Hindawi
Security and Communication Networks
Volume 2018, Article ID 2507293, 12 pages
https://doi.org/10.1155/2018/2507293

Research Article

Street-Level Landmark Evaluation Based on Nearest Routers

Ruixiang Li, Yuchen Sun, Jianwei Hu, Te Ma, and Xiangyang Luo

State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou 450001, China

Correspondence should be addressed to Xiangyang Luo; luoxy_ieu@sina.com

Received 10 May 2018; Accepted 5 July 2018; Published 18 July 2018

Academic Editor: Lianyong Qi

High reliable street-level landmarks are the basis of IP geolocation, but landmark evaluation methods having been proposed cannot evaluate the street-level landmarks. Therefore, in this paper, a street-level landmark evaluation method based on nearest router is proposed. The location organization declared is regarded as an area not a point. Firstly, the declared location of preevaluated landmark is verified by IP location databases. Secondly, the preevaluated landmarks are grouped according to their nearest router. Then, the distance constraint is obtained using delay value between landmark and its nearest router by delay-distance correlation. And relation model is established among distance constraint, organization's region radius, and distance between two landmarks. Finally, the reliability value of landmarks is calculated in each group based on relational model and binomial distribution. Landmarks evaluation experiment is taken based on 7082 preevaluated landmarks, and the results show that geolocation errors decrease obviously using evaluated landmarks. The mean error of 100 targets in Shanghai is reduced from 7.832km to 2.185km.

1. Introduction

The Internet of Things is an important part of a new generation of information technology, and it is an extended network based on the Internet. The core supporting technology of the Internet of Things is cloud computing. The cloud computing model enables real-time dynamic management and intelligent analysis of millions of items in the Internet of Things. IP geolocation is a technology to determine the geographical location of IP network entities. It has wide application prospects in the field of Internet of Things, such as ensuring that data flow within the privacy protection in cloud data sharing [1–3], improving the secret security in mobile cloud when secret keys distributed [4, 5], supporting location-based cloud services[6–8], and targeting covert communication subjects [9–11]. Landmark-based IP geolocation is an effective means to determining the region of the Internet entity. High reliable street-level landmarks are the basis of IP geolocation. Existing evaluation methods of landmarks are divided into evaluation of web-based landmarks and evaluation of the IP location databases.

A method, mining landmarks from web, has been proposed [12]. In this method, IP.cn database was used to evaluate the city location of the landmarks, and major Chinese ISP databases were used to assess the province location. On the basis of [12], Jiang H et al. [13] exclude the cloud services landmarks according to cloud service providers' IP address (such as Amazon AWS, Rackspace Hosting, and other top cloud providers), exclude hosting landmarks whose organizations in Whois do not contain specific keywords such as "university", "academy", or "institution", and use MaxMind database to verify landmarks' city location. E-GeoTrack algorithm [14] based on voting strategy and implicit information in nearest router was proposed to evaluate Internet forum landmarks obtained from a web page. Above methods can only evaluate the accuracy of city location, and there are large errors in fine-grained geolocation using evaluated landmarks. Wang Y et al. [15] obtain the web landmarks on the basis of paper [12] and verify the landmarks as follows: (1) if the zip code of landmark is inconsistent with the zip code entered during the query, the landmark will be deleted, (2) visit the website by IP address and domain name respectively, and if the contents, or heads (distinguished by <head> and </head>), or titles (distinguished by <title> and </title>) returned are different, the web site is considered to be hosted on a shared host or

Table 1: Calculating *LWV* of a DNS name.

url \ location	Loc_a	Loc_b	Loc_c	Loc_d
dns_a/url1	0.64	-	0.96	0.89
dns_a/url2	0.64	-	0.95	0.89
dns_a/url3	-	0.57	0.95	0.86
LWV of dns_a	0.43	0.19	0.95	0.88

on CDN ("Content Delivery Network"), and the landmark is deleted, and (3) delete the landmarks whose domains are the same but zip codes are different. This method deletes some hosting, shared hosting, and CDN landmarks. But to street-level geolocation, it is also necessary for evaluation further on this basis.

Different IPs in the same IP block may correspond to widely distributed geographical locations; thus, the location information in the foreign commercial IP location databases is very rough [16, 17]. Based on [16], Backstrom et al. [18] proposed to determine user's location based on their social network and use this location to improve IP location database. Poese et al. [19] studied IP location databases such as MaxMind, IP2LOCATION, IPInfoDB, and HostIP, found that the vast majority of IP data in United States were inconsistent with IP initial allocation data, and pointed that the city-level location of IP in these databases was inaccurate. In [20, 21], the authors analysed the distribution characteristics of IP address blocks of different geographic granularity in mainland China, compared the data consistency between multiple IP location database, verified the accuracy of the IP location database using a small number of existing landmarks, and improved the accuracy of the database by clustering IP blocks at different locations. For improving the accuracy of database, location data collected from search engine logs [22], IP allocation strategy of ISPs, urban areas, and population [23] were introduced into the database evaluation methods. Using those methods improve the accuracy at city level, but there are inadequacies in evaluating fine-grained (e.g., street level) locations.

In the absence of reliable street-level landmarks, there are significant errors in the evaluation of street-level landmarks based on street-level geolocation methods. And there are no other effective street-level landmark evaluation methods currently. In view of this reality, a street-level landmark evaluation method based on the nearest router is proposed. In this paper, the location of the organization is a region, and the radius of region is the maximum value of the distance calculated by latitude and longitude between the organizations' declared location and the location of network entity.

The method includes three steps. Firstly, multiple databases are used to verify the city-level location of the original candidate landmarks obtaining from web pages based on voting strategy. If the zip code information is obtained, it will be further used to verify. Then, route paths to candidate landmarks are obtained by using traceroute commend. Candidate landmarks are grouped based on the nearest router. And the relationship model between radius of organization, geographical distance, and network distance constraints is established within two candidate landmarks in the same group. Combining the binomial distribution, the reliability value of each landmark is calculated by relationship model.

The rest of this paper is organized as follows. The related work is introduced in Section 2. The principles and steps of street-level landmark evaluation method based on nearest routers are elaborated in detail in Section 3. Section 4 briefly analyses the street-level landmark evaluation method. The experimental results are given in Section 5. Finally, this paper is concluded in Section 6.

2. Related Work

Structon [12] is a webpage-based landmark mining method and the key idea is that web pages are embedded with rich geographic information (such as province/state, city, and zip code). The geographic information extracted from web page can be mapped onto the IP address of the web server. As a result, landmarks (network entities whose IP and geographic address have been known) are obtained. The main steps of Structon are as follows.

Firstly, according to HTML tags, each HTML file is parsed into multiple blocks, and each block is treated as a string roughly. For each block, a regular expression is used to extract geographical location information. Web servers for the same domain name are collected, and the location weight vector (*LWV*) for the domain name is calculated as Table 1.

Based on the weights of different urls of dns_a in different locations, the *LWV* values of dns_a in different location are calculated.

Then, the IP address and its corresponding location weight vector are taken as input. The multistage inference algorithm is used to increase the coverage and accuracy of the IP location database. The multistage inference algorithm includes three parts as follows.

Part 1. Location calculation of /24 IP segment: the location probability distribution function (*PDF*) is calculated by the *LWV* of each IP in /24 IP segment. The highest probability location is regarded as the geographical location of all IP in /24 IP segment. An example is shown in Table 2.

As shown in Table 2, the Loc_b is regarded as the IP segment geographical location.

Part 2. Error correction based on majority voting. If most of the IP subsegments have the same geographic location

Table 2: The *PDF* of the segment.

IP \ location	Loc_a	Loc_b	Loc_c	Loc_d	Loc_e
61.155.111.42	0.003	0.004	0.003	0.24	-
61.155.111.44	-	0.02	-	-	-
61.155.111.70	-	0.77	-	-	0.13
Location *PDF*	0.26%	68%	0.26%	20.5%	11%

and the remaining IP subsegments are at other locations, the entire IP segment is assumed to be in the same location.

Part 3. Inference based on AS and BGP information: the BGP routing table shows that some ASs only contain a small number of IP addresses. Therefore, these ASs are small ISPs, and these small ISPs are likely to be located in the same province or city. In these small ASs, if some IP are in the same location L, it can be well inferred that the entire ISP is also located in L.

Finally, IP address location tables of a major Chinese ISP and ip.cn database are used to verify the accuracy of inferred results.

Structon is a method using web page information to obtain landmarks. After location inference, a large amount of landmarks can be obtained. However, the landmarks verification strategies are rough, which only verify the provincial and city-level locations and cannot satisfy the requirements of street-level landmarks. At the same time, the database (such as ip.cn) is not completely accurate. The reliability of the verification results cannot be guaranteed if only using single database. After Structon, the researchers use multiple database to verify the provincial and city-level locations and use zip codes for further verification [13–15]. Those strategies improved the accuracy of landmarks obviously but did not reach the requirement of street-level landmarks evaluation.

Based on [12], a street-level landmarks evaluation method is proposed in this paper. Landmarks are collected from web, and IP location databases were used to verify the data consistency between declared location and databases' results. If the zip codes about original candidate landmarks were got, that would be used to verify the location further. Landmarks after location verification are named candidate landmarks. Then, the route paths from probe to candidate landmarks are got, and in the router path, the minimum single-hop delay from the nearest router to the candidate landmark is obtained. According to the same nearest router, candidate landmarks are divided into some group which can be considered as a set. Finally, in each set, we get distance constraint by delay-distance correlation and delay from nearest router to candidate landmark and calculate the distance between any two candidate landmarks by their latitude and longitude. According to the relationships among distance constraint, distance, and organization's region radius, all subsets satisfying the relationships can be found, and the reliable values of candidate landmarks in each subset can be calculated based on binomial distribution and initial probability. All groups are evaluated, and all landmarks whose reliable value is greater than reliability threshold can be selected out.

3. Methods

Aiming at the deficiencies of existing landmark evaluation methods in evaluating street-level landmarks, a street-level landmark evaluation method based on the nearest router is proposed. The precondition of this method is that if two terminal landmarks have the same nearest router, they are close in geographical location.

The method flowchart is shown in Figure 1. This method is mainly divided into three parts: location verification, landmarks grouping, and group landmarks evaluation. In location verification, candidate landmarks whose locations in multiple IP location databases are consistent with declared city were selected from original candidate landmarks. If zip code of original candidate landmarks is obtained, zip code needs to be used for verification further. In landmarks grouping, route paths to candidate landmarks were measuring many times, and final route path for one candidate landmark was got by merging route paths. The candidate landmarks with the same nearest router are grouped. In each group, if there are at least two landmarks for one institution, the one whose delay from nearest router to candidate landmark is minimal will be reserved. In group landmark evaluation, relationship model among distance constraint, region radius, and distance is established, and all subsets in which any two elements satisfy the relationship model will be found. In each subset, in the basis of initial probability and binomial distribution, the reliability values of candidate landmarks are calculated.

The main steps of the method are as follows:

Input: original candidate landmark and multiple IP location databases

Output: reliable landmarks with reliability value

Step 1 (original candidate landmarks acquisition). According to the landmark acquisition method mentioned in [12], information such as address, server domain, and zip code of companies, universities, and governments is obtained from web pages. Public mail server will be excluded, and server domain and address are converted to IP (IP_i) and latitude (lat_i) and longitude (lng_i), respectively. Original candidate landmark is marked as data pair ($< IP_i, lat_i, lng_i >$). If a domain name corresponds to n IPs, n data pairs will be marked whose IPs are different only. That means an organization may correspond to multiple candidate landmarks.

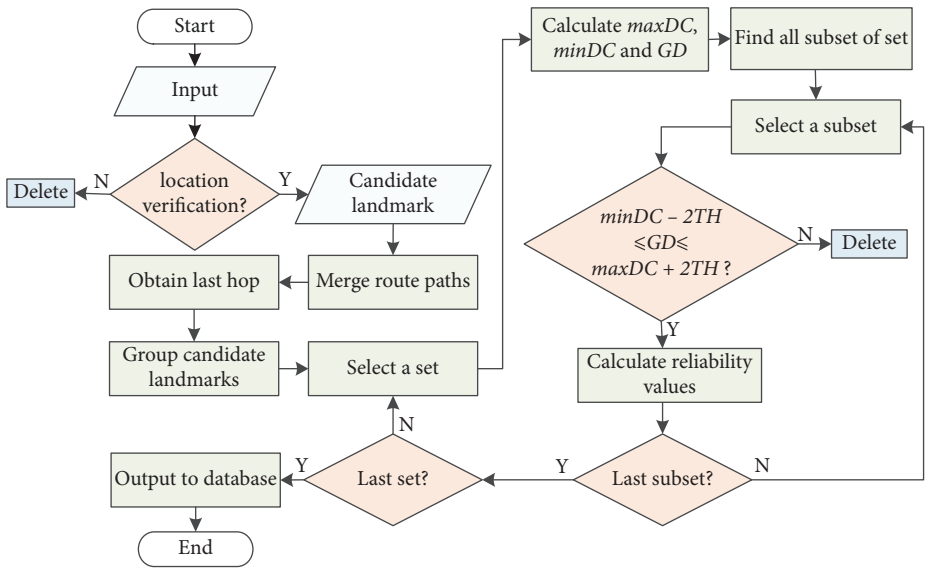


Figure 1: Method flowchart.

Step 2 (location verification). City address of original candidate landmark is searched from IP location databases such as IPIP, IP.cn, and Baidu. If the query results of databases and declared address are the same, the original candidate landmark will be retained. If zip code of original candidate landmark has been obtained, that will be used to verify landmark address further, and original candidate landmark (called "candidate landmark") whose claimed location belongs to the zip code area will be retained.

Step 3 (routing information acquisition). Use traceroute command to get route path from probe to candidate landmarks during small network delay fluctuation period repeatedly, and merge paths to obtain more detailed routing information and the minimum single-hop delay.

Step 4 (candidate landmark groupings). The candidate landmarks were grouped by the nearest common routers. Each group is called an evaluation landmark set. In a set, if multiple candidate landmarks belong to same organization, the one whose single-hop delay to nearest router will be retained, and other candidate landmarks will not participate in the evaluation process further, but the reliability value is the same as the retained one. If single-hop delay value to nearest router is greater than 1ms, the value may be regarded as inaccurate measurement one and the candidate landmark will not be evaluated further.

Step 5 (candidate landmark group evaluation). In an evaluation landmark set, the relation model among distance constraint, distance, and radius of region was built. All subsets satisfying relation model need to be found. This issue can be converted to solve complete subgraph problem in an undirected graph. The construction method of graph is as follows: the elements in the set are mapped to the vertices in the undirected graph. And if two elements in set satisfy the relation model, there is edge between the two vertices. The subset consists of all vertices in one complete subgraph. Obtaining all subsets is our goal. According to the IP allocation strategy of ISPs, the number of vertices of the graph is less than 64 generally. So, the issue can be solved in acceptable time.

In a subset, the reliability (recorded as "p_{re}") of evaluated landmarks is calculated based on initial reliability and binomial distribution.

$$p_{re} = 1 - \prod_{i=1}^{k} (1 - p_i) \tag{1}$$

where k is the number of elements in subset and p_i is the initial reliability value of ith element.

Step 6 (reliable landmark storage). Repeat Step 5 to evaluate all groups. When a candidate landmark appears in multiple subsets and may have multiple reliability values, the final reliability of the landmark evaluated is the maximum value. The reliability threshold is set as α, which means evaluated landmark is reliable when its reliability value is not less than α. And store the reliable landmark into result database.

Then, the process of merging the route paths and building the relation model are explained in detail.

Merging the Route Paths. Use traceroute command to get route path during small network delay fluctuation period repeatedly, and merge route paths. Figure 2 is an example of route paths merging.

The route path from probe (S) to candidate landmark (L) contains four routers (R1, R2, R3, and R4). In the first measurement, the IP addresses of R2, R3, and R4 are

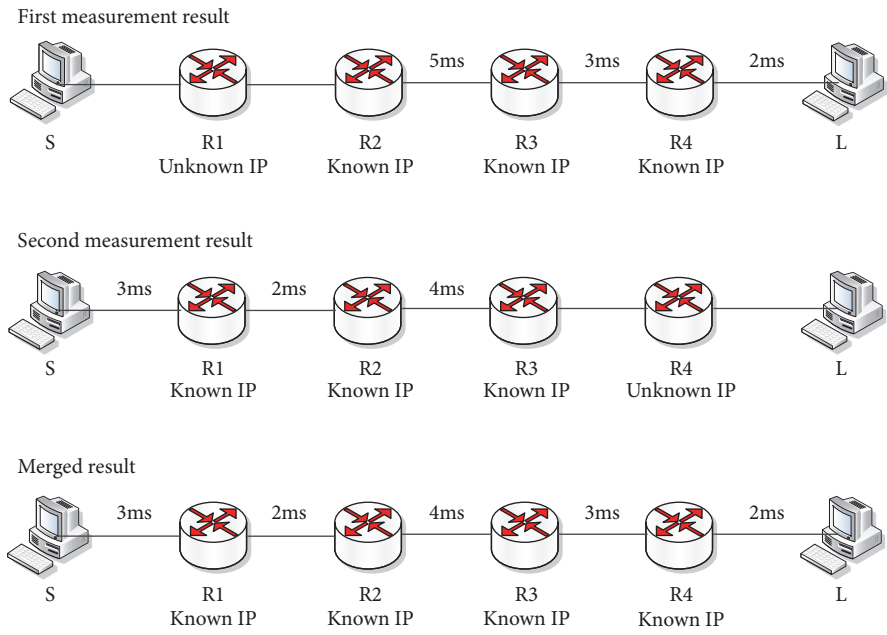


Figure 2: Example of route paths merging.

obtained, and the delay between R2 and R3 is 5ms. And, in the second measurement, the IP addresses of R1, R2, and R3 are obtained, and the delay between R2 and R3 is 4ms. Finally, the merged results are that the IP addresses of four routers are obtained and the delay of R2 and R3 is 4ms.

Building the relation model: for a subset, the geographical distance (GD) for two candidate landmarks L_i and L_j is calculated by latitude and longitude. According to the single-hop delay (t) from nearest router to candidate landmark, the distance constraint (DC) between candidate landmark and nearest router can be got by

$$DC = v * t \tag{2}$$

where v is the propagation velocity of the electromagnetic wave ($v = 20km/ms$) and t is the single-hop delay. Then, the minimum distance constraint (recorded as "min DC") and the maximum distance constraint (recorded as "$maxDC$") between two candidate landmarks can be obtained.

$$\min DC = v * \left|t_1 - t_2\right| = |DC1 - DC2| \tag{3}$$

$$\max DC = v * (t_1 + t_2) = DC1 + DC2 \tag{4}$$

The trilateral relation among GD, min DC, and max DC is established as inequality (5).

$$\begin{aligned} \min DC &\le GD \\ GD &\le \max DC \end{aligned} \tag{5}$$

Adding the radius of region (recorded as "TH"), relation model among distance constraint, region radius, and distance between L_i and L_j is established as inequality (6).

$$\begin{aligned} \min DC - 2TH &\le GD \\ GD &\le \max DC + 2TH \end{aligned} \tag{6}$$

4. Analysis of Method

In this section, the relation model among distance constraint, distance and region radius, and reliability calculation strategy will be discussed.

4.1. Relation Model. Organization location is a region, but online maps return the latitude and longitude which is a point when you translate organization address. And the given point deviates from the location of IP entity. When establishing the relationship between distance and distance constraint, we should consider the radius of organization region, as shown in Figure 3.

S is the nearest common router of L_1 and L_2, and P_1 is the latitude and longitude location given by online map. TH is the radius of organization region which IP entity of L_1 locates in. The distance constraint between S and L_1 is named $DC1$. Analysis of L_2 is the same as L_1. The distance of P_1 and P_2 is named GD. Therefore, the minimum value of the distance between IP entity of L_1 and IP entity of L_2 is $GD - 2TH$ (from P_1'' to P_2'), and the maximum value is $GD + 2TH$ (from P_1' to P_2''). According to (1) and (2), the values of max DC and min DC can be calculated.

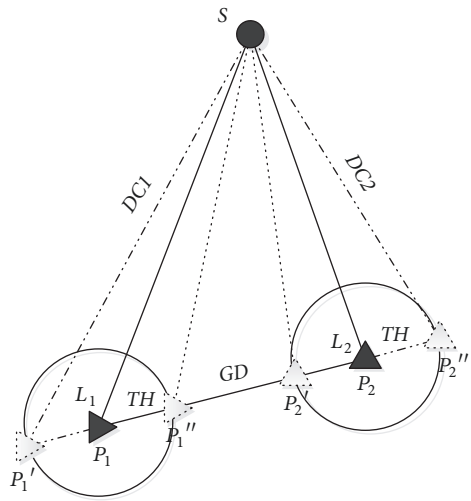


Figure 3: Relationship among distance, distance constraint, and radius.

Based on the triangular trilateral relationship, there are

$$\begin{aligned} \min DC &\le GD - 2TH \\ GD - 2TH &\le \max DC \end{aligned} \tag{7}$$

and

$$\begin{aligned} \min DC &\le GD + 2TH \\ GD + 2TH &\le \max DC \end{aligned} \tag{8}$$

That is equivalent to

$$\begin{aligned} \min DC + 2TH &\le GD \\ GD &\le \max DC + 2TH \end{aligned} \tag{9}$$

and

$$\begin{aligned} \min DC - 2TH &\le GD \\ GD &\le \max DC - 2TH \end{aligned} \tag{10}$$

Combining inequality (9) and inequality (10), we obtain the range of GD, which is the relation model (inequality (6)).

$$GD \in [\min DC - 2TH, \max DC + 2TH] \tag{11}$$

According to (11), for L_1 and L_2 after evaluation, the smaller the TH is, the closer the locations are.

4.2. Reliability Calculation Strategy. The information in the web page lacking effective verification deceives information receivers. Based on this reality, if the organization corresponds to multiple IP addresses in the same set, only the candidate landmark with the smallest single-hop delay value to the nearest router is retained. Therefore, the landmarks may be considered that are not related to each other.

For any candidate landmark L_i in a subset (marked as "C") of evaluation landmark set, the probability that L_i locates in organization region is p_i denoted as

$$P(L = True) = p_i, \quad p_i \in [0, 1] \tag{12}$$

Then, the probability that L_i does not locate in organization region is $1 - p_i$ denoted as

$$P(L_i = False) = 1 - P(L_i = True) = 1 - p_i \tag{13}$$

When there are n elements in C, the probability that all elements are not in the their region is denoted as P_c.

$$P_c = \prod_{i=1}^{n} P(L_i = False) = \prod_{i=1}^{n} (1 - p_i) \tag{14}$$

Since $p_i \in [0, 1]$, there is

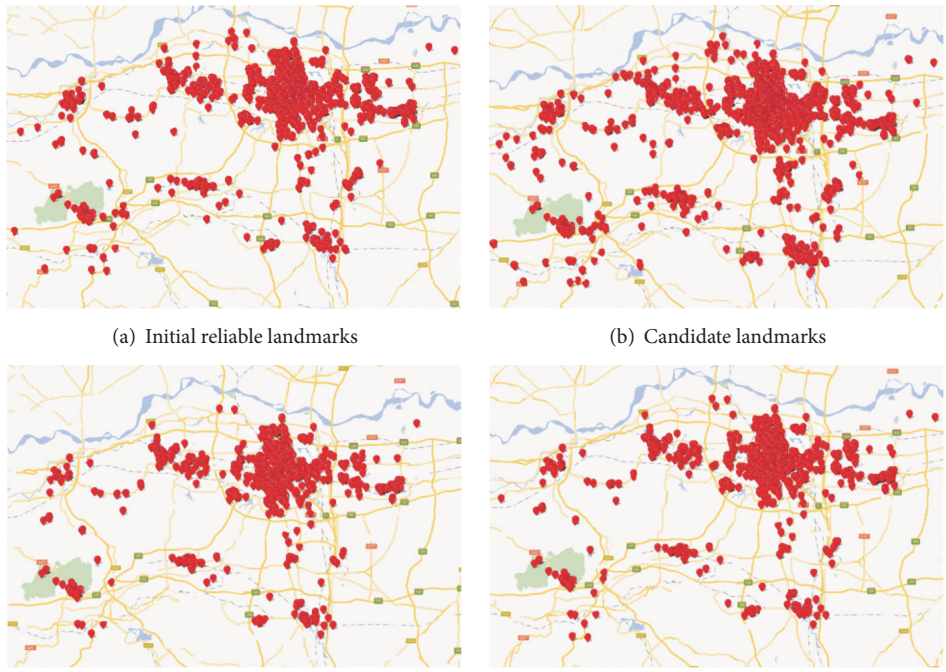
$$P_c \le 1 - p_j \tag{15}$$

where $p_j = P(L_j = True)$ and $p_j \in C$. Only when $\forall L_k \in C, k \ne j, p_k = P(L_k = True) = 0$, "=" is taken.

In each set, one organization only retains one candidate landmark. So, it can be considered that the elements in the set are not related to each other. When any two elements in subset satisfy inequality (6), the probability that our method mistakes the candidate landmarks as reliable landmarks is P_c, which is smaller than (or equal to) single candidate landmark misjudged as reliable. And the more elements in the subset, the higher reliability of evaluated landmarks.

5. Experimental Results

In order to verify the validity of our method, the feasibility verification experiment and the street-level landmark evaluation experiment were carried out.



(a) Initial reliable landmarks (b) Candidate landmarks

(c) Evaluated Landmarks (TH=3) (d) Evaluated Landmarks (TH=5)

Figure 4: Landmarks distribution.

Before conduction the experiments, with 7 days as a cycle and 30 minutes as a time period, we make a statistical analysis of the network stability in each time period within four cycles. Finally, the results show that, in China, the period (called "suitable period") from 22:30 to 06:30 (next day) is suitable for network measurement.

5.1. Feasibility Verification Experiment. 500 reliable street-level landmarks in Zhengzhou (called "initial reliable landmarks") were selected, and the distribution of initial reliable landmarks is shown in Figure 4(a). According to the distribution characteristics of initial reliable landmarks in geographical space, 500 street-level different locations were generated using a random function around the initial reliable landmark locations. 500 online IPs were selected which relate to Zhengzhou city in Baidu, IPIP, and IP.cn database. We generated 500 landmarks based on 500 online IPs and 500 street-level locations. The generated landmarks may be considered unreliable. In this experiment, the dataset (called "candidate landmarks") containing 500 reliable landmarks and 500 generated landmarks is shown in Figure 4(b), and the initial reliability values of all landmarks in dataset are set to 50%.

Performing 50 times routing measurements for each candidate landmark in suitable period, and merging the route paths, the delay values from nearest router to candidate landmarks were obtained. According to nearest router, candidate landmarks will be grouped. In this experiment, the reliability threshold was set to 75% and TH value was set to 3 and 5, respectively. When TH is 3, 416 evaluated landmarks (shown in Figure 4(c)) are evaluated. The ratio of evaluated landmarks to initial reliable landmarks is 83.2%. And when TH=5, 435 evaluated landmarks (shown in Figure 4(d)) are evaluated and the ratio is 87%.

Comparing Figures 4(a), 4(c), and 4(d), the evaluated landmarks are more concentrated in geospatial distribution than initial reliable landmarks. And compared with TH=3, the difference in geospatial distribution of evaluated landmarks (TH=5) is not obvious.

The experimental results show that street-level landmark evaluation method can obtain reliable landmarks from candidate landmarks effectively. The reasons that the method fail to obtain all initial reliable landmarks from candidate landmarks may be as follows.

(1) The delay measurement results are inaccurate. Relation model is established based on delay. For two candidate landmarks, if one's single-hop delay value increases, the value of min DC increases that may lead to $GD < \min DC$.

(2) The distribution of initial reliable landmarks is dispersed. If the distribution of initial reliable landmarks is dispersed, GD increases for two candidate landmarks that may lead to $GD > \max DC$.

5.2. Street-level Landmark Evaluation Experiment. The data are collected from Internet web pages, including organization name, location, web home page, and zip code. During the collecting process, the IPs with stable features, such as time servers, mail servers, and ftp servers, may be got more attention. The IP address is obtained based on DNS services, and latitude and longitude are obtained by online map

Table 3: The number of landmarks.

City	Before location verification	After location verification	After evaluation (TH=3)	After evaluation (TH=5)
Beijing	4658	1072	392	456
Shanghai	8966	3289	1227	1341
Shenzhen	6605	1857	783	869
Xiamen	3702	864	236	301

services. According to the famous mail server providers' IP addresses, part of obtained data is deleted. Eventually, the numbers of original candidate landmarks in Beijing, Shanghai, Shenzhen, and Xiamen are 4658, 8966, 6605, and 3702, respectively. Baidu, IPIP, IP.cn databases, and zip codes are used for location verification. For privacy protection, the obtained data only retain the IP address and information of latitude and longitude. After location verification, the numbers of candidate landmarks in Beijing, Shanghai, Shenzhen, and Xiamen are 1072, 3289, 1857, and 864, respectively. In addition, data set in this experiment includes 400 reliable street-level landmarks also (100 landmarks in each city). Reliable landmarks are used to verify the evaluated landmarks and do not participate in the evaluation process.

The initial reliability of all candidate landmarks is set to 50%, and reliability threshold is set to 75%. 50 times routing measurements are performed for each candidate landmark in suitable period. Candidate landmarks are evaluated when TH=3 and TH=5, respectively. The number of landmarks in each city is shown in Table 3.

The number of candidate landmarks has been greatly reduced after location verification because of servers hosting. There are two reasons that the number of landmarks decreased after evaluation. One reason is that some of the landmarks are not reliable and do not satisfy the relation model. Another reason is that the method proposed in this paper cannot evaluate the landmarks whose nearest router connect with a landmark only, although this landmark may be reliable. According to inequality (6), the larger the value of TH is, the more landmarks rest after evaluation. When the value of TH increases, the range of GD is expanded, meaning that the allowance error distance from the organization center to server location and allowance error delay from nearest router to landmark increase. Therefore, the larger the TH values are, the more landmarks are obtained according to relation model. When TH=3, the numbers of evaluated landmarks in Beijing, Shanghai, Shenzhen, and Xiamen are 392, 1227, 783, and 236, respectively, accounting for 36.57%, 35.15%, 42.16%, and 27.31% of candidate landmarks and 8.42%, 13.69%, 11.85%, and 6.37% of original candidate landmarks. When TH=5, the numbers are 456, 1341, 869, and 301, respectively, accounting for 42.54%, 40.77%, 46.8%, and 34.84% of candidate landmarks and 9.79%, 14.96%, 13.16%, and 8.13% of original candidate landmarks. The distribution of candidate landmarks and evaluated landmarks of each city is shown in Figure 5 (candidate landmarks marked as "c-landmarks" and evaluated landmarks marked as "e-landmarks").

In each city, candidate landmarks (obtaining method is similar to Structon [12], called "before evaluation" landmarks) and evaluated landmarks are used to locate 100 reliable street-level landmarks using SLG algorithm [15], respectively. The relationship between the mean error and the number of landmarks is shown in Figure 6.

In Figure 6, the mean error using evaluated landmarks to locate 100 reliable street-level landmarks is smaller than using the same number of candidate landmarks in the same city. The mean error of geolocation using landmarks evaluated by TH=3 is slightly less than the value caused by the landmarks evaluated by TH=5. The geolocation accuracy is affected by the accuracy of landmarks on the one hand. The base of street-level geolocation is street-level landmarks. On the other hand, the geolocation accuracy is affected by the number of reliable landmarks. The accuracy of geolocation increases when the number of landmarks increases. In Figure 6, when the number of reliable landmarks increases from 0 to 300, the geolocation error decreases rapidly. When the number of reliable landmarks is greater than 300, the number of landmarks increases, and the speed of the geolocation error decreases gradually. When all candidate landmarks are used for geolocation, the mean error in Beijing, Shanghai, Shenzhen, and Xiamen is 9.624 km, 7.832 km, 7.634 km, and 8.994 km, respectively. But the value is 3.98km, 2.185km, 2.234km, and 5.237km respectively, when TH=3, and the value is 3.943 km. 2.198 km, 2.241 km, and 4.473 km, respectively, when TH=5.

In Beijing, Shanghai, Shenzhen, and Xiamen city, when all candidate landmarks and evaluated landmarks are used to locate 100 reliable street-level landmarks using SLG algorithm. The relationship between geolocation error and cumulative probability is shown in Figure 7.

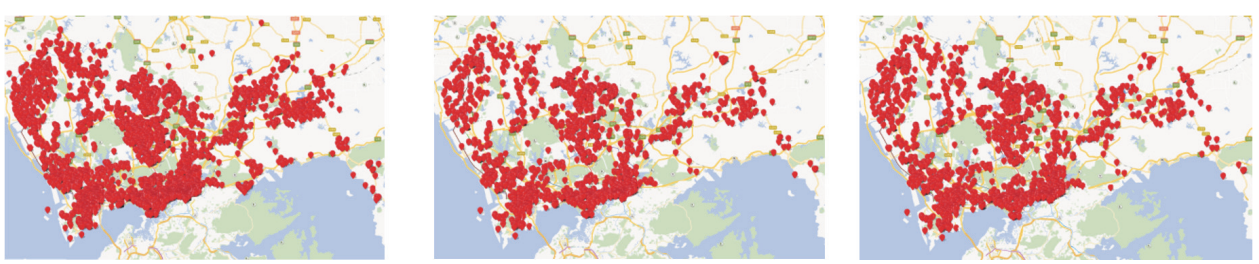
Figure 7 shows that the location accuracy of using evaluated landmarks is significantly higher than the value of using candidate, and the difference in location accuracy between TH=3 and TH=5 is not obvious. In Beijing, Shanghai, and Shenzhen, comparing with pre-landmarks, the probability of geolocation error less than 5km increases by more than 35%, and the probability of geolocation error less than 10km increases by 20%, when evaluated landmarks are used. In Xiamen, the probability of geolocation error less than 5km and 10km increases by 20% and 15%, respectively. The main

(a-1) c-landmarks (a-2) e- landmarks(*TH*=3) (a-3) e-landmarks(*TH*=5)

(a) Beijing

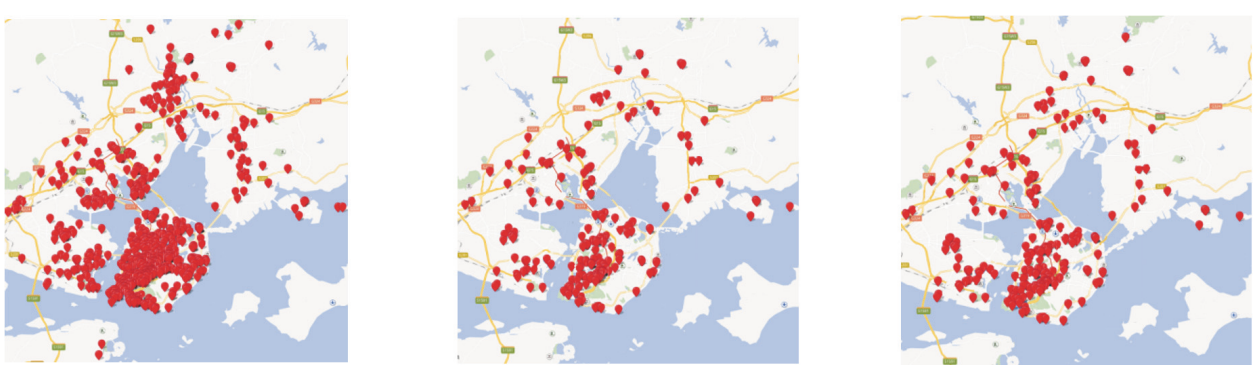
(b-1) c-landmarks (b-2) e-landmarks (*TH*=3) (b-3) e-landmarks (*TH*=5)

(b) Shanghai



(c-1) c-landmarks (c-2) e-landmarks (*TH*=3) (c-3) e-landmarks (*TH*=5)

(c) Shenzhen



(d-1) c-landmarks (d-2) e-landmarks (*TH*=3) (d-3) e-landmarks (*TH*=5)

(d) Xiamen

Figure 5: Landmarks distribution of each city.

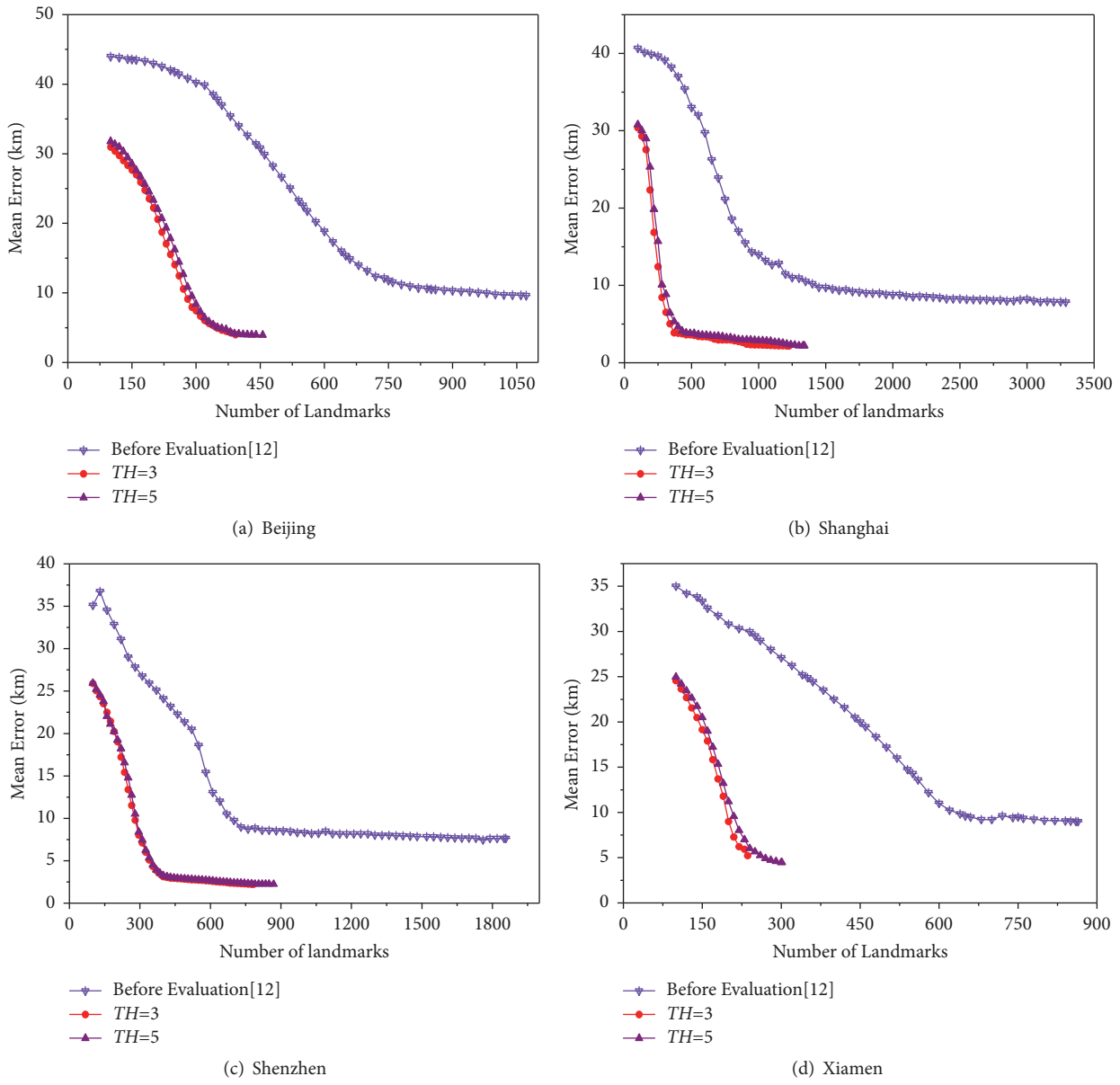


Figure 6: Relationship between the mean error and the number of landmarks.

reason for the low increasing percentage is that the number of evaluated landmarks is less than former three cities.

These experimental results show that comparing with candidate landmarks, using landmarks evaluated by our method for street-level geolocation can improve the geolocation accuracy.

6. Conclusions

Landmark-based street-level IP geolocation has a high application prospect in the field of Internet of Things. In view of the deficiencies of current methods in street-level landmark evaluation, a street-level landmark evaluation method based on the nearest router is proposed. Using the fact that "locations of terminal landmarks with the same nearest router are close to each other in geographical location", the relation model among distance constraint, distance, and region radius. In the basis of the relation model, the goal of evaluating street-level landmark is achieved. Meanwhile, the reliability values of evaluated street-level landmarks are calculated by initial reliability and binomial distribution. The experimental results show that the landmarks evaluated by this method increase the geolocation accuracy. Effectiveness of this method is affected by accuracy of delay value, landmarks distribution, and anonymous nearest router. In the future, the delay value acquisition method and street-level landmark evaluation method when nearest router is anonymous are the focus of work.

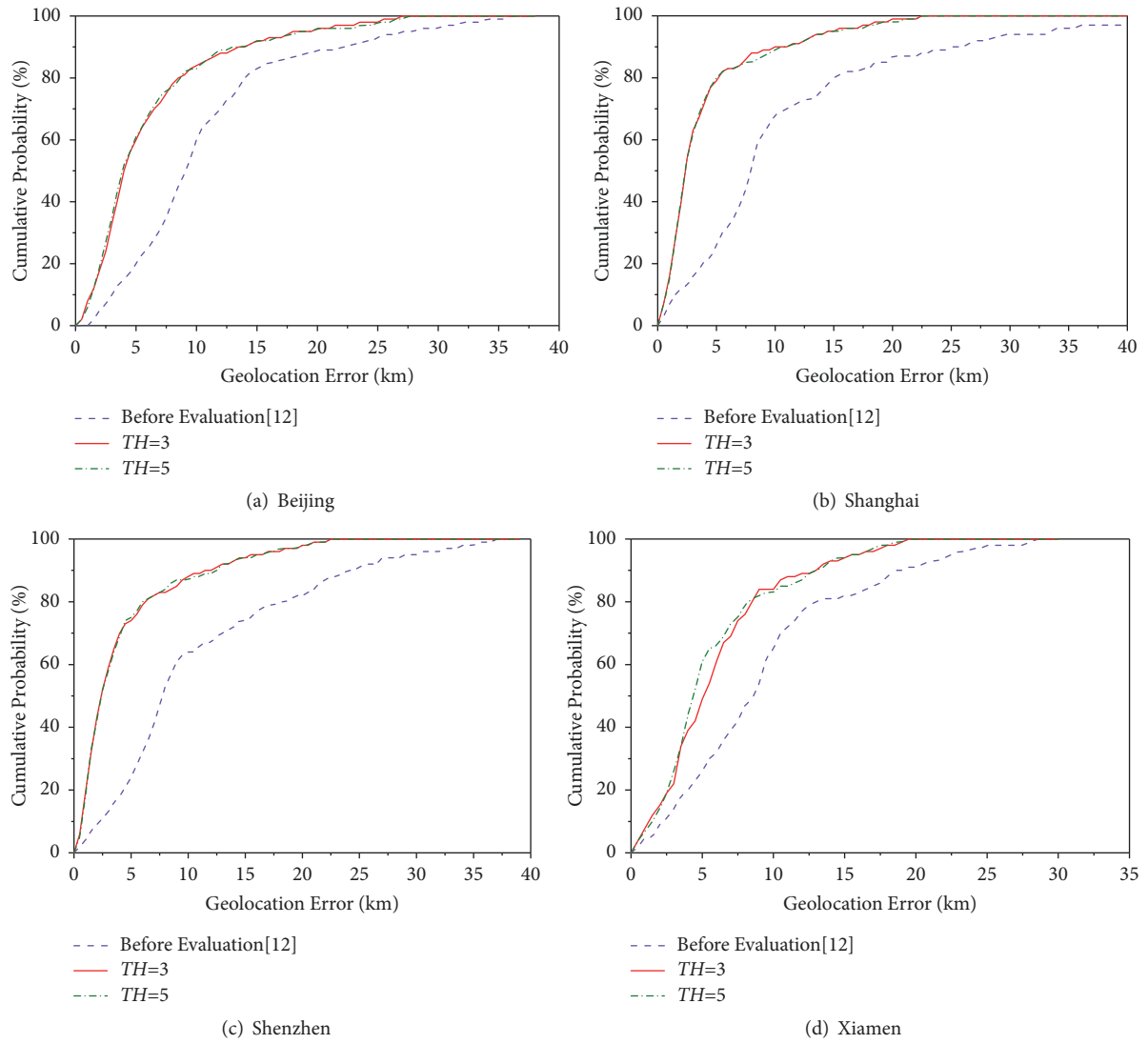


Figure 7: Relationship between geolocation error and cumulative probability.

Data Availability

The data in this article is mainly got from website pages (such as "http://tool.bridgat.com", "http://www.71ab.com/"). The DNS service is used to convert the domain name into IP, and the Baidu map API ("http://lbsyun.baidu.com/index.php?title=webapi/guide/webservicegeocoding") is used to convert the address information to latitude and longitude. Baidu ("http://lbsyun.baidu.com/index.php?title=webapi/ip-api"), IPIP ("https://www.ipip.net/"), and IP.cn ("https://ip.cn/") databases are used to verify candidate landmarks' city location.

Disclosure

Part of the paper was represented in the following conference: http://www.icccsconf.org/%E6%8E%A8%E8%8D%90SCI%E6%9C%9F%E5%88%8A%E8%AE%BA%E6%96%87%E5%88%97%E8%A1%A8.pdf.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The work presented in this paper is supported by the National Key R&D Program of China [nos. 2016YFB0801303 and 2016QY01W0105], the National Natural Science Foundation of China [nos. U1636219, 61602508, 61772549, U1736214, and 61572052], Plan for Scientific Innovation Talent of Henan Province [no. 2018JR0018], and the Key Technologies R&D Program of Henan Province [no. 162102210032].

References

[1] M. Gondree and Z. N. J. Peterson, "Geolocation of data in the cloud," in *Proceedings of the 3rd ACM Conference on Data and*

Application Security and Privacy, CODASPY 2013, pp. 25–36, New York, NY, USA, February 2013.

[2] E. Schmieders, A. Metzger, and K. Pohl, "A Runtime Model Approach for Data Geo-location Checks of Cloud Services," in *Service Oriented Computing and Applications*, vol. 8831 of *Lecture Notes in Computer Science*, pp. 306–320, Springer, Berlin, Germany, 2014.

[3] Z. Cai, H. Yan, P. Li, Z.-A. Huang, and C. Gao, "Towards secure and flexible EHR sharing in mobile health cloud under static assumptions," *Cluster Computing*, vol. 20, no. 3, pp. 2415–2422, 2017.

[4] L. Yang, Z. Han, Z. Huang, and J. Ma, "A remotely keyed file encryption scheme under mobile cloud computing," *Journal of Network and Computer Applications*, vol. 106, pp. 90–99, 2018.

[5] J. Xu, L. Wei, Y. Zhang, A. Wang, F. Zhou, and C. Gao, "Dynamic fully homomorphic encryption-based merkle tree for lightweight streaming authenticated data structures," *Journal of Network and Computer Applications*, vol. 107, pp. 113–124, 2018.

[6] X. Xu, W. Dou, X. Zhang, C. Hu, and J. Chen, "A traffic hotline discovery method over cloud of things using big taxi GPS data," *Software: Practice and Experience*, vol. 47, no. 3, pp. 361–377, 2017.

[7] Z. Zhou, W. Dou, G. Jia et al., "A method for real-time trajectory monitoring to improve taxi service using GPS big data," *Information and Management*, vol. 53, no. 8, pp. 964–977, 2016.

[8] X. Xu and W. Dou, "An assistant decision-supporting method for urban transportation planning over big traffic data," in *Proceedings of the International Conference on Human Centered Computing*, pp. 251–264, Phnom Penh, Cambodia, 2014.

[9] Y. Ma, X. Luo, X. Li, Z. Bao, and Y. Zhang, "Selection of rich model steganalysis features based on decision rough set α-positive region reduction," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1-1, 2018.

[10] Y. Zhang, C. Qin, W. Zhang, F. Liu, and X. Luo, "On the fault-tolerant performance for a class of robust image steganography," *Signal Processing*, vol. 146, pp. 99–111, 2018.

[11] X. Luo, X. Song, X. Li et al., "Steganalysis of HUGO steganography based on parameter recognition of syndrome-trellis-codes," *Multimedia Tools and Applications*, vol. 75, no. 21, pp. 13557–13583, 2016.

[12] 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, Rio de Janeiro, Brazil, April 2009.

[13] H. Jiang, Y. Liu, and J. N. Matthews, "IP geolocation estimation using neural networks with stable landmarks," in *Proceedings of the IEEE International Conference on Computer Communications Workshops*, pp. 170–175, San Francisco, Calif, USA, 2016.

[14] G. Zhu, X. Luo, F. Liu, and J. Chen, "An Algorithm of City-Level Landmark Mining Based on Internet Forum," in *Proceedings of the IEEE International Conference on Network-Based Information Systems*, pp. 294–301, Taipei, Taiwan, September 2015.

[15] Y. Wang, D. Burgener, M. Flores et al., "Towards street-level client-independent IP geolocation," in *Proceedings of the USENIX Conference on Networked Systems Design and Implementation*, pp. 365–379, Boston, Mass, USA, 2011.

[16] S. S. Siwpersad, B. Gueye, and S. Uhlig, "Assessing the geographic resolution of exhaustive tabulation for geolocating internet hosts," in *Proceeding of the Springer-Verlag International Conference on Passive and Active Network Measurement*, pp. 11–20, Cleveland, Ohio, USA, 2008.

[17] Y. Shavitt and N. Zilberman, "A geolocation databases study," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 10, pp. 2044–2056, 2011.

[18] L. Backstrom, E. Sun, and C. Marlow, "Find me if you can: Improving Geographical Prediction with Social and Spatial Proximity," in *Proceedings of the ACM International Conference on World Wide Web*, pp. 61–70, ACM, Raleigh, NC, USA, April 2010.

[19] I. Poese, S. Uhlig, M. A. Kaafar et al., "IP geolocation databases: unreliable?" *ACM SIGCOMM Computer Communication Review*, vol. 41, no. 2, pp. 53–56, 2011.

[20] H. Li, P. Zhang, Z. Wang et al., "Changing IP geolocation from arbitrary database query towards multi-databases fusion," in *Proceedings of IEEE Symposium on Computers and Communications*, pp. 1150–1157, Heraklion, Greece, July 2017.

[21] H. Li, Y. He, R. Xi et al., "A Complete evaluation of the chinese IP geolocation databases," in *Proceedings of the IEEE International Conference on Intelligent Computation Technology and Automation*, pp. 13–17, Nanchang, China, June 2015.

[22] O. Dan, V. Parikh, and B. D. Davison, "Improving IP geolocation using query logs," in *Proceedings of the 9th ACM International Conference on Web Search and Data Mining, WSDM 2016*, pp. 347–356, San Francisco, Calif, USA, February 2016.

[23] D. Komosny, M. Voznak, S. U. Rehman et al., "Location accuracy of commercial IP address geolocation databases," *Information Technology and Control*, vol. 46, no. 3, pp. 333–344, 2017.