. Perturbation Augmented PBG (PAPBG), inspired by Stochastic Resonance [1], *enhances* the background traffic signature by introducing a controlled amount of "perturbation" traffic into the network using a **perturber**.

The technique works as follows. One of the probe nodes, acting as perturber, sends large ICMP echo request packets (e.g. of size 100 bytes each) to all the landmarks and the target at a rate, say 50 packets per second. This corresponds to signal traffic of 40 Kbps to each landmark and target. The remaining probe nodes send regular small ICMP request packets (of size 30 bytes each) at a nominal rate of 5 packets per second for 100 seconds to measure the RTT sequences. These probe nodes then run PBG algorithms on the measured RTT sequences to give the best matching landmark. Thus, PAPBG is essentially PBG with an additional perturber which introduces a controlled amount of perturbation traffic in the network for better differentiation of PMFs.

3. EXPERIMENTS AND RESULTS

We first collected 30 training data sets to empirically choose the 'best values' for ϕ and a. Each dataset consists of synchronous RTT sequences collected from the two probe nodes at College Park and Siver Spring over 20 landmarks in our testbed. We collected RTT squences at a rate of 5 samples per second for 100 seconds from each landmark per probe node. We explored three penalty functions, Logarithmic, Linear and Exponential, and 100 values of $a \in [0, 1]$. For each combination of the two parameter values we used a leave-one-out [6] approach to run PBG on the 30 training data sets. The best performance (minimum mean geolocation error) was obtained for exponential penalty function ($\phi(s_{min}) = 2^{s_{min}}$), and $a = 0.9$ for Comcast network and $a = 0.95$ for Verizon network.

To evaluate the performance of PBG we collected 50 additional datasets with the same setup as mentioned above. Using the same leave-one-out approach and parameter values discussed above, we ran PBG computations on this data. To compare the performance of PBG we used an existing measurement based geolocation technique, Constraint Based Geolocation (CBG) [5], to geolocate targets on these datasets as well.

The mean error obtained with CBG was 15.39 miles for Comcast network and 18.06 miles for Verizon network, which is worse than the mean pairwise distance between the landmarks on the two networks. Compared to this, our PBG gives a mean error of 2.13 miles for Comcast network and 4.34 miles for Verizon network. Further, on an average PBG matches the target to the geographically closest landmark in majority of the cases (>50%). Note that if we randomly select one of the landmarks as target's location estimate, the mean error obtained is 7.62 miles for Comcast network and 8.76 miles for Verizon network. Thus, existing techniques perform worse than 'random selection', while PBG geolocates the target to within $2-4$ miles of its actual location.

For PAPBG we collected additional datasets for 5 perturbation intensites: 10, 20, 30, 40 and 50 Kbps per destination node. We used the probe node at Potomac as perturber and collected 50 datasets for each intensity using the other two probe nodes. We achieved the best performance for perturbation intensity of 30 Kbps; the mean error reduces to 1.2 miles for Comcast network and 3.4 miles for Verizon network. Beyond 30 Kbps we enter a region of diminishing returns and no more gains in performance are obtained. Thus by sending an additional 30 Kbps traffic to each of 20 nodes for 100 seconds, PAPBG reduces geolocation error by approximately $20-40\%$.

4. REFERENCES

[1] R. Benzi, A. Sutera, and A. Vulpiani. The mechanism of stochastic resonance. *Journal of Physics A: Mathematical and General*, 14(11):L453, 1981.

[2] C. J. Bovy, H. T. Mertodimedjo, G. Hooghiemstra, H. Uijterwaal, and P. Mieghem. Analysis of end-to-end delay measurements in Internet. In *Proc. Passive and Active Measurement Workshop (PAM 2002)*, Fort Collins, CO, USA, 2002.

[3] D. D. Clark, C. Partridge, R. T. Braden, B. Davie, S. Floyd, V. Jacobson, D. Katabi, G. Minshall, K. K. Ramakrishnan, T. Roscoe, I. Stoica, J. Wroclawski, and L. Zhang. Making the world (of communications) a different place. *SIGCOMM Comput. Commun. Rev.*, 35(3):91–96, 2005.

[4] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2000.

[5] B. Gueye, A. Ziviani, M. Crovella, and S. Fdida. Constraint-based geolocation of Internet hosts. *IEEE/ACM Transactions on Networking*, 14(6):1219–1232, Dec. 2006.

[6] R. Kohavi. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. pages 1137–1143. Morgan Kaufmann, 1995.

[7] S. Steiniger, M. Neun, and A. Edwardes. Foundations of Location Based Services. *Lecture Notes on LBS*, 2006.

[8] B. Wong, I. Stoyanov, and E. G. 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. USENIX Association.