ISSN:0975-9646

Review of Different IP Geolocation Methods and Concepts

Jayaprabha Bendale , Prof. J. Ratanaraj Kumar

G.S.Moze College of Engineering, Balewadi, Pune-45.
University Of Pune, Pune, India.

Abstract: Most of the methods used for IP geolocation are taking the advantages from the geolocation and hence this resulted into efficient way to the many end users. The examples of such approach is the applications in which online content access constrained to the particular geographic area as well as cloud computing because some companies must ensure their virtual machines stay in an appropriate geographic region. In this review paper, we present the study over basics of geolocation, different IP geolocation based techniques with their detailed methodology of working, disadvantages of existing techniques of IP geolocation. The methods, which we have reviewed in this paper, are Geo-Track, Geo Cluster, and GeoPing. These methods are used for finding out the exact of geographic location of the internet hosts.

Keywords: IP, geolocation, geographic location, Geo Track, GeoPing, Geo Cluster.

I. INTRODUCTION

Since from last decade, the problem of IP location mapping is gaining the more research interest due its wide range of applications now days. This problem is one of the challenge research domains for researchers. This kind of services allows big and interesting class of location-aware applications for Internet hosts. For example GPS system which is used now days very commonly in mobile handsets. Another application called web service which is used to send the information related to the local events, regional weather etc based on the user location. This application uses the prior information of user location. This application also classifies the users depending current location of users. Every application is having its own needs over the required location information resolution.

There are many real time security sensitive applications those are based on use of Geolocation. The real time applications of providing the online contents like Hulu [13], Real Media [22], Pandora [20], and BBC iPlayer [22] etc. are having limits on their content sharing to specific geographical areas because of security reasons. Thses applications first identify the location of end user, then based on that location information such applications takes the decision whether to allow user to view the content or not. Before allowing a client to view the content, they determine the client's location from its IP address and allow access only if the client is in a permitted jurisdiction. In addition to this, the internet application websites restrict their access to various applications depending on the end users risk legal repercussions or location. As per this, such commercial application depends on geolocation in order to limit their online services to end users.

As we know that IP geolocation is an active area of research since from the last decade, but all the existing geolocation methods considering a benign target which is not trying to intentionally misguide the user, as well as there has been very less work done over geolocating malicious targets.

In [12], the author Caste lluccia et al. applies Constraint-Based Geolocation (CBG) to the problem of geolocating fast-flux hidden servers that use a layer of proxies in a botnet [5] to conceal their location. In [18], Muir and Oorschot presents the limitations of passive geolocation techniques (e.g., whois services) and present a technique for finding the IP address of a machine using the Tor anonymization network [28].

During this survey paper, we are taking the review of different methods presented for IP location mapping problem by various researchers. These methods having their different characteristics and observations over the Internet like hierarchical addressing as well as correlation between delay and distance. For practical analysis of such methods, there are different kinds of datasets available publically by researchers. In below section II we are first taking the review of geolocation and its different algorithms are described. In section III we are presenting the related work over the IP geolocation, different methods, and their shortcoming are presented. Finally, conclusion is made based on above discussions.

II. SURVEY OF GEOLOCATION

2.1 Background Geolocation

IP Geolocation IP address to a given geographic location to solve the problem of determining the. Solution to varying degrees of granularity can be expressed; for most applications, the IP is located in the city, the result is either a city or longitude and latitude where the goal is to determine accurate returning to situated. Geolocation of the two main approaches to the host or IP of the database location mapping to use either active network measurement Measurement-based geolocation algorithms [1, 2, 3, 4, 5, and 6] leverage a set of geographically distributed landmark hosts with known locations to locate the tar-get IP. These landmarks measure various network properties, such as delay, and the paths taken by traffic between themselves and the target. These results

are used as input to the geolocation algorithm, which uses them to determine the target's location using methods such as: constraining the region where the target may be located (geolocalization) [1, 5]; iterative force directed algorithms [6], machine learning [1] and constrained optimization [2].
Geolocation algorithms mainly rely on ping [7] and trace route [7] measurements. Ping measures the round-trip time (RTT) delay between two machines on the Internet, while trace route discovers and measures the RTT to routers along the path to a given destination. We classify measurement-based geolocation algorithms by the type of measurements they use to deter-mine the target's location. We refer to algorithms that use end-to-end RTTs as delay-based [8, 11, 6] and those that use both RTT and topology information as topology-aware algorithms [3, 5].
An alternative measurement-based Geolocation IP Geolocation database space-using mapping to these databases can be either proprietary or public. Public database contains regional Internet Registry (for example, ARIN [3], [23] RIPE) administered by Quova. [6] and Max-[4] owns IP database companies such as geographic location mappings to provide the exact method to build while these databases is not They sometimes whois services, the DNS loc records and autonomous system (AS) [2] is based on a combination of numbers. Do Registries and databases to be coarse grained, usually returning the IP address of the Head Office of the registered organization with Geolocation Database miss leading location.

Table 1: Average accuracy of measurement-based geolocation algorithms.

Class	Algorithm	Average accuracy (km)
Delay-based	GeoPing [19]	150 km (25th percentile); 109 km (median) [30]
	CBG [12]	78-182
	Statistical [31]	92
	Learning-based [9]	407-449 (113 km less than CBG [12] on their data)
Topology-aware	TBG [14]	194
	Octant [30]	35-40 (median)
Other	Geo Track [19]	156 km (median) [30]

DNS loc [8] is an open standard that DNS administrators, DNS server IP location information location information to create a publicly available database allows to effectively increase. However, it has gained widespread use. Since loc DNS database are not authenticated and the content of the IP addresses are set by the owners themselves it is poorly suited for security-sensitive applications. More research to improve the accuracy of measurement-based geolocation algorithms has gone to; As a result, they provide reliable results. Table 1 recently reported Pro-generate geolocation algorithms shows average accuracies. Reported based on AC-curacies, we believe that the current geolocation algorithms a machine to a country or place within the jurisdiction are adequately accurate. In particular, CBG [11] and Octant [5] appear to offer accuracies well within the size of most countries and may even be able to place users within a metropolitan area. Measurement-based geoloca-tion is particularly appealing for secure geolocation be-cause if a measurement can reach the target (e.g., using application layer measurements [10]), even if it is behind a proxy (e.g., SOCKS or HTTP proxy), the effectiveness of proxying will be diminished.

2.2 Delay Based Geolocation Methods

Delay-based geolocation algorithms target end-to-end network IP geolocate to use delay measurements delay-based geolocation to execute, geographical distance and delay network connection between the need to calibrate it every historical sites, Lee, all other sites is done by having Ping. Since destinations known geographic locations, To delay a network, dij, geographical distance, gij, mapping function can then derive a celebrated landmark Lj where I 6 = j [11]. Each milestone performs this calibration and network delay geographic distance develops its own mapping. After calibrating its distance-to-delay function, it then pings the target IP. Using the distance-to-delay function, the landmark can then transform the ob-served delay to the target into a predicted distance to the target. All landmarks perform this computation to triangulate the location of the target.
Delay-based geolocation is the underlying assumption that delay network operates well under the geographical distance is correlated with. However the network is made up of Queuing delay, processing, [14] de-transmission and propagation traveled where only network traffic to propagation time is related to the distance to and other components. Adding to the noise measured delay network load Vary, depending on network traffic when the host a direct ("as the crow flies") does not take the path that perception is also violated. These indirect paths are referred to as "circuitous" routes [5]. There are many proposed methods for delay-based geolocation, including GeoPing [4], Statistical Geoloca-tion [6], Learning-based Geolocation [9] and CBG [12].
These algorithms differ in how they express the distance-to-delay function and how they triangulate the position of the target. GeoPing is based on the observation that hosts that are geographically close to each other will have de-lay properties similar to the landmark nodes [4]. Statistical Geolocation develops a joint probability density function of distance to delay that is input into a force-directed algorithm used to geolocate the target [6]. In contrast, Learning-based Geolocation utilizes a Native Bayes framework to geolocate a target IP given a set of measurements [9]. CBG has the highest reported accuracy of the delay-based algorithms, with a mean error of 78-182 km [12]. The remainder of this section therefore focuses on CBG to model and evaluate how an adversary can influence delay-based geolocation techniques.
CBG [12] establishes the distance-delay function, de-scribed above, by having the landmarks ping each other to derive a set of points (g_{ij}, dij) mapping geographic distance to network delay. To mitigate the effects of congestion on network

delays, multiple measurements are made and the landmarks to calibrate their distance-to-delay mapping use the 2.5-percentile of network delays. Each milestone that is closest to, but not below, a linear ("best line") function computes the set points. The distance between each historical and target IP "best line" function is inferred by each landmark is an implied circle around where the tar-get IP can be located. Intersection of circle of target IP all destinations to be in lIe predict the results of this process since. A viable area where the target can be located, the CBG determines the centroid of the area and as a result assumes Geolocation. A mean error 182 km in America and in Europe 78 km of they also find that practical target IP 105 km2 is located in North America, Europe could be where area 104 km2.

2.3 Topology Aware Geolocation Methods

Delay-based geolocation correlating measured distances between destinations depending on the delay with. We saw previously, these correlations or mapping to create overlapping confidence regions landmark-to-apply to target delay; Practical overlap in the area, and its predicted target centroid. when the difference-landmark and landmark-to-target delay similarly (e.g., end of winding paths) are correlated with physical distance due to the distance from the resulting delay targeted relationships can significantly deviate from earlier computing correlations. Topology-aware geolocation addresses this problem by limiting the impact of circuitous end-to-end paths; specifically, it localizes all intermediate routers in addition to the target node, which results in a better estimate of delays. Starting from the landmarks, the geolocation algorithm iteratively estimates the location of all intermediate routers on the path between the landmark and the target. This is done solely based on single-hop link delays, which are usually significantly less circuitous than multi-hop end-to-end paths, enabling topology-aware geolocation to be more resilient to circuitous network paths than delay-based geolocation.

There are two first proposed topology-aware topology-based geolocation methods, (TBG) [3] and Octant [5]. These methods differ in how they geolocate intermediate routers. On the contrary, the Octant "geolocalization" CBG [11], where intermediate routers and target spaces of their delayed depending on the sites and other intermediate routers are constrained to specific areas of the framework leverages TBG and the goal for the intermediate routers IP [3] [5]. These delays are mapped into distances using a convex hull rather than a linear function, such as the best line in CBG to improve the mapping between distance and delay.

Octant other geolocation algorithms improve performance on many optimization advantages. These include taking into account both positive and negative obstacles; Network path with fixed delay accounting for, and barriers based on latency measurement weight decreasing. Wong et al. Their plan CBG, 35-40 km [30] search outperforms with accuracies average. In addition, viable returned by Octant returned from those very small are CBG. They further their plan even after performing 15 places to flatten a small number of destinations with strong oversight.

When analyzing and evaluating attacks on topology-aware geolocation, we consider a generic geolocation framework. Intermediate routers are localized using constraints generated from latencies to adjacent routers. The target is localized to a feasibility region generated based on latencies from the last hop(s) before the target, and the centroid of the region is returned.

III. Literature Review Of Existing Methods

There has been much work on the problem of locating hosts in wireless environments. The most well known among these is the Global Positioning System (GPS) [5]. However, GPS is in active indoors. There have been several systems targeted specifically at indoor environments, including Active Badge [8], Active Bat [9], and RADAR [1]. As we discuss later, our GeoPing technique uses a variant of RADAR's NNSS algorithm. However, in general these techniques are specific to wireless networks and do not readily extend to the Internet.

In the Internet context, an approach that has been used to determine user/host location is to seek the users input (e.g. cookies, registering on a service etc.). However, such approaches are likely to be (a) burdensome on the user, (b) ineffective if the user uses a client other that the one where the cookie is stored, and (c) prone to errors due to the (possibly deliberate) inaccuracies in the location information provided by an individual user.

There are several ways of building an IP address to geo-graphic location mapping service [34]. Many existing approaches and proposals for solving the problem can be broadly classified into the following categories:

1. Incorporating location information (e.g. latitude/longitude) is in the Domain name Service.
2. Using the Whois [7] database to determine the location of the organization to which an IP address was allocated.
3. Using the trace route tool and mapping the router labels in the path to geographic locations.
4. Leveraging an existing and extensive content distribution network which have a wide enough reach to build an exhaustive tabulation of IP address ranges and their corresponding location. Examples of such an approach are Akamai's Edges cape [14] and Digital Island's trace-ware [6]. Since the algorithms employed by these ser-vices are proprietary, it is hard for us to compare them to our research e ort.

2.1 Existing Approaches and Their Shortcomings

The DNS-based approach was proposed in RFC 1876 [14]. This work defines the format of a new Resource Record (RR) for the DNS and reserves a corresponding DNS type mnemonic (LOC) and numerical code (4). The DNS-based approach faces deployment problems since it requires a modification of the record structure of the DNS records. It also poses a burden on the administrators with

the task of entering the LOC records and there is no easy way of verifying whether the location entered is correct and trustworthy.

An approach used widely in many tools is to query Whois servers [7]. Tools such as IP2LL, Allwhois, and Net-Geo [13] use the location information recorded in the Whois database to infer the geographic location of a host. A list of publicly accessible Whois servers is available at [4].

The main problems with Whois based approaches are:

1. Unreliable Database: The domain name maintainers do not insist on keeping the database accurate and current.
2. One record for Large Chunks: A large block of IP addresses may be allocated to a single entity and there will be only one entry in the Whois database for that whole chunk. For example, the 8.0.0.0/8 IP address block is allocated to BBN Planet and a query to ARIN Whois database returns the location as Cambridge.
3. Web Hosting and Domain Name Transfers: Many web sites may be registered to one location but hosted in a different location Domain name transfers are not always reflected in the Whois database. Several commercial products use trace route [10] as the basic tool for tracing the geographic path to a given IP ad-dress. These products include Visual Route, Neotrace, GeoBoy, What Route and Gtrace. The basic idea in any trace route-based tool is to perform a trace route from a source to the target IP address and map the router labels (i.e., the DNS names associated with a router's network interfaces) along the path to their geographic locations using airport codes, city codes and country codes.

Some of the problems of trace route-based approaches are:

1. Unavailability of Router Labels: There are many routers, which do not have router labels making it impossible to decipher their location.
2. Multiple Naming Conventions: Each ISP uses its own naming conventions for labeling their routers, which makes the problem of translating router labels to geography challenging.
3. Cities with same name: There are many instances where multiple cities have the same name (Eg. 21 different cities named Bloomington in the US). Given a code for one such city name, it may be hard to associate it with one of the many cities with that name.

2.2 Fundamental Limitation due to Proxies

Many Web clients are behind proxies (or rewalls). So the \client" IP address seen by the external network would actually correspond to the proxy, (e.g., a caching proxy on a university campus). An example of the latter is the AOL network [13], which has clients all over the U.S., but has a centralized cluster of proxies at one location (Virginia). Figure 1 shows the distribution of the distance between the AOL proxies and clients.

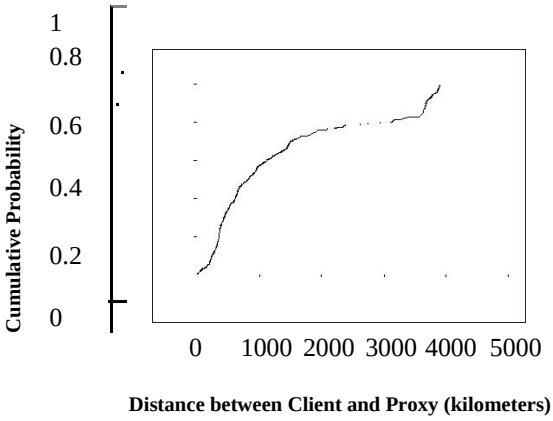

Figure 1: Distribution of distance between AOL proxies and clients.

Proxies impose a fundamental limitation of all location mapping approaches that depend on client IP address. This includes approaches based on Whois, trace route (e.g., GeoTrack), and network delay measurements (e.g., GeoPing). Not only are these schemes unable to determine the true location of a client, they are also oblivious to the error. Our Geo Cluster technique is an exception in that it is able to automatically tell when its location estimate is likely to be erroneous. We discuss this is more detail in Section 6.2.

IV. Conclusion And Future Work

This review paper discussing the one of interesting research problem of finding the geographical location of end user based on their IP address. The survey over the geolocation is presented in details with their different methods we have studied the different geolocation methods and point their limitations for future work in this research domain. We have defined the IP geolocation problem with their different algorithms presented so far such as delay based geolocation algorithms and topology aware geolocation algorithms. The future work we suggest to work on improvement to the existing geolocation methods in order to overcome their limitations.

References

1. ERIKSSON, B., BARFORD, P., SOMMERS, J., AND NOWAK, R. A learning based approach for IP geolocation. In Proceedings of the Passive and Active Measurement Workshop (April 2010).
2. [12] GUEYE, B., ZIVIANI, A., CROVELLA, M., AND FDIDA, S. Constraint based geolocation of Internet hosts. IEEE/ACM Transactions on Networking 14, 6 (December 2006).
3. [14] KATZ-BASSET, E., JOHN, J., KRISHNAMURTHY, A., WETHERALL, D., ANDERSON, T., AND CHAWATHE, Y. towards IP geolocation using delay and topology measurements. In Proceedings of the ACM SIGCOMM Internet Measurement Conference (October 2006).
4. Pandora Internet radio, 2010. http://www.pandora.com.
5. WONG, B., STOYANOV, I., AND SIRER, and E. G. Octant: A comprehensive framework for the geolocalization of Internet hosts. In of the fourth Symposium on Networked Systems Design and Implementation (NSDI) (Cambridge, MA, April 2007).

6. YOUNG, I., MARK, B., AND RICHARDS, D. Statistical geolocation of Internet hosts In Proceedings of the 18th International Conference on Computer Communications and Networks (Au-gust 2009).
7. CROVELLA, M., AND KRISHNAMURTHY, B. Internet Measurement: Infrastructure, Traffic and Applications.
8. ERIKSSON, B., BARFORD, P., SOMMERS, J., AND NOWAK, R. A learning based approach for IP geolocation In Proceedings of the Passive and Active Measurement Workshop (April 2010).
9. GARFINKEL, T., PFAFF, B., CHOW, J., ROSENBLUM, M., AND BONEH, D. Terra: A virtual machine-based platform for trusted computing. In Proceedings of the 19th ACM Symposium on Operating Systems Principles (SOSP) (October 2003).
10. GILL, P., ARLITT, M., LI, Z., AND MAHANTI, A the flattening Internet topology: Natural evolution, unsightly barnacles or contrived collapse? In Proceedings of the Passive and Active Measurement Workshop (April 2008).
11. [12] GUEYE, B., ZIVIANI, A, CROVELLA, M., AND FDIDA, S. Constraint-based geolocation of Internet hosts. IEEE/ACM Transactions on Networking 14, 6 (December 2006).
12. Hulu - watch your favorites. Anytime for free 2010. Http: //www.hulu.com/.
13. KATZ-BASSET, E., JOHN, J., KRISHNAMURTHY, A., WETHERALL, D., ANDERSON, T., AND CHAWATHE, Y. towards IP geolocation using delay and topology measurements. In Proceedings of the ACM SIGCOMM Internet Measurement Conference (October 2006).
14. KUROSE, J., AND ROSS, K. Computer networking a top down approach featuring the Internet.
15. Maxine geolocation and online fraud prevention, 2010. http://www.maxmind.com.