Street-level Landmark Mining Algorithm Based on Radar Search

Xiaonan Liu, Wen Yang, Meijuan Yin, Fenlin Liu, Chenshu Yun

State Key Laboratory of Mathematical Engineering and Advanced Computing, China
nine_day@163.com, byygxs015@126.com, raindot_ymj@163.com, fenlinliu@vip.sina.com, 402410849@qq.com

Abstract

Street-level landmarks are the important foundation for high-precision IP geolocation, which is of significant value in network security. As online map-based landmark mining algorithms are constrained by the online map service itself, a street-level landmark mining algorithm based on radar search is proposed in this study. Initially, the region in which landmarks will be mined is divided into smaller sub-regions. Then, the radar search service of an online map is used to perform a recursive query request for each sub-region, and street-level candidate landmarks in the sub-region are obtained. Finally, IP geolocation databases and the street-level geolocation algorithm are used to evaluate the landmarks and retain reliable ones. Landmark mining experiments of two groups were conducted to verify the algorithm. Experimental results show that the number of obtained reliable landmarks of the proposed method increases by 4.1 times, the landmark coverage area increases by 59% and the average geolocation error is reduced from 9.94 km to 4.33 km compared with the existing online maps-based method, and the proposed algorithm can also obtain more candidate landmarks as well as reliable landmarks and got lower mean error of geolocation results than the state-of-art landmark mining algorithms based on other web resources.

Keywords: Street-level landmarks, Landmark mining, IP geolocation, Online map, Radar search

1 Introduction

IP geolocation refers to the task of finding the geographic location of an Internet host [1-2]. Obtaining the geolocation information of Internet hosts is valuable for many applications in network security, for instance, protecting social network privacy, tracing network attack and intrusion detection [3-4]. Hence, the technology has a broad range of potential applications in the civil, military, and security fields. At present, IP geolocation approaches consist of three main types: database query-based method, data mining-based method and network measurement-based method The database query-based geolocation, such as NetGe [5] and RIPE IPmap active geolocation [6], queries IP location database on the Internet, such as Maxmind[1], IP2location[2], NetAcuity[3], GeoIP[4], IPIP[5], OpenIPMap[6], Aiwen[7], IPcn[8], Whois[9] and etc, and makes comparative analysis to determine the geographic location of IP. This type of methods is highly efficient and easy to implement, but the precision of locations obtained is not high. Locations of country level is reliable, while the accuracy of city level locations is hard to ensure [7-8]. Gharaibeh et al. [9] and Dan et al. [10] respectively tested several of these IP databases, the results show that the accuracy of city locations in these databases needs to be improved. The data mining-based geolocation mines and analyzes the associated geographic location for an IP address from different kinds of network data with IP address and geographic location information, such as the social network, mobile phone application, DNS, login log, and etc. The representative algorithm is Checkin-Geo [11], DRoP (DNS-based Router Positioning) [12], R-DNS [13], GQL (Geolocation using Query Logs) [10], and so on. In addition, a few methods such as HLOC [14] assesses the reliability of associated geographic locations by a small number of network measurements. For this type of approaches, the overhead of measurement is low, but a mass of basic data is needed, which can be implemented by

[1] https://www.maxmind.com
[2] https://www.ip2location.com
[3] https://www.digitalelement.com
[4] https://geoip.com
[5] https://www.ipip.net
[6] https://openipmap.ripe.net
[7] https://www.ipplus360.com
[8] https:// www.ipcn.co.uk
[9] https:// www.whois.com

*Corresponding Author: Meijuan Yin; E-mail: raindot_ymj@163.com
DOI: 10.3966/160792642021032202005

collaborating with large Internet companies. The network measurement-based method firstly measures the delays between the probe sources and the target IP to be located, or obtains the network topology information. Then The transformation relationship or correspondence between the time delay or topological structure and geographical distance is analyzed based on the geographic location information of the probe source or the location reference point, which is IP entity with known and stable geographic location, named landmark. At last, the location of target IP is estimated by multi-point location, structure matching or optimization. The representative algorithm is GeoTrack [15], CBG (Constrained-Based Geolocation) [16], TBG (Topology-Based Geolocation) [4], SLG (Street-Level Geolocation) [17], NNBG (Neural Networks-Based Geolocation) [18], Geolocation base on RIPE altas [19], and Active Geolocation [20]. The location result of these methods is relatively accurate as it estimates the location of target IP by network measurement and combining with reference point locations. In addition, some researches have integrated multiple methods simultaneously for IP geolocation. For example, Fanou et al. [21] successively used IP location databases, reverse DNS resolution, network delay measurement, landmark reference locating and other geolocation techniques to locate infrastructure IPs to improve the accuracy of geolocation results. In above methods, the network measurement-based approach is the current research hotspot in this field, among which the landmark-based approaches have the highest precision [22-23]. However, the accuracy of the results mainly depends on the density and accuracy of the landmarks [17]. Hence, the key to high-precision Internet entity geolocation is obtaining high-quality landmarks.

A landmark refers to an Internet entity that has a stable IP identity and a known geographic location, usually represented as a two-tuple <the IP address, the geographic location>. According to the precision of the geographic location, landmarks can be divided into two types, that is city-level landmarks and street-level landmarks, for instance <IP1, "Zhengzhou, Henan Province, China"> and <IP2, "latitude 34.46, longitude 113. 40">. Landmark mining is extracting mass landmark information from various network resources to provide location reference points for IP geolocation.

City-level landmark mining methods mainly consist of IP geolocation database-based methods, webpage-based methods, and Internet forum-based methods. IP geolocation database-based methods [24] select landmarks from blocks of IP addresses with the same geographic location in several IP geolocation databases. These approaches can obtain a large number of landmarks. However, there may be many non-active IP addresses in an IP address block, which limits the effectiveness of IP geolocation based on these landmarks. Guo et al. [25] proposed Structon, a webpage based landmark mining method, for acquiring a large number of city-level landmarks. This method extracts geographic locations information from web pages and employs majority voting on these locations to determine the location of the server for the website. Then, they associate the location and IP address of the web server to build a seed landmark. Finally, they extend the location of the seed landmarks to the /24 IP segment where the IP addresses are located and correct the location of the landmarks using AS and BGP information. Zhu et al. [26] proposed a city-level landmark mining method based on Internet forums. They mine users' IP addresses from city-themed forums and infer their city by the name of the forum to create plenty of city-level landmarks.

Street-level landmarks are an important basis for high-precision geolocation algorithms. The main street-level landmark mining methods are those based on different web resources, such as online maps, yellow pages, service ports, etc. Wang et al. [17] proposed the Comprehensive Landmark Mining Algorithm (CLMA), which is based on online maps. This method extracts the domain names and geographic locations of organizations in the given region using text search service of an online map, such as Google Maps (or local search service in some online maps, such as Baidu Maps), and then maps their domain names to IP addresses to generate candidate landmarks. Finally, the method verifies the candidate landmarks to obtain reliable landmarks using webpage test requests and multi-branch verification. However, as the number of landmarks obtained by the text search service is limited, some landmarks may be missed, especially in areas with many organizations. Moreover, the method verifies the candidate landmarks according to whether the IP address and the domain name of the candidate landmark link to the same webpage, which may cause mis-selected and mis-deleted landmarks. Ma et al. [27] proposed an algorithm based on yellow pages. This method extracts the locations and Web or Email service domains of institutions by using regular expression from yellow pages, and the corresponding IPs of domains are parsed. Then those IPs with reliable locations of both city information and street-level information obtained by SLG geolocation algorithm are landmarks. The method can effectively correct the mis-deletion and mis-evaluation of some landmarks by the existing typical landmark obtaining algorithm. However, as the yellow pages and their corresponding regular expressions of this method are specified manually, it is difficult to obtain landmarks quickly in different regions of the world. Li et al. [28] proposed the landmarks acquisition method based on SVM (Support Vector Machine) classifiers of service ports. This method obtains the characteristics of open ports for known services by using an optimal SVM classifier and classifies IPs with some known services in some given regions. And then according to the domain

names corresponding to the IPs, the relationships between the classified server IPs and their reliable locations obtained by querying online maps are established. The number of obtained street-level landmarks with a low geolocation error is increased substantially for this method. But the process of detecting open ports is time consuming, and some landmarks obtained by this method are not reliable as the locations returned by online maps are sometimes incorrect. In above methods, the online map-based method has obvious advantages. Online maps, such as Google Maps, Bing Maps and Baidu Maps, contain huge amounts of formatted POI data, including the exact geographic location of organizations and domain names of their web service. Massive landmarks of web server type with stable performance and open ports (web service ports 80 or 8080) can be mined and it is easy to acquire global street-level landmarks in bulk from online maps.

We focus on methods of street-level landmark mining. As online map-based methods are limited by the map local search service itself, a radar search-based street-level landmark mining algorithm (RSLM) is proposed in this study. The algorithm takes advantage of the radar search service, such as that of Google Maps, (or search in bounds service in some online maps, such as Baidu Maps), which can specify the search center point and size of a circular area to mine all available landmarks in an online map. First, RSLM divides the given administrative region into square sub-regions of an appropriate size and then performs a recursive radar search on the circumscribed circle of each sub-region to obtain a list of all the organization identifications (IDs) within it. Second, RSLM searches for the domain name and location of each organization using its ID and converts the domain names into IP addresses using DNS queries to build candidate landmarks. Finally, landmarks with large location errors are eliminated using IP location databases and reliable street-level landmarks are selected using the SLG algorithm [17].

2 RSLM Algorithm

2.1 Problem Definition

Landmark mining is extracting massive and stable IP addresses and their probably associated geographic location from various network resources with both IP addresses and geographic location information usually by data mining, and identifying IPs with reliable location among them by evaluating the reliability of associated geographic location based on correlation analysis or network measurement. The obtained landmarks are usually represented as <an IP address, its geographic location> two-tuples, which will be taken as location reference points for IP geolocation.

Landmark mining is similar to IP geolocation in determining the geographic location of IP addresses. But they are two different problems. IP geolocation usually determines the geographic location for given target IP address(es), while landmark mining aims to automatically obtain a large number of stable IP addresses, as well as the location information of these IP addresses, without any given IP addresses. And in order to meet the global IP geolocation needs, it is expected to acquire landmarks covering all regions of the world for landmark mining. In addition, landmark mining usually aims to obtain the geographic location of the entity with stable IP, such as routers and servers, so that the obtained landmarks have a long validity period to be taken as the reference points for locating IPs whose geographical location are unknown. Whereas, IP geolocation aims to determine the current geographic location of the target IP (or IPs), whether it is a stable IP address or a dynamically allocated address.

2.2 Framework and Steps of RSLM

The framework of the RSLM algorithm is shown in Figure 1.

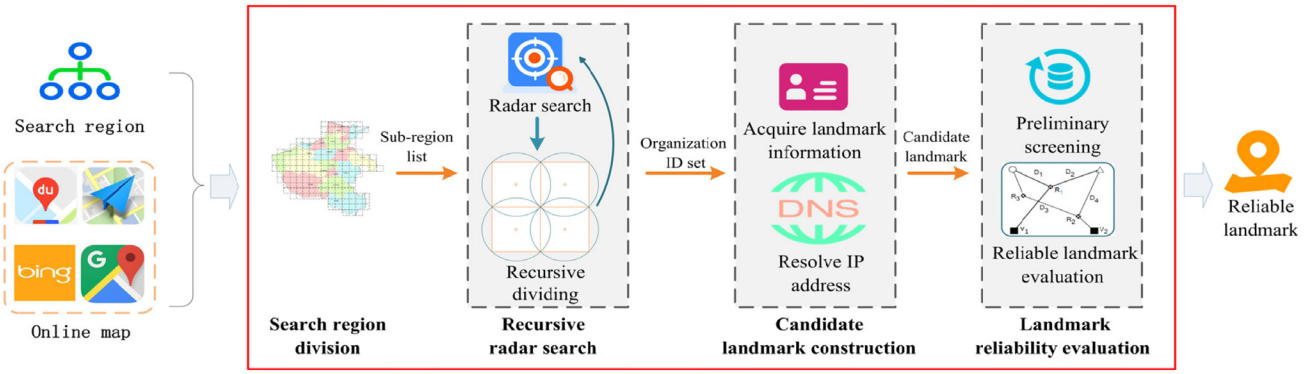


Figure 1. Framework of RSLM

RSLM comprises four parts: search region division, recursive radar search, candidate landmark construction, and landmark reliability evaluation. In the search region division, the search region is divided into square sub-regions of the same size to define sub-region search set $\mathcal{T}$. Then, in the recursive radar search, the

radar search service of an online map, such as the Google Maps, is used to obtain all the organization IDs in every sub-region in $\mathcal{T}$ to yield organization ID set $\mathcal{O}$. In the candidate landmark construction, the service of the online map and a DNS service are used to obtain the landmark information according to the organization IDs in $\mathcal{O}$ to give candidate landmark set $\mathcal{L}_c$. In landmark reliability evaluation, the IP geolocation database and SLG algorithm are used to evaluate the reliability of the landmarks in $\mathcal{L}_c$, and the reliable landmarks are selected to form set $\mathcal{L}$.

The details of each step in RSLM are given as follows:

Step 1: Search region division

Step 1.1: Initial search sub-region division. A point in the search region (denoted as A) is selected. Then, a square S with side length $D=\sqrt{2}\,R$, where R refers to the initial radar search radius, is defined using the point as its center and added to the ordered set $\mathcal{T}$ of search sub-regions as the first element.

Step 1.2: Peripheral sub-region expansion: The four neighboring squares that share an edge with S to the east, west, south, and north of S are denoted as S_1, S_2, S_3, and S_4 respectively. Then, $S'$is added to $\mathcal{T}$ if $S' \notin \mathcal{T} \wedge S' \cap A \neq \emptyset$, otherwise, S' is discarded.

Step 1.3: Step 1.2 is executed on each new element in $\mathcal{T}$ until there are no new elements that can be added to $\mathcal{T}$.

Figure 2 shows the process of search region division in Henan Province. The initial search region S_0 is constructed from one point in Henan Province, and then the peripheral sub-region expansion is started from S_0. Accordingly, we obtain sub-regions S_1, S_2, S_3, and S_4. The intersection of the four sub-regions and the territory of Henan Province (A) is not empty. Hence, the four sub-regions are added to the ordered set of search region $\mathcal{T}$, i.e., $\mathcal{T}=\{S_0, S_1, S_2, S_3, S_4\}$. We then obtain four new sub-region S_1', S_2', S_3', and S_4', when we expand the peripheral sub-regions of S_1. The intersection of the four sub-regions and A is not empty; S_2', is originally S_0, which is already in $\mathcal{T}$. Hence, we just add S_1', S_3', and S_4' to $\mathcal{T}$. Then, we have $\mathcal{T}=\{S_0, S_1, S_2, S_3, S_4, S_1', S_3', S_4'\}$.

Step 2: Recursive radar search

Step 2.1: For any sub-region $S_k \epsilon \mathcal{T}$ ($k \in N^+$, $k \leq |\mathcal{T}|$), the radar search service of an online map is called with the radar search center R_c and radius R_r equal to S_k and R, respectively, to obtain organization ID set $\mathcal{D}$. If $|\mathcal{D}|=M \wedge R_r>100$m, the algorithm proceeds to Step 2.2; otherwise, the organization ID in $\mathcal{D}$ is added to $\mathcal{O}_k$, the organization ID set of S_k. Where M refers to the upper bound of the number of organization IDs that can be returned by the radar search.

Step 2.2: As shown in Figure 3, the current region is divided into the four equal square sub-regions, denoted as S_1'', S_2'', S_3'', and S_4''.

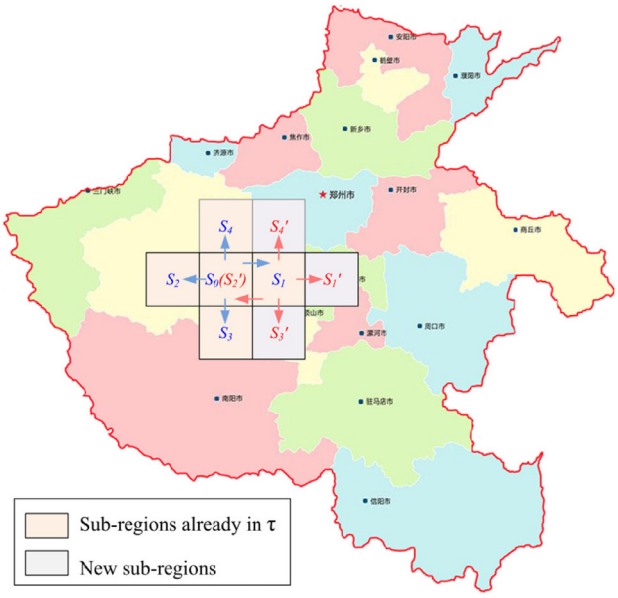


Figure 2. Search region division

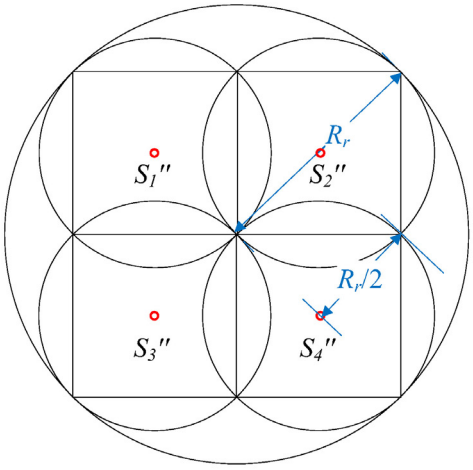


Figure 3. Sub-region division

Step 2.3: For $\forall S'' \epsilon \{S_1'', S_2'', S_3'', S_4''\}$, the radar search service of online map is called with R_c as the geometric center of $S''$and $R_r=R_{r/2}$ to yield the final organization ID set $\mathcal{D}$.

Step 2.4: The organization IDs in $\mathcal{D}$ are added to $\mathcal{O}_k$ and we remove any duplicates, when $|\mathcal{D}|<M \vee R_r \geq 100$m. Otherwise, Step 2.2 is executed again.

Step 2.5: After all the sub-regions in $\mathcal{T}$ have been traversed, the organization IDs of all sub-regions are combined and duplicates are removed to obtain organization ID set $\mathcal{O}$, i.e., $\mathcal{O}=\cup_{k \in N\ k \leq |\mathcal{T}|} \mathcal{O}_k$.

Step 3: Candidate landmark construction

Step 3.1: Landmark information acquisition. For $\forall d' \epsilon \mathcal{O}$, the online map service is called to acquire the domain name, latitude, and longitude of the corresponding organization of d'. These data are added to landmark information set $\mathcal{F}$, if they are complete. Finally, landmark information set $\mathcal{F}=\{<dom_i, loc_i>|i \in N^+\}$ corresponding to all organization IDs in $\mathcal{O}$ is obtained, where dom_i refers to the domain name of

the i-th landmark and loc_i indicates the latitude and longitude of the i-th landmark.

Step 3.2: IP address resolution. For $\forall<dom_i, loc_i> \epsilon \mathcal{F}$, the IP address list $\{ip_1, ip_2, ..., ip_k\}$ of dom_i is acquired through a DNS request. If k=1, the single IP address and loc_i are associated to construct a candidate landmark that is added to candidate landmark set $\mathcal{L}_c$, i.e., $\mathcal{L}_c=\mathcal{L}_c \cup \{<ip_1, loc_i>\}$. If k>1, k candidate landmarks are constructed by associating every IP address and loc_i, and then adding them to $\mathcal{L}_c$, i.e., $\mathcal{L}_c=\mathcal{L}_c \cup \{<ip_1, loc_i>, <ip_2, loc_i>, ..., <ip_k, loc_i>\}$. This yields candidate landmark set $\mathcal{L}_c$.

Step 4: Landmark reliability evaluation

Step 4.1: Preliminary screening. For any candidate landmark $<ip_j, loc_k> \in \mathcal{L}_c$, the city in which ip_j is located is searched using multiple IP geolocation databases, and the corresponding city list $\{city_1, city_2, ..., city_n\}$ is obtained. Then, the city with the most occurrences (say, $city_v$) in the list is returned using majority voting. If loc_k is in $city_v$, Step 4.2 is executed. Otherwise, $<ip_j, loc_k>$ is deleted, i.e., $\mathcal{L}_c=\mathcal{L}_c \setminus \{<ip_j, loc_k>\}$.

Step 4.2: Reliability evaluation. (1) SLG geolocation. For any candidate landmark $<ip_j, loc_k> \in \mathcal{L}_c$, ip_j is set to be the geolocation target, and IP addresses with known locations in the same city are used as the reference. The candidate landmark is located by the SLG algorithm [16] and the geolocation result is denoted as loc_{SLG}. (2) Reliable landmark evaluation. Threshold τ is set according to the IP geolocation precision requirement, and then the distance between the geolocation result and the declared location is calculated, that is, $d_g=dis(loc_{SLG}, loc_k)$. The candidate landmark is added to reliable landmark set $\mathcal{L}$ i.e., $\mathcal{L}=\mathcal{L} \cup \{<ip_j, loc_k>\}$, when $d_g<\tau$. Finally, the reliable landmark set $\mathcal{L}$ is returned.

3 Analysis of the RSLM Algorithm

In this section, we analyze the effectiveness and advantages of the RSLM algorithm by describing the rationale behind dividing the search region to obtain candidate landmarks. We also analyze the effectiveness of landmark evaluation.

3.1 Rationale for the Divide-and-search Approach

For convenience, a large number of organizations deploy web servers in their own company, so the location of the IP address of their web server should identify their geographical location. An online map records website domain names, the geographical locations of organizations, and provides rich resources for landmark mining. CLMA uses the text search service of an online map to search each administrative region of a target region to obtain candidate landmarks. A text search service is provided through an application programming interface (API) by an online map. In this service, the district(county) level administrative region is the input and the organization IDs in that region are the output. Every search with the same input will get the same output, and the upper bound of the number of results is N, so we can only obtain a very limited number of landmarks (the top N) using text search.

In the search region division step, the RSLM algorithm divides the target administrative region into square sub-regions of width D and performs a recursive radar search in each sub-region to obtain candidate landmarks. A radar search service is also an API provided by an online map, with the center and radius of a circular area as the input and the IDs of organizations in the area as the output. We can search fora point of interest in a given circular area of an online map using radar search and obtain at most M organization IDs in the area. For some sub-regions with high organization density, the results of radar search may exceed the upper bound, so RSLM algorithm adopts recursive radar search to recursively segment the sub-region into much smaller sub-regions. Hence, we can then perform a radar search with a smaller radius to ensure that all candidate landmarks in areas with high organization density area are obtained.

Taking Google Maps[10] as an example, the upper bound N of the number of organizations returned by text search currently is 60, which means the CLMA method can only obtain a maximum N=60 candidate landmarks for each administrative region (district/ county). In contrast, the minimum radius of a radar search currently is 0.1 km and the maximum number M of organizations that can be returned is 60 at present. In fact, the number of organizations obtained within the range corresponding to the minimum radar search radius (currently 0.01km^2) will not exceed the upper limit M (currently 60) in most cases in online maps, so the RSLM method can obtain the information of almost all organizations within each administrative region (district/county).

3.2 Landmark Evaluation Effectiveness

In landmark reliability evaluation, we perform a preliminary screening of landmarks at city level and then use the SLG algorithm to locate the candidate landmark with IP addresses that have accurate geographical locations. Finally, we regard candidate landmarks that meet $d_g=dis(loc_{SLG}, loc_k) <\tau$ as reliable landmarks. In this section, we analyze the effectiveness of the landmark evaluation algorithm.

Using the SLG algorithm, we can obtain the possible location loc_{SLG} of a candidate landmark T and its geolocation error e. This determines the location range

[10] https://developers.google.com/maps/docuumentation/javascript/places?hl=zh-cn#rada r_search_requests.

of candidate landmark T, which is a circle with center loc_{SLG} and radius e. According to the distance d_g between the locations given by the candidate landmarks and the locations acquired by SLG, we can determine the location error E_T of reliable landmarks, that is, $E_T \leq e+d_g \leq e+\tau$, as shown in Figure 4. The geolocation errors of the SLG geolocation algorithm on a residential network and online map data set are 2.25 km and 2.11 km, respectively (i.e., the landmark location error e is 2.25 km — the bigger value). If the threshold of location error τ is set to2.5 km, then the error of a reliable landmark is upper bounded by 4.75 km, which can meet the demands of street-level landmark geolocation (10 km in most cases).

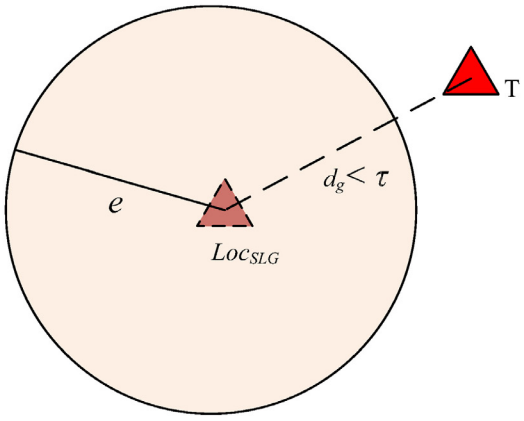


Figure 4. Geolocation error of landmarks

In contrast, to evaluate landmark reliability, CLMA mainly relies on whether the IP address and the domain name of a candidate landmark can be used to open the same webpage. Candidate landmarks obtained from shared hosting or cloud services as well as landmarks with accurate locations that cannot be accessed by IP addresses may be mistakenly deleted. Meanwhile, candidate landmarks from CDN networks and server hosts might be incorrectly selected as reliable landmarks. RSLM uses the geolocation algorithm to evaluate the reliability of candidate landmarks directly; therefore, the resulting landmark location is relatively accurate. Hence, RSLM should be able to correct some landmarks that are incorrectly evaluated by CLMA.

4 Experiments and Results

In the experiments in this study, we used a VPS server (CPU: Intel(R) Core(TM) E5 2620, memory: 8G, hard disk: 500G) deployed in Los Angeles, USA, to implement our algorithm. In order to verify the effectiveness of our method RSLM, we conducted two groups of comparative experiments. In the first group, RSLM is compared with the existing online maps-based method CLMA to verify which one of radar and text search services of the same online map is better. In the second group, RSLM is compared with the other two existing street-level landmark mining methods based on other web resources, the yellow pages-based landmark mining method (YPLM) and the services ports SVM classifier-based landmark mining method (SVMLM), to verify the advantage of mining landmarks based on online maps. In both groups of experiments, all the online maps-based methods mined candidate landmarks by calling the Google Maps API, and SLG algorithm is used to conduct IP geolocation to verify the effectiveness of reliable landmarks mined by all the comparative methods used as the reference points.

4.1 Landmark Mining Experiment Based on Online Maps

We conducted landmark mining and evaluation experiments for Taiwan and Hong Kong using CLMA and RSLM, then analyzed the number and distribution of landmark as well as erroneous landmark correction. Then, SLG geolocation experiments were carried out with the reliable landmarks used as the reference points.

4.1.1 Landmark Mining Results

RSLM was performed for Taiwan with an initial radar search radius R=5km and for Hong Kong with R=2km. All 53 organization types provided by Google Maps, including universities, hospitals, hotels, and government departments, were retrieved. First, we acquired all the organization IDs for both regions using the radar search service of Google Maps. Second, we built candidate landmarks by acquiring organization information and parsing the IP addresses of domains, and then excluded landmarks without domain names or invalid ones with IP addresses that could not be resolved. Third, we performed preliminary screening on candidate landmarks using majority voting with IP2location[11], Maxmind[12], DBIP[13], Baidu[14], IPIP[15], and only retained candidate landmarks with city information that was consistent with the information provided by these databases. Finally, we located the candidate landmarks using the SLG algorithm, and retained reliable landmarks with an upper error threshold τ=2.5km. The numbers of candidate and reliable landmarks obtained are listed according to city in Table 1.

We performed CLMA in Taiwan and Hongkong with the administrative divisions information collected from Baidu Baike, for which the smallest granularity is district (county). The number of candidate and reliable landmarks obtained by CLMA are listed by city in Table 2.

[11] https://www.ip2location.com/
[12] https://www.maxmind.com/
[13] https://db-ip.com/db/
[14] http://lbsyun.baidu.com/index.php?title=webapi/ip-api
[15] https://www.ipip.net/ip.html

Table 1. Number of landmarks obtained by RSLM for each city in the target regions

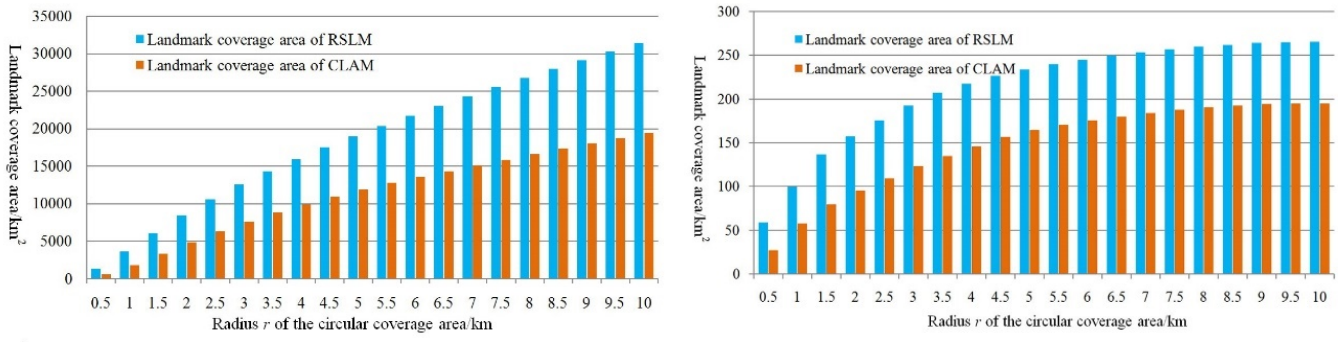
City	Candidate landmark	Reliable landmark
Hong Kong	19,375	1,891
Taibei	25,059	1,842
Jilong	4,098	356
NewTaibei	21,712	1,105
Liangjiang	1,192	75
Yilan	1,560	43
Xinzhu(D)	4,613	146
Taoyuan	11,129	927
Miaoli	5,096	293
Xinzu	2,824	194
Taizhong	18,885	1209
Zhanghua	4,250	662
Nantou	1,309	32
Jiayi	2,416	122
Jiayi (D)	5,399	140
Yunlin	4,114	434
Tainan	17,474	354
Gaoxiong	23,900	652
Penghu	1,249	242
Jinmen	834	93
Pingdong	7,116	251
Taidong	8,264	260
Hualian	7,522	133

Table 2. Number of landmarks obtained by CLMA for each city in the target regions

City	Candidate landmark	Reliable landmark
Hong Kong	3,362	323
Taibei	4,127	487
Jilong	3,160	164
NewTaibei	3,233	46
Liangjiang	182	12
Yilan	678	5
Xinzhu(D)	3,295	83
Taoyuan	3,532	437
Miaoli	4,280	70
Xinzu	606	84
Taizhong	4,015	636
Zhanghua	3,621	413
Nantou	690	6
Jiayi	591	53
Jiayi (D)	4,278	85
Yunlin	3,433	384
Tainan	2,383	34
Gaoxiong	5,897	451
Penghu	814	120
Jinmen	143	28
Pingdong	3,197	85
Taidong	1,991	51
Hualian	2,775	46

Table 1 and Table 2 show that RSLM obtains 18,010 candidate landmarks and 9,565 reliable landmarks for the cities in Taiwan as well as 19,375 candidate landmarks and 1,891 reliable landmarks in Hong Kong. In contrast, CLMA obtained 56,921 candidate landmarks and 3,780 reliable landmarks for the cities in Taiwan and 3,432 candidate landmarks and 323 reliable landmarks in Hong Kong. Hence, the number of candidate landmarks obtained by RSLM for Taiwan and Hong Kong are 3.4 and 5.9 times higher, respectively than those obtained by CLMA. Moreover, the number of reliable landmarks obtained by RSLM for Taiwan and Hong Kong are 2.53 and 5.1 times of those obtained by CLMA respectively. These results demonstrate that (as described in Section 3.1) RSLM's recursive radar search obtains more candidate landmarks. Accordingly, the probability of obtaining more reliable landmarks is increased.

The distribution of landmarks determines the area that can support IP geolocation and is simply referred to as the landmark coverage area. More widely distributed landmarks can support effective IP geolocation over a larger area. Using the reliable landmarks of Taiwan and Hong Kong listed in Table 1 and Table 2, we show the overall coverage area with respect to the circular coverage area of radius r around each landmark in Figure 5.



(a) Taiwan (b) Hongkong

Figure 5. Overall landmark coverage area of RSLM and CLMA with respect to individual landmark coverage

Figure 5 shows that the reliable landmark coverage area obtained by RSLM is much larger than that obtained by CLMA (Taiwan>59%, Hong Kong>36%) for different values of *r*. This indicates that the landmark distribution of RSLM is more reasonable, and thereby more conducive to IP geolocation.

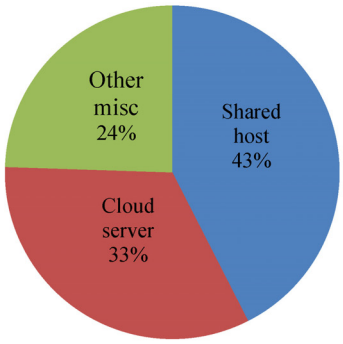
4.1.2 Landmark Correction Results

We evaluated candidate landmarks obtained by CLMA in Taiwan and Hong Kong using the landmark reliability evaluation step of RSLM, and then counted the number of unreliable landmarks selected by CLMA (referred to as mis-selected landmarks) and the landmarks deleted by CLMA that were reliable (referred to as mis-deleted landmarks). The result is shown in Table 3.

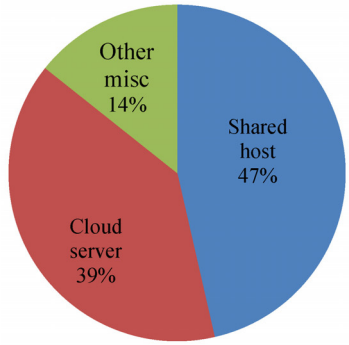
Table 3. Incorrect CLMA landmarks

City	Reliable landmarks	Mis-deleted landmarks	Mis-selected landmarks
Taiwan	3780	294	421
Hong Kong	323	28	36

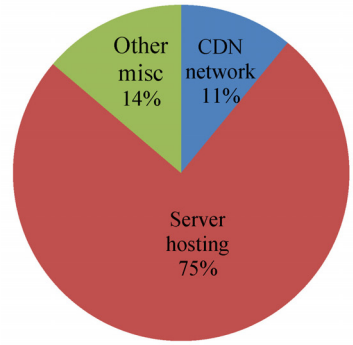
Figure 6 shows the types of mis-deleted and mis-selected landmarks obtained by CLMA.



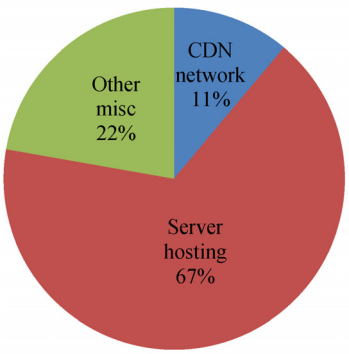
(a) Mis-deleted landmarks in Taiwan



(b) Mis-deleted landmarks in Hong Kong



(c) Mis-selected landmarks in Taiwan



(d) Mis-selected landmarks in Hong Kong

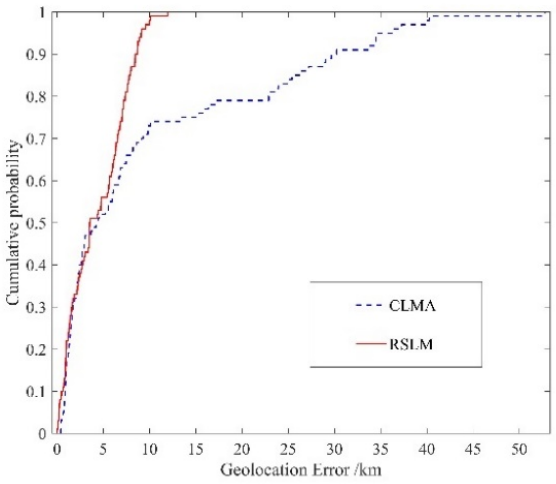
Figure 6. Types of incorrect CLMA landmarks

As Table 3 and Figure 6 show, RSLM corrects many mis-deleted landmarks of CLMA, such as those for shared host and cloud servers, as well as some mis-selected landmarks, such as CDN network and server hosted landmarks. This demonstrates that the conclusion of the analysis in Section 3.2, that RSLM selects reliable landmarks with errors within the acceptable range from the perspective of SLG geolocation algorithm and can effectively correct some of the Mis-deleted and Mis-selected landmarks of CLMA.

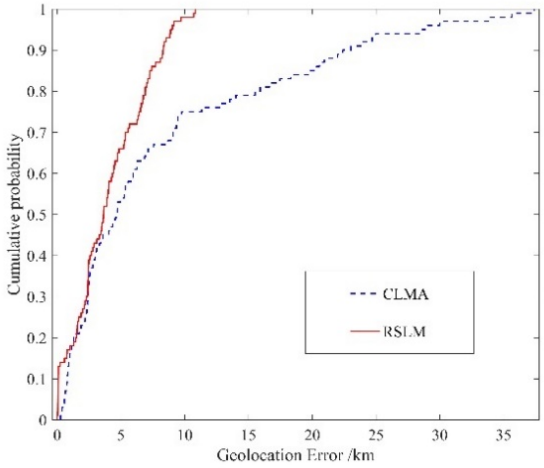
4.1.3 Landmark Geolocation Results

We choose 100 IP addresses with known geographical locations as geolocation targets in Taiwan and Hong Kong and then performed SLG geolocation algorithm using the reliable landmarks obtained by each of the two algorithms. The cumulative probability of geolocation errors is shown in Figure 7.

We can infer from the results in Figure 7 that the average error of landmarks mined by the RSLM algorithm (Hong Kong 4.03 km, Taiwan 4.33 km) is far below that of CLMA (Hong Kong 8.28 km, Taiwan 9.94 km). In addition, RSLM yields a much smaller maximum error (RSLM: Hong Kong 10.85 km, Taiwan 11.94 km; CLMA: Hong Kong 37.42 km, Taiwan 52.79 km). Correspondingly, the geolocation based on landmarks obtained by RSLM is better than that based on CLMA, which indicates that landmarks obtained by RSLM are more reliable. This is a result of the higher number of landmarks and wider landmark distribution of RSLM.



(a) Taiwan



(b) Hongkong

Figure 7. Geolocation error of reliable landmarks

4.2 Landmark Mining Experiment Based on Different Web Resources

We conducted landmark mining and evaluation experiments for six cities in China and America using RSLM, YPLM and SVMLM, then analyzed the number of landmarks excavated and the geolocation effectiveness of reliable landmarks. In order to avoid the influence of different landmark assessment methods on the quantity and quality of reliable landmarks mined, SLG algorithm, which is the same as in RSLM and YPLM, is also used to identify reliable landmarks in the landmark assessment step for SVMLM.

4.2.1 Landmark Mining Results

The numbers of candidate and reliable landmarks obtained by RSLM, YPLM and SVMLM were listed in Table 4. It took 7 days and 8 hours to obtain the landmarks in the 6 cities using the proposed algorithm, and took 9 days and 1 hours using YPLM algorithm, and took 47 days and 17 hours when using SVMLM algorithm.

As we can see in Table 4, the proposed algorithm RSLM can obtain more candidate landmarks as well as reliable landmarks identified by SLG algorithm.

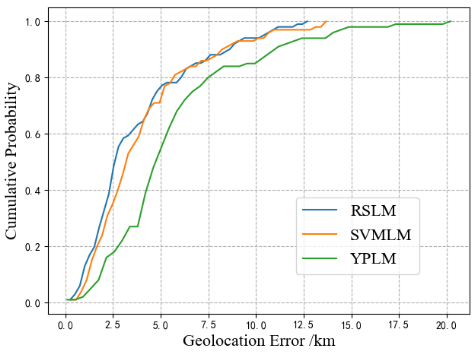
Table 4. Number of landmarks obtained by RSLM YPLM and SVMLM from 6 cities in China and America

	City	Beijing	Zhengzhou	Hongkong	Taibei	SanDigo	Honolulu
RSLM	Candidate landmarks	230393	38695	19375	25059	36248	10997
	Reliable landmarks	6658	721	1891	2242	631	406
YPLM	Candidate landmarks	5699	5385	3258	2802	10758	8592
	Reliable landmarks	1556	342	612	580	407	325
SVMLM	Candidate landmarks	102156	7111	17331	114803	3299	1251
	Reliable landmarks	2351	464	907	821	103	82

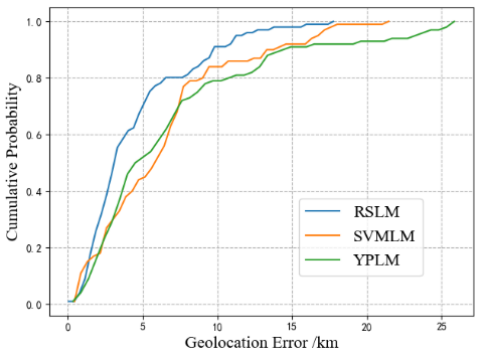
4.2.2 Landmark Geolocation Results

For the purpose of verifying the geolocation effect, we located the IP addresses with known location by SLG algorithm with the landmarks gained by RSLM, YPLM and SVMLM. Firstly, we select 100 IP addresses with known locations as the geolocation target in the 6 cities respectively. Secondly, we use these reliable landmarks obtained by three algorithms as the reference points of the SLG algorithm, and locate the given 100 IP addresses with known locations in these cities. Finally, the cumulative probabilities of geolocation errors for three methods in different cities are shown in Figure 8. The abscissa in the figure is the geolocation error, and the ordinate is the cumulative probability.

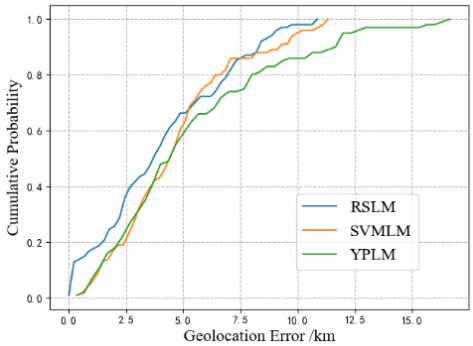
As can be seem in Figure 8, the mean error of geolocation results for the given target 100 IP address in Beijing, Zhengzhou, Hongkong, Taibei, SanDigo and Honolulu is 3.72 km, 4.39 km, 4.03 km, 3.58 km, 4.30 km and 3.88 km respectively, which is clearly lower than that of YPLM (5.59 km, 6.82 km, 5.30 km, 5.42 km, 4.68 km and 4.35 km respectively) and SVMLM (4.02 km, 6.22 km, 4.63 km, 4.82 km, 5.22 km and 4.96 km respectively). It verified that the street-level landmarks obtained by the proposed



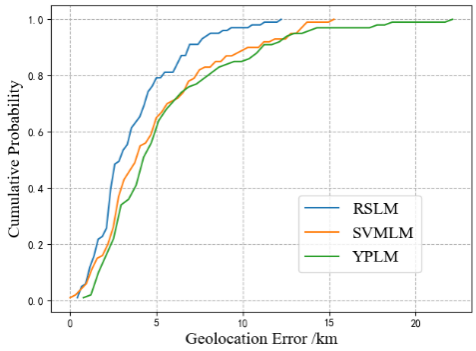
(a) Beijing



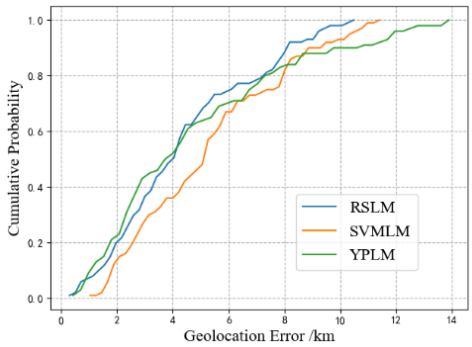
(b) Zhengzhou



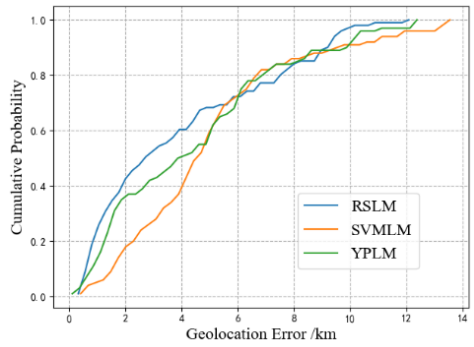
(c) Hongkong



(d) Taibei



(e) SanDigo



(f) Honolulu

Figure 8. Geolocation errors of reliable landmarks

algorithm RSLM are of higher reliability than the state-of-art algorithms.

5 Conclusion

This paper proposed RSLM, a street-level landmark mining method based on online map radar search. First, RSLM divides the target administrative region into square sub-regions of the same size, and then performs recursive radar search to obtain all the organization IDs using the API of an online map. Second, candidate landmarks are constructed using the landmark information acquired by the organization ID. Finally, unreliable landmarks are filtered out using preliminary screening and SLG. Experiments in Taiwan and Hongkong demonstrate that the algorithm outperforms the existing online map-based approach with respect to landmark quantity, distribution, and reliability. However, this algorithm is only applicable to mining web server-type landmarks. Its effectiveness in the underdeveloped areas of the Internet would not be as obvious. In the future, we will focus on the mining of terminal-type landmarks.

References

[1] J. Taylor, J. Devlin, K. Curran, Bringing location to IP Addresses with IP Geolocation, *Journal of Emerging Technologies in Web Intelligence*, Vol. 4, No. 3, pp. 273-277, August, 2012.

[2] J. A. Muir, P. C. V. Oorschot, Internet geolocation, *ACM*

Computing Surveys, Vol. 42, No. 1, pp. 1-23, December, 2009.

[3] C. Huang, D. A. Maltz, J. Li, A. Greenberg, Public DNS system and Global Traffic Management, *Proceedings IEEE INFOCOM*, Shanghai, China, 2011, pp. 2615-2623.

[4] E. Katz-Bassett, J. P. John, A. Krishnamurthy, D. Wetherall, T. Anderson, Y. Chawathe, Towards IP geolocation using delay and topology measurements, *Proceedings of the 6th ACM SIGCOMM on Internet Measurement – IMC'06*, Rio de Janeriro, Brazil, 2006, pp. 71-84.

[5] D. Moore, R. Periakaruppan, J. Donohoe, K. Claffy, Where in the world is netgeo.caida.org? *International Networking Conference (INET)*, Yokohama, Japan, 2000.

[6] B. Du, M. Candela, B. Huffaker, A. C. Snoeren, K. Claffy, RIPE IPmap active geolocation: mechanism and performance evaluation, *SIGCOMM Computer Communication Review*, Vol. 50, No. 2, pp. 3-10, April, 2020.

[7] Y. Shavitt, N. Zilberman, A Geolocation Databases Study, *IEEE Journal on Selected Areas in Communications*, Vol. 29, No. 10, pp. 2044-2056, December, 2011.

[8] I. Poese, S. Uhlig, M. A. Kâafar, B. Donnet, B. Gueye, IP Geolocation Databases: Unreliable? *ACM SIGCOMM Computer Communication Review*, Vol. 41, No. 2, pp. 53-56, April, 2011.

[9] M. Gharaibeh, A. Shah, B. Huffaker, H. Zhang, R. Ensafi, C. Papadopoulos, A Look at Router Geolocation in Public and Commercial Databases, *ACM Internet Measurement Conference*, London, UK, 2017, pp. 463-469.

[10] O. Dan, V. Parikh, B. D. Davison, Improving IP Geolocation Using Query Logs, *Proceedings of the 9th annual ACM International Conference on Web Search and Data Mining*, San Francisco, USA, 2016, pp. 347-356.

[11] H. Liu, Y. Zhang, Y. Zhou, D. Zhang, X. Fu, K. K. Ramakrishnan, Mining checkins from location-sharing services for client-independent IP geolocation, *IEEE Conference on Computer Communications*, Toronto, ON, Canada, 2014, pp. 619-627.

[12] B. Huffaker, M. Fomenkov, K. Claffy, DRoP: DNS-based Router Positioning, *ACM SIGCOMM Computer Communication Review*, Vol. 44, No. 3, pp. 5-13, July, 2014.

[13] O. Dan, V. Parikh, B. D. Davison, Distributed Reverse DNS Geolocation, *Proceedings of the IEEE International Conference on Big Data*, Seattle, WA, USA, 2018, pp. 1581-1586.

[14] Q. Scheitle, O. Gasser, P. Sattler, G. Carle, HLOC: Hints-Based Geolocation Leveraging Multiple Measurement Frameworks, *Proceedings of the IEEE International Conference on Network Traffic Measurement and Analysis*, Dublin, Ireland, 2017, pp. 1-9.

[15] V. N. Padmanabhan, L. Subramanian, An investigation of geographic mapping techniques for internet hosts, *ACM SIGCOMM Computer Communication Review*, Vol. 31, No. 4, pp. 173-185, August, 2001.

[16] B. Gueye, A. Ziviani, M. Crovella, S. Fdida, Constraint-based geolocation of internet hosts, *Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement – IMC'04*, Sicily, Italy, 2004, pp. 288-293.

[17] Y. Wang, D. Burgener, M. Flores, A. Kuzmanovic, C. Huang, Towards Street-Level Client-Independent IP Geolocation, *8th USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, Boston, MA, USA, 2011, pp. 365-379.

[18] H. Jiang, Y. Liu, J. N. Matthews, IP geolocation estimation using neural networks with stable landmarks, *IEEE Conference on Computer Communications Workshops*, San Francisco, California, USA, 2016, pp. 170-175.

[19] M. Candela, E. Gregori, V. Luconi, A. Vecchio, Using RIPE Atlas for Geolocating IP Infrastructure, *IEEE Access*, Vol. 7, pp. 48816-48829, April, 2019.

[20] Z. Weinberg, S. Cho, N. Christin, V. Sekar, P. Gill, How to Catch when Proxies Lie: Verifying the Physical Locations of Network Proxies with Active Geolocation, *In Proceedings of the Internet Measurement Conference (IMC'18)*, *Association for Computing Machinery,* Boston, MA, USA, 2018, pp. 203-217.

[21] R. Fanou, G. Tyson, E. L. Fernandes, P. Francois, F. Valera, A. Sathiaseelan, Exploring and Analysing the African Web Ecosystem, *ACM Transactions on The Web*, Vol. 12, No. 4, pp. 1-26, November, 2018.

[22] Z. Wang, Y. Chen, H. Wen, L. Zhao, L. Sun, Discovering Routers as Secondary Landmarks for Accurate IP Geolocation, *IEEE 86th Vehicular Technology Conference (VTC-Fall)*, Toronto, Canada, 2017, pp. 1-5.

[23] J. Chen, F. Liu, X. Luo, F. Zhao, G. Zhu, A Landmark Calibration-Based IP Geolocation Approach, *EURASIP Journal on Information Security*, Vol. 2016, Article No. 4, pp. 1-11, January, 2016.

[24] D. Komosny, M. Voznak, S. Ur Rehman, Location Accuracy of Commercial IP Address Geolocation Databases, *Information Technology and Control*, Vol. 46, No. 3, pp. 333-344, September, 2017.

[25] C. Guo, Y. Liu, W. Shen, H. J. Wang, Q. Yu, Y. Zhang, Mining the Web and the Internet for Accurate IP Address Geolocations, *INFOCOM 2009: IEEE 28th Conference on Computer Communications*, Rio de Janeiro, Brazil, 2009, pp. 2841-2845.

[26] G. Zhu, X. Luo, F. Liu, J. Chen, An Algorithm of City-Level Landmark Mining Based on Internet Forum, *18th International Conference on Network-Based Information Systems*, Taipei, Taiwan, 2015, pp. 294-301.

[27] T. Ma, F. Liu, X. Luo, M. Yin, R. Li, An Algorithm of Street-Level Landmark Obtaining Based on Yellow Pages, *Journal of Internet Technology*, Vol. 20, No. 5, pp. 1415-1428, September, 2019.

[28] R. Li, Y. Liu, Y. Qiao, T. Ma, B. Wang, X. Luo, Street-Level Landmarks Acquisition Based on SVM Classifiers, *Computers Materials & Continua*, Vol. 59, No. 2, pp. 591-606, January, 2019.

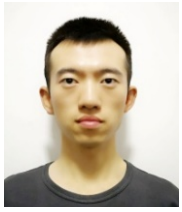
Biographies

Xiaonan Liu was born in Liaoning Province, China. He was conferred a Ph.D. in computer science by State Key Laboratory of Mathematical Engineering and Advanced Computing, China, in 2016. And now he is an associate professor of the laboratory. His research interests include binary translation, quantum computing and information security.

Wen Yang was born in Henan Province, China at June, 1986. He received his B.S. degree in information security from National University of Defense Technology, China, in 2008, where he is currently pursuing a master's degree. His research interests include network entity geolocation, data analysis, and information security.

Meijuan Yin was born in Anhui Province, China, 1977. She was conferred a Ph.D. in computer application by State Key Laboratory of Mathematical Engineering and Advanced Computing, China, in 2012. She is now an associate professor of the laboratory. Her current research interests include data mining, social network analysis, and information security.

Fenlin Liu received the Ph.D. degree from Northeast University, in 1998. He is currently a Professor in State Key Laboratory of Mathematical Engineering and Advanced Computing. His research interests include network topology and network geolocation. He received the Found of Innovation Scientists and Technicians Outstanding Talents of Henan Province of China.



Chenshu Yun received the B.S. degree from State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou, China, in 2018. His research interest includes network security and network geolocation.

Appendix

The following is the radar search sampling experiment based on Google Maps.

For each kind of organization (hotel, government restaurant), we conduct the sampling experiment on six cities in China and America, including Beijing, Zhengzhou, Taipei, Hong Kong, Honolulu and San Diego. The number of organizations of each kind can be got by calling the radar search API of Google Maps with sampling points as parameters. The details steps of the experiment are as follows.

Step 1: Using Geocoding API, the longitude and latitude of the northeast corner and southwest corner of the administrative region of a city can be obtained according to its name, the rectangle with the line segment formed by which as the diagonal contains the city. Then the sampling of the city can be carried out in the rectangle.

Step 2: Randomly select the longitude of n (n is set as 30 in our experiments) non-repeating points on the edge parallel to the latitude line in the rectangle and the latitude of n non-repeating points on the edge parallel to the longitude line in the rectangle. Thus n*n non-repeating sampling points in the rectangle can be got through combining the longitudes and latitudes obtained in the previous step.

Step 3: The number of organizations of each sampling point can be obtained by calling radar search API, with search radius set as the minimum value (currently is 0.1 km), each kind institution of three types and the latitude and longitude of the sampling point as parameters.

The number of organizations returned by radar search sampling experiments with different organization types in each city was counted, and the results are shown in below table.

Table 1. Statistical results of organization number of sampling point

Org. Type	City Name	Statistical results of organization number			
		Mode	Mean	Maximum	Proportion of M (60) in all numbers
Hot.	Honolulu	0	3.86	60	0.010
	San Diego	0	2.81	60	0.004
	Beijing	2	3.75	60	0.016
	Zhengzhou	2	3.20	60	0.006
	Taipei	0	6.60	60	0.052
	Hong Kong	1	6.02	60	0.051
Gov.	Honolulu	0	2.78	60	0.009
	San Diego	2	3.17	60	0.007
	Beijing	2	3.64	60	0.018
	Zhengzhou	2	3.04	60	0.003
	Taipei	0	7.94	60	0.062
	Hong Kong	1	3.51	60	0.017
Res.	Honolulu	0	0.17	33	0
	San Diego	0	0.08	7	0
	Beijing	0	0.14	27	0
	Zhengzhou	0	0.10	19	0
	Taipei	0	0.62	40	0
	Hong Kong	0	0.23	53	0

It can be seen from Table 1 in the appendix, that the number of organizations returned by radar search in all samples in the six cities is very small (no more than 2) in most cases, and the number in the average case does not exceed 8. For two high-density organizations hotel and government, in most of six cities only about 1% cases hit the upper limit of M (currently 60), while in Hong Kong and Taipei with very high-density organizations still only about 5% cases hit the upper limit M. For restaurant with a small density, all cases in all the six cities did not reach the upper limit of M.

From the above, we can conclude that taking Google Maps as an example, the number of organizations obtained within the range corresponding to the minimum radar search radius (currently 0.01 km^2) will not exceed the upper limit M (currently 60) of the organizations number returned by radar search in most cases in online maps.