Research Article

An Optimization Method for the Geolocation Databases of Internet Hosts Based on Machine Learning

Ting Wang, Ke Xu, Junde Song, and Meina Song

PCN&CAD Center, Beijing University of Posts and Telecommunications, Beijing 100876, China

Correspondence should be addressed to Ting Wang; wangting2031@126.com

Received 26 January 2015; Revised 3 April 2015; Accepted 4 May 2015

Academic Editor: Haranath Kar

Copyright © 2015 Ting Wang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In order to improve the accuracy and robustness of geolocation (geographic location) databases, a method based on machine learning called GeoCop (Geolocation Cop) is proposed for optimizing the geolocation databases of Internet hosts. In addition to network measurement, which is always used by the existing geolocation methods, our geolocation model for Internet hosts is also derived by both routing policy and machine learning. After optimization with the GeoCop method, the geolocation databases of Internet hosts are less prone to imperfect measurement and irregular routing. In addition to three frequently used geolocation databases (IP138, QQWry, and IPcn), we obtain two other geolocation databases by implementing two well-known geolocation methods (the constraint-based geolocation method and the topology-based geolocation method) for constructing the optimized objects. Finally, we give a comprehensive analysis on the performance of our method. On one hand, we use typical benchmarks to compare the performance of these databases after optimization; on the other hand, we also perform statistical tests to display the improvement of the GeoCop method. As presented in the comparison tables, the GeoCop method not only achieves improved performance in both accuracy and robustness but also enjoys less measurements and calculation overheads.

1. Introduction

With rapid development of cloud computing and cloud storage, the cloud is becoming a popular medium for storing and computing data. As data is stored on Virtual Machines, which are deployed on a cloud provider's infrastructure, cloud users give up direct control of their data in exchange for faster on-demand resources and shared administrative costs. Typically, they specify the QoS (Quality of Service) requirements of their outsourced data in a SLA (Service Level Agreement), including not only common QoS requirements such as delay but also location restrictions which define the regional access control about the cloud resource [1]. The aspects referred to above are all related with the geographic locations of Internet hosts. To fulfil the demand of the SLA, cloud providers should choose the proper servers for reducing the delay before an Internet host gets the required cloud resource or for making a decision about whether this Internet host can get access to the cloud resource according to his geographic location [2]. The granularity of geographic location used above is always coarse-grained, such as country, province, or city.

Cloud providers always establish a geolocation database for collecting the mapping relationships between IP addresses of Internet hosts and their geographic locations. The sources of geolocation information in the database are obtained either by implementing existing geolocation methods or directly from the public IP geolocation databases [3]. There occur some problems which resulted from the geolocation methods and the database itself for affecting the accuracy of the geolocation database of Internet hosts. The geolocation methods can be divided into two different types depending on their underlying methodologies for collecting geolocation information: registration-based geolocation methods and measurement-based geolocation methods [4]. The former set of methods use previously registered data for gaining the information on the geolocation locations of respective IP addresses. In general, these methods provide accurate location information. However, in some cases, their errors are large enough for entire blocks of IP addresses owing to the fact that their precision greatly depends on the resolution and reliability of the previously registered data they utilize [5]. The latter ones utilize active delay and topology measurements

to overcome the aforementioned limitations; but because of queuing delays and circuitous routes, the additive noise produces some inherent inaccuracy and unpredictability into the measurements [6–9]. Moreover, owing to the large number of available Internet hosts and the change in IP assignment, it is hard for the geolocation database to maintain and continuously update the geolocation information of Internet hosts. Because of the drawbacks of geolocation methods and the database itself, the accuracy of geolocation databases still remains limited [10].

This research aims to improve the accuracy and robustness of the geolocation databases of Internet hosts and to analyze the properties of routing policy in China from a view point of geographic location. In this paper, we propose a method based on machine learning (GeoCop) for optimizing the geolocation databases of Internet hosts, which combines network measurement, machine learning method, and routing policy for deriving the geolocation model of Internet hosts. The proposed GeoCop method makes the geolocation results less prone to imperfect measurements and irregular routing. To demonstrate the accuracy and the robustness of the GeoCop method, we not only compare the performance of the geolocation databases after optimization but also perform statistical tests to display the improvement of the GeoCop method. It is presented in the evaluation that the proposed method is effective. The rest of this paper is organized as follows. Section 2 summarizes previous work on measurement-based geolocation methods and analyzes the problems of these methods. Section 3 briefly evaluates the existing geolocation databases of Internet hosts. In Section 4, the GeoCop method is described in detail. The experiments and results are presented in Section 5 and Section 6 includes the conclusions of the paper.

2. Related Work

Measurement-based geolocation methods always leverage a set of geographically distributed landmarks with known geographic locations for geolocating the targets. These landmarks make use of network measurement tool for obtaining various network properties, such as Internet delay and topology information [11, 12]. Then, we classify the existing measurement-based methods into two types depending on the network properties they used: delay-based geolocation methods and topology-based geolocation methods.

2.1. Delay-Based Geolocation Methods. Most delay-based geolocation methods geolocate the target by exploiting the relationship between Internet delay and geographic distance and are differentiated only by the way they express the distance to delay function and triangulate the geographic location of the target [13]. IP2Geo [14] is included among the first for suggesting a delay-based approach to approximate the geographic distance of Internet hosts. Youn et al. [15] presented a statistical geolocation scheme. They apply kernel density estimation to delay measurements among a set of landmarks and estimate the target location by maximizing the likelihood of the distances from the target to the landmarks.

Maziku et al. [16] proposed an Enhanced Learning Classifier approach for estimating the geolocation of Internet hosts with increased accuracy. They reduced average error distance in the geolocation of Internet hosts by extracting six features from network measurements. Arif et al. [17] used bivariate kernel density estimation for approximating joint probability distributions of the distance and delay. Eriksson et al. [18] reduced IP geolocation to a machine learning classification problem and used a Naive Bayes framework for increasing geolocation accuracy.

Gueye et al. [19] proposed a more mature approach called CBG (Constraint-based Geolocation), which used several delay constraints for inferring the geographic location of an Internet host by a triangulation-like method. For each landmark, they used distance-to-delay relationships between landmarks for deriving a maximum distance bound for a given delay from this landmark to the target. They drew a circle centered at this landmark based on the distance bound. Then, the intersection of the circles derived from all of the landmarks formed a convex region. It was assumed that the target resided in the convex region and the centroid of this convex region was the target location.

2.2. Topology-Based Geolocation Methods. In the geolocation for the target, the topology-based geolocation methods also leverage the network topology in addition to the relationship between Internet delay and geographic distance. Laki et al. [13] increased geolocation accuracy by decomposing the overall path-wise packet delay to link-wise components. Guo et al. [20] used web mining together with network measurement to geolocate IP address with significantly better accuracy. Tian et al. [21] performed a large-scale topology mapping and geolocation study for China's Internet. They developed a heuristic approach for clustering the interfaces in a hierarchical ISP (Internet Service Provider) and applied it to the hierarchical structure of the major ISPs in China's Internet. Shavitt and Zilberman [22] introduced a novel approach for generating POP (Point of Presence) level map and then combined the geolocation information of all the IP addresses in a POP from the geolocation database for assigning geographic locations to the POP. Biswajit et al. [23] proposed a classification-based method to improve the accuracy of the geolocation for the datacenters in commercial cloud providers. Wong et al. [24] presented a novel geolocation framework called Octant for the geolocation of Internet hosts, which gained its accuracy and precision by taking advantage of both positive and negative constraints.

Katz-Bassett et al. [25] presented a topology-based geolocation method called TBG (Topology-based Geolocation) for estimating the geographic location of arbitrary Internet host. They converted the data of Internet delay along with the information of network topology into a set of constraints for the geolocation. They first obtained the maximum distance bound based on the maximum transmission speed of packets in fiber and then further refined the region using interrouter delays along the path from the landmarks to the target. At last, they got the geographic location of the target through a global optimization that minimized average location errors for the target and the routers.

FIGURE 1: Geolocation misleading behavior.

Owing to the utilization of the relationship between Internet delay and geographic distance, both of the delay-based methods and topology-based methods may not really work in the two conditions: (1) the delay from the target to the landmark deviates much from the normal value; (2) one malicious Internet host, known as adversary, tries to disguise his geographic location by tempering with the delay measurements. Then, the geographic location of the Internet host appears to be wrong, which is presented as an instance in Figure 1.

For simplicity, we assume that there are only three landmarks. The delay from each of the landmarks to the target is denoted by t_1 , t_2 , and t_3 , and the black arcs C_1 , C_2 , and C_3 are the circles drawn by these landmarks while geolocating the target. The region enclosed by the arcs is the feasible region of the target location, and the geographic location of the centroid is the geolocation result for the target (black dot). When the delays t_1 and t_2 are increased to $t_1 + \Delta t_1$ and $t_2 + \Delta t_2$, respectively, the circles C_1 and C_2 change into C'_1 and C'_2 that are presented by the red dotted lines. Then the centroid of the enclosed region changes into a wrong location, so the target appears at a wrong location (red dot). Consequently, more accurate and robust geolocation estimation requires further improvement for the existing geolocation methods to offset network measurement errors.

3. Geolocation Databases

In this section, we briefly evaluate the geolocation databases of Internet hosts, which are currently available for cloud providers. Owing to the limitation in the number of ground truth nodes, we also use cross validation for evaluating the accuracy of these geolocation databases in addition to the normally ground truth nodes-based validation. We consider two kinds of geolocation databases in this study. The first kind is three existing geolocation databases (IP138, QQWry, and IPcn) [21], which are well known in the Chinese

TABLE 1: Accuracy rates for the ground truth nodes.

	IP138	QQWry	IPcn	CBG	TBG
Province	0.801	0.789	0.795	0.832	0.856
City	0.731	0.704	0.712	0.795	0.802

Internet community. The geographic locations returned by these databases generally have two granularities: the province granularity and the city granularity. The second kind is two geolocation databases contained by implementing two existing geolocation methods (CBG and TBG). We first randomly select a set of 3×10^4 IP addresses S_x from peers collected by crawling the Xun Lei DHT [26], which is a popular P2P download acceleration application in China. Then, we take advantage of CBG and TBG to geolocate these IP addresses and convert the form of geolocation results from longitude and latitude to province and city. At last, these geolocation results are collected for constructing the TBG geolocation database and the CBG geolocation database. We define $S_{\text{province}} = \{\text{province}_1, \dots, \text{province}_k\}$ and $S_{\text{city}} = \{\text{city}_1, \dots, \text{city}_k\}$ as the set of all the geographic locations in the granularities of province and city, respectively, and they are collectively referred to as S_{loc} .

3.1. Ground Truth Nodes-Based Validation. We leverage the numerous IDCs (Internet Datacenters) located in many cities of China for collecting a set of 1500 nodes S_g and use these nodes as the ground truth nodes for evaluating the accuracy of the geolocation databases. We define the accuracy rate a_g as the fraction of the cases for which the database provides the correct geographic location information of the ground truth nodes:

$$a_g = \frac{t_{\text{right}}^i}{t}, \quad (1)$$

where t_{right} is the number of ground truth nodes whose geolocation information in the database is the same as the actual geographic locations, and t is the number of the ground truth nodes. Table 1 presents the accuracy rates for the five databases at the province and city granularities. As observed from Table 1, all the accuracy rates are not high enough, and the accuracy rates in the granularity of city are lower than those in the granularity of province.

3.2. Cross Validation. Considering the limitation of the ground truth nodes, we also use cross validation to complement the evaluation for the accuracy of these geolocation databases. For an Internet host, if the geographic locations from the five databases are the same, it is most likely that the geographic location is correct; else, we have a low level of confidence on the geolocation information. We define the coverage rate a_{re} as another criterion for the accuracy of the geolocation databases, which is the fraction of the cases for which different databases have the same geolocation information for the IP addresses in S_x . The

TABLE 2: Coverage rates for different numbers of comparison databases.

Comparison number	2	3	4	5
Province	0.755	0.701	0.692	0.653
City	0.650	0.632	0.601	0.599

higher the coverage rate is, the more accuracy the databases have:

$$a_{re} = \frac{\sum_1^{C_n^x} r_x}{C_n^x}, \quad (2)$$

$$r_x = \frac{t_{repeat}}{t},$$

where x is the number of chosen comparison databases, n is the number of the total comparison databases, t_{repeat} is the number of IPs that have the same geographic locations in different comparison databases, and t is the number of all the IP addresses in S_x . Table 2 presents the coverage rates for different numbers of comparison databases. As observed from the table, all the coverage rates are not high enough, and the coverage rates in the granularity of city are lower than those in the granularity of province.

From the evaluation for the geolocation databases of Internet hosts, we observe that there is still a lot of room for these databases to improve. The goal of this paper is to develop a method for optimizing the geolocation databases of Internet hosts in order to improve the accuracy and robustness of the geolocation for China's Internet hosts. The following section explains the process of the GeoCop method in details.

4. The GeoCop Method

In this section, we propose the GeoCop method that utilizes machine learning method in the network measurement data for optimizing the geolocation databases of Internet hosts. Section 4.1 describes the process of the collection required for network measurement data. Section 4.2 describes the definitions of two new network measurement metrics. Section 4.3 describes the analysis of the network measurement metrics. Section 4.4 describes the data processing for the network measurement metrics of the router IPs in the network measurement data. Section 4.5 describes the geolocation model for edge routers. Section 4.5 describes the geolocation model for Internet hosts.

4.1. Network Measurement Data Collection. To generate the set of network measurement data, we perform traceroute measurements in China's Internet with a number of PlanetLab nodes in China. PlanetLab is a scalable and universal network measurement platform, which consists of 1315 nodes at 629 sites around the globe. The process of traceroute measurements is described as follows.

(1) Pick n effective Internet hosts with known geographic locations in the geolocation database which needs to be optimized. As the targets in the traceroute measurements, the

FIGURE 2: The relationship between the number of PlanetLab nodes and the increment of routers.

FIGURE 3: The distribution of landmarks (PlanetLab nodes) in use.

Internet hosts need to be chosen evenly throughout every city in China. In this paper, $n = 1, 141, 815$, and the targets are uniformly distributed over 34 provinces and 595 cities in China.

(2) Select k effective PlanetLab nodes to be landmarks according to the relationship between the number of PlanetLab nodes and the increment of routers, which is presented in Figure 2.

The straight line defined as a least square linear fit presents a positive correlation between itself and the observed values, with the absolute value of the ACC (Accuracy Correlation Coefficient) being approximately 0.79. ACC denotes the fitting degree of fitted curve and observed values. It is observed that the increment of IPs Δn roughly follows a linear distribution presented by the equation shown as follows:

$$\Delta n = -3051.8295 \times k + 75136.5048. \quad (3)$$

We consider the theoretical value of x when the increment is zero as the number of PlanetLab nodes in use, which is $75136.5048/3051.8295 = 25$. The distribution of the used PlanetLab nodes is illustrated in Figure 3.

(3) Send traceroute requests from each landmark to all of the targets for $a = 10$ times; this will result in a set of $k \times n \times a$ traceroute measurements.

4.2. The Definitions of Network Measurement Metrics. Listing all the interfaces along the routing path from the landmark to the target, traceroute is used to learn the routing path between two devices in the Internet. According to the statistics in both the dataset of traceroute measurements and the geographic locations of the targets, we introduce two new definitions on network measurement metrics called geographic location degree d_y and synchronization frequency $t_{IP_y}^{loc_x}$.

Definition 1. A measurement unit m includes all the interfaces in the traceroute measurements from a landmark to a target except the landmark and the target. Each measurement unit corresponds to a destination location loc_x , which is the geographic location of the target:

$$m = \left\{ \begin{array}{c} IP_{h1}^1, IP_{h1}^2, \dots, IP_{h1}^n \\ IP_{h2}^1, IP_{h2}^2, \dots, IP_{h2}^n \\ \vdots \\ IP_{hn}^1, IP_{hn}^2, \dots, IP_{hn}^n \end{array} \right\} \rightarrow loc_x, \quad (4)$$

where IP_{hn}^n denotes one of the measured interfaces on the n th hop of the routing path.

Definition 2. Geographic location degree d_y denotes the total number of different geographic locations which are corresponded to all of the measurement units including IP_y :

$$d_y = \sum_{loc_x \in S_{loc}} \delta_{IP_y}^{loc_x}, \quad (5)$$

where S_{loc} denotes a set of some certain geographic locations. If there is a measurement unit including IP_y and corresponding to loc_x at the same time, then the value of $\delta_{IP_y}^{loc_x}$ is one; otherwise, it is zero. d_y is classified into two categories: geographic location degree in the level of province d_y^p and geolocation location degree in the level of city d_y^c depending on the granularities of geographic location.

Definition 3. Synchronization frequency $t_{IP_y}^{loc_x}$ denotes the total number of measurement units, which meet the two conditions: (a) including IP address IP_y and (b) the geographic location of the target is loc_x :

$$t_{IP_y}^{loc_x} = \sum_{m \in S_m} \delta, \quad \delta = \begin{cases} 1, & (IP_y \in m) \cap (loc_x \in m) \\ 0, & \text{otherwise,} \end{cases} \quad (6)$$

where S_m denotes the set of $k \times n$ measurement units, constructed in Section 4.1.

Definition 4. $T_{IP_y}^{S_{loc}}$ denotes the synchronization frequency vector of IP_y corresponding to a set of geographic locations S_{loc} :

$$T_{IP_y}^{S_{loc}} = T_{IP_y} \times e_{S_{loc}}, \quad T_{IP_y} = [t_{IP_y}^{loc_1}, \dots, t_{IP_y}^{loc_k}], \quad (7)$$

where T_{IP_y} denotes the synchronization frequency vector of IP_y , including all the synchronization frequencies related with IP_y , $e_{S_{loc}}$ is a vector with all zeros except the entries whose corresponding geographic locations belong to the set of S_{loc} , which are ones, and S_{loc} denotes a set of geographic locations meeting both the condition $t_{IP_y}^{loc_x} > 0$ and the location restrict in the SLA. Figure 4 presents an instance for the construction of $T_{IP_y}^{loc}$.

4.3. The Analysis of Network Measurement Metrics. The geographic locations of the targets in the analysis of network measurement metrics are initialized by the geolocation results of the targets in the existing geolocation database; but on account of the wrong geolocation results in the existing geolocation databases, the actual statistics of network measurement metrics must have certain disparity with the theoretical values. In this section, we analyze the impact of the inaccuracy of geolocation databases on the analysis of network measurement metrics.

4.3.1. The Theoretical Comparison between Theoretical and Practical Values. Taking the routing policy of China Internet into consideration, we categorize a typical routing path into four sections, which are presented in Figure 5(a): the routers along the subpath from the landmark to the first backbone router (edge routers 1), the routers along the subpath from the first backbone router to the core backbone router (backbone routers 1), the routers along the subpath from the core backbone router to the last backbone router (backbone routers 2), and the routers along the subpath from the last backbone router to the target (edge routers 2). Theoretically, if $k \times n$ is sufficiently large, the corresponding statistics of d_y and $t_{IP_y}^{loc_x}$ associated with routers on different subpaths of the routing path lead us to draw useful conclusions, as presented in Figure 5(b). In the first two sections, the value of d_y is the same as the number of all the different geographic locations where the targets are located in China, and the distribution of $t_{IP_y}^{loc_x}$ is approximately uniform. In the third section, the value of d_y is the same as the number of the geographic locations which are corresponding to the measurement units including the backbone routers 2, and the distribution of $t_{IP_y}^{loc_x}$ is approximately uniform. In the last section, the value of d_y is one, and the values of $t_{IP_y}^{loc_x}$ are zero except only one geographic location, which is one; but on account of the wrong geolocation results in the existing geolocation databases and the accuracy rates of these databases which are always above 50%, the actual statistics of d_y and $t_{IP_y}^{loc_x}$ must have certain disparity with the theoretical values, which is presented in Figure 5(c).

FIGURE 4: An instance for the construction of T_{IP}^{loc} .

FIGURE 5: The theoretical comparison of two network measurement metrics.

4.3.2. The Statistical Analysis of Practical Values. Figure 6 presents the distributions of geographic location degrees d_y^p and d_y^c for the router IPs in the whole network measurement dataset. As the figure reveals, the distributions of all the geolocation databases have the same trend. Both d_y^p and d_y^c have fat-tailed distributions. The values of d_y^p and d_y^c for most of router IPs fall into the value of one and decrease rapidly with the increase of d_y^p and d_y^c .

Compared with regular graphs, the distributions of d_y^p and d_y^c are seriously distorted. The maximum metric and the

minimum metric, in general, cannot describe the properties of the network measurement metrics correctly. For one of the distributions in Figure 6(a) and the corresponding distribution in Figure 6(b), we take the logarithm of the values on x -axis and y -axis and then do piecewise least square fitting for them. The results are presented in Figure 7.

For these router IPs meeting the condition $d_y = 1$, the distribution of the synchronization frequencies is a straight line which is perpendicular to x -axis. So we only analyze the distributions of the synchronization frequencies for the left router IPs in the network measurement dataset, which

FIGURE 6: The distributions of geographic location degree.

FIGURE 7: The logarithm distributions of geographic location degree.

is presented in Figure 8. Figures 8(a) and 8(b) present the distributions in the granularity of province and city, respectively. x -axis denotes the sequence of geographic locations while sorting them according to the values of $t_{IP_y}^{loc_x}$. y -axis denotes the percentage of the router IPs for a given $t_{IP_y}^{loc_x}$.

z -axis denotes router IP. In the black area, the ratio of the largest $t_{IP_y}^{loc_x}$ to the total number of measurement units including IP_y is in the range of $[0.7 \sim 1]$. In the dark gray area, the ratio of the largest $t_{IP_y}^{loc_x}$ to the total number of measurement units including IP_y is in the range of $[0.3 \sim 0.7]$.

FIGURE 8: The distributions of the synchronization frequency.

FIGURE 9: The routing paths between three landmarks and three targets.

In the light gray area, the ratio of the largest $t_{IP_y}^{loc_x}$ to the total number of measurement units including IP_y is in the range of [0~0.3]. As observed in Figure 8, for most of the router IPs in analysis, the distribution of the synchronization frequency $t_{IP_y}^{loc_x}$ conforms to the power law distribution, and the distribution of the other router IPs conforms to approximate uniform distribution.

Combined with the theoretical and statistical analysis, we obtain the following conclusions: (a) if the distribution of synchronization frequency $t_{IP_y}^{loc_x}$ conforms to the power law distribution, then the corresponding router IP belongs to the set of edge routers 2; (b) if the distribution of synchronization frequency $t_{IP_y}^{loc_x}$ conforms to approximate uniform distribution, then the corresponding router IP belongs to the set of backbone routers 2; (c) if the values of d_y meet the conditions $d_y^p = \text{length}(S_{\text{province}})$ and $d_y^c = \text{length}(S_{\text{city}})$, then the router IP belongs to the set of edge routers 1 and backbone routers 1. In the following sections, after carrying out the data processing of the network measurement metrics, we classify the routers in order to find and geolocate the set of edge

routers 2. Then we geolocate the Internet hosts according to the geolocation results of edge routers 2.

4.4. The Data Processing of Network Measurement Metrics

4.4.1. The Construction of Synchronization Frequency Matrix. Filtering the set of backbone routers 1 and edge routers 1 from the IP list obtained from the whole traceroute measurement dataset, we obtain a set of different IPs S_{IP} . According to the synchronization frequency $t_{IP_y}^{loc_x}$ of each route IP in S_{IP} , we get the synchronization frequency matrix of S_{IP} , denoted as $T_{n \times k}$:

$$T_{n \times k} = \begin{bmatrix} t_1^{loc_1} & \dots & t_1^{loc_k} \\ \vdots & \ddots & \vdots \\ t_n^{loc_1} & \dots & t_n^{loc_k} \end{bmatrix}. \quad (8)$$

An instance for the construction of synchronization frequency matrix is presented as shown in Figure 9.

Figure 9 presents the routing paths between three landmarks and three targets with a result of 9 measurement units.

TABLE 3: Synchronization frequency.

	IP ₁₁	IP ₁₂	IP ₁₃	IP ₂₁	IP ₂₂	IP ₂₃	IP ₃₁	IP ₃₂	IP ₃₃
Location ₁	2	3	3	1	2	2	1	1	0
Location ₂	1	1	0	1	3	3	1	1	0
Location ₃	1	2	0	1	2	2	1	3	3

We get the synchronization frequency $t_{IP_y}^{loc_x}$ of each IP in the measurement units, which is presented in Table 3.

Then the related synchronization frequency matrix is

$$T_{3 \times 3} = \begin{bmatrix} 2 & 3 & 3 & 1 & 2 & 2 & 1 & 1 \\ 1 & 1 & 0 & 1 & 3 & 3 & 1 & 1 \\ 1 & 2 & 0 & 1 & 2 & 3 & 1 & 3 \end{bmatrix}. \quad (9)$$

4.4.2. The Construction of Conditional Probability Matrix. $P(IP_y, loc_x)$ denotes the probability that router IP_y and location loc_x are in the same measurement unit. Applying Bayes Theorem $P(A, B) = P(A | B)P(B) = P(B | A)P(A)$ to $P(IP_y, loc_x)$, we obtain

$$\begin{aligned} P(IP_y, loc_x) &= P(loc_x | IP_y) P(IP_y) \\ &= P(IP_y | loc_x) P(IP_y). \end{aligned} \quad (10)$$

$P(IP_y)$ denotes the probability of measurement unit including router IP_y which occurred in the whole measurement unit set S_m . $P(loc_x | IP_y)$ denotes the conditional probability of a geographic location for a given IP. According to the aforementioned equation, it follows that

$$\begin{aligned} P(loc_x | IP_y) &= \frac{P(IP_y, loc_x)}{P(IP_y)} \\ &= \frac{P(IP_y, loc_x)}{P(IP_y, loc_1) + \dots + P(IP_y, loc_x) + \dots + P(IP_y, loc_n)} \\ &= \frac{t_{IP_y}^{loc_x} / n_m^{loc_x}}{\sum_{loc \in S_{loc}} t_{IP_y}^{loc} / n_m^{loc}} = \frac{t_{IP_y}^{loc_x}}{\sum_{loc \in S_{loc}} t_{IP_y}^{loc}} = \frac{t_{IP_y}^{loc_x}}{T_{IP_y}^{S_{loc}}}. \end{aligned} \quad (11)$$

With the value of $P(loc_x | IP_y)$, we restate the synchronization frequency matrix $T_{n \times k}$ as conditional probability matrix:

$$P_{n \times k} = \begin{bmatrix} P(loc_1 | IP_1) & \dots & P(loc_k | IP_1) \\ \vdots & \ddots & \vdots \\ P(loc_1 | IP_n) & \dots & P(loc_k | IP_n) \end{bmatrix}. \quad (12)$$

4.5. The Geolocation Model for the Edge Routers

4.5.1. Finding the Edge Routers. In this section, we classify these router IPs in S_{IP} into two categories: backbone router 2 IPs (hereinafter referred to as backbone routers or backbone router IPs) and edge router 2 IPs (hereinafter referred to as edge routers or edge router IPs) in order to get the set of edge router IPs on the last subpath of the routing path. Let us define λ as the minimum threshold of $\max P(loc_x | IP_y)$, which is classified into the minimum threshold in the level of province λ_p and the minimum threshold in the level of province λ_c depending on different granularities of geographic location. Owing to the discrepancies between both the characteristics of routing policy in different countries and the accuracies of different geolocation databases, λ_p and λ_c may not be the same with respect to different countries or different geolocation databases. Consider

$$\lambda_p = \frac{1}{\arg \max (n_x^p / (N - n_1^p))}, \quad (13)$$

$$\lambda_c = \max \frac{1}{\arg \max ((n_x^c - t_x^p) / (N - n_1^c - t^p))},$$

where N is the total number of IPs in S_{IP} ; n_x^p and n_1^p are the numbers of IPs when the value of d_y^p is 1 and x , respectively; n_x^c and n_1^c are the numbers of IPs when the value of d_y^c is 1 and x , respectively; t^p is the number of IPs which meet the condition $\max P(loc_x | IP_y) < \lambda_p$; t_x^p is the number of IPs which meet the conditions $\max P(loc_x | IP_y) < \lambda_p$ and $d_y^c = x$.

The classification of the router IPs in S_{IP} consists of the following steps:

- (1) Moving the IPs from the conditional probability matrix $P_{n \times k}$ when the conditions $\max P(loc_x | IP_y) < \lambda_p$ and $\max P(loc_x | IP_y) < \lambda_c$ are met, we have m IPs left.
- (2) Sorting the values in every row from the largest to the smallest and extracting t columns in the front of every row, we generate a new matrix $P_{m \times t}$. Consider

$$t = \arg \max \frac{n_x^p}{N - n_1^p},$$

$$P_{m \times t} = \begin{bmatrix} P_1^1 & \dots & P_1^t \\ \vdots & \ddots & \vdots \\ P_m^1 & \dots & P_m^t \end{bmatrix}. \quad (14)$$

- (3) Dividing the values in every row by the value of the first column, which is denoted as v_x^1 , we have the matrix $E_x(v_x^1)P_{m \times t}$:

$$E_x(v_x^1)P_{m \times t} = \begin{bmatrix} \frac{P_1^1}{v_1^1} & \dots & \frac{P_1^t}{v_1^1} \\ \vdots & \ddots & \vdots \\ \frac{P_m^1}{v_m^1} & \dots & \frac{P_m^t}{v_m^1} \end{bmatrix}. \quad (15)$$

(4) Moving the first column, we get the matrix $P_{m \times (t-1)}$:

$$P_{m \times (t-1)} = \begin{bmatrix} \frac{P_1^2}{v_1^1} & \dots & \frac{P_1^t}{v_1^1} \\ \vdots & \ddots & \vdots \\ \frac{P_m^2}{v_m^1} & \dots & \frac{P_m^t}{v_m^1} \end{bmatrix}. \quad (16)$$

(5) Define the vector models for routers in different granularities of geographic location. The vector model for the edge router IP, the theoretical distribution whose synchronization frequency conforms to the power law distribution, is defined as $[0, 0, \dots, 0]$. And the vector model for the backbone router IP, the theoretical distribution whose synchronization frequency conforms to uniform distribution, is defined as $[1, 1, \dots, 1]$.

(6) Calculate the Weighted Euclidean Distance between every row and the two vector models. The formulation of Weighted Euclidean Distance is

$$d = \sqrt{(x_1 - s_1)^2 + \frac{1}{2}(x_2 - s_2)^2 + \dots + \frac{1}{n}(x_n - s_n)^2}. \quad (17)$$

We define d_e as the distance between a row and the vector model for the edge router IP and d_b as the distance between a row and the vector model for the backbone router IP. Every row corresponds to a router IP in the traceroute measurement dataset. The classification principals for the router IPs are illustrated as follows: (a) if $d_e > d_b$, then the router IP belongs to the set of edge routers; (b) if $d_b > d_e$, then the router IP belongs to the set of backbone routers. Then the router IPs in the traceroute measurement dataset are classified into a set of edge routers S_e and a set of backbone routers S_b .

4.5.2. The Geolocation Model for Edge Router. The geographic location of edge router IP_y is determined by the one which maximizes the conditional probability function $P(\text{loc}_x | IP_y)$. The model of geolocation for edge router is established as follows:

$$y = \arg \max P(\text{loc}_x | IP_y). \quad (18)$$

4.6. The Geolocation Model for Internet Host. Find the set of router IPs in all the measurement units including IP_y , and the set is denoted as S_r . The geolocation of Internet hosts is classified into two cases depending on whether S_r and S_e have common router IPs. The first one is discussed in Case 1, and the second one is discussed in Case 2.

4.6.1. Case 1: $S_r \cap S_e \neq \emptyset$. Wipe off all the router IPs which are not included in S_e from S_r . Construct the location probability

matrix $L_{m \times n}$ with the geographic location of every router IP in S_r . Consider

$$L_{m \times n} = \begin{bmatrix} \delta_1^{\text{loc}_1} & \dots & \delta_1^{\text{loc}_n} \\ \vdots & \ddots & \vdots \\ \delta_m^{\text{loc}_1} & \dots & \delta_m^{\text{loc}_n} \end{bmatrix}^T \quad \delta = \begin{cases} 1, & \text{loc}_x = y \\ 0, & \text{others.} \end{cases} \quad (19)$$

Summing over the values of every row in a given column of the matrix, we have

$$L_{m \times n} = [N(\text{loc}_1), \dots, N(\text{loc}_n)], \quad (20)$$

$$N(\text{loc}_x) = \sum_{i=1}^{i=m} L_{i,x}.$$

The model of geolocation for Internet host IP_y is established as follows:

$$y = \arg \max N(\text{loc}_x). \quad (21)$$

The procedure of geolocation for Internet hosts in Case 1 is presented in Figure 10.

4.6.2. Case 2: $r_{\text{last}} \cap r_{\text{route}} = \emptyset$. Cluster the IPs geolocated in Case 1 according to the geographic locations. IPs in the same geographic location are put into the same cluster. And C_{loc} denotes the set of IPs in a certain cluster.

Based on the instantaneous delay measurements, every IP gets a delay vector $D_i = (t_{i1}, t_{i2}, \dots, t_{ik})$. Then calculate the average delay vector for each cluster C_{loc} , which is expressed as follows:

$$D_a = (t_{a1}, t_{a2}, \dots, t_{ak}), \quad (22)$$

$$t_{ax} = \frac{\sum_{IP \in C_{\text{loc}}} t_{ix}}{|C_{\text{loc}}|},$$

where k and $|C_{\text{loc}}|$ are the numbers of landmarks and IPs in the cluster C_{loc} , respectively. We use cosine similarity to calculate the similarity between the delay vector of IP_y and the average vector of C_{loc} :

$$\begin{aligned} \text{sim}(IP_y, C_{\text{loc}}) &= \frac{D_a \cdot D_i}{\|D_a\| \|D_i\|} \\ &= \frac{\sum_{j=1}^k d_{aj} \times d_{ij}}{\sqrt{\sum_{j=1}^k (d_{aj})^2} \times \sqrt{\sum_{j=1}^k (d_{ij})^2}}. \end{aligned} \quad (23)$$

The geographic location of Internet host IP_y is the same as the cluster which has the minimum cosine similarity. The geolocation model for Internet host IP_y is established as follows:

$$y = \arg \min \text{sim}(IP_y, C_{\text{loc}}). \quad (24)$$

FIGURE 10: The procedure of geolocation for Internet hosts in Case 1.

TABLE 4: Accuracy rates for the ground truth nodes after optimization.

	IP138	QQWry	IPcn	CBG	TBG
Province	0.982	0.981	0.979	0.989	0.993
City	0.978	0.973	0.975	0.982	0.991

TABLE 5: Coverage rates for different numbers of comparison databases after optimization.

Comparison number	2	3	4	5
Province	0.982	0.981	0.978	0.986
City	0.976	0.975	0.973	0.957

5. Experiments and Results

In this section, we evaluate the performance of the GeoCop method from three aspects: accuracy, robustness, and efficiency. Section 5.1 evaluates the GeoCop method on the aspect of accuracy. Section 5.2 evaluates the GeoCop method on the aspect of robustness. Section 5.3 evaluates the GeoCop method on the aspect of efficiency.

5.1. Accuracy Evaluation. In this section, we not only use the same empirical evaluation methods as presented in Section 3 to evaluate the accuracy of these geolocation databases after optimization but also compare the improvement of the geolocation databases statistically in order to ensure the significance of the differences of performance between the geolocation databases with and without the optimization.

5.1.1. Empirical Evaluation. As observed in Table 4, the accuracy rates of the geolocation databases in Table 4 are higher than those in Table 1. It is clear that databases after optimization are more accurate. In the original databases, the accuracy rates in the granularity of province are much higher than the granularity of city. In the databases after optimization, the accuracy rates in the granularity of province are still higher than the granularity of city, but the differences are very small. As observed in Table 5, the coverage rates are

higher than the results in Table 2. It is presented that the GeoCop method improves the accuracy of the geolocation databases as a whole.

5.1.2. Statistical Tests. To perform the statistical tests, we divide the IP addresses in S_g into 10 identical portions and calculate the accuracy rate of each portion. Then all data are analyzed using SPSS (Statistical Package for the Social Science) 22.0 statistical software (IBM Corporation, Somers, NY). The significance of the differences of performance between the geolocation databases with and without optimization is tested by paired-samples t -test and Wilcoxon test. For all analyses, $P < 0.05$ is considered significant. Taking the geolocation database IP138 for example, the results of the two statistical tests are presented in Tables 6 and 7, respectively.

As shown in Table 6, the value of $P = \text{sig. (2-tailed)}$ is 0.000, which is smaller than 0.05, so the results of paired-samples t -test reject the null hypothesis with significant level 5%. The difference of performance between the geolocation databases IP138 with and without the optimization is significant. And the value of average accuracy rate, which is 0.801, is smaller than that after optimization, which is 0.982. It is presented that the accuracy rate after optimization is higher than that before optimization.

As shown in Table 7, the value of $P = \text{sig. (2-tailed)}$ is 0.02, which is smaller than 0.05, so the results of the Wilcoxon test reject the null hypothesis with significant level 5%. The difference of performance between the geolocation databases IP138 with and without the optimization is significant. And the mean rank of positive ranks, which is 5.50, is larger than that of negative ranks, which is 0.00. It is presented that the accuracy rate after optimization is higher than that before optimization.

Owing to the limitation of space or the following statistical tests, we just present the important results in Table 8. T and Z denote the results of paired-samples t -test and Wilcoxon test, respectively.

As shown in Table 8, we can observe that all the values of P are smaller than 0.05, so the results of the statistical tests reject the null hypothesis with significant level 5%. For each

TABLE 6: The results of paired-samples t -test for the geolocation database IP138.

(a) Paired-samples statistics								
	Mean		N		Std. deviation		Std. error mean	
Pair 1								
Before	.80100		10		.002000		.000632	
After	.98200		10		.010770		.003406	
(b) Paired-samples test								
Paired differences								
	Mean	Std. deviation	Std. error mean	95% confidence interval of the difference		t	df	Sig. (2-tailed)
				Lower	Upper			
Pair 1								
Before – after	-.181000	.011353	.003590	-.189121	-.172879	-50.416	9	.000

TABLE 7: The results of Wilcoxon test for the geolocation database IP138.

(a) Ranks			
	N	Mean rank	Sum of ranks
After – before			
Negative ranks	0 ^a	.00	.00
Positive ranks	10 ^b	5.50	55.00
Ties	0 ^c		
Total	10		
(b) Test statistics ^a			
	After – before		
Z	-2.812 ^b		
Asymp. sig. (2-tailed)	.005		

^aAfter < before.
^bAfter > before.
^cAfter = before.

geolocation database, the difference of performance between the geolocation databases with and without the optimization is significant. According to the means in the results of paired-samples t -tests and the mean ranks in the results of Wilcoxon tests, we can observe that the accuracy rates of the geolocation databases after optimization are higher than those before optimization. It has been proven that the GeoCop method improves the accuracy of the geolocation databases by comparing the improvement of the GeoCop method statistically.

5.2. Robustness Evaluation. In this section, we compare the GeoCop method with two existing geolocation methods (CBG and TBG) to evaluate the performance of robustness from three aspects: the dramatic increment in the delay, the accuracy rate of the geolocation database before optimization, and the landmark distribution.

5.2.1. Aspect 1: The Dramatic Increment in the Delay. We evaluate the robustness of the GeoCop method in two scenarios. One is about the delay noise introduced as a result of queuing delays and circuitous routes. To simulate the scenario, we add increment ranging from 0 ms to 1 ms to the delays from the landmarks to the ground truth node. Figures 11(a) and 11(b) present the accuracy rates of the geolocation databases depending on the increment of delay. As observed in Figure 11, the accuracy rates of both CBG and TBG before optimization decrease with the increment of delay, but the accuracy rates after optimization remain constant.

The other one is about the delay-based misleading behaviors by adversary who tampers with the delay to fake his geographic location. This is realized by making the delay appear larger than the actual one. Let $d_{i,t}$ be the minimum delay between the landmark L_i and the target D_t , and let $d_{i,\tau}$ be the minimum delay between the landmark L_i and the forged target D_τ . The delay we add to each traceroute measurement to D_t is $\Delta d = d_{i,\tau} - d_{i,t}$ ($d_{i,\tau} > d_{i,t}$), and then $d_{i,t}$ appears as $d_{i,\tau}$. Figures 11(c) and 11(d) present the success rates for the misleading behaviors depending on how far the adversary attempts to move the target. As presented in Figures 11(c) and 11(d), if the distance of attempted move in the misleading behaviors is small, it is difficult for the measurement-based geolocation methods before optimization to find the adversaries. But, after optimization, whatever the distance is, the adversaries hardly succeed in implementing the misleading behaviors. The plots in Figure 11 indicate the hypothesis that the GeoCop method is robust on the aspect of the dramatic increment in the delay.

5.2.2. Aspect 2: The Accuracy Rate of Geolocation Database before Optimization. Figures 12(a) and 12(b) present the accuracy rate of ground truth nodes after optimization depending on the accuracy rate of geolocation database before optimization with the granularity of province and city, respectively. As presented in Figure 12(a), whatever the granularity is of province or of city, when the accuracy rate of the geolocation database with the granularity of province before optimization decreases but still remains above 50%, the accuracy rate of ground truth nodes after optimization

TABLE 8: The results of statistical tests for different geolocation databases.

	IP138		QQWry		IPcn		CBG		TBG	
	t	$P(t)$	t	$P(t)$	t	$P(t)$	t	$P(t)$	t	$P(t)$
Province	-50.416	0.000	-48.997	0.000	-41.007	0.000	-28.632	0.000	-11.528	0.000
City	-30.631	0.000	-64.365	0.000	-82.348	0.000	-58.111	0.000	-69.899	0.000
	Z	$P(Z)$	Z	$P(Z)$	Z	$P(Z)$	Z	$P(Z)$	Z	$P(Z)$
	t	$P(t)$	t	$P(t)$	t	$P(t)$	t	$P(t)$	t	$P(t)$
Province	-2.812	0.002	-2.805	0.002	-2.805	0.002	-2.803	0.002	-2.803	0.002
City	-2.803	0.002	-2.805	0.002	-2.803	0.002	-2.812	0.002	-2.805	0.002

TABLE 9: The results of statistical tests for different geolocation methods in different aspects.

	Aspect 1								Aspect 3			
	Scenario 1				Scenario 2							
	CBG		TBG		CBG		TBG		CBG		TBG	
	t	$P(t)$	t	$P(t)$	t	$P(t)$	t	$P(t)$	t	$P(t)$	t	$P(t)$
Province	-6.410	0.000	-8.288	0.000	17.014	0.000	20.607	0.000	-9.979	0.000	-10.343	0.000
City	-10.102	0.000	-11.165	0.000	13.067	0.000	14.054	0.000	-21.747	0.000	-16.353	0.000
	Z	$P(Z)$	Z	$P(Z)$	Z	$P(Z)$	Z	$P(Z)$	Z	$P(Z)$	Z	$P(Z)$
Province	-2.934	0.001	-2.934	0.001	-2.934	0.001	-2.934	0.001	-2.934	0.001	-2.937	0.001
City	-2.934	0.001	-2.934	0.001	-2.934	0.001	-2.934	0.001	-2.936	0.001	-2.936	0.001

remains constant, but once it goes below 50%, the accuracy rate of ground truth nodes after optimization decreases sharply. As presented in Figure 12(b), the accuracy rate of ground truth nodes after optimization remains constant when the accuracy rate of geolocation database with the granularity of city before optimization decreases but still remains above 40%. But once it goes below 40%, the accuracy rate of ground truth nodes with the granularity of province after optimization still remains constant, while the accuracy rate with the granularity of city decreases sharply.

5.2.3. Aspect 3: Landmark Distribution. Figure 13 plots the accuracy rates of ground truth nodes depending on 10 different kinds of landmark distributions. As presented in Figure 13, the accuracy rates of both CBG and TBG before optimization are different when the landmark distribution changes, but the accuracy rates after optimization remain at a high level. It is observed that measurement-based methods after optimization will not be affected by the distribution of landmarks.

5.2.4. Statistical Tests. As shown in Table 9, we can observe that all the values of P are smaller than 0.05, so the results of the statistical tests reject the null hypothesis with significant level 5%. For each geolocation method, the difference of performance between the geolocation databases with and without optimization is significant. According to the means in the results of paired-samples T tests and the mean ranks in the results of Wilcoxon tests, we can observe the following: (1) in scenario 1 of aspect 1 and the aspect 3, the accuracy rates of the geolocation methods after optimization are higher than those before optimization; (2) in scenario 2 of aspect 1, the success rates of adversary after optimization are lower than those before optimization. It has been proven that the GeoCop method improves the robust of the geolocation

methods by comparing the improvement of the GeoCop method statistically.

5.3. Measurement and Computation Overheads. Figure 14 plots the cumulative distribution function for the numbers of edge routers and Internet hosts whose geographic locations vary with time. As presented in Figure 14, the change in the geographic locations of edge routers is less than that of Internet hosts. It is observed that, if the geolocation methods adopt the optimization by the GeoCop method, the calculation frequency for the updating of the geolocation results of Internet hosts in the geolocation databases can be decreased. Table 10 presents the measurement and computation overheads of different geolocation methods before and after optimization for one time and t times. It is observed that the measurement overheads of both CBG and TBG after optimization are the same as those before optimization. The computation overheads after optimization are less than those before optimization, especially when the update times increase.

6. Conclusion

In this paper, a novel method based on machine learning (GeoCop) is proposed for optimizing the geolocation databases of Internet hosts in the cloud environment. The geolocation model for Internet host is derived from network measurement with the supplement of machine learning and routing policy, making the geolocation less prone to imperfect measurements and irregular routings. This work also involves theoretical analysis on the drawbacks of the existing geolocation methods as well as statistical analysis on the accuracy of the existing geolocation databases. In comparison with three frequently used geolocation databases and two well-known measurement-based methods, the performance

TABLE 10: The measurement and computation overheads of different geolocation methods.

Overhead	Geolocation method			
	CBG	CBG + GeoCop	TBG	TBG + GeoCop
Measurement (one time)	$n * k * 10$	$n * k * 10$	$n * k * 10$	$n * k * 10$
Measurement (t times)	$n * k * 10 * t$	$n * k * 10 * t$	$n * k * 10 * t$	$n * k * 10 * t$
Computation (one time)	n	$n + 1$	n	$n + 1$
Computation (t times)	$t * n$	$\frac{t}{2} * n + t * 1$	$t * n$	$\frac{t}{2} * n + t * 1$

FIGURE 11: Robustness evaluation on the aspect of delay.

FIGURE 12: Robust evaluation on the aspect of accuracy rate before optimization.

FIGURE 13: Robust evaluation on the aspect of landmark distribution.

FIGURE 14: The changes in the numbers of edge routers and Internet hosts.

of the GeoCop method is validated on the PlanetLab network test bed from three aspects: accuracy, robustness, and efficiency. And we not only use the typical benchmarks to compare the performance of the GeoCop method but also perform statistical tests to display the improvement of the GeoCop method. As presented in the comparison tables, the GeoCop method achieves improved performance in both accuracy and robustness with less measurements and calculation overheads. Future work will be focused on a more robust method for the geolocation of Internet hosts and more diversity behaviors for misleading the geographic location of an Internet host.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work is supported by the National Key Project of Scientific and Technical Supporting Programs of China (Grant no. 2013BAH10F01, Grant no. 2013BAH07F02, and Grant no. 2014BAH26F02), the Research Fund for the Doctoral Program of Higher Education (Grant no. 20110005120007), Beijing Higher Education Young Elite Teacher Project (Grant no. YETP0445), and the Fundamental Research Funds for the Central Universities and Engineering Research Center of Information Networks, Ministry of Education.

References

- [1] K. D. Bowers, M. Van Dijk, A. Juels, A. Oprea, and R. L. Rivest, "How to tell if your cloud files are vulnerable to drive crashes," in *Proceedings of the 18th ACM Conference on Computer and Communications Security*, pp. 501–514, October 2011.
- [2] K. Benson, R. Dowsley, and H. Shacham, "Do you know where your cloud files are?" in *Proceedings of the 3rd ACM Workshop on Cloud Computing Security Workshop (CCSW'11)*, pp. 73–82, October 2011.
- [3] Y. Shavitt and N. Zilberman, "A geolocation databases study," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 10, pp. 2044–2056, 2011.
- [4] J. Bendale and J. R. Kumar, "Review of different IP geolocation methods and concepts," *International Journal of Computer Science and Information Technologies*, vol. 5, no. 1, pp. 436–440, 2014.
- [5] M. Zhang, Y. Ruan, V. Pai et al., "How DNS misnaming distorts Internet topology mapping," in *Proceedings of the USENIX Annual Technical Conference*, p. 34, 2006.
- [6] Z. Dong, R. D. W. Perera, R. Chandramouli, and K. P. Subbalakshmi, "Network measurement based modeling and optimization for IP geolocation," *Computer Networks*, vol. 56, no. 1, pp. 85–98, 2012.
- [7] P. Gill, Y. Ganjali, B. Wong et al., "Where's that IP? Circumventing measurement-based IP geolocation," in *Proceedings of the USENIX Security Symposium*, pp. 241–256, 2010.
- [8] S. Laki, P. Mátray, P. Hágá, T. Sebok, I. Csabai, and G. Vattay, "Spotter: a model based active geolocation service," in *Proceedings of IEEE INFOCOM*, pp. 3173–3181, April 2011.
- [9] S. Laki, P. Mátray, P. Hágá, I. Csabai, and G. Vattay, "A model based approach for improving router geolocation," *Computer Networks*, vol. 54, no. 9, pp. 1490–1501, 2010.
- [10] I. Poesse, S. Uhlig, M. A. Kaafar, B. Donnet, and B. Gueye, "IP geolocation databases: unreliable?" *ACM SIGCOMM Computer Communication Review*, vol. 41, no. 2, pp. 53–56, 2011.
- [11] P. Mátray, P. Hágá, S. Laki et al., "On the network geography of the internet," in *Proceedings of the IEEE INFOCOM*, pp. 126–130, 2011.
- [12] P. Mátray, P. Hágá, S. Laki, G. Vattay, and I. Csabai, "On the spatial properties of internet routes," *Computer Networks*, vol. 56, no. 9, pp. 2237–2248, 2012.
- [13] S. Laki, P. Mátray, P. Hágá, I. Csabai, and G. Vattay, "A detailed path-latency model for router geolocation," in *Proceedings of the 5th International Conference on Testbeds and Research Infrastructures for the Development of Networks & Communities and Workshops (TridentCom '09)*, pp. 1–6, IEEE, Washington, DC, USA, April 2009.
- [14] V. N. Padmanabhan and L. Subramanian, "An investigation of geographic mapping techniques for Internet hosts," in *Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications (SIGCOMM '01)*, pp. 173–185, ACM, San Diego, Calif, USA, August 2001.
- [15] D. I. Youn, B. L. Mark, and D. Richards, "Statistical geolocation of internet hosts," in *Proceedings of the 18th International Conference on Computer Communications and Networks (ICCCN '09)*, pp. 1–6, August 2009.
- [16] H. Maziku, S. Shetty, K. Han, and T. Rogers, "Enhancing the classification accuracy of IP geolocation," in *Proceedings of the IEEE Military Communications Conference (MILCOM '12)*, pp. 1–6, IEEE, Orlando, Fla, USA, November 2012.
- [17] M. J. Arif, S. Karunasekera, S. Kulkarni, A. Gunatilaka, and B. Ristic, "Internet host geolocation using maximum likelihood estimation technique," in *Proceedings of the 24th IEEE International Conference on Advanced Information Networking and Applications (AINA' 10)*, pp. 422–429, April 2010.
- [18] B. Eriksson, P. Barford, J. Sommers, and R. Nowak, "A learning-based approach for IP geolocation," in *Proceedings of the 11th International Conference on Passive and Active Measurement (PAM '10)*, pp. 171–180, 2010.
- [19] B. Gueye, A. Ziviani, M. Crovella, and S. Fdida, "Constraint-based geolocation of internet hosts," *IEEE/ACM Transactions on Networking*, vol. 14, no. 6, pp. 1219–1232, 2006.

- [20] C. Guo, Y. Liu, W. Shen, H. J. Wang, Q. Yu, and Y. Zhang, "Mining the web and the internet for accurate IP address geolocations," in *Proceedings of the 28th Conference on Computer Communications (INFOCOM '09)*, pp. 2841–2845, IEEE, April 2009.
- [21] Y. Tian, R. Dey, Y. Liu, and K. W. Ross, "China's internet: topology mapping and geolocating," in *Proceedings of the IEEE Conference on Computer Communications (INFOCOM '12)*, pp. 2531–2535, March 2012.
- [22] Y. Shavitt and N. Zilberman, "A structural approach for pop geolocation," in *Proceedings of the Infocom Workshop on Network Science for Communications*, 2010.
- [23] B. Biswajit, S. Sachin, and R. Tamara, "Classification based IP geolocation approach to locate data in the cloud datacenters," in *Proceedings of the ASE BIGDATA/SOCIALCOM/CYBER SECURITY Conference*, Stanford, Calif, USA, May 2014.
- [24] B. Wong, I. Stoyanov, and E. G. Sirer, "Octant: a comprehensive framework for the geolocalization of Internet hosts," in *Proceedings of the 4th USENIX Conference on Networked Systems Design & Implementation (NSDI '07)*, p. 23, 2007.
- [25] E. Katz-Basnett, J. P. John, A. Krishnamurthy, D. Wetherall, T. Anderson, and Y. Chawathe, "Towards IP geolocation using delay and topology measurements," in *Proceedings of the 6th ACM SIGCOMM on Internet Measurement Conference*, pp. 71–84, October 2006.
- [26] P. Dhungel, K. W. Ross, M. Steiner et al., "Peer-assisted download acceleration on a massive scale," Tech. Rep., Polytechnic Institute of NYU, 2011.