IP Geolocation based on identification routers and local delay distribution similarity

Fan Zhao¹ | Xiangyang Luo¹ | Yong Gan² | Shuodi Zu¹ | Qingfeng Cheng¹ | Fenlin Liu¹

¹State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou 450001, China
²School of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou 450001, China

Correspondence
Xiangyang Luo, State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou 450001, China. Email: luoxxy_ieu@sina.com

Funding information
National Natural Science Foundation of China, Grant/Award Number: U1636219, 61379151, 61572445 and 61572052; National Key R&D Program of China, Grant/Award Number: 2016YFB0801303 and 2016QY01W0105; Plan For Scientific Innovation Talient of Henan Province Grant/Award Number: 2018JR0018

Summary
IP geolocation is usually used in fog computing to avoid high latency and discriminate malicious requests by judging the location of users. Existing delay measurement-based IP geolocation approaches are not applicable to the network that has hierarchical topology and weak connectivity, and the precision of the classical Street-Level Geolocation (SLG) method will decrease dramatically when the common routers are anonymous. In this paper, an IP geolocation method based on identification routers and local delay distribution similarity is proposed. The target IP's location at city-level is firstly derived by matching its routing path with the identification routers that only forward packets to the same city. After that, the target IP's local delay between the nearest common router and the target IP is gathered, and the landmarks' are obtained at the same time. Finally, the location of the landmark that has the most similar local delay distribution with the target IP is taken as the geolocation result. Theoretical analysis and experimental results show that the proposed method can derive reliably geolocation results at city-level for the target IP in the network with hierarchical architecture. Moreover, the geolocation accuracy of classical SLG method is improved obviously when the common routers are anonymous.

KEYWORDS
anonymous router, city-level geolocation, identification router, IP geolocation, local delay distribution similarity, street-level geolocation

1 | INTRODUCTION

Fog computing is proposed to enable computing directly at the edge of the network, which can provide elastic resources and services to end users through fog nodes.¹,² IP geolocation technology, which aims to obtain a network device’s location only with its IP address, has attracted extensive attention of researchers,³ and it plays an important role in fog computing. For example, judging the geographical locations of users using IP geolocation technology could help the fog computing service providers to select nearby fog nodes for users such as to improve the quality of service.⁴ With IP geolocation technology, the service providers can authenticate users and identify malicious requests based on their location. In addition, with the continuous improvement of IP geolocation technology, it is expected that it will play a role in geolocating hidden servers in Tor networks⁵,⁶ and data in Cloud Computing.⁷,⁸

In recent decade years, many methods have been proposed in IP geolocation. Existing IP geolocation methods can be divided into two categories according to the granularity of geolocation results: city-level geolocation and street-level geolocation.

The city-level geolocation methods aim to obtain the location of target IP at city-level granularity and can be further categorized into two types: measurement-based and database-based. The measurement-based geolocation method is the focus of current research in IP geolocation. They usually geolocate the target IP based on delay, topology, or both, simultaneously. These methods include GeoPing,³ CBG (Constraint-based...
Geolocation), Learning-based Geolocation (LBG), Topology-based Geolocation (TBG), Octant, and so on. The GeoPing is proposed in. It utilizes the delay measured from multiple probing hosts to construct the delay vectors for the target and the landmarks (IPs with known geographical locations) and takes the location of the landmark who has the most similar delay vector with the target as the target’s estimated location. CBG algorithm tries to find a linear conversion relationship between delay and geographical distance. It converts the delay measured from several probing hosts to distance constraints according to the conversion relationship and determines the target’s location based on multilateration. TBG algorithm is proposed in the work of Katz-Bassett et al. It estimates the target’s location into a convex region based on the delay get from probing hosts and geolocates the target and intermediate routers at the same time by minimizing the average error. Octant algorithm calculates both positive and negative constraints of each probing host, that is, constraints on where the target can and cannot be, and uses geographic information of routers along the path to the target by undns tool. Then, Monte-Carlo algorithm is used to select the best point from the estimated region as the target’s location. LBG algorithm reduces the IP geolocation to a machine-learning classification problem. It considers a set of lightweight measurements from a set of probing hosts to the target and then classifies the location of the target based on the most probable geographic region given probability densities learned from a training set. In addition, we have proposed a city-level geolocation method based on routing feature, which judges the target IP’s location by using the characteristics of the network that has hierarchical topology. Besides the measurement based geolocation method, another method to obtain the target’s city-level geographical location is to query the IP geolocation databases. There are several IP geolocation databases that provide query service on the Internet such as Whois, Akamai, Maxmind, Quova, IP2location, IP138, etc. Researchers evaluate the accuracy and reliability of some of these databases. The works of Shavitt and Zilberman and Poese et al point out that the query result is credible at national-level granularity, but it is not reliable at city-level. In addition, the coverage of these databases are usually concentrated in several popular countries, which makes them not appropriate to provide generic IP geolocation query services. The performances of 3 popular databases (the database names are not given) are evaluated in Dan et al using the test data of 10 countries, and the results show that the average correct rate at city-level granularity is less than 70%.

Street-level geolocation for a target is more difficult than city-level geolocation. Existing street-level geolocation methods include Street-Level Geolocation (SLG), Checkin-Geo, Geo-NN, etc. SLG is one of the typical methods, and its geolocation process can be summarized in 3 steps. Firstly, the multilateration in CBG is used to geolocate the target in a large area. Secondly, a large number of landmarks are mined from the Internet in this area. The relative delay is ingeniously exploited in multilateration again to further refine the area. Finally, the location of the landmark who has the minimum relative delay with the target is regarded as the final geolocation result. Social networks are now widely used. Checkin-Geo uses large amounts of users’ social data to geolocate the target. Clustering and classification methods are used to accurately extract the user’s geographical location from check-in in mobile social network as well as the IP address from information of login in fixed device. Then, the geographical location and the IP address are associated by the same user account. In Geo-NN algorithm, a RBF (radial basis function) trained from measured delay is used to estimate the region where the target is located. Then, a multi-layer perceptron neural network trained from the data collected from that region is used to geolocate the target to a street-level geographical location.

Researchers have proposed a series of IP geolocation methods in the past decades as mentioned above. However, existing measurement-based city-level geolocation methods are not suitable for the network with weak connectivity or hierarchical architecture, and the query results of several databases are often inconsistent. The geolocation accuracy of the SLG method is significantly reduced when the common routers are anonymous. In order to solve these problems, an IP geolocation method based on identification routers and local delay distribution similarity is proposed. This method geolocates the target at city-level granularity, utilizing the characteristic of network with hierarchical architecture and weak connectivity, and determines the target’s street-level location by using the relationship between the routing paths and delay distribution.

The rest of the paper is organized as follows. Section 2 points out the deficiency of existing city-level geolocation methods, when applied to the network with hierarchical topology and weak connectivity, and Street-Level Geolocation (SLG) method, when the common routers are anonymous. The framework and main steps of the proposed method are presented in Section 3. Section 4 analyzes the geolocation method we proposed. The method’s performance is evaluated by experiments in Section 5. Conclusions are drawn and future directions are outlined in Section 6.

2 PROBLEM FORMULATION

This section points out the defects of existing city-level geolocation methods and the flaw in SLG method when the common routers are anonymous.

2.1 Analysis of existing typical city-level geolocation methods

Measurement-based geolocation methods such as CBG and Octant can geolocate the target IP at city-level granularity in strongly connected network. In addition, the city where the target is located can also be obtained by querying the IP geolocation databases. Determining the target
IP's location at city-level granularity can not only satisfy the needs of many Internet service providers for providing online services but also is very important for street-level geolocation.

Nevertheless, city-level geolocation technology is still worth researching. One reason is that the query results from several different IP geolocation databases are often inconsistent, which leads to the fact that we cannot determine the true geographical location of the target. The other reason is that existing city-level geolocation methods based on measurement need to be improved because of their low reliability. For example, multilateration with geographic distance constraints is used in CBG algorithm as illustrated in Figure 1A, where the probing hosts are denoted as $L_1$, $L_2$, and $L_3$. $g_i$ is the real geographic distance, $r_i$ is the additive geographic distance distortion, and the target is denoted as $\tau$. Figure 1B shows the conversion between measured delay and geographic distance, where “baseline” is used to obtain an absolute physical lower bound on the RTT and “bestline” is closest to, but below all data points (distance, RTT) and has a non-negative intercept for each probing host.

However, the work of Li et al.\textsuperscript{23} points out that “though traditional IP-geolocation mapping schemes that depend on strong delay-distance correlation work well for rich connected Internet regions, they are not suitable for moderately connected Internet regions.” Actual tests also prove that the geographical distance converted from delay based on the “bestline” in CBG is much larger than actual geographical distance, which results in big geolocation error. Similarly, the converted geographical distance in TBG will be much lose if the correlation between delay and geographical distance is weak in the network. Also, it is difficult to get the “negative constraints” for Octant in the weakly connected network. LBG algorithm geolocates the target in the region with the maximum probability based on the actual measurement results. Although it does not consider the connectivity of the distance and delay, its geolocation accuracy is still need to be improved.

Thus, the proper conversion relationship between delay and distance is not easy to get for the network with weak connectivity. In addition, in the network with hierarchical architecture, the communication between two hosts that is located in different region or network operator usually relays on the top of exchange centers to forwarding the packets. This results in a great deviation between the packet propagation distance and the actual geographical distance between these two hosts. This further makes it difficult to obtain the exact conversion relationship between the measured delay and the geographical distance. In this situation, the geolocation performances of the above methods are seriously affected.

### 2.2 Analysis of traditional street-level geolocation method

SLG (Street-Level Geolocation) is a typical three-tier high-precision geolocation method based on the idea that gradually narrows the estimated geographic scope of the target IP. The geolocation process of SLG is present in Figure 2.

SLG method can be summarized into 3 tiers. In the first tier, the delay of the target is measured from lots of probing hosts and converted to geographical distance constraints. And multilateration in CBG algorithm is used to geolocate the target to a coarse-grained region. In the second tier, more landmarks in the coarse-grained region are obtained by web data mining, and the landmarks connected with the target through common routers are found out by paths detection. In addition, the relative delay is ingeniously exploited in this tier. As shown in Figure 3, $D_1$ represents the delay between the router $R_1$ and the landmark $L_1$, and $D_2$ denotes the delay between the router $R_1$ and the target $T$, then $(D_1 + D_2)$ is called the relative delay between $L_1$ and $T$. The minimum relative delay between the landmarks and the target is converted to distance constraints, and multilateration is used again to further refine the target’s location. In the third tier, the number of landmarks is increased again, and the landmark whose relative delay is the minimum is selected as the estimated location of the target. In the process of geolocation, the tier 2 can be repeated multiple times according to requirement.

Facts proved that SLG method has good applicability and high geolocation accuracy. However, there is still a flaw in SLG that it can only recognize the common routers that are non-anonymous in tier 3, but the true nearest common routers cannot be found once they are anonymous. Then,
selecting the landmark that has the minimum relative delay with the target’s may lead to large errors. As illustrated in Figure 3, when the routers \((R_1, R_2, \text{and } R_3)\) located within the dashed box are anonymous, the SLG will not be able to identify the real nearest common router and is very probably not going to select \(L_1\) as the estimated location of the target based on the principle of minimum relative delay. In other words, SLG can hardly select the landmark nearest to the target utilizing the minimum relative delay principle when the common routers are anonymous.

To solve these problems, a geolocation method based on identification routers and local delay distribution similarity is proposed in this paper. The method derives the city where the target is located according to the identification routers of that city and takes the location of the landmark whose local delay distribution is the most similar with the target as the final geolocation result.

### 3.1 Framework and main steps

The proposed geolocation method based on identification routers and local delay distribution similarity includes 2 parts: city-level geolocation and street-level geolocation. The framework is displayed in Figure 4, and the main steps are as follows.
FIGURE 4  Framework of proposed geolocation method

Input: a target IP, landmarks, candidate cities
Output: the target IP’s street-level geographical location

1. Topology discovery for all landmarks located in candidate cities. Traceroute tool is used to obtain the routing paths of all landmarks (denoted as Landmark Set 1) located in all candidate cities and the target IP from several probing hosts. Then, we can get the topology of landmarks in Landmark Set 1.

2. Identification routers discovery. The IPs of the routers that forward packets to only one city are found out by examining the landmarks’ routing paths. We regard these IPs as identification IPs of that city. In this step, each candidate city can determine its unique identification routers. The algorithm of identification router discovery is described in detail in Section 3.2.

3. Routing path matching and city discrimination. The routing path of the target IP collected in step 1 is matched with all the candidate cities’ identification IPs. If the target’s routing path contains an identification IP, the city corresponding to this IP is taken as the city-level geolocation result of the target.

4. Topology discovery within a city. Traceroute tool is used again to get the routing paths of the landmarks (denoted as Landmark Set 2) located in the city determined in step 3, that is, the city where the target is located. Then, landmarks that are connected to the target through the nearest common routers are screened and retained, and we denote these landmarks as Landmark Set 3.

5. Local delay distribution obtaining. The local delay of the target and landmarks in Landmark Set 3 are obtained by measuring and calculating. The process of local delay measurement and calculation is described in detail in Section 3.3. The histogram is used to get the local delay distribution of the target and landmarks in Landmark Set 3 by statistically analyzing the data.

6. Similarity calculation. Relative entropy is used to evaluate the similarity between two distributions. The relative entropy of local delay distributions of the target and each landmark in Landmark Set 3 is calculated, and the location of the landmark whose local delay distribution is the most similar to the target’s is taken as the geolocation result.

The key of proposed geolocation method includes 3 steps: identification routers discovery, local delay distribution obtaining, and similarity calculation. They will be described in detail as follows.

3.2 Identification router discovery

In our previous work, we have proposed an identification routers search algorithm based on decision tree in the work of Zhao et al. However, this algorithm cannot derive all identification routers, and the efficiency needs to be improved. This paper proposes an identification router discovery algorithm based on forward path lookup. Compared with the algorithm in the aforementioned work, this algorithm is simpler and more efficient. The forward path is the routing path from probing host to destination host, and the backward path is opposite. The process of identification IPs discovery algorithm is shown in Figure 5. The main steps are as follows:

Input: Candidate city set and topology of landmarks in landmark set 1.
Output: The identification IPs of each candidate city

I. Take a landmark routing path. The routing path of the first landmark in landmark set 1 is extracted.
II. Take the 1st hop of the landmark routing path. The first hop IP in the landmark routing path is taken as the first IP to be examined.
III. Landmark extraction. The landmarks (in landmark set 1) whose routing paths contain the examined IP are extracted.
IV. Location consistency judgment. The location consistency of the landmarks extracted in II is judged as yes if all the landmarks are located in the same city, then the examined IP is stored in identification IP database and turn to VI.
V. Take the forward hop IP. The forward hop IP of the current examined IP is taken as the next IP to be examined and turn to III.
### FIGURE 5  Process of identification IPs discovery

VI. Check whether all landmarks are examined or not. Landmarks in landmark set 1 are checked. If all of them are examined, then the algorithm goes to end or goes to II.

Finally, we can get the identification IPs of each candidate city.

#### 3.3 Local delay distribution obtaining

Here, we describe in detail the measurement and calculation of local delay and the method of obtaining its distribution.

1. Measurement and Calculation. Local delay measurement is one of the key steps in proposed method. The set of landmarks (marked as $L = \{L_1, L_2, \ldots, L_n\}$) connected to the target by the nearest common routers is determined by topology discovery and analysis. The local delay of landmarks in $L$ is obtained by measuring the delay between the nearest common router, and the target's is obtained in the same way. The delay of the landmarks and target is measured several groups in a period of time. The measurements in a group are repeated several times at the same time, and the minimum delay is reserved. We take Figure 3 as an example to elaborate how to measure and calculate the local delay. The delay of $L_i$ ($i = 1, 2, 3, 4$), $T$ and $R_t$ is denoted as $L_i, T_t, a n d R_t$, and the local delay of $L_i$ is calculated by $L_i, T_t - R_t$. Similarly, the local delay of $T$ is obtained by $T_t - R_t$. Ultimately, a large number of local delays of $L_i$ and $T$ are obtained.

2. Histogram Statistics. In order to avoid the influence of outliers, we remove the abnormal values and use the histogram to gather the local delay distributions of the landmarks and target.

   For each landmark or target, its local delays in ascending order are denoted as $(t_1, t_2, \ldots, t_h)$, where $h$ is the number of measurements. We divide the local delay into $m$ numbers of intervals and gather statistics of the count of local delay that presents in each interval, which are denoted as $M_1, M_2, M_3, \ldots, M_m$. The frequency of every interval is given by $M_1/h, M_2/h, \ldots, M_m/h$. We use the frequency histogram of an IP to represent its local delay distribution. For example, the frequency histogram of 125.41.72.1 is shown in Figure 6.

#### 3.4 Calculation of distribution similarity

Usually, the similarity between two probability density distribution $P$ and $Q$ is measured by calculating the relative entropy that is also called Kullback-Leibler divergence, referred to as KLD. Typically, $P$ and $Q$ represent the true and theoretical distribution of data, respectively. The relative entropy is zero when $P$ and $Q$ are the same, and the relative entropy would be increased with the increasing of the difference between two distributions.

Let $P_T(X)$ be the local delay distribution of the target $T$ and $Q_L(X)$ be the local delay distribution of the landmark $L_i$ that has the nearest common router with the target. Then, the similarity between them is calculated by Equation (1):

$$D_{KL} (P_T (X) || Q_L (X)) = \sum_{x \in X} P_T (x) \log \frac{P_T (x)}{Q_L (x)}$$

(1)

where $X$ is the range of local delay, which is determined according to actual measurements.
Finally, the similarity between the target and the local delay distribution of each landmark is compared, and the landmark $\hat{l}_i$, whose local delay distribution is the most similar with the target's, is selected as the estimated location of the target as shown in Equation (2):

$$\hat{i} = \arg \min_{i} D_{KL}(P_T(X) || Q_i(X)).$$  \hspace{1cm} (2)

For example, the relative entropies between the target 222.137.96.1 and the landmarks connected with the target by the nearest common router 218.29.102.1 are shown in Table 1. According to the calculation results of the relative entropy in Table 1, the location of landmark 125.41.72.1 is taken as the estimated location of the target 222.137.96.1.

### 4 ANALYSIS OF PROPOSED METHOD

The key of proposed method to geolocating the target at city-level granularity is the existence of identifying IPs in each candidate city, and they can be found out by the proposed algorithm of identification router discovery. The premise that the target’s location at street-level granularity can be determined is that there is a strong correlation between the similarity of routing paths and delay distributions. The proposed method is analyzed from the following three aspects: the existence of identification IPs, the performance of the geolocation algorithm based on identification IPs, and the relationship between the similarity of routing paths and delay distributions.

#### 4.1 Existence analysis of identification router

The topology, device deployment, routing strategy, and so on usually have their own characteristics in a network with strict hierarchical architecture such as China’s Internet. Moreover, only a small number of routing and forwarding devices are deployed between two different levels in the network with hierarchical architecture. In this paper, we deploy a probing host in Beijing and collect the routing paths of 1653 landmarks located...
in 7 cities including Luoyang, Jiaozuo, Xinxiang, Kaifeng, Xuchang, and Hebi in Henan province. We count the number of duplicate IPs per hop of all landmarks' routing paths. Figure 7 shows the distribution of statistics.

Figure 7 indicates that the number of duplicate IPs decreases significantly on the 9th hop. Therefore, it can be inferred that there is a hierarchical structure between the network where the probing host is located and the network of Henan province, and the 9th hop or its neighborhood connects the two levels. Figure 8 displays the statistical results of the duplicate IPs' number per hop of landmarks in 7 cities.
**TABLE 2** The IPs on the 10th hop of routing paths of the landmarks in Jiaozuo, Hebi, and Kaifeng city

| Jiaozuo City     | Hebi City      | Kaifeng City   |
|------------------|----------------|---------------|
| 61.168.193.38    | 219.158.101.222| 61.168.192.90 |
| 61.168.194.26    | 219.158.102.2  | 61.168.192.94 |
| 61.168.194.30    | 61.168.192.114 | 61.168.193.138|
| 61.168.192.86    | 61.168.192.254 | 61.168.194.122|
| 61.168.194.18    | 61.168.192.70  | 61.168.194.58 |
| 61.168.194.22    | 61.168.194.250 | 61.168.194.138|
| 61.168.192.94    | 61.168.194.90  | 61.168.194.142|
| 61.168.253.114   | 61.168.194.94  | 61.168.194.254|
| 61.168.192.142   | 61.168.254.226 | —             |
| 61.168.194.142   | —              | —             |
| 61.168.192.138   | —              | —             |
| 61.168.192.246   | —              | —             |

**FIGURE 9** The percentage of landmarks containing the processed IPs on the 10th hops in all landmarks in each city

Figure 8 illustrates that the number of duplicate IPs decreases significantly on the 10th hop in each city. Moreover, experimental results show that the IPs on the 10th hop is different from each other. Thus, we speculate that the identification IPs may on the 10th hop of the routing paths. The IPs on the 10th hop of routing paths of the landmarks in Jiaozuo, Hebi, and Kaifeng city are shown in Table 2.

Furthermore, IPs on the 10th hop of landmarks’ routing paths in 7 cities are processed as follows. For each city, the IPs that present with low frequency (we set it to 5 in this paper) or in the routing paths of landmarks located in different cities (for example, 61.168.253.114 appears in the landmarks routing paths located in Jiaozuo and Xinxiang cities) are removed. Then, we calculate the percentage of the number of landmarks containing the reserved IPs on the 10th hop and the number of all landmarks for each city. The calculation results are shown in Figure 9.

Figure 9 indicates that the IPs on the 10th hop appears on more than 92% landmarks’ routing paths. Therefore, the IPs on the 10th hop or their neighbors are likely to be the identification IPs, which also proves the existence of the identification routers.

### 4.2 Performance analysis of city-level geolocation based on identification router

Compared with the typical delay-based geolocation methods, the geolocation method based on identification routers has the following advantages:

1. No multiple probing hosts are required.

The methods based on delay measurement usually convert delay to distance constraint first and then geolocate the target based on multilateration. Therefore, these methods require multiple probing hosts and can give reliable geolocation results only when the probing hosts are located around the target. The proposed method can get more reliable results just using one probing host. The number of probing hosts and geolocation error of typical geolocation methods are displayed in Table 3. Planet-Lab is a global research network that consists of 1353 nodes with accurate geographical location. Planet-Lab nodes are usually used as probing hosts or landmarks in IP geolocation.

---

1. [https://www.planet-lab.org/](https://www.planet-lab.org/)
TABLE 3  Comparison of different city-level geolocation methods

| Methods   | The Number of Probing Hosts | Geolocation Results                  |
|-----------|-----------------------------|--------------------------------------|
| CBG       | 57 Planet-Lab nodes         | US: Median error < 100 km            |
|           |                             | Western Europe: Median error < 25 km |
| TBG       | 68 Planet-Lab nodes         | Median error: 67 km                  |
| Octant    | 51 Planet-Lab nodes and 53 traceroute servers | Median error: 35 km                |
| LBG       | 78 Planet-Lab nodes         | Average accuracy: 112 km             |
| Proposed  | ≥ 1                         | City-level accuracy >92%             |

2. Only the topology information is needed

The key for typical delay-based geolocation methods is measuring the delay between two hosts accurately. The delay is usually composed of transmission delay, propagation delay, processing delay, and queuing delay. Although the propagation delay is completely related to the length of the transmitted channel, however, in the actual network environment, the propagation delay is not always dominant in the measured delay. The effect that comes from the processing and queuing delays of the intermediate routers is difficult to be ignored, and the delay variation often occurs. On the contrary, the network routing protocols are required to be stable in the design phase, and their routing strategies are often maintained in a considerable period of time. Changes in topology and network traffic only results in small changes in routing strategies. Consequently, the routing path between two hosts is usually stable for a particular network.

4.3  Analysis of the relation between routing path and delay distribution

Inspired by the work of Eriksson et al.,10 the following assumptions are made in this paper: the more similar the forwarding paths of the two packets are, the more similar the delay distribution is. In order to verify this hypothesis, the following method is used.

The number of IP in landmark set is set to be \( m \), and the relative entropy between two delay distributions of IPs is calculated to measure their similarity. For a given IP, denoted as \( I_{0} \), the relative entropy between \( I_{0} \) and the rest \( m - 1 \) IPs is calculated and sorted from small to large, which is denoted as \( (d_{1}, I_{1}), (d_{2}, I_{2}), \ldots, (d_{m-1}, I_{m-1}) \) (where \( d_{1} \leq d_{2} \leq \cdots \leq d_{m-1} \)). \( n \) IPs are randomly selected from \( m - 1 \) IPs, and their rankings in the relative entropy sequence are found out denoted as \( (r_{1}, r_{2}, \ldots, r_{n}) \). Formula (3) is used to measure the average similarity between \( I_{0} \) and the \( n \) IPs,

\[
d_{\text{ave}} = \frac{r_{1} + r_{2} + \cdots + r_{n}}{mn},
\]  

(3)

where the smaller \( d_{\text{ave}} \) is, the higher the average similarity is.

In this paper, we randomly select 10 landmarks from landmark set which contains 505 landmarks (\( m = 505 \)). For each of these 10 IPs, two groups of IPs (each group contains 5 IPs, that is, \( n = 5 \)) whose routing paths are similar and dissimilarity to it are separately selected from the reserved. The relative entropy of the 10 IPs and the two groups are recorded as \( d_{\text{ave}1} \) and \( d_{\text{ave}2} \). The calculation results are presented in Table 4.

It can be seen from Table 4 that the delay distribution similarity between the eight IPs in this table and the IPs that have similar routing paths with them is higher. These experimental results show that delay distribution of two IPs is similar if their routing paths are similar.

TABLE 4  The calculation results of the average similarity between randomly selected IP and other IPs with different routing paths

| Sequence | Randomly Selected IPs | \( d_{\text{ave}1} \) | \( d_{\text{ave}2} \) |
|----------|-----------------------|-----------------------|-----------------------|
| 1        | 219.155.24.1          | 0.0814                | 0.1932                |
| 2        | 61.168.2.181          | 0.0763                | 0.1898                |
| 3        | 125.41.24.1           | 0.188                 | 0.397                 |
| 4        | 125.46.152.1          | 0.0966                | 0.1407                |
| 5        | 125.41.208.1          | 0.1102                | 0.4186                |
| 6        | 125.47.17.1           | 0.1864                | 0.4746                |
| 7        | 182.119.232.1         | 0.3305                | 0.2136                |
| 8        | 125.41.72.1           | 0.2424                | 0.3085                |
| 9        | 61.52.203.1           | 0.2915                | 0.3881                |
| 10       | 61.163.144.1          | 0.2051                | 0.1937                |
5 | EXPERIMENTS

In order to evaluate the performance of the proposed method, the related experiments are carried out. This section first introduces the settings of our experiment and then gives the experimental results along with detail analysis.

5.1 | Experiment setting

The distribution of experimental cities, dataset, probing hosts, and measurement tools are introduced as follows.

5.1.1 | Distribution of cities

In order to evaluate the performance of proposed method in city-level geolocation for the target IP, we carry out the experiment in China Unicom Network. The candidate cities include Zhengzhou, Luoyang, Jiaozuo, Xinxiang, Kaifeng, Xuchang, and Hebi city in Henan province of China. The geographical distribution of these cities is shown in Figure 10. The straight line distances between every pair of these cities are shown in Table 5.

5.1.2 | Dataset

The landmark set used in this paper consists of 2 parts. One is used to evaluate the performance of city-level geolocation in proposed method, and another is used to assess the effect of street-level geolocation.

The method based on identification routers we proposed only requires landmarks with location at city-level granularity. Landmarks used for evaluating our method and the learning-based geolocation approach in the work of Eriksson et al.\textsuperscript{10} are sampled from QQWry. QQwry database maintained by China’s Internet users contains a total of about more than 440000 IP address segments with geographical locations at city-level granularity. Sampling approach is as follows: First, selecting IP address segments that belong to China Unicom and located in above seven cities. Second, querying these IP address segments from databases IP138 and IPcn and reserving the records with consistent query results. Next, in order to make the measurement with low load, high efficiency, and insensitive, we randomly select about 1000 IPs in each city. Finally, we get their paths by traceroute program developed based on ICMP, and the reachable IPs are used in this experiment at last. The number of landmarks in each city can be seen in Figure 10.

| City   | Zhengzhou | Jiaozuo | Xinxiang | Kaifeng | Xuchang | Luoyang | Hebi   |
|--------|-----------|---------|----------|---------|---------|---------|-------|
| Zhengzhou | —        | 62.9    | 67.6     | 65.8    | 81.9    | 107     | 126.8 |
| Jiaozuo | 62.9      | —       | 63.0     | 110.5   | 143     | 95.5    | 112.1 |
| Xinxiang | 67.6      | 63.0    | —        | 67.8    | 142.2   | 152.4   | 59.5  |
| Kaifeng | 65.8      | 110.5   | 67.8     | —       | 97.4    | 172.7   | 105.8 |
| Xuchang | 81.9      | 143     | 142.2    | 97.4    | —       | 143.7   | 195.8 |
| Luoyang | 107       | 95.5    | 152.4    | 172.7   | 143.7   | —       | 206.9 |
| Hebi   | 126.8     | 112.1   | 59.5     | 105.8   | 195.8   | 206.9   | —     |

FIGURE 10 | Geographical distribution of experimental cities

TABLE 5 | Straight line distances between every pair of experimental cities (km)
It is difficult to obtain the landmarks with accuracy known geographical location. In order to test the performance of proposed method at street-level granularity, 505 landmarks with accuracy location in Zhengzhou are selected. The geographical distribution of these landmarks is displayed in Figure 11. The Leave-One-Out Cross-Validation is used in our experiments. That is, in each geolocation test, only one of the IPs is taken as the target, and the reserved are regarded as landmarks.

5.1.3 Probing hosts and measurement tools

The probing host is deployed in Beijing when assessing the accuracy of city-level geolocation of proposed method. In order to reproduce the comparison method proposed in the work of Eriksson et al.,\textsuperscript{10} we also deploy probing hosts in Zhengzhou, Nanjin, Xuzhou, and Chengdu city.

Traceroute technology is used to get the routing paths from the probing hosts to all landmarks and the targets. Traceroute tool is an effective way for observing how the packets flow in the Internet.\textsuperscript{26} In the phase of local delay distribution acquisition, the Ping program developed based on Winpcap is used to measure delays, and the measurement accuracy is microsecond. The landmarks and targets are measured 72 groups and each group contain 10 times. The minimum value of each group is taken as the final measured delay.

5.2 City-level geolocation experiment

In the experiments, a city is regarded as the smallest geolocation unit, and the above 7 cities are taken as candidate cities. The method in the aforementioned work\textsuperscript{10} takes the district (county) as the unit. It does not try to look for relationship between delay and distance but takes the delay, hops, and population of a specific region (district or county) as its characteristics. For the network with weak connectivity and hierarchical network topology, the method proposed in the same work\textsuperscript{10} is also feasible, so we compare it with our approach in terms of correct rate. The test and comparison results are shown in Figure 12.

Figure 12 indicates that the correct rate of geolocation at city-level granularity of proposed method is higher than 92%, and that of Learning-based method in Zhengzhou, Jiaozuo, Xinxiang, Kaifeng, and Hebi city are higher than 70% but about 50% in Luoyang and Xuchang city. The learning-based
method does not utilize the linear transformation relationship between delay and distance and distinguishes different region by the delay, hop counts, and populations. However, as a matter of fact, these factors could not be considered as accurate features of a region due to the bypass when communicating between two regions far away in the network with hierarchical network topology, which leads to the unsatisfactory geolocation results.

In contrast, the routing paths and geographic location have relatively closer relationship for the network with hierarchical topology because of the stabilization of routing paths. The proposed method cannot return a completely correct geolocation result due to the insufficient number of landmarks, the coverage ratio of the candidate cities, and anonymous routers. As shown in Figure 12, the experimental results demonstrate that the correct rate of proposed method is higher than that of the method proposed in the work of Eriksson et al.10

5.3 Street-level geolocation experiment

Table 6 gives the geolocation results of 15 IPs selected randomly. The columns of this Table include the target IPs, the landmarks used to estimate the target location, the geolocation errors of the proposed method and traditional SLG, and ranking of errors of the selected landmark to all landmarks that are connecting with the target by common routers.

Table 6 presents that the geolocation accuracy of proposed method is higher or equivalent (the bold line in the table) compared with the traditional SLG method. However, due to the limited number of measurements, the measured delays cannot describe the delay distribution of the landmarks and targets completely. This results in that the geolocation accuracy of a few targets is lower than that of the traditional SLG method.

The geolocation results of 505 targets are presented by the cumulative probability distribution of errors as shown in Figure 13.
Figure 13 exhibits that the median error of proposed method for 505 targets is about 4.3 km, and that of the traditional SLG is 5.8 km. On the whole, the proposed method has more accuracy geolocation results.

6 | CONCLUSION

In order to overcome the shortcomings of existing geolocation methods based on delay measurement and the traditional SLG method, the geolocation method based on identification routers and local delay distribution similarity is proposed in this paper. Then, the method is analyzed from three aspects, including the existence of identification routers, the performance of city-level geolocation, and the relationship between routing path and delay distribution. The experiments show that the proposed method improves the reliability of geolocation results at city-level granularity compared with LBG algorithm, and improves the geolocation accuracy of the traditional SLG when the nearest common routers are anonymous. Although the proposed method achieves better results, it still needs to be improved. In the next step, we will study how to reduce the measurement overhead while obtaining the local delay distribution of landmarks and the targets.

ORCID

Fan Zhao http://orcid.org/0000-0002-4691-8963

REFERENCES

1. Yi S, Li C, Li Q. A survey of fog computing: concepts, applications and issues. In: Proceedings of the 2015 Workshop on Mobile Big Data (Mobidata ’15); 2015; Hangzhou, China.
2. Yi S, Qin Z, Li Q. Security and privacy issues of fog computing: a survey. In: Proceedings of the 10th International Conference on Wireless Algorithms, Systems, and Applications (WASA 2015); 2015; Qufu, China.
3. Padmanabhan VN, Subramanian L. An investigation of geographic mapping techniques for Internet hosts. ACM SIGCOMM Comput Commun Rev. 2001;31(4):173-185.
4. Ivan S, Wen S, Huang X, Luan H. An overview of fog computing and its security issues. Concurr Comput Pract Exp. 2015;28(10):2991-3005.
5. Ling Z, Luo J, Wu K, Yu W, Fu X. TorWard: discovery, blocking, and traceback of malicious traffic over Tor. IEEE Trans Inf Forensics Secur. 2015;10(12):2515-2530.
6. Ling Z, Luo J, Wu K, Fu X. Protocol-level hidden server discovery. In: Proceedings of the 32nd IEEE International Conference on Computer Communications (INFOCOM); 2013; Turin, Italy.
7. Li J, Zhang Y, Chen X, Xiang Y. Secure attribute-based data sharing for resource-limited users in cloud computing. Comput Secur. 2018;72:1-12. https://doi.org/10.1016/j.cose.2017.08.007
8. Li P, Li J, Huang Z, Gao C, Chen W, Chen K. Privacy-preserving outsourced classification in cloud computing. Cluster Comput. 2017;1:1-10. https://doi.org/10.1007/s10586-017-0849-9
9. Gueye B, Ziviani A, Crovella M, Fdida S. Constraint-based geolocation of internet hosts. IEEE/ACM Trans Netw. 2006;14(6):1219-1232.
10. Eriksson B, Barford P, Sommersy J, Nowak R. A learning-based approach for IP geolocation. In: Proceedings of the 11th International Conference on Passive and Active Measurement (PAM 2010); 2010; Zurich, Switzerland.
11. Katz-Bassett E, John JP, Krishnamurthy A, Wetherall D, Anderson T, Chawathe Y. Towards IP geolocation using delay and topology measurements. In: Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement; 2006; Rio de Janeiro, Brazil.
12. Wong B, Stoyanov I, Sirer EG. Octant: a comprehensive framework for the geolocation of Internet hosts. In: Proceedings of the 4th USENIX Symposium on Networked Systems Design & Implementation (NSDI); 2007; Cambridge, MA.
13. Spring N, Mahajan R, Wetherall D. Measuring ISP topologies with Rocketfuel. ACM SIGCOMM Comput Commun Rev. 2002;32(4):133-145.
14. Zhao F, Song Y, Liu F, Ke K, Chen J, Luo X. City-level geolocation based on routing feature. In: Proceedings of the 29th International Conference on Advanced Information Networking and Applications (AINA); 2015; Gwangju, South Korea.
15. Shavitt Y, Zilberman N. A geolocation databases study. IEEE J Sel Areas Commun. 2011;29(10):2044-2056.
16. Poesio I, Uhlig S, Kaafar MA, Donnet B, Gueye B. IP geolocation databases: unreliable? ACM SIGCOMM Comput Commun Rev. 2011;41(2):53-56.
17. Dan Q, Parikh V, Davison BD. Improving IP geolocation using query logs. In: Proceedings of the 9th ACM International Conference on Web Search and Data Mining (WSDM); 2016; San Francisco, CA.
18. Wang Y, Burgener D, Flores M, Kuzmanovic A, Huang C. Towards street-level client-independent IP geolocation. In: Proceedings of the 8th USENIX Conference on Networked Systems Design and Implementation (NSDI); 2011; Boston, MA.
19. Liu H, Zhang Y, Zhou Y, Zhang D. Mining checks from location-sharing services for client-independent IP geolocation. In: Proceedings of the 33rd IEEE International Conference on Computer Communications (INFOCOM); 2014; Toronto, Canada.
20. Jiang H, Liu Y, Matthews JN. IP geolocation estimation using neural networks with stable landmarks. In: Proceedings of the 35th IEEE on Computer Communications Workshops (INFOCOM WKSHPS); 2016; San Francisco, CA.
21. Wen S, Haghighi MS, Chen C, Xiang Y, Zhou WL, Jia WJ. A sword with two edges: propagation studies on both positive and negative information in online social networks. IEEE Trans Comput. 2015;64(3):640-653.
22. Jiang J, Wen S, Yu S, Xiang Y, Zhou W. Identifying propagation sources in networks: state-of-the-art and comparative studies. IEEE Commun Surv Tutor. 2017;19(1):465-481.
23. Li D, Chen J, Guo C, et al. IP-geolocation mapping for moderately connected internet regions. IEEE Trans Parallel Distrib Syst. 2013;24(2):381-391.
24. Prieditis A, Chen G. Mapping the internet: geolocating routers by using machine learning. In: Proceedings of the 4th International Conference on Computing for Geospatial Research & Application (COM.Geo); 2013; San Jose, CA.

25. Tian Y, Dey R, Liu Y, Ross KW. China’s internet: topology mapping and geolocating. In: Proceedings of the 31st IEEE International Conference on Computer Communications (INFOCOM); 2012; Orlando, FL.

26. Huffaker B, Plummer D, Moore D, Claffy AK. Topology discovery by active probing. In: Proceedings of the applications and the Internet (SAINT) Workshops; 2002; Nara, Japan.

**How to cite this article:** Zhao F, Luo X, Gan Y, Zu S, Cheng Q, Liu F. IP Geolocation based on identification routers and local delay distribution similarity. *Concurrency Computat Pract Exper*. 2019;31:e4722. [https://doi.org/10.1002/cpe.4722](https://doi.org/10.1002/cpe.4722)