Research Article

A Street-Level IP Geolocation Method Based on Delay-Distance Correlation and Multilayered Common Routers

Shichang Ding ^{1,2}, Fan Zhao ¹, and Xiangyang Luo ¹

¹State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou 276800, China

²Institute of Computer Science, University of Gottingen, Gottingen 37075, Germany

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

Received 10 November 2020; Revised 7 December 2020; Accepted 26 December 2020; Published 12 January 2021

Academic Editor: Jinwei Wang

Copyright © 2021 Shichang Ding et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The geographical locations of smart devices can help in providing authentication information between multimedia content providers and users in 5G networks. The IP geolocation methods can help in estimating the geographical location of these smart devices. The two key assumptions of existing IP geolocation methods are as follows: (1) the smallest relative delay comes from the nearest host; (2) the distance between hosts which share the closest common routers is smaller than others. However, the two assumptions are not always true in weakly connected networks, which may affect accuracy. We propose a novel street-level IP geolocation algorithm (Corr-SLG), which is based on the delay-distance correlation and multilayered common routers. The first key idea of Corr-SLG is to divide landmarks into different groups based on relative-delay-distance correlation. Different from previous methods, Corr-SLG geolocates the host based on the largest relative delay for the strongly negatively correlated groups. The second key idea is to introduce the landmarks which share multilayered common routers into the geolocation process, instead of only relying on the closest common routers. Besides, to increase the number of landmarks, a new street-level landmark collection method called WiFi landmark is also presented in this paper. The experiments in one province capital city of China, Zhengzhou, show that Corr-SLG can improve the geolocation accuracy remarkably in a real-world network.

1. Introduction

Thanks to the rapid growth of mobile multimedia services like online video and remote conferencing on smart mobile devices (e.g., cellphone and tablets), the fifth-generation (5G) mobile and wireless communication systems are in great demand all over the world [1–3]. How to manage the trust relations between users and multimedia content providers in the 5G network is an important problem [4]. Previous works like [4] point out that the geographical locations of users or content providers are important information for detecting unauthenticated or malicious devices. IP geolocation can find the geographical location of Internet hosts as well as smart devices based on its IP address [5]. Besides authentication, IP geolocation also helps in identifying the geographical location of cyberattacks or online frauds for law enforcement organizations and government agencies [6].

Existing IP geolocation methods can be categorized into two kinds by accuracy: city-level IP geolocation and street-level IP geolocation. City-level IP geolocation aims to find the city where the target IP is located. The median error distance of main city-level IP geolocation methods is between tens and hundreds of kilometers. After obtaining the city-level location information, street-level IP geolocation methods can be used to find the specific street, community, or organization where the target IP is located, of which the median error distance is usually less than 10 kilometers.

City-level IP geolocation has developed into a relatively mature stage. The main methods include GeoPing [7], CBG [8], TBG [9], Octant [10], Structon [11], GeoGet [12], Chen-Geo [13], PLAG [14], Yuan-Geo [15], and RNBG [5]. This paper mainly discusses street-level IP geolocation methods. There are three main street-level IP geolocation methods: Checkin-Geo [16], Geo-NN [17], and Wang-Geo [18]. Besides these three methods, IP databases can also provide the

street-level location for a small proportion of IP addresses, which could hardly satisfy people's demand [19, 20].

Checkin-Geo is a data-mining-based IP geolocation method. It first builds the relationship between "user name" and "smartphone location" from the check-in [16, 21] information collected on specific smartphone applications. Then it builds the relationship between "user name" and "PC IP" from the login information collected on corresponding PC applications. Based on the two relationships, it finally gets the relationship between "PC IP" and "smartphone location" to geolocate target IP. The problem of Checkin-Geo is that these two relationships are hard to get for most researchers or companies. The relationships are usually only possessed by a minority of Internet giants, which can access one user's location and IP address from both mobile applications and PC applications. For example, Tencent (the largest instant messaging service provider in China, <http://www.tencent.com>) can get one user's corresponding data from his/her QQ clients on smartphones and PCs. In this paper, we mainly focus on the measurement-based street-level IP geolocation methods.

Geo-NN and Wang-Geo methods are both measurement-based IP geolocation methods. These methods estimate the location of one target IP mainly based on the network delays between probing hosts, target IP, and known landmarks [17, 18]. Compared with Geo-NN, Wang-Geo also leverages network topology data during the geolocation. Therefore, it can get better results if the number of probing machines is limited. In particular, if there is only one probing machine in the city of target IP, only Wang-Geo can still estimate the location, so Wang-Geo is a low-cost method for large-scale real-world deployment.

Wang-Geo is initially designed for richly connected networks such as the American network [18]. In moderately or weakly connected networks such as the Chinese network [22], the accuracy of Wang-Geo can be seriously affected [16]. In richly connected networks, the correlation between network delay and geographical distance is strong [22]. The network delay usually increases as the distance increases. In weakly connected networks, the correlation is usually weak, and there is no clear relationship between network delay and geographical distance. We aim to improve the performance of Wang-Geo in a weakly connected network.

Why is the accuracy of Wang-Geo method affected in a weakly connected network? There are two important assumptions for Wang-Geo method: (1) for one host, the smallest delay comes from the host which has the smallest geographical distance to it; (2) the distances between hosts which share the closest common routers are usually smaller than the distances between hosts which share the other common routers. However, through measurements, we find that these two assumptions do not always hold for weakly connected networks. This may be the reason that the accuracy of Wang-Geo method decreases dramatically in weakly connected networks.

Besides the two assumptions, sufficient and widely distributed street-level landmarks are also one of the foundations of Wang-Geo method. However, as more and

more websites are deployed in cloud services, the geographical location of a website is not necessarily related to its company. Therefore, the number of classical web landmarks that Wang-Geo method relies on becomes more and more limited. This is also an ineligible influence on the performance of Wang-Geo method in recent years.

To obtain more accurate measurement-based IP geolocation results in weakly connected networks, this paper proposes an IP geolocation algorithm based on relative-delay-distance correlation and multilayered common routers—Corr-SLG. Corr-SLG includes two parts: landmark collection and IP geolocation.

In the stage of landmark collection, besides the classical web landmark collection method, we present a new street-level landmark collection method called WiFi (Wireless Fidelity) landmark. This method collects landmarks based on the diversely distributed WiFi access points and the accurate geographical location information of smartphones.

In the stage of IP geolocation, to find out which landmark is nearest to the target IP in a weakly connected network, Corr-SLG divides landmarks into three groups based on the relative-delay-distance correlation. In the group where delay-distance correlation is strongly positive (i.e., near to 1), the landmark which has the smallest delay to target IP is selected as the candidate landmark (i.e., the landmark nearest to target IP). In the group where delay-distance correlation is strongly negative (i.e., near to -1), the landmark which has the largest delay to target IP is selected as the candidate landmark. In the third group, the landmark is randomly selected as the candidate landmark. To introduce more landmarks that are near to target IP into the selection procedure of candidate landmarks, Corr-SLG selects candidate landmarks not only from the closest common router layer but also from the other common router layers. The experiments in one province capital city of China, Zhengzhou, show that Corr-SLG can increase the accuracy of street-level IP geolocation by about 38.59%.

The rest of the paper is organized as follows. In Section 2, we show that the two key assumptions of Wang-Geo are not always true in two real-world networks. Then, we present the landmark collection method in Section 3. Section 4 introduces IP geolocation algorithm of Corr-SLG. Section 5 shows the experiment results. The paper is concluded in Section 6.

2. Two Assumptions of Wang-Geo Method

Wang-Geo method is a typical street-level measurement-based IP geolocation method. In this section, we first introduce the basic principles of Wang-Geo method. Then, we will test whether the two assumptions are true in two real-world networks.

2.1. Basic Principles of Wang-Geo Method. Wang-Geo method maps target IP to the closest landmark by three steps. The first two steps actually use a modified version of one classical city-level IP geolocation method—CBG [8]—to

find the city or region where the target IP is located. This paper mainly discusses the street-level IP geolocation, which is mainly done in Step 3. Step 3 is shown as follows:

- (1) As shown in Figure 1, a probing host P measures the delays and router paths to the landmarks ($L1, L2, L3$) and the target IP T . In this paper, the probing host means a computer which can be used by researchers to measure the delay between it and other computers, and its geographical location is also known by the researchers.
- (2) Find the closest common router between T and each landmark. The closest common router between T and $L1$ is $R1$; the closest common router between T and $L2$ is also $R1$; the closest common router between T and $L3$ is $R3$.
- (3) Calculate the relative delay [18] between T and each landmark. Wang-Geo method only calculates relative delay based on the closest common router. The relative delay between T and $L1$ is $d_1 + d_2$; the relative delay between T and $L2$ is $d_2 + d_3$; the relative delay between T and $L3$ is $d_2 + d_4 + d_5 + d_7 + d_8$.
- (4) The landmark which has the smallest relative delay to the target IP is chosen as the candidate landmark. In this paper, the landmark estimated to be nearer to the target IP than the others in a group of landmarks is called a *candidate landmark*.
- (5) If there is only one probing host, the target IP is mapped to the location of the only one candidate landmark; if there is more than one probing host, the target IP is mapped to the location of the candidate landmark which has the smallest relative delay.

The relative delay will be underestimated if the delay to one closest common router is overestimated. This will affect the candidate landmark selection. This kind of router is called “inflating router” [18]. Wang-Geo method divides the geographical distance by the relative delay between two landmarks to discover inflating routers. All the measured data associated with inflating routers have to be discarded.

We collect 80 web landmarks, respectively, in Zhengzhou, one province capital city of China, and Toronto, the capital of Canada. We calculate the relative delays between web landmarks in each city and simply check the closest common routers whose corresponding relative delays are negative. The results show that at least 50% of the closest common routers in both two cities are inflating routers. This may be caused by the data processing policy of routers inside cities. Many routers tend to give a low priority to the packets which aim to measure the delay to these routers. This phenomenon may forbid some landmarks which are near to target IP to be selected as candidate landmarks if their routers are inflating. This means that Wang-Geo method has to give up a considerable amount of landmarks because the common routers between them and target IP are inflating, which may affect the accuracy of Wang-Geo method.

FIGURE 1: Common router and relative delay.

2.2. Motivation. Wang-Geo method is actually based on the two important assumptions: (1) for one host, the smallest delay comes from the host which has the smallest geographical distance to it, so Wang-Geo method selects the landmark which has the smallest relative delay as a candidate landmark; (2) the distance between hosts which share the closest common routers is usually smaller than that between hosts which share the other common routers, so Wang-Geo method only selects a candidate landmark from landmarks which share the closest common routers with target IP and ignores all the other landmarks. In a richly connected network environment, such as PlanetLab dataset used by [18], where these two assumptions are true, Wang-Geo method can achieve accurate geolocation results. However, as we explain in this section, the two assumptions may not hold for a weakly connected network.

2.2.1. Does the Smallest Delay Always Come from the Closest Landmark? We can judge whether this assumption is true by the delay-distance correlation. Delay-distance correlation is the first-order linear correlation coefficient between delay and distance [23]. To calculate the delay-distance correlation of a certain network, we need to measure network delays and direct geographical distances between a certain group of hosts in this network. The delay of each pair of hosts is measured many times and only the minimum one is selected. The distance of each pair of hosts is calculated based on [24]. Assume that the variance of delay is V_{delay} , the variance of distance is V_{distance} , and the covariance of delay and distance is $\text{cov}(\text{delay}, \text{distance})$; then, the delay-distance correlation Corr can be calculated by the following formula [12]:

$$\text{Corr} = \frac{\text{cov}(\text{delay}, \text{distance})}{\sqrt{V_{\text{delay}} \times V_{\text{distance}}}}. \quad (1)$$

The range of Corr is $[-1, 1]$. If Corr of a group of hosts is positive and near to 1, the delay is strongly positively correlated with the distance, which means that delay increases as distance increases and decreases as distance decreases. In this circumstance, the smallest delay has a large opportunity to come from the closest landmark. However, if Corr is strongly negative and near to -1 , the largest delay usually comes from the closest landmark; if Corr is close to 0, no matter whether it is negative or positive, there is no clear relationship between delay and distance [22].

In [18], researchers find that the smallest relative delay often comes from the nearest landmarks in the USA. Thus, Wang-Geo method selects candidate landmarks based on the smallest relative delay. We measure the relative delay and distance between 80 web landmarks, respectively, in Zhengzhou and Toronto. The relative-delay-distance relationships of landmarks in the two cities are shown in Figures 2 and 3.

The above Corr is called collective Corr in this paper. Collective Corr is calculated based on the delays and distances between a group of hosts and usually represents the general network characteristics of a certain area. If the collective Corr is strong, the relative delays increase as the distances increase, and the smallest delays often come from the nearest landmark. This network is called a richly connected network in this paper. Otherwise, the network is referred to as moderately or weakly connected networks. From Figures 2 and 3, we can see that there is no clear relationship between the smallest relative delay and the smallest distance in both Zhengzhou and Toronto. The absolute values of Corr of landmarks in the two cities are both between -0.1 and 0.1 (i.e., weak). Therefore, actually there is no clear relationship between relative delay and distance in weakly connected networks. The collective Corr is weak in many cities (e.g., Zhengzhou and Toronto). This may be because, between two hosts inside a city, the delay caused by geographical distance may only make up a small proportion of the whole delay. The main part of the whole delay consists of queuing delay and processing delay in routers, which has little relationship with distance.

Though the collective Corr is weak in many cities, we find that the individual Corr of a small number of landmarks could be much stronger. Individual Corr is calculated based on the delays and distances between only one specific host and the other hosts. It only represents the network characteristics of a certain host. Figures 4 and 5 show the individual Corr of web landmarks in Zhengzhou and Toronto. It can be seen that although the absolute value of most individual Corr is under 0.2, there is still a small number of landmarks whose individual Corr is much stronger. For these landmarks, the smallest or largest delay is still probably related to the smallest distance. This encourages us to divide landmarks into different groups by individual Corr and apply different candidate landmark selection strategies in different groups.

FIGURE 2: The relative-delay-distance relationship of landmarks in Zhengzhou.

FIGURE 3: The relative-delay-distance relationship of landmarks in Toronto.

2.2.2. Is the Distance between Hosts That Share the Closest Common Routers Always Smaller than That between Hosts That Share the Other Common Routers? In this paper, multilayered common routers mean that all the common routers are shared by two hosts. Besides the closest common routers used in Wang-Geo method, multilayered common routers also include the other common routers. As shown in Figure 1, R_1 , R_2 , and R_3 are the multilayered common routers of T and L_1 ; R_2 and R_3 are the other common routers of T and L_1 .

Wang-Geo method only selects candidate landmarks from the landmarks which share the closest common routers. It actually assumes that the distance between hosts that share the closest common routers is always smaller than that between hosts which share the other common routers. This assumption

FIGURE 4: The Individual Corr of landmarks in Zhengzhou.

FIGURE 5: The Individual Corr of landmarks in Toronto.

is usually true in cities which have an abundant amount of public IP addresses. In this kind of cities, the landmarks which share the closest common routers are very likely to come from one organization because it has a great amount of public IP addresses. These hosts are closer to each other than hosts which belong to different organizations.

However, organizations in many cities only possess a very small number of public IP addresses. The hosts inside these organizations usually use private IP addresses. Existing landmark collection can hardly get all the public IP addresses owned by one organization. Accordingly, the hosts which share the closest common routers and the other common routers are all from different organizations, which means that there is no clear difference between these two kinds of hosts.

Figure 6 shows the distribution of landmarks which share the closest common routers and the other common routers with one target IP. The other common routers in

Figure 6 are the second closest common routers. In Figure 6, the landmarks which share the closest common routers with the target IP also cover most areas of Zhengzhou. In Figure 6, the distance between landmarks that share the closest common routers is even larger because their number is smaller. This encourages us to introduce the landmarks which share multilayered common routers into the selection of candidate landmarks.

Based on the above analysis, we can conclude that, in many cities, the collective Corr is weak. Therefore, always selecting the landmark which has the smallest delay as the candidate landmark may cause a large error distance. Moreover, the distance between landmarks that share the closest common routers with target IP may not always be the smallest. In fact, besides the two assumptions, there is still an important influencing factor—enough street-level landmarks—which we will discuss in the next section. These three problems are the main motivation of the proposed Corr-SLG.

3. Landmark Collection

A street-level landmark consists of two main components: a public IP address and a street-level geographical location. The number and distribution of street-level landmarks directly influence the accuracy of IP geolocation algorithms. In this section, we will introduce how to collect street-level landmarks in Corr-SLG.

3.1. Web Landmark Collection. Currently, the main method to collect street-level landmarks is the web landmark collection method proposed in Wang-Geo method [18]. Its basic idea is to discover the organizations which own a website and distribute it in the city or area where the target IP lies. The IP of a web landmark is the IP of a website server, and the location is the geographical address of the organization. This method is one of the key contributions of Wang-Geo method and one of the most important reasons for its high precision. The detailed process of the web landmark collection method can be found in paper [18].

However, web landmark has two flaws: (1) both the number and distribution of web landmarks are limited; (2) it becomes more and more difficult to find enough web landmarks in recent years. It is easy to explain the first flaw. The organizations which own a website are usually only abundant in several metropolises. In addition, most of the organizations which own a website usually distribute it in several certain areas of the city, like Central Business District. As for the second flaw, not all organizations put their website servers in their own buildings. In fact, as the cloud services develop, more and more organizations deploy their websites in cloud services, which will inevitably reduce the number of available web landmarks. To provide more street-level landmarks, we try to present a new street-level landmark collection method in this paper.

3.2. WiFi Landmark Collection. In recent years, both the number and distribution of WiFi access points increase very fast. Many public places like banks, supermarkets,

FIGURE 6: The distribution of the landmarks which share the closest common routers and the second closest common routers in (a) Zhengzhou and (b) Toronto.

and hotels provide WiFi access points for people to go online. Inspired by this phenomenon, this paper presents a new way to discover street-level landmarks, the WiFi landmark collection method, which is shown in Figure 7.

In one public place, we connect to the WiFi which the public place provides to make the smartphone get online. Then, the WiFi landmark collection software installed on the smartphone will measure the router path to one Internet server. The IP address of a WiFi landmark is the first public IP address of the measured router path from the smartphone to the server. The location information of the WiFi landmark is the Global Positioning System (GPS) location of the smartphone. Besides IP and location, we also need to record the router path to the server, the Service Set Identifier (SSID) of the WiFi access point, the geographical address, and the name of the public place.

After landmark discovery, a number of landmarks have to be removed because the GPS location may be far away from the real geographical location of the IP. If the first IP address of the router path to the server is a public IP address, the distance between the real geographical location of the IP and the GPS location of the smartphone is limited by the scope of the public place, which can usually be ignored. However, if the first public IP address appears on the other hops of the path, the distance may be much further. We have to check all landmarks in the following steps:

- (1) If the first public IP address appears on the first or the second hop of the router path, the landmark can be preserved; the others are discarded.
- (2) Then, check all the preserved landmarks. Find out the landmarks which have the same IP. The WiFi access points of these landmarks use the same router to go online. If all these landmarks belong to one public place, these landmarks can be replaced by one single landmark. The IP of this landmark is the IP of the

FIGURE 7: WiFi landmark collection method.

replaced landmarks, and the location is the location of the public place to which the replaced landmarks belong. If these landmarks belong to different public places, but the maximum distance between them is under 50 m, then these landmarks can be replaced by one single landmark, too. The IP of this landmark is the IP of the replaced landmarks while the location is the average of GPS locations of the replaced landmarks. If the distance is larger than 50 m, all of the landmarks which share the same IP are discarded.

WiFi landmarks are very reliable because the distance between the real geographical location of a public IP address and the GPS location of the smartphone is limited. Furthermore, WiFi access points are widely distributed in many cities. In fact, many families and companies also use WiFi access points. The main shortcoming of WiFi landmarks is its high collection cost. Currently, it should be used as a supplementary measure if the web landmarks are insufficient in certain areas. In this paper, we mainly use WiFi landmarks as the target IP which needs to be located in experiments.

4. IP Geolocation Algorithm of Corr-SLG

In this section, we will illustrate the IP geolocation algorithm of Corr-SLG. This paper mainly focuses on street-level IP geolocation, so here we assume that before IP geolocation, we already know the city where the target IP is located and get enough street-level landmarks of this city. The geolocation result of Corr-SLG is different on each probing host. The final result is the average of the geolocation result of each probing host. Accordingly, the following introduction of IP geolocation algorithm is on one single probing host if there is no special instruction.

4.1. Extracting Multilayered Common Routers. First, the probing host measures the router paths to all landmarks, and then it measures the delays to all the landmarks and the middle routers. The delay between each pair of hosts is measured many times, and only the minimum one is selected.

Second, multilayered common routers between each pair of landmarks are extracted (multilayered common routers are all common routers between two hosts). At last, we can get all the common routers of the landmark dataset. For each common router, we find the corresponding landmarks dataset, which are the landmarks whose router paths include the common router. Though there are some landmarks included by different corresponding landmarks dataset, the landmark datasets of different common routers are usually different. In this paper, we refer to a common router and its landmark dataset as “a layer.” The relationship between one common router and its landmark dataset is shown in Figure 8.

Each pair of landmarks usually share at least one common router, the probing host. However, in some special circumstances (e.g., when the preceding hops of router path are all anonymous routers), a landmark may not share any common routers with the other landmarks. Because both Wang-Geo method and Corr-SLG need to geolocate target IP by common routers, in this circumstance, we can add a temporary virtual common router before all the paths of landmarks. The delay between the probing host and the virtual common router is zero. However, the fundamental way to solve this problem is to get enough landmarks or change the probing hosts.

4.2. Calculating Individual Corr. In this part, we need to calculate the individual Corr of each landmark. Before this, first, we need to measure the relative delay between landmarks in each layer. For two landmarks L_i and L_j in one layer, assume that the delay between the probing host and L_i is d_i , the delay between the probing host and L_j is d_j , the delay between the probing host and the common router of this layer is d_r , and the relative delay between L_i and L_j in this layer is $d_i + d_j - 2d_r$. Thus, the relative delay may be different even for the same pair of landmarks if they are in different layers, and so is the individual Corr of each landmark. The individual Corr of one landmark is calculated based on its relative delay and distance to the other

FIGURE 8: Multilayered common router and corresponding landmarks.

landmarks in the same layer (the formula of Corr is included in Section 2). For one landmark, its individual Corr in different layers is usually different.

If the number of landmarks in one layer is less than 5, there is no need to calculate the individual Corr, because the absolute value of the individual Corr tends to be too large.

4.3. Searching for the Best Combination of Parameters. Before the IP geolocation algorithm of Corr-SLG can geolocate an unknown target IP, we need to set three key parameters to make the geolocation result as accurate as possible. This part is responsible for searching for the best combination of parameters. In this part, we geolocate the landmark dataset and try to find out the combination of parameters whose corresponding median error distance is the least. This part is the critical process of Corr-SLG and consists of three steps.

4.3.1. Selecting Candidate Landmarks in Each Layer. In this step, a landmark is treated as a target IP and the other landmarks are used to geolocate it. Candidate landmarks are selected from the landmarks which share the common routers with target IP.

For the target IP, first, we need to check its router path and extract all the common routers its path includes. Then, each layer will choose its own candidate landmarks, respectively. As shown in Figure 9, R is one common router that belongs to the target IP T . $L_i (L1, L2, L3, \dots)$ are landmarks of this layer. Corr-SLG divides the landmarks of this layer into three groups based on two key parameters, C_a and C_b . If individual Corr is larger than C_a , the landmark belongs to Group A; if individual Corr is less than C_b , the landmark belongs to Group B; the other landmarks belong to Group C. The range of C_a is $[0, 1]$. The range of C_b is $[-1, 0]$.

FIGURE 9: Candidate landmark selection strategies of Corr-SLG.

The candidate landmark selection strategy is different in each group: the candidate landmark of Group A is the landmark which has the smallest relative delay to target IP; the candidate landmark of Group B is the landmark which has the largest relative delay to target IP; and the candidate landmark of Group C is selected randomly. If the number of landmarks in one layer is less than 5, all of them are chosen as candidate landmarks. In this way, each layer can select at most 3 candidate landmarks or at least 1 candidate landmark.

Wang-Geo method has to give up all the data associated with “inflating routers.” Because Corr-SLG selects candidate landmarks in each layer and the influence of the inflating router is the same for all the landmarks in the same layer, there is no need to discard any router or landmarks even when the relative delay is negative. This can help increase accuracy.

4.3.2. Discarding Outliers. In this step, we need to discard landmarks that may be wrongly selected as candidate landmarks. For example, landmarks which do not have an individual Corr or are randomly selected from Group C may be very far from target IP.

The wrongly selected candidate landmarks can be discarded by detecting outliers. We gather the candidate landmarks from each layer. If the number of all candidate landmarks is more than 2, we can use LOF (local outlier factor) algorithm [25] to detect the outliers. Candidate landmarks are ordered by LOF value in ascending order. The LOF value of outliers is bigger than the others. Another key parameter R is used here to control the number of outliers that will be discarded. Only the first $R\%$ of all candidate landmarks will be kept. The range of R is $[1, 100]$.

If there are no more than 2 candidate landmarks, there is no need to detect outliers. The geolocation result is the average of the location of all remaining candidate landmarks.

4.3.3. Finding Minimum Median Error Distance. Only after the three key parameters are set, Corr-SLG can get a geolocation result. This paper searches for the best combination of parameters by finding the minimum median error distance of the landmark dataset. At the first time to geolocate the landmark dataset, C_a is 0, C_b is -1 , and R is 1. After getting the geolocation result of all landmarks, the median error distance is calculated. Then, both C_a and C_b are increased by 0.1 while R is increased by 1 at one time. All the landmarks are geolocated again using new combination parameters. At last, all combinations of three parameters and their corresponding median error distance are gathered. The best combination of parameters is the one that has the minimum median error distance.

4.4. Geolocating the Unknown Target IP. First, the probing host measures the router path to target IP and then measures the delay to target IP and middle routers. The delay between each pair of hosts is measured many times, and only the minimum one is selected. Then, probing host can geolocate target IP based on the best parameters in the same way as geolocating the landmark. The only difference is that the parameters to geolocate the target IP are already determined. If there is only one probing host, its geolocation result is the estimated location of the target IP; if there is more than one probing hosts, the average of the geolocation results of all probing hosts is the estimated location of the target IP.

5. Evaluation

To test whether Corr-SLG can increase the accuracy of street-level IP geolocation, this paper makes experiments in one province capital city of China, Zhengzhou.

5.1. Experiment Dataset. The previous experiments in Wang-Geo method [18] and Checkin-Geo [16] use web landmarks as landmark dataset. To keep consistent with previous work, this paper also uses web landmarks as landmark dataset. WiFi landmarks with known locations are used as the target IP dataset. This paper discovers 3104 websites based on organization names in Zhengzhou. After the websites which may not put their servers inside the organizations are discarded, 181 web landmarks are retained. 163 WiFi landmarks are collected in the way illustrated in Section 3. In this experiment, only one probing host located in our lab is used to reduce the deployment cost.

5.2. Searching for Best Parameters. The probing host searches for the best parameters based on the landmark dataset. When $C_a = 0.45$, $C_b = 0.35$, and $R = 0.35$, the median error distance of Corr-SLG for the landmark dataset is the smallest, 3.34 km. The median error distance of Wang-Geo

method for the landmark dataset is 8.95 km. The best parameters of Corr-SLG can decrease 63.27% of the median error distance. Figure 10 shows the cumulative probability of error distances of Corr-SLG and Wang-Geo method on the landmark dataset.

5.3. Geolocating Target IP Dataset. Then, the probing host geolocates the target IP dataset based on the best parameters. The median error distance of Corr-SLG for target IP dataset is 4.82 km. The median error distance of Wang-Geo method is 7.85 km. Corr-SLG can decrease 38.59% of the median error distance. Figure 11 shows the cumulative probability of error distances of Corr-SLG and Wang-Geo method for target IP dataset. Based on the experiment results, Corr-SLG can increase the accuracy of street-level IP geolocation by about 38.59% in Zhengzhou.

From Figures 10 and 11, we can see that, at first, the performance of Corr-SLG is much better than Wang-Geo and, in the end, Corr-SLG is slightly better. This means the following: (1) for targets with smaller error distances, Corr-SLG is more accurate than Wang-Geo; (2) for targets with larger error distances, Corr-SLG is almost similar to Wang-Geo.

This phenomenon is caused by the following reasons. There are 3 kinds of hosts: (1) the smallest delay comes from the shortest distance; (2) the smallest delay comes from the longest distance; and (3) the smallest delay comes from random distance. Corr-SLG is much better than Wang-Geo on the second kind of hosts and equal to Wang-Geo on the first kind of hosts. For the last kind of hosts, both Wang-Geo and SLG are not accurate.

Please note that, in cumulative probability, the smaller error distances are shown (added) first. For Wang-Geo, the hosts with smaller error distance are only of the first kind. For Corr-SLG, the hosts with smaller error distance include both the first and second kinds of hosts. Therefore, at first, there are much more hosts with smaller error distances for SLG than those for Wang-Geo. However, for both Corr-SLG and Wang-Geo, the error distances for the third kind of hosts are large. Moreover, they are all much larger than the previous 2 kinds of hosts. Hence, in the end (for hosts with larger error distance), the performance of Corr-SLG seems similar to Wang-Geo.

The main reason for the phenomenon is that Corr-SLG and Wang-Geo are both based on the delay-distance relationship. However, the delay-distance relationship is not clear for the third groups of hosts. How to improve the accuracy for the third kind of hosts remains as our future work.

Besides, we can also see that the increased degree of the target IP dataset is clearly less than that of the landmark dataset. That is because the best parameters for the landmark dataset may not be exactly suitable for the target IP dataset. If there are several kinds of hosts in one city and the difference between their network characteristics is relatively significant,

FIGURE 10: Error distances of Corr-SLG and Wang-Geo method for landmark dataset.

FIGURE 11: Error distances of Corr-SLG and Wang-Geo method for target IP dataset.

such as hosts belonging to different ISP (Internet Service Provider), it is suggested that using the landmark dataset whose network characteristic is similar to the target IP may achieve better accuracy.

6. Conclusion

To improve the performance of street-level IP geolocation in weakly connected networks, this paper proposes a measurement-based street-level IP geolocation algorithm called Corr-SLG. First, this method introduces the landmarks associated with multilayered common routers into candidate landmark selection. This aims to make sure that most landmarks near to target IP have a chance to participate in candidate landmark selection. Second, Corr-SLG selects candidate landmarks in each layer of the common router to avoid the influence of inflating routers. Third, it divides landmarks into three groups based on individual *Corr* and uses different candidate landmark selection strategies in different groups. This can achieve better accuracy in weakly connected networks where the smallest delay may not come from the closest landmark. Last but not least, we present a new street-level landmark collection method called WiFi landmark which is inspired by rapid-growth WiFi service. WiFi landmarks can be of great help in cities where the number of web landmarks is insufficient. In this work, we find that there are three different kinds of relative-delay-distance correlation through measurements in two real inside-city networks. This finding helps to improve the accuracy of Corr-SLG. However, we still lack theoretic explanations as to what the reason for different relative-delay-distance correlation is. In our future work, we will carry out network measurements in more different cities in various countries and try to explain the cause of different relative-delay-distance correlation. This can help us to extend Corr-SLG to more cities and have better understanding of inside-city network.

Data Availability

The datasets of this work are available from the corresponding author on reasonable request.

Conflicts of Interest

The authors declare no conflicts of interest in publishing this article.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. U1636219, 61602508, 61772549, U1736214, and U1804263), the National Key R&D Program of China (Nos. 2016YFB0801303 and 2016QY01W0105), and the Zhongyuan Science and Technology Innovation Leader Talent Project (No. 214200510019).

References

- [1] X. Zhang and Q. Zhu, "Information-centric virtualization for software-defined statistical qos provisioning over 5g multimedia big data wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 8, pp. 1721–1738, 2019.
- [2] X. Ge, H. Wang, R. Zi, Q. Li, and Q. Ni, "5g multimedia massive mimo communications systems," *Wireless Communications and Mobile Computing*, vol. 16, no. 11, pp. 1377–1388, 2016.
- [3] Y. Kang, C. Kim, D. An, and H. Yoon, "Multipath transmission control protocol-based multi-access traffic steering solution for 5g multimedia-centric network: design and testbed system implementation," *International Journal of Distributed Sensor Networks*, vol. 16, no. 2, Article ID 1550147720909759, 2020.
- [4] Y. Zhang, R. Deng, E. Bertino, and D. Zheng, "Robust and universal seamless handover authentication in 5g hetnets," *IEEE Transactions on Dependable and Secure Computing*, vol. 99, p. 1, 2019.
- [5] C. Liu, X. Luo, F. Yuan, and F. Liu, "Rnbg: a ranking nodes based ip geolocation method," in *Proceedings of the IEEE INFOCOM 2020-IEEE Conference On Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 80–84, IEEE, Toronto, ON, Canada, 2020.
- [6] Z. Wang, H. Li, Q. Li, W. Li, H. Zhu, and L. Sun, "Towards ip geolocation with intermediate routers based on topology discovery," *Cybersecurity*, vol. 2, no. 1, pp. 1–14, 2019.
- [7] V. N. Padmanabhan and L. Subramanian, "An investigation of geographic mapping techniques for internet hosts," *ACM SIGCOMM Computer Communication Review*, vol. 31, no. 4, pp. 173–185, 2001.
- [8] B. Gueye, A. Ziviani, M. Crovella, and S. Fdida, "Constraint-based geolocation of internet hosts," *IEEE/ACM Transactions on Networking*, vol. 14, no. 6, pp. 1219–1232, 2006.
- [9] E. Katz-Bassett, J. P. John, A. Krishnamurthy, D. Wetherall, T. Anderson, and Y. Chawathe, "Towards ip geolocation using delay and topology measurements," in *Proceedings Of the 6th ACM SIGCOMM Conference on Internet Measurement*, pp. 71–84, ACM, Rio de Janeiro, Brazil, 2006.
- [10] B. Wong, I. Stoyanov, and E. G. Sirer, "Octant: a comprehensive framework for the geolocalization of internet hosts," in *Proceedings of NSDI*, vol. 7, p. 23, Cambridge, MA, USA, 2007.
- [11] 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 IEEE INFOCOM*, pp. 2841–2845, Rio De Janeiro, Brazil, 2009.
- [12] D. Li, J. Chen, C. Guo et al., "Ip-geolocation mapping for moderately connected internet regions," *IEEE Transactions on Parallel and Distributed Systems*, vol. 24, no. 2, pp. 381–391, 2013.
- [13] J. Chen, F. Liu, T. Wang, X. Luo, F. Zhao, and G. Zhu, "Towards region-level ip geolocation based on the path feature," in *Proceedings of the Advanced Communication Technology (ICACT), 2015 17th International Conference*, pp. 468–471, IEEE, Phoenix Park, Korea, 2015.
- [14] M. Fotouhi, A. Anand, and R. Hasan, "Plag: practical landmark allocation for cloud geolocation," in *Proceedings of the 2015 IEEE 8th International Conference on Cloud Computing*, pp. 1103–1106, IEEE, Washington, DC, USA, 2015.
- [15] F. Yuan, F. Liu, R. Xu, Y. Liu, and X. Luo, "Network topology boundary routing ip identification for ip geolocation," in *Proceedings of the International Conference On Artificial Intelligence And Security*, pp. 534–544, Springer, Cairo, Egypt, 2020.
- [16] H. Liu, Y. Zhang, Y. Zhou, D. Zhang, X. Fu, and K. Ramakrishnan, "Mining checkins from location-sharing services for client-independent ip geolocation," in *Proceedings of IEEE INFOCOM*, pp. 619–627, Toronto, ON, Canada, 2014.
- [17] H. Jiang, Y. Liu, and J. N. Matthews, "Ip geolocation estimation using neural networks with stable landmarks," in

- Proceedings of 2016 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 170–175, IEEE, San Francisco, CA, USA, 2016.
- [18] Y. Wang, D. Burgener, M. Flores, A. Kuzmanovic, and C. Huang, “Towards street-level client-independent ip geolocation,” in *Proceedings of NSDI*, vol. 11, p. 27, Boston, MA, USA, 2011.
 - [19] Y. Shavitt and N. Zilberman, “A geolocation databases study,” *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 10, pp. 2044–2056, 2011.
 - [20] I. Poesse, S. Uhlig, M. A. Kaafar, B. Donnet, and B. Gueye, “IP geolocation databases,” *ACM SIGCOMM Computer Communication Review*, vol. 41, no. 2, pp. 53–56, 2011.
 - [21] D. Wu, J. Shi, and N. Mamoulis, “Density-based place clustering using geo-social network data,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 5, pp. 838–851, 2017.
 - [22] S. Ding, X. Luo, M. Yin, Y. Liu, and F. Liu, “An ip geolocation method based on rich-connected sub-networks,” in *Proceedings of the Advanced Communication Technology (ICACT), 2015 17th International Conference*, pp. 176–181, IEEE, Phoenix Park, Korea, 2015.
 - [23] A. Ziviani, S. Fdida, J. F. De Rezende, and O. C. M. Duarte, “Toward a measurement-based geographic location service,” in *Proceedings Of Passive And Active Network Measurement*, pp. 43–52, Springer, Juan-les-Pins, France, 2004.
 - [24] GeographicLib, 2020, <https://sourceforge.net/projects/geographiclib/>.
 - [25] M. M. Breunig, R. T. Ng, and J. Sander, “Lof,” *ACM SIGMOD Record*, vol. 29, no. 2, pp. 93–104, 2000.