SLAC-PUB-16463

Adaptive Geolocation of Internet Hosts

Raja A.A. Khan
National University of Sciences and Technology, Pakistan
11mseerakhan@seecs.edu.pk

Anjum Naveed
Faculty of CS and IT
University of Malaya, Malaysia
anjum@um.edu.my

R. Les Cottrell
SLAC National Accelerator Laboratory, USA
cottrell@slac.stanford.edu

***Abstract*—IP based geolocation is a widely used geolocation technique because of its ability to geolocate the hosts where GPS or other techniques become ineffective or unavailable. Measurement-based geolocation techniques utilize landmarks to make end-to-end delay measurements and compute the host location based on delay to distance mappings. Fewer landmarks and/or inaccurate delay to distance mapping leads to large error margins. In this research an Adaptive-IP-Geolocation (AIG) technique is proposed. AIG is based on PingER and PerfSonar worldwide deployments. Based on the analysis of PingER data, AIG uses two tier approach where tier one landmarks identify the region of a target host. This is followed by geolocation of the target host using regional landmarks only. A variable alpha is introduced for delay to distance conversion. Results show that AIG outperforms previous techniques with the error margin reduced to 25 km or less for the majority of the hosts in the tested region.**

I. Introduction

Numerous applications rely on location of the Internet hosts either for proper functionality or for improved performance. Examples for such location enabled applications include: automatic language selection for displaying content, region specific advertising and content delivery restrictions, localized delivery of news, location aware fraud detection etc. GPS [1] is frequently used to acquire geographic location of the hosts. However, GPS signals are not always available, specifically, in buildings, in tunnels and under the densely shaded areas. All popular applications including the widely used GoogleMaps and Maps in general use IP geolocation in one form or the other.

The process of finding the geographic location of an internet host from the IP address is termed as IP Geolocation. There are several research challenges for measurement based geolocation. First of all, the mapping between delay and distance is not straight-forward and varies from one geographic region to another depending on network traffic conditions, types of links and route directness. Secondly, measurement based techniques are significantly affected by the number of landmarks and their relative distance from the target host. In general, fewer landmarks that are also distant from the target host result in higher error.

This research is based on inferences drawn from real data collected using the worldwide PingER and PerfSonar infrastructures. The infrastructures consist of approximately 300 measurement points that perform delay measurements to approximately 950 active and passive hosts. The geographic location of all nodes is known, providing a huge data set of ground truth from approximately 285000 host pairs. To the best of our knowledge, no other research uses such an intensive infrastructure of real Internet hosts. Analysis of the collected data and the inferences on data have lead us to propose improvements on delay to distance mapping. This research contributes to IP based geolocation on three aspects.

Firstly, the observation that a distant landmark results in a poor location estimate, is used and the landmarks as well as targets are grouped into geographical regions. Secondly, a variable α is introduced for delay to distance mapping. α can be seen as a measure of route directness we refer to as the $Directivity$. Given that different regions have different network connectivity, a localized value of α is used for each region. Thirdly, the value of alpha to be used for the delay to distance mapping is chosen at the run time depending upon the value of delay. Results are presented for one region, showing significant improvement in accuracy of geolocation.

The paper is organized as follows. Section II reviews the related work. Section III provides the details of PingER and PerfSonar infrastructure and the relevant details of collected data. The proposed geolocation technique AIG is explained in Section IV. Empirical evaluation of AIG is carried out in Section V. Section VI concludes the paper.

II. Geolocation of Internet Hosts

This section is divided into two categories of non-measurement based and measurement Based techniques. The shortcomings of these techniques are also highlighted.

A. Non-measurement based techniques

Use of additional DNS records can be a possible technique to infer geographic locations of Internet hosts as proposed by C. Davis [2]. The adoption of this method is not very practical since it requires continuous changes in the records that have to be done manually. Certain tools like Cello [3], netgeo [4] and WBG [5] obtain the location information by querying the Whois databases. The information in these databases, however, may be erroneous or out of date. There are geolocation services that are based on comprehensive tabulation between range of IP addresses and their geographic locations; amongst these are projects GeoURL [6], Networld [7] and commercial services Geonetmap [8], Geopoint [9]. The problem with this approach is the difficultly to manage the tabulation and to keep it up to date. In practice some of these do very well for the majority of end hosts. However they typically fail badly for

This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, under Contract No. DE-AC02-76SF00515.

routers usually identifying them at the parent organization site. They also often fail for web sites that have a proxy at another location. Thus having an independent mechanism (such as a measurement based approach) for validating the location is very important.

GeoCluster [10] divides the IP range into clusters in such a manner that all hosts within a cluster are likely to share similar location. This technique finds the location of the complete cluster from the location of a few hosts. However, the technique relies on incomplete information that is most likely also inaccurate. On the similar lines Prieditis et al. [11] have proposed the use of machine learning for IP geolocation of routers without using landmarks.

B. Measurement Based Techniques

Shortest Ping (SPing) [10] is one of the first techniques that makes use of Internet delay measurements from landmarks to locate a host. An improved version of Shortest Ping is GeoPing [10], which improves accuracy by introducing passive landmarks. These landmarks are hosts with known geographic location that can only respond to pings. Both SPing and GeoPing use the coordinates of the closest landmark as the estimated location of the target. GeoGet [12] is a recently proposed scheme on the similar lines. The results can be highly inaccurate given the fact that the closest landmark may still be at a considerable distance from the target. Techniques such as Virtual Landmarks [13], GNP [14] and Vivaldi [15] have been proposed using coordinate systems to address the estimation of network closeness between Internet hosts. However, distance in such problems refers to the network delay between Internet hosts and the coordinates assigned do not represent the geographic location, whereas for geolocation, the actual geographic coordinates are desired.

Gueye et al. [16] proposed an approach for IP geolocation named as Constraint-based Geolocation (CBG). CBG uses multilateration to estimate the geographic location of Internet hosts. However, there is no definite correlation between geographic distance and delay. This is because the network delay in the Internet is affected by many factors, such as issues of triangle inequality [17], the absence of great-circle paths and queuing delays [18]. Landa et al. [19] have computed the expected errors for the metrics that are estimated from delay measurements using experiments. Given these facts, CBG results in a large error when the target is far from landmarks. Eriksson et al. [20] have proposed improvement on CBG by redefining the possible location of target within the constrained region. Their approach eliminates the non-populous regions and improves accuracy. Spotter [21] uses the ground truth of 23,000 nodes to derive probabilistic properties of landmark delay to distance mappings. The authors have claimed improved accuracy compared to CBG.

Katz-Bassett et al. [22] proposed Topology Based Geolocation (TBG) in 2006. TBG is claimed to be an improvement over CBG in terms of accuracy since it forms and uses topology constraints in addition to the delay constraints of CBG. However, TBG achieves accurate geolocation only if the traceroutes provide sufficient structural constraints on the target. This is not always the case specifically when most of the routers are configured not to respond to pings. Wang et al. [23] have extended TBG in an effort to reduce overhead. Similar to TBG, Wong et al. [24] proposed Octant, which also uses negative constraints.

Fig. 1: Landmarks in Europe

III. Data set: PingER, PerfSonar and PlanetLab Infrastructure

For this research, we have used the world-wide deployment of PingER, PerfSonar and PlanetLab network monitoring nodes. There are 6 PingER, 52 PlanetLab, 31 PerfSONAR nodes in North America. Similarly, Europe has 4 PingER, 40 PlanetLab, 14 PerfSONAR nodes while Pakistan contains 30 PingER nodes. The monitoring node deployment in Europe is shown in Figure **??**. Every monitoring node sends 10 ping messages to every other node in the network at an interval of 30 minutes. The response is collected by the monitoring node. The collected data is transferred to a central location once every day. The following inferences have been drawn from the data.

(i) With reference to geolocation, empirically an RTT value less than 6ms is not useful as it indicates the same location for the two hosts. Similarly, values larger than 80 ms are not useful for geolocation and result in increased error. (ii) RTT does not go beyond a certain value, irrespective of the geographic location of the two hosts. Such measurements are not used for the computation of α or for dynamic geolocation. (iii) A larger value of RTT not only indicates a distant node but also may indicate a more indirect route and vice versa. This information can be used to dynamically adjust the delay to distance mapping. (iv) For every value of RTT, absolute maximum and absolute minimum distances can be associated with the value. The range of possible distances is bounded by the two values using minimum and maximum α value in the region.

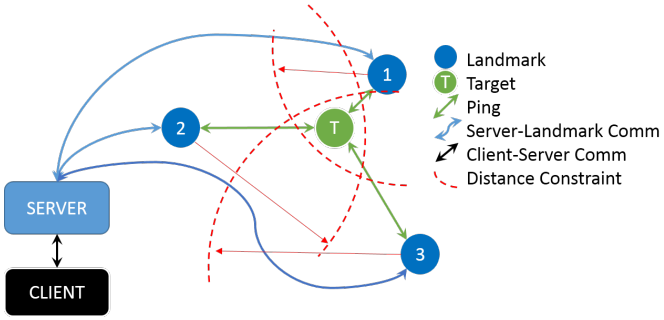

Fig. 2: Geolocation using Multilateration

IV. Adaptive IP Geolocation

This section explains the Adaptive IP Geolocation in detail. The section starts with the description of multilateration, which is the basic technique employed by AIG. The use of two distance based constraints is also explained. This is followed by the explanation of the three major contributions of this research. Finally, the implementation details of the proposed technique are given.

A. Two constrained multilateration

Multilateration is the technique of geolocating the Internet hosts using delay measurements. The technique measures the round trip propagation delays from a set of hosts, termed landmarks, to the target host. Landmarks are special servers with known geographic locations that can ping a target host upon request. The delay measurements are converted to maximum distance estimates of the target from specific landmarks using the well known constant of the speed of light in the fiber (2/3 times the speed of light). The estimated distances, also known as distance constraints, are used to draw virtual circular disks around the landmarks. The area of intersection of all virtual disks is considered as the area of interest where the target can possibly be located as shown in figure 2. Ideally, the area of interest should be a single point, which should be the location of the target. However, this is never the case. Therefore, the centroid of the area of interest is estimated as the location of the target.

The size of the area of interest is proportional to the margin of error in estimated location and the actual location. To reduce the size of the area of interest, an additional distance constraint is used. For every value of RTT, a minimum possible distance can be associated such that for the given RTT, smaller distance than the minimum distance is not possible. Therefore, two concentric virtual circles can be drawn around each landmark with the area of interest being area of intersection of all landmarks and lying only between two circles of each landmark. This is shown in figure 3.

B. World Regions

AIG divides the globe into multiple regions, as shown in figure 4. The regions have been created based on the global Internet connectivity. Geolocation of a host is a two phase process. The first phase identifies the region of the target while the second phase estimates the location of the target within the region. The landmarks are divided into two sets. Two landmarks near the geographic center of each region are selected as tier1 landmarks. These landmarks are used in first phase for region identification of the target. The region of the landmark returning least RTT value is selected as the region of the target. All landmarks including tier1 landmarks are considered tier2 landmarks. The two tier approach has also been used by Gueye et al. [25] for GeoPing; however, worldwide division of regions has not been done before.

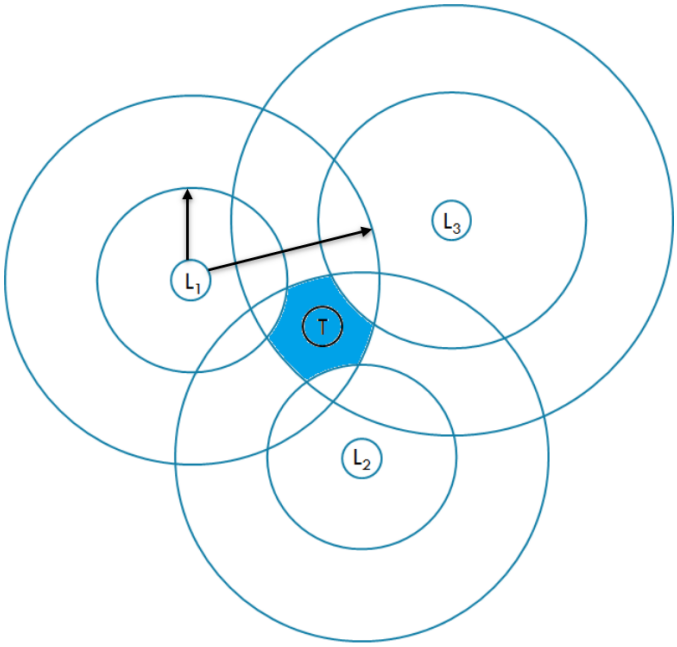

Fig. 3: AIG Technique with 3 landmarks

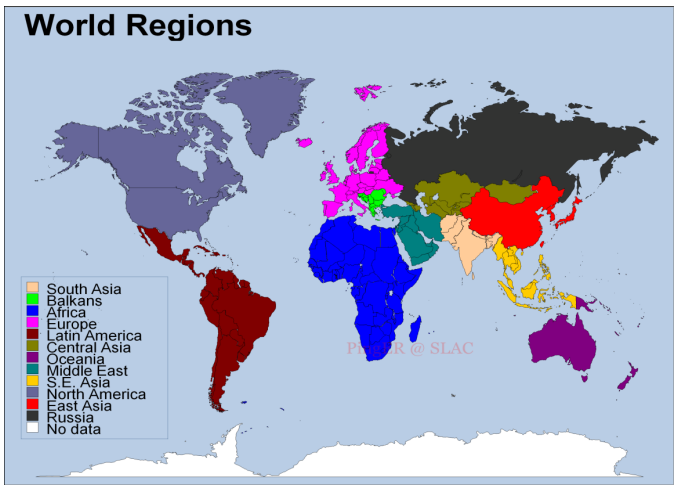

Fig. 4: AIG geographic regions of the World

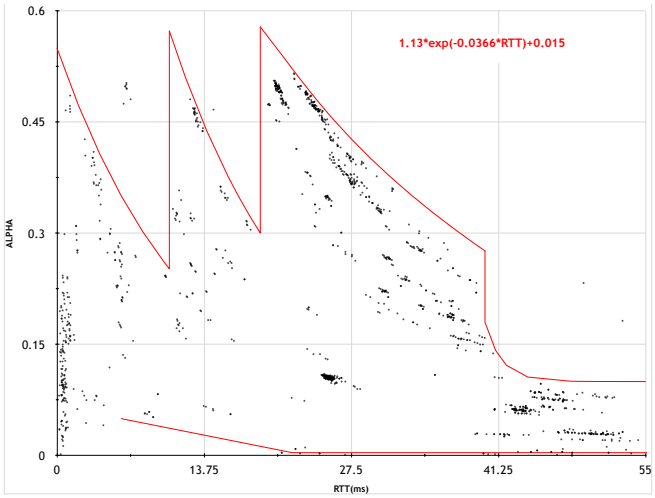

Fig. 5: Alpha Analysis for Pakistan

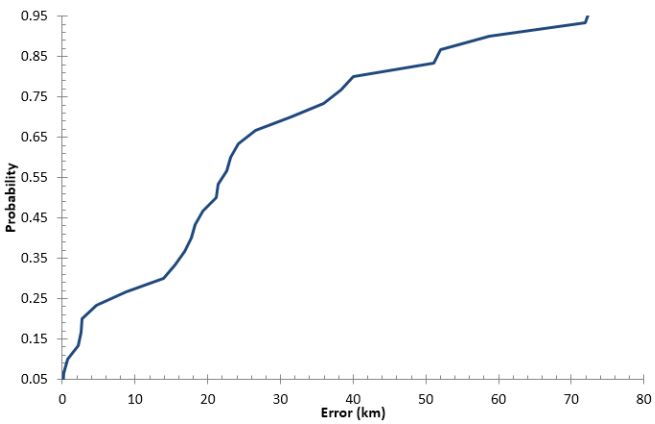

Fig. 6: North America CDF

C. Effective Speed of Light

The speed of signal propagation in optical fibre is 2/3 times the speed of light $\approx 200km/ms$. Since the round trip time (RTT) is approximately double the one way network delay, the maximum distance D_{max} between any two hosts for a given RTT is:

$$D_{max}(km) = 100(km/ms) \times RTT(ms)$$

The actual distance D however, can be anything between $0 - D_{max}$ since there is hardly ever a direct optical fiber connection between hosts going along a great circle path. Furthermore, packets experience queuing and processing delays due to intermediate routers in the path, affecting the value of RTT. Therefore, first AIG adjusts the RTT by subtracting a fixed minimum region dependent queuing and processing delay. Secondly, the effective speed of light is considered. α is defined as a relative measure of speed required to translate RTT measurement into the great circle path distance.

$$D(km) = \alpha \times 100km/ms \times RTT(ms) \quad (1)$$

where $0 \leq \alpha \leq 1$. A larger value of α indicates arelatively direct path while smaller value indicates more indirect route. Hence we refer to α as the $Directivity$. PingER data has been used to derive α values.

D. Adaptive α

Proper selection of the α value can significantly improve the accuracy of delay to distance mapping which in turn can reduce the error margin of geolocation of targets. However, given that the RTT is largely dependent on network topology, the value of α will vary for every source destination pair. In order to find the optimum value of α, the real PingER data is considered where values of RTT between landmarks are available. Figures 5 shows the scatter plot of the alpha values plotted against the RTT values for the landmarks in Pakistan.

The exponentially decreasing trend in the values of α is very obvious in the plot. The maximum value of α can be represented as piecewise exponentially decreasing function as shown in the plot. Similar results have been achieved for other regions. The analysis for other regions is skipped for the save of brevity.

V. Results and Discussion

In this section we look at the results of geolocation using AIG for North America and Pakistan. We also present a discussion on the significance of the convex hull and the area of intersection.

A. North America

The CDF plot of geolocation errors using AIG for North America region are shown in Figure 6. The graph clearly shows that AIG can locate 65% of the targets with an error margin of less than 25 km. We have compared AIG with SOI and CBG. The maximum error in case of AIG (73 km) is significantly lower than CBG (708 km) and SOI (801 km). The detailed results are skipped for the sake of brevity.

B. Pakistan

The CDF plots of geolocation errors using AIG for Pakistan region are shown in Figure 7. Similar to the North Mmerican region, AIG can predict the location of the targets with significant accuracy. 65% of the targets have an error margin of less than 8 km while a maximum error for the reported targets was less than 15 km. The high accuracy of AIG in the Pakistan region is attributed to the number of landmarks available. We observe that the error margin significantly decreases as the landmark density is decreased. The results presented use the landmark density of 18.75 landmarks per million square kilometers. For half the landmark density, a decrease of approximately 20% targets is observed for the reported error of 8 km or less. This indicates that on one side AIG performs effective geolocation while at the same time it indicates that appropriate number of landmarks can significantly improve the results.

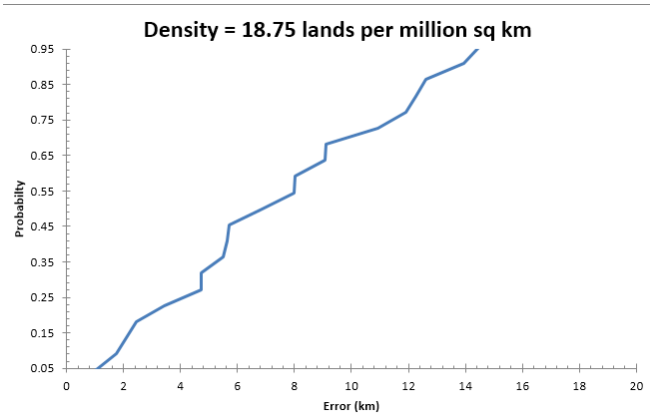

Fig. 7: Pakistan CDF

C. Convex-Hull

The lower constraint of AIG enables the technique to provide acceptable results even when the target is outside the convex-hull of landmarks. In such scenarios, AIG shows profound improvement over other measurement based geolocation approaches. In order to better illustrate this scenario, one of the test cases is shown in figure 8. The actual location of the target is shown with the green marker, AIG's intersection region and estimate is shown in red and CBG in shown in blue. The lower constraint of the landmark in Salt Lake City made the AIG intersection region much smaller than that of CBG. The geolocation error using AIG is 22.3km where as for CBG, the error is 356.2km.

D. Area of Intersection

The area of the intersection region provides a measure of confidence and accuracy of the geolocation. This can enable location-aware applications to judge if the result is suitable/accurate enough for use. The CDF of the intersection areas of AIG and CBG are shown in Figure 9. It clearly shows that AIG is capable of providing much smaller intersection regions, this is due to the additional lower constraint and also due to the adaptive nature of alpha selection.

VI. Conclusion

We have proposed an adaptive internet geolocation technique. The technique uses adaptive and variable α values depending on the region and RTT value of measurements. The evaluation shows the AIG achieved less than 8 km of error for 65% targets in Pakistan while achieving a less than 25 km error for the same percentage of hosts in North America. Comparison with the state of the art indicates that AIG outperforms existing techniques by a huge margin. In future, we plan to extend the technique by using dynamic α values that are adjusted during measurements. We further plan to conduct a detailed study of the impact of the number of landmarks on geolocation. We will compute the sufficient number of landmarks required for effective geolocation.

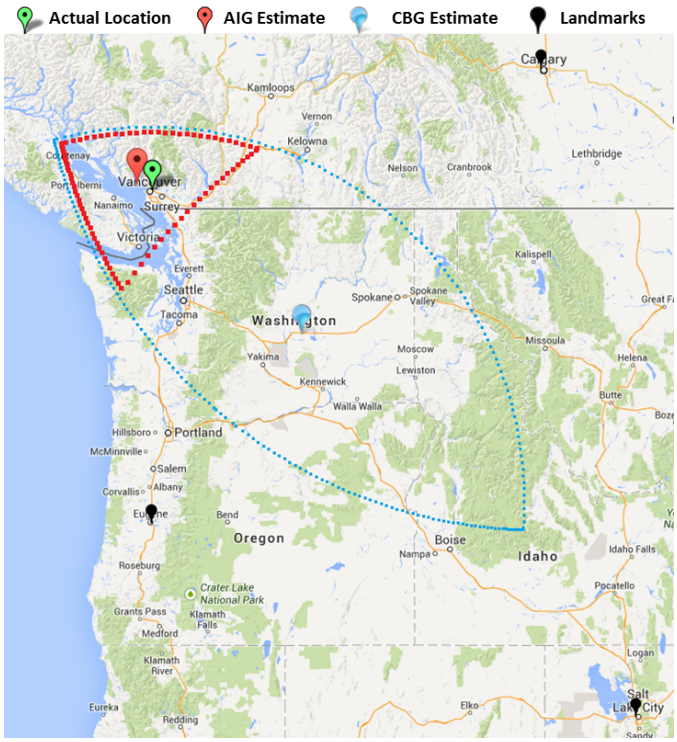

Fig. 8: Example of target outside convex-hull

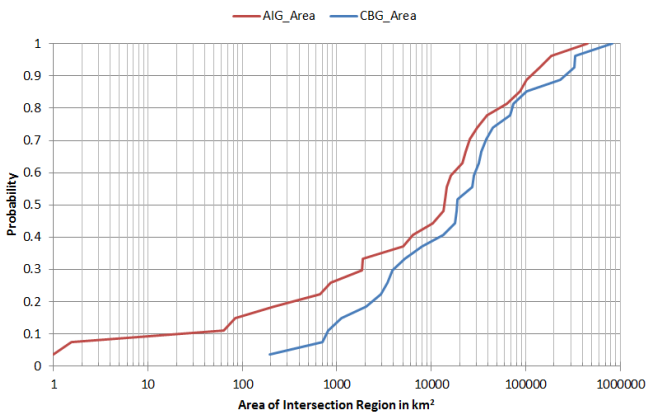

Fig. 9: CDF of Intersection Area

References

[1] P. Enge and P. Misra, "Special issue on global positioning system," in *Proc. IEEE, vol. 87, no. 1*, Jan 1999, pp. 3–5.

[2] 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.

[3] Ip address to latitude/longitude. univ. illinois, urbana-champaign. [Online]. Available: http://cello.cs.uiuc.edu/cgi-bin/slamm/ip2ll/

[4] D. Moore, R. Periakaruppan, J. Donohoe, and K. Claffy, "Where in the world is netgeo.caida.org?" in *Proc. INET 2000 Conf., Yokohama, Japan*, Jul 2000.

[5] P. Endo and D. Sadok, "Whois based geolocation: A strategy to geolocate internet hosts," in *Advanced Information Networking and Applications (AINA), 2010 24th IEEE International Conference on*, April 2010, pp. 408–413.

[6] Geourl. [Online]. Available: http://www.geourl.org/

[7] Net world map. [Online]. Available: http://www.networldmap.com/

[8] Geonetmap. geobytes, inc. [Online]. Available: http://www.geobytes.com/GeoNetMap.htm

[9] Geopoint. quova inc. [Online]. Available: http://www.quova.com/

[10] V. N. Padmanabhan and L. Subramanian, "An investigation of geographic mapping techniques for internet hosts," in *Proc. ACM SIGCOMM, San Diego, CA*, Aug 2001, pp. 173–185.

[11] A. Prieditis and G. Chen, "Mapping the internet: Geolocating routers by using machine learning," in *Computing for Geospatial Research and Application (COM.Geo), 2013 Fourth International Conference on*, July 2013, pp. 101–105.

[12] D. Li, J. Chen, C. Guo, Y. Liu, J. Zhang, Z. Zhang, and Y. Zhang, "Ip-geolocation mapping for moderately connected internet regions," *Parallel and Distributed Systems, IEEE Transactions on*, vol. 24, no. 2, pp. 381–391, Feb 2013.

[13] L. Tang and M. Crovella, "Virtual landmarks for the internet," in *Proc. ACM Internet Measurement Conf. 2003,Miami, FL*, Oct 2003, pp. 143–152.

[14] T. S. E. Ng and H. Zhang, "Predicting internet network distance with coordinates-based approaches," in *Proc. IEEE INFOCOM, New York*, Jun 2002, pp. 170–179.

[15] F. Dabek, R. Cox, F. Kaashoek, and R. Morris, "Vivaldi: A decentralized network coordinate system," in *Proc. ACMSIGCOMM 2004, Portland, OR*, Aug 2004, pp. 15–26.

[16] B. Gueye, A. Ziviani, M. Crovella, and S. Fdida, "Constraint-based geolocation of internet hosts," *IEEE/ACM Trans. Networking, vol. 14, no. 6*, pp. 1219–1232, 2006.

[17] S. Banerjee, T. G. Griffin, and M. Pias, "The interdomain connectivity of planetlab nodes," in *Proc. Passive and Active Measurement Workshop (PAM 2004), Antibes Juan-les-Pins, France*, Apr 2004.

[18] L. Subramanian, V. N. Padmanabhan, and R. Katz, "Geographic properties of internet routing," in *Proc. USENIX 2002, Monterey, CA*, Jun 2002, pp. 243–259.

[19] R. Landa, R. Clegg, J. Araujo, E. Mykoniati, D. Griffin, and M. Rio, "Measuring the relationships between internet geography and rtt," in *Computer Communications and Networks (ICCCN), 2013 22nd International Conference on*, July 2013, pp. 1–7.

[20] B. Eriksson, P. Barford, B. Maggs, and R. Nowak, "Posit: A lightweight approach for ip geolocation," *SIGMETRICS Perform. Eval. Rev.*, vol. 40, no. 2, pp. 2–11, Oct. 2012. [Online]. Available: http://doi.acm.org/10.1145/2381056.2381058

[21] S. Laki, P. Matray, P. Haga, T. Sebok, I. Csabai, and G. Vattay, "Spotter: A model based active geolocation service," in *INFOCOM, 2011 Proceedings IEEE*, April 2011, pp. 3173–3181.

[22] E. Katz-Bassett, J. P. John, A. Krishnamurthy, D. Wetherall, T. Anderson, and Y. Chawathe, "Towards ip geolocation using delay and topology measurements," in *Internet Measurement Conference (IMC2006)*, 2006.

[23] Y. Wang, D. Burgener, M. Flores, A. Kuzmanovic, C. Huang, "Towards street-level client-independent ip geolocation," in *In Proc. USENIX NSDI*, 2011.

[24] Bernard Wong, Ivan Stoyanov and Emin Gn Sirer, "Octant: A comprehensive framework for the geolocalization of internet hosts," in *Proc. Symposium on Networked System Design and Implementation, Cambridge, Massachusetts*, Apr 2007, pp. 411–414. [Online]. Available: http://www.cs.cornell.edu/~bwong/octant/

[25] B. Gueye, A. Ziviani, S. Fdida, J. F. de Rezende, O. C. M. B. Duarte, "Two-tier geographic location of internet hosts," in *In Proc. 7th IEEE International Conference on High Speed Networks and Multimedia Communications (HSNMC)*, Jul 2004.