Intelligent Automation & Soft Computing
DOI:10.32604/iasc.2020.011902
Article

Tech Science Press

City-Level Homogeneous Blocks Identification for IP Geolocation

Fuxiang Yuan, Fenlin Liu, Chong Liu and Xiangyang Luo*

School of Cyberspace Security, PLA Strategic Support Force Information Engineering University, Zhengzhou, 450001, China
*Corresponding Author: Xiangyang Luo. Email: luoxy_ieu@sina.com
Received: 04 June 2020; Accepted: 27 June 2020

Abstract: IPs in homogeneous blocks are tightly connected and close to each other in topology and geography, which can help geolocate sensitive target IPs and maintain network security. Therefore, this manuscript proposes a city-level homogeneous blocks identification algorithm for IP geolocation. Firstly, IPs with consistent geographic location information in multiple databases and some landmarks in a specific area are obtained as targets; the /31 containing each target is used as a candidate block; vantage points are deployed to probe IPs in the candidate blocks to obtain delays and paths, and *alias* resolution is performed. Then, based on the analysis of paths of all IPs in blocks as well as last-hop routers of paths, conditions are set to identify homogenous blocks, and the city-level location of each homogenous block is analyzed based on the identification of city topology boundary IPs. Finally, the size of each homogeneous block is expanded step by step and the new block is identified until the largest city-level homogeneous block containing each target IP is identified. Experiments are conducted in many cities in China and the US. Results show that the proposed algorithm has a good effect on the identification of city-level homogeneous blocks, and the location accuracy of IPs in homogeneous blocks is about 99.4%. When the identified homogenous blocks are applied to target IP geolocation, the average geolocation accuracy of probing reachable target IPs is about 95.7%; when applied to landmark expansion, the number of landmarks can be greatly increased, thereby the success rate of existing geolocation algorithm such as SLG is improved.

Keywords: Homogeneous block; IP geolocation; landmark expansion; network measurement; network security

1 Introduction

Today, despite the continuous development and maturity of the Internet, the network security situation is still very serious. The Internet is facing attacks from botnets, Trojans, viruses and Distributed Denial of Service (DDoS), etc. Accurately geolocate sensitive target IPs used in various types of attacks is of great significance for attack prevention, judgment, and forensics [1–4].

Obtaining a large number of IPs that are tightly connected, close in topology and geographically attributed to the same city is expected to help quickly and accurately geolocate target IPs [5,6]. IPs within an ordinary block seem to have the above characteristics. However, different levels of blocks have

different sizes, IPs in a block may be scattered in topology, and the locations are not necessarily close. Geographically, a large number of blocks cover far beyond the range of a city, so IPs in a block may belong to multiple cities. In addition, for the purpose of privacy protection, the Internet Service Providers (ISPs) do not disclose the geographic locations of blocks, and the accuracy of the block locations given by existing databases is difficult to guarantee [7,8]. These problems make it impossible to simply use ordinary blocks to obtain a large number of IPs that can be applied for geolocation.

Some existing studies have mentioned the concept of homogeneous block and indicated that relative to the same source IP, IPs within a homogeneous block have the same or highly similar paths and are tightly connected. These IPs are close to each other in the network topology and their geographic locations should be very close [9]. Therefore, identifying homogenous blocks belonging to a city is very important for geolocating sensitive target IPs, tracing malicious network behaviors, sensing network situation, and maintaining network security [10–14].

There are few existing studies on homogenous block identification. Many classic studies in the field of network measurement and IP geolocation often do not focus on the identification of homogeneous blocks, and simply treat fixed-size blocks (such as /24s) as basic units and homogeneous blocks. For example, when probing a target network, blocks with certain sizes are used as units and only a small amount of IPs within each block are probed to reduce the measurement task [15]. When analyzing the network topology, /24s are used as units and the topology is generated based on the aggregated BGP prefix information [10,16]. When geolocating target IPs, fixed-size blocks are considered homogeneous, and a target IP's location is estimated based on the location of the block to which the IP belongs [17,18]. Some existing public databases such as IP2 Location and Maxmind so assume that some blocks are homogeneous and based on this give the locations of IPs within a block. In addition, homogeneous blocks are also widely used in related researches on network services, such as optimizing network services based on homogeneous blocks, judging and using nearby servers to provide services to customers, thereby improving the quality of service [15,19,20].

From the above introduction we can see the importance of homogenous blocks for IP geolocation. Existing studies do not have much in-depth discussion on the identification of homogeneous blocks, but simply treat fixed-size blocks (such as /24s) as homogeneous blocks and use them, which may affect network characteristics analysis and target IP geolocation. In addition, many classic studies have pointed out that in the real network environment, although the path from a source IP to a target IP is relatively stable for a period of time, load balancing is widespread. Due to this factor, paths from a source IP to IPs in a block may still be different, which has a greater impact on the identification of homogeneous blocks [21–23]. Therefore, it is necessary to design an algorithm that can accurately identify homogenous blocks belonging to a city under the real network environment, and then provide support for the target IP geolocation. This manuscript conducts research on these issues and the main work is as follows.

1. The meaning and significance of homogeneous blocks are summarized; problems in the identification of homogenous blocks in existing studies and the impact of these problems on network measurement and IP geolocation are elaborated.
2. A city-level homogeneous blocks identification algorithm for IP geolocation is proposed. Based on the statistical analysis of the entire paths and the last-hop routers in paths of IPs in a large number of blocks, conditions for identifying homogenous blocks are set; combined with the identification of city topology boundary IPs in paths, the location of each homogeneous block is analyzed.
3. Experiments are designed to verify the effectiveness of the proposed algorithm. Homogeneous blocks belonging to many cities in China and the US are identified, and the effect of obtained homogeneous blocks on target IP geolocation, landmark expansion, and existing geolocation methods are tested; results show that the proposed algorithm can help IP geolocation.

The rest of the manuscript is structured as follows. In Section 2, the homogenous block as well as its significance are introduced, and then problems in existing studies are analyzed. In Section 3, the main steps of the proposed algorithm are given and the key steps such as probe data processing, homogeneity identification, and homogenous block location analysis are elaborated in detail. In Section 4, the experimental settings are given, and then the effect of the proposed algorithm and the help for landmark expansion as well as IP geolocation are verified. In Section 5, the full text is summarized and future work is pointed out.

2 Problems Description

In this section the meaning as well as significance of homogenous blocks are briefly introduced first, and then problems in existing relevant studies are analyzed.

2.1 Homogeneous Block

Generally, a large number of IPs are aggregated into blocks and allocated to different networks by ISPs. This means that a particular network will be assigned one or more blocks, and blocks often vary in size according to different needs. IPs located in different networks or in different locations of the same network often exhibit different characteristics. These IPs are difficult to be close to each other. A homogenous block is a special type of block. Although the existing studies do not clearly give its specific meaning, some studies indicate that IPs within a homogeneous block are tightly connected and have the same or very similar network characteristics, especially in terms of route. Relative to the same source IP, paths of IPs in a homogeneous block are usually the same or highly similar. In topology, these IPs often have the same location, and the geographical location should also be very close [9,10].

Homogeneous blocks are of great importance for network measurement, network characteristics analysis, and target IP geolocation. For example, when measuring a network, some of the IPs in homogeneous blocks can be selected as the probe targets. This will reduce the workload of network measurement and increase efficiency. As IPs within a homogeneous block are close, when geolocating a target IP, the location of the target can be given according to the location of the homogeneous block containing the target. In addition, many existing geolocation algorithms such as SLG [24], LENCR [25], etc. require a large number of landmarks to successfully geolocate targets, the number of landmarks is one of the key factors that determine the success of geolocation [26]. Homogenous blocks to which landmarks belong can be analyzed and IPs within the blocks are used as landmarks, which is expected to expand the number of landmarks and provide support for existing geolocation algorithms.

2.2 Problems in Some Existing Related Studies

In view of the characteristics of homogeneous block, many classic studies related to network measurement and IP geolocation often use this special block, but there is no specific in-depth analysis and discussion of it. These studies often assume that fixed-size blocks are homogeneous. For example, as mentioned in the Introduction, /24s are often regarded as homogeneous blocks and only one IP is taken from each /24 to probe [15]. IPs in a /24 are considered to be in the same city and the location of the target IP is given based on the location of the /24 to which the target belongs [18]. Although /24 is a relatively common prefix, if there is no in-depth study on the identification of homogeneous blocks and /24s are directly used as homogeneous blocks, this will affect the accuracy of network measurement and the success rate of IP geolocation.

In addition to the above problem, when applying the homogeneous block mentioned in existing studies to IP geolocation, there are also the following two problems.

1. Some existing studies believe that paths of IPs in homogeneous blocks are the same or highly similar relative to the same source. This view may affect the application of homogeneous blocks in IP geolocation. This is because on the one hand, some studies such as [21,22] show that in the actual Internet environment, due to the prevalence of load balancing, although the path from a source IP to a target IP is often stable for a period of time, the paths from a source IP to IPs in a block are often different. Through analysis of probe results of a large number of homogeneous blocks, it is found that this phenomenon does exist. On the other hand, although the condition that paths of IPs in a block are identical is more conducive to geolocation, the condition is sufficient but unnecessary. If the paths of IPs in a homogeneous block are the same, or paths with load balancing can be aggregated, or the last-hop routers of the paths are the same, or the route table entries contain relationships such as inclusion and crossover, so that the locations of IPs are close, the homogeneous block is expected to be applied to IP geolocation.
2. The homogeneous block mentioned in the existing studies may be difficult to apply directly to IP geolocation. As mentioned above, a large number of IPs with tight connections, similar locations in topology and belonging to the same city can be applied to geolocation. Although existing studies indicate that IPs in homogeneous blocks are topologically and geographically close to each other, there is no guarantee that these IPs are located in the same city and they may be scattered among adjacent cities. Therefore, only homogeneous blocks belonging to a city, that is, city-level homogeneous blocks can be used for IP geolocation.

Through analysis, it can be seen that the identification of city-level homogeneous blocks that suitable for IP geolocation still needs further study.

3 The Proposed Algorithm

In view of the above problems, a city-level homogeneous block identification algorithm for IP geolocation is proposed. IPs in a specific area are regarded as targets, and /31s containing the targets are selected as candidate blocks. Based on the analysis of paths of all IPs in blocks as well as last-hop routers of paths, conditions are set to identify homogenous blocks. Based on the identification of city topology boundary IPs, the city-level location of each homogenous block is analyzed. The size of each homogeneous block is expanded step by step and the new block is identified. A large number of city-level homogeneous blocks in a specific area can be obtained through the proposed algorithm, which can provide support for IP geolocation. In this section, the main steps of the proposed algorithm will be described in detail.

3.1 Main Steps

The proposed algorithm mainly includes IP set construction, candidate block selection, candidate block analysis, multi-source probe and other steps. The detailed steps are as follows. The principle framework is shown in Fig. 1.

Input: An IP set of a specific area.

Output: A large number of city-level homogeneous blocks in the specific area.

Step 1: IP set construction. For a specific area, by querying multiple location databases, IPs with consistent locations in multiple databases are selected. These IPs and the existing landmarks in the area together form the set $\boldsymbol{U}$ of the area. The IPs in $\boldsymbol{U}$ are scattered in different blocks.

Step 2: Candidate block selection. Let $\forall IP_i \in \boldsymbol{U}$, IP_i is any IP that has not been identified in $\boldsymbol{U}$. In order to analyze the city-level homogenous block to which IP_i belongs, IP_i is used as a target. The initial block containing IP_i is selected as the candidate homogeneous block, which is denoted as /b, $1 \leq b \leq 31$, and $b = 31$ at this time.

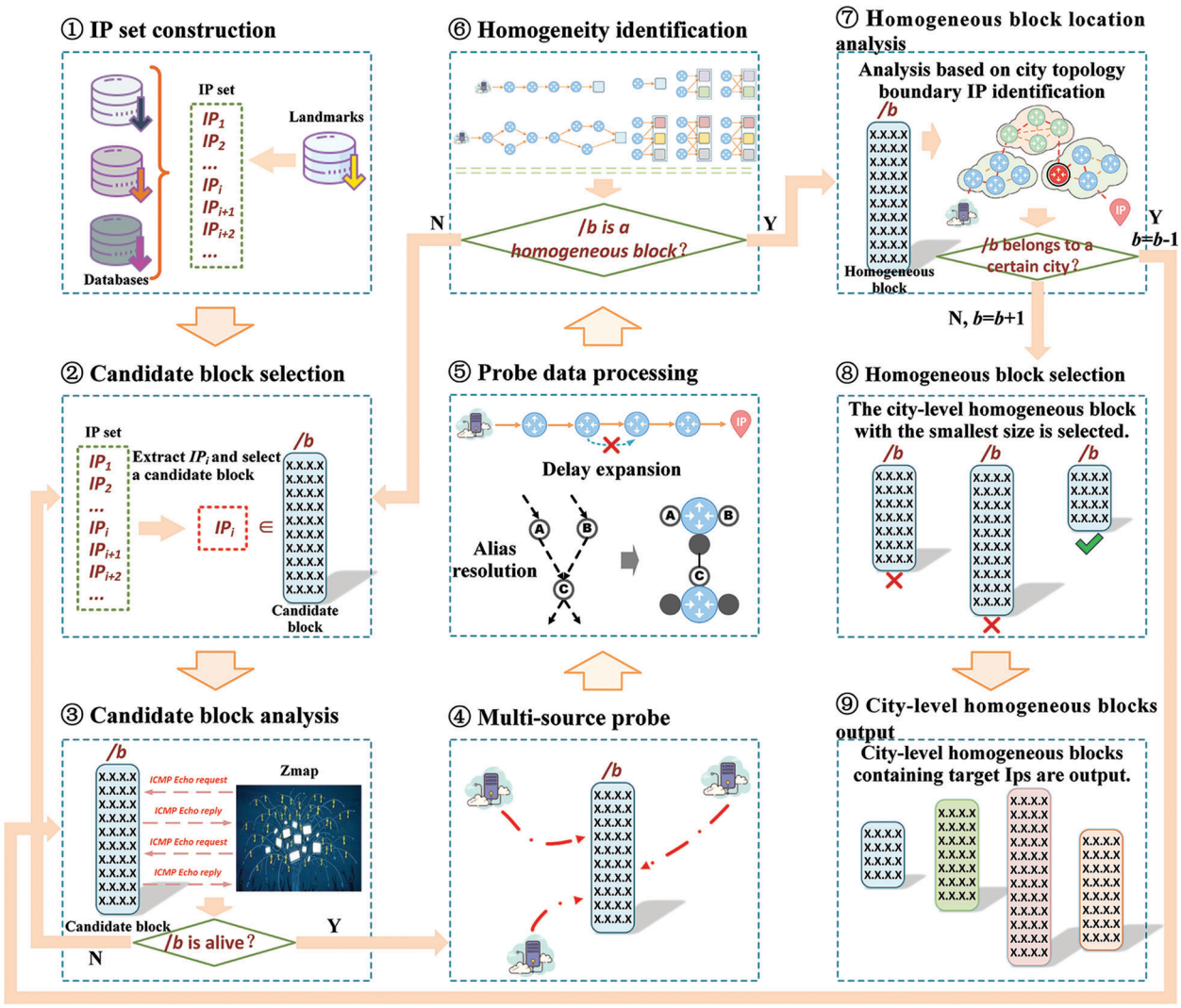

Figure 1: The principle framework

Step 3: Candidate block analysis. Zmap is used to scan /*b*. ICMP Echo request packets are sent to each IP within /*b*, and whether /*b* is alive is analyzed according to the response. If an ICMP Echo reply packet can be returned by each IP in /*b*, that is, a response packet can be returned for the request packet, then /*b* is considered to be alive and Step 4 is executed continuously. Otherwise, /(*b*+1) containing IP_i is output as the largest city-level homogeneous block (if $b = 31$ at this time, there is no homogeneous block containing IP_i), and then return to Step 2.

Step 4: Multi-source probe. Vantage points (here after VPs) S_1~S_n are deployed around the specific area. Using traceroute, IPs in the candidate block /*b* are probed multiple times to obtain data such as paths and single-hop delays.

Step 5: Probe data processing. For paths and delays of /*b*, noise in the data should be removed, such as links between IPs with extremely low frequency, or inflated delays and the corresponding links. Route IPs belonging to the same router are merged by *alias* resolution.

Step 6: Homogeneity identification. According to the probe data of /b from S_1~S_n, the discriminant conditions are used to analyze whether /b is a homogeneous block. According to the data obtained from S_x ($1 \le x \le n$), if /b is judged as a homogeneous block, the algorithm will be continued. Otherwise, /(b + 1) containing IP_i is output as the largest city-level homogeneous block (if b = 31 at this time, there is no homogeneous block containing IP_i), and then return to Step 2.

Step 7: Homogenous block location analysis. After the above steps, /b is judged as a homogeneous block, that is, IPs in /b are close to each other in the topology. The city-level location of /b is analyzed by performing city topology boundary IP identification on the path of any IP in /b. If /b belongs to a certain city, it will be expanded to /(b – 1), let $b = b - 1$, and return to Steps 3 to analyze the new block. Otherwise, /(b + 1) containing IP_i is output as a city-level homogeneous block (if b = 31 at this time, there is no homogeneous block containing IP_i), and then continue.

Step 8: Homogeneous block selection. After the iteration of Step 2~7, for each VP, a city-level homogeneous block containing IP_i can be obtained. Therefore, for the n blocks obtained from S_1~S_n, the one with the smallest size, that is, with the largest prefix is selected as the largest city-level homogeneous block to which IP_i belongs. Then return to Step 2.

Step 9: City-level homogeneous block output. After the above steps, the largest city-level homogeneous block to which each IP in $\boldsymbol{U}$ belongs can be obtained. All these homogenous blocks whose city-level locations have been determined are output as the final result of this algorithm.

In the above steps, probe data processing, homogeneity identification, and homogenous block location analysis are the key parts of this algorithm, which will be described in detail in Sections 3.2, 3.3, and 3.4, respectively.

3.2 Probe Data Processing

In order to better analyze the homogeneity of the candidate block, probe data such as paths, delays and routing IPs obtained are processed as follows:

- The path between two IPs in the network is relatively stable for a period of time, and the link between any two hops is often fixed. Some extremely low frequency links cannot reflect the route relationship between two hops, so they are removed.
- The inflated single-hop delay is meaningless for estimating the distance between two IPs. Therefore, only those minimum single-hop delays are kept. Single-hop delays within all cities in a specific area (VPs and landmarks belong to the same city) are calculated, and the maximum single-hop delay are obtained as the threshold D. Links between IPs with single-hop delays greater than D are removed.
- Typical *alias* resolution algorithms such as MIDAR [27], TreeNET [28], etc. are used to resolve a large number of route IPs within the paths. IPs belonging to the same router are merged to obtain IP-level, router-level nodes and links.

3.3 Homogeneity Identification

IPs in a homogeneous block are relatively close to each other in topology, but geographically, it is uncertain whether these IPs belong to the same city. Only the homogeneous blocks belonging to a certain city, that is, city-level homogeneous blocks, can be used for IP geolocation. Therefore, in order to obtain a city-level homogeneous block, the first step is to ensure that the block is homogeneous. In this section, how to determine whether the candidate block /b is a homogeneous block will be introduced in detail.

/b containing the target IP_i is composed of two /(b + 1)s. It can be seen from Section 3.1 that one of the /(b + 1)s containing IP_i has been judged as a homogenous block. However, /b still needs to be judged whether it satisfies the homogeneity. After probe data processing, /b is analyzed for the entire paths or the last-hop

routers. When at least one of the following conditions is met, IPs in /*b* are considered to be tightly connected and have similar locations in topology, that is, /*b* is a homogeneous block. The conditions are as follows.

(1) Entire paths analysis. Paths from S_x to /*b* are analyzed. If one of the following two conditions is met, /*b* is considered to be a homogeneous block:

a) There is only one path from S_x to /*b*. As shown in Fig. 2a, the light blue square indicates /*b*. Paths of all IPs in /*b* are exactly the same, in such a situation IPs are tightly connected, and the locations in the topology must be similar.

b) There are multiple paths from S_x to /*b*, but after merging different paths, different nodes and links form one or more rhombic or polygonal structures. As shown in Fig. 2b, due to the existence of load balancing, multiple paths appear between S_x and /*b*. Nodes *A* and *C* are load balancers. These paths diverge when passing *A* and *C*, and converge after a few hops. Although paths are different, IPs in /*b* are also tightly connected and still have similar locations in the topology.

(2) Last-hop routers analysis. The last-hop routers of paths of /*b* are analyzed. If one of the following two conditions is met, /*b* is considered to be a homogeneous block:

a) /*b* has a unique last-hop router. As shown in Fig. 3a, the last-hop routers of all IPs in /*b* are the same. At this time, all IPs are tightly connected and they must have similar locations in the topology.

b) /*b* has multiple different last-hop routers, but when the IPs in /*b* are sorted by numbers, the route table entries on the different last-hop routers have an "inclusive" or "crossover" relationship, making links between different routers and IPs in /*b* constitute one or more 'Z', "Bi-fan" [3], or a combination of both. As shown in Fig. 3b~3f, squares with different colors from top to bottom in the box represent the child blocks formed when the IPs are arranged in order. Delays between routers and IPs are lower than the single-hop delay threshold within a city. Under these circumstances, IPs in /*b* are tightly connected and have very close locations in the topology.

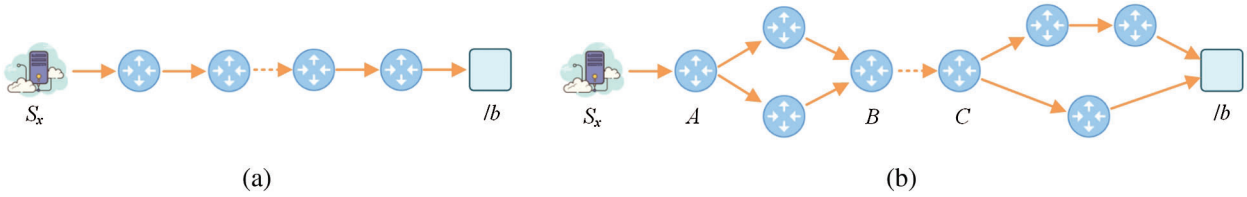

Figure 2: Entire paths analysis. (a) Only one path. (b) Multiple paths

3.4 Homogenous Block Location Analysis

When probing a target IP, a VP is often far away from the target, thus the path can be roughly divided into three parts: The path segment of the city to which the source IP belongs, the backbone network path segment, and the path segment of the city to which the target IP belongs. The first route IP that can indicate the path has entered the city to which the target belongs is the topology boundary IP of the target city. By judging the city to which the boundary IP belongs, the city to which the target IP belongs can be obtained. Therefore, in order to obtain city-level homogeneous blocks that can be used for IP geolocation, in this section, based on the difference in the single-hop delays of the path, city topology boundary IPs in paths of /*b* are identified. Then these boundary IPs are compared with boundary IPs in paths of landmarks to determine the location of /*b*. Details are as follows.

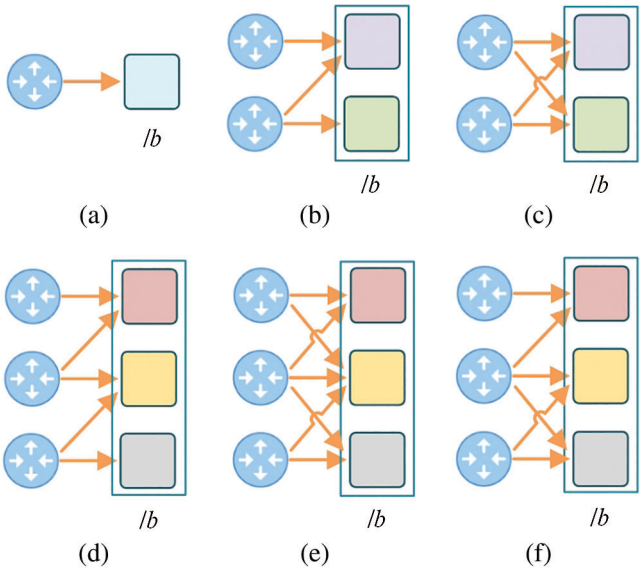

Figure 3: Last-hop routers analysis. (a) Only one last hop. (b) One Z. (c) One Bi-fan. (d) Multiple Zs. (e) Multiple Bi-fans. (f) Combination

(1) Boundary IP identification and extraction. IPs in /*b* and landmarks in a specific area are used as targets, boundary IPs are extracted from the paths as follows.

1. Suppose that the path P obtained from S_x to the target IP_{n+1} contains a total of n hops of intermediate route IPs. The $(n+1)$th hop is the target. From backward to forward, determine which IP is the city topology boundary IP. Let $j = n$, and regard the jth hop, that is, IP_j as the candidate boundary IP to be identified.
2. Suppose that the delay between the jth hop and the $(j+1)$th hop is T_j. To determine whether IP_j is a boundary IP, the single-hop delay T_j and T_{j-1} are extracted.
3. T_j and T_{j-1} are compared with the threshold D (obtained in Section 3.2):

- If $T_j < D$ and $T_{j-1} > D$, this means that the distance between IP_j and IP_{j+1} is less than the maximum distance between adjacent IPs in a city. IP_{j+1} is the internal IP of the city, so IP_j is also the internal IP. The distance between IP_{j-1} and IP_j is more than the maximum distance between adjacent IPs in the city, so IP_{j-1} is the external IP of the city. Therefore, IP_j is the first route IP that the path passes into the city, that is, the boundary IP. As shown in Fig. 4, the short lines with arrows indicates delays between adjacent hops, and the length of the line indicates the value of the delay. For IP_j, T_{j-1} indicated by the red short line is greater than D, and T_j indicated by the green short line is less than D, then IP_i can be identified as the boundary IP.
- If $T_j < D$ and $T_{j-1} < D$, this means that the distances between IP_j and IP_{j+1}, IP_{j-1} and IP_j are less than the maximum distance between adjacent IPs in a city. IP_{j+1} is the internal IP, so IP_j and IP_{j-1} are also the internal IPs. IP_j is not a boundary IP, but whether IP_{j-1} is a boundary IP still needs to be judged. Let $n = n - 1$ and return 1) to continue analysis.
- If $T_j > D$ and $T_{j-1} > D$, $T_j > D$ and $T_{j-1} < D$, this means that the single-hop delays may be inaccurate due to network congestion (according to statistics, it is found that a few hops close to the target often belong to the target city, and the single-hop delays should be less than the internal delay threshold of the city). Therefore, it is necessary to re-probe to obtain accurate delays, and then continue the above analysis.

(2) Location analysis of /*b*. Suppose that through the above steps, all the non-duplicate boundary IPs extracted from paths of /*b* form a set $\boldsymbol{X}$, and all non-duplicate boundary IPs extracted from the paths of landmarks of *m* different cities in a specific area constitute sets $\boldsymbol{Y_1}$, $\boldsymbol{Y_2}$, …, $\boldsymbol{Y_{m-1}}$, and $\boldsymbol{Y_m}$, respectively. $\boldsymbol{X}$ is compared with $\boldsymbol{Y_1}$, $\boldsymbol{Y_2}$, …, $\boldsymbol{Y_{m-1}}$, and $\boldsymbol{Y_m}$. If $\exists\ \boldsymbol{Y_k}$, $1 \le k \le m$, $\boldsymbol{X} \subseteq \boldsymbol{Y_k}$, this means that the path of every IP in /*b* passes into the city identified by the boundary IP in $\boldsymbol{Y_k}$, that is, /*b* belongs to the *k*th city. Otherwise, IPs in /*b* belong to different cities.

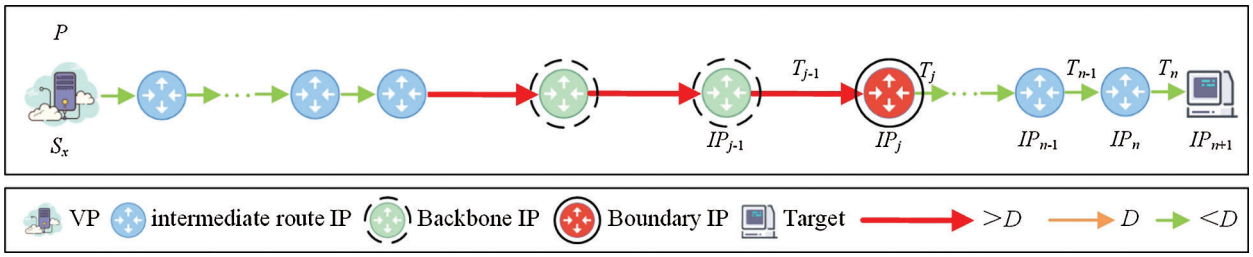

Figure 4: Boundary IP identification based on single-hop delay analysis

4 Experiments and Results

In order to verify the effect of the proposed algorithm, a number of IPs in different cities in China and the US are used to perform identification experiments. The experimental settings are described in Section 4.1. Identification tests of homogeneous blocks are conducted and results are analyzed in Section 4.2. In order to verify the help of the homogenous block for IP geolocation, landmark expansion, and existing geolocation methods, target IP geolocation tests are conducted in Section 4.3.

4.1 Experimental Settings

In this section, blocks with location and AS information are obtained from several databases such as IPIP, IPcn, IP2Location, Maxmind, and Chacuo. A large number of IPs within these blocks are selected for probe. Those reachable IPs and some landmarks are used as targets to analyze. VPs are mainly deployed in Beijing, Shanghai, Los Angeles, Washington, etc. 150 times of probe are performed on these targets for a period of 6 months. The detailed experimental settings are shown in Tab. 1.

4.2 Homogeneous Block Identification Tests

In this section, the identification results of city-level homogenous blocks in different cities are analyzed first, and then the effectiveness of the identification is verified by comparison with existing databases.

4.2.1 Homogeneous Block Identification Results

For target IPs in different cities such as Beijing, Shanghai, Guangzhou, etc., the largest city-level homogenous block to which each target belongs is identified. The identification results are shown in Tab. 2. In order to avoid the table being too wide and make the table more standardized, Beijing, Shanghai, Guangzhou, Zhengzhou, Wuhan, New York, Chicago, Washington, Los Angeles, and Miami in the table are abbreviated as BJ, SH, GZ, ZZ, WH, NY, C, W, LA, and M.

From the statistical results in Tab. 2, it can be seen that the largest city-level homogeneous blocks to which approximately 99.1% of the target IPs belong are /31s or larger. The proportion of IPs belonging to /24s is only about 18.4% on average. Most city-level homogeneous blocks are smaller than /24, and some are even larger, which suggests that it is inaccurate to directly treat the /24 as a homogeneous block. The largest city-level homogeneous block to which a target IP belongs still needs to be analyzed.

Table 1: Experimental settings

Country	City	Number of targets (include some landmarks)	Locations of VPs	Times of probe
CN	Beijing	13328	Zhengzhou, Beijing, Shanghai, Guangzhou, Hangzhou, Tianjin, Chengdu, Xian, Jinan, Wuhan	150
	Shanghai	12805		
	Guangzhou	12443		
	Zhengzhou	11661		
	Wuhan	11312		
US	New York	23492	New York, Chicago, Atlanta, Washington, Miami, Seattle, Los Angeles, Dallas, Silicon Valley, Phoenix	
	Chicago	12565		
	Washington	13669		
	Los Angeles	14335		
	Miami	23633		

Table 2: City-level homogeneous blocks identification results

Blocks	The number of city-level homogeneous blocks with different sizes in different cities									
	BJ	SH	GZ	ZZ	WH	NY	C	W	LA	M
/20	0	0	0	0	0	705	251	547	430	945
/21	133	0	0	233	113	1644	628	957	573	1182
/22	267	1152	747	583	792	1879	754	410	1003	2127
/23	813	2049	1991	1283	1018	940	2262	1230	1577	1418
/24	**3599**	**3457**	**2737**	**2799**	**2149**	**3759**	**2010**	**1640**	**1434**	**2600**
/25	2132	1665	1742	2332	1584	3289	1759	2050	1864	3309
/26	1200	640	1369	2099	1244	2114	1885	2324	2294	3545
/27	1333	896	1120	700	1357	2819	1005	1367	1434	2836
/28	1466	1024	622	583	1697	2584	754	957	1147	1418
/29	1066	1409	871	805	792	1644	503	683	1290	2127
/30	786	256	747	117	339	1492	377	820	860	1418
/31	493	154	436	0	90	470	245	492	340	444
Total	13288	12702	12382	11534	11175	23339	12433	13477	14246	23369
Target IP	13328	12805	12443	11661	11312	23492	12565	13669	14335	23633

4.2.2 Effect Verification of City-Level Homogeneous Block Identification

The existing databases can provide locations of a large number of IPs. However, some existing studies have shown that in some cases, different databases cannot give a consistent location for the same IP, and the reliability of a single database needs to be further improved, which is also one of the reasons why IP

geolocation still needs to be studied. In network measurement and IP geolocation, it is generally considered that when the IP location information given by multiple databases is consistent, the location is reliable. In order to verify the effect of city-level homogenous block identification, the locations of a large number of IPs obtained based on the city-level homogenous blocks are compared with the locations given by multiple IP location databases. For cities in China, TaobaoIP, IPIP, and IPcn are used for comparison, and for cities in the US, Maxmind, IP2Location, and Hostip are used. The comparison results are shown in Tabs. 3 and 4.

Table 3: Comparison result 1

Database	Proportion of IPs in city-level homogeneous blocks whose locations are consistent with that given by multiple databases (%)				
	Beijing	Shanghai	Guangzhou	Zhengzhou	Wuhan
TaobaoIP, IPIP, IPcn	98.6	99.3	98.9	98.2	98.0

Table 4: Comparison result 2

Database	Proportion of IPs in city-level homogeneous blocks whose locations are consistent with that given by multiple databases (%)				
	New York	Chicago	Washington	Los Angeles	Miami
Maxmind, IP2Location, Hostip	98.7	99.1	99.4	98.8	98.5

It can be seen from Tabs. 3 and 4 that the locations of IPs given by city-level homogeneous blocks have a high consistency with the locations given by multiple databases, the highest up to 99.4%, which suggests that the locations of city-level homogeneous blocks are relatively accurate. When geolocating a target IP, a more reliable location of the target IP can be given based on the location of the city-level homogeneous block to which the target IP belongs. In addition, the analysis of city-level homogeneous blocks can be used to calibrate IPs' locations that are inconsistent in multiple databases. This indicates that the city-level homogeneous blocks identified by the proposed algorithm can be applied to IP geolocation.

4.3 Target IP Geolocation Tests

In order to verify the help for IP geolocation, in this section, a large number of existing landmarks are used as targets, and city-level homogeneous blocks are used to geolocate these targets. Then, IPs in the homogeneous blocks are used as landmarks to analyze the help for landmark expansion. Finally, in order to verify the help of homogeneous blocks for the existing geolocation methods, the original landmarks and the expanded landmarks based on homogeneous blocks are used for geolocation, and the results are compared.

4.3.1 Target IP Geolocation Results

Using the proposed algorithm, while obtaining a large number of city-level homogeneous blocks, the locations of these blocks are also given. These blocks can be used to geolocate target IPs. Confirm whether a target IP belongs to an acquired block, if so, the city-level location of the block is used as the location of the target. If not, the proposed algorithm is used to analyze the largest homogenous block to which the target IP belongs and the location of the target IP is given according to the location of the

homogenous block. In order to further verify the identification effect on homogeneous blocks of the proposed algorithm, and the help for IP geolocation, landmarks of different cities are used as targets, the geolocation tests are performed. When the correct location of each target can be given, the geolocation is considered to be accurate. The results are shown in Tab. 5.

Table 5: Geolocation results using homogeneous blocks

Target	Target IP geolocation results in different cities									
	BJ	SH	GZ	ZZ	WH	NY	C	W	LA	M
Number	6883	7356	4451	3328	1766	3747	4435	3405	5056	3559
Number of targets accurately geolocated	6601	7084	4304	3142	1692	3492	4262	3282	4899	3395
Geolocation accuracy (%)	95.9	96.3	96.7	94.4	95.8	93.2	96.1	96.4	96.9	95.4

From the results in Tab. 5, it can be seen that the city-level homogeneous blocks identified by the proposed algorithm have relatively high geolocation accuracy for the targets, with a maximum accuracy of 96.9% and an average of 95.7%. After analyzing the probe data of the targets that cannot be accurately geolocated, it is found that almost all these targets are unreachable, which shows that the city-level locations of blocks obtained by the proposed algorithm are relatively reliable. Target IPs can be accurately geolocated at the city level based on the homogeneous blocks when the IPs are reachable.

4.3.2 Landmark Expansion Results

Using IPs belonging to homogeneous blocks as landmarks can greatly increase the number of landmarks available in each city. Therefore, the landmarks of different cities are further expanded based on homogeneous blocks. The results are shown in Tab. 6. It can be seen that the maximum number of landmarks in a city can be increased by 101 times, and the average is about 52 times. The success rate of target IP geolocation is expected to be increased with a large number of landmarks.

Table 6: Landmarks expansion based on homogeneous blocks

Landmark	The number of landmarks in different cities									
	BJ	SH	GZ	ZZ	WH	NY	C	W	LA	M
Original landmarks	411	235	398	205	168	433	347	266	312	446
Expanded landmarks	20551	13308	13396	20637	13876	28623	12271	11393	10084	10799
/20	0	0	0	0	0	3012	0	2936	0	0
/21	1201	0	0	0	0	3573	0	0	0	0
/22	1562	3014	741	7014	756	2998	1437	0	1431	0
/23	1014	1245	2002	1988	1254	612	306	983	1579	2302
/24	6985	4010	3785	5215	7005	7835	4010	2780	2055	3180
/25	6004	3121	3316	3974	3012	5133	2996	2701	1799	1645
/26	1617	1025	1765	1911	1243	3522	2088	819	1301	1415
/27	1333	365	721	238	344	1077	527	414	997	717

Table 6 (continued).

Landmark	The number of landmarks in different cities									
	BJ	SH	GZ	ZZ	WH	NY	C	W	LA	M
/28	302	196	513	86	123	421	332	342	355	677
/29	209	115	127	86	45	116	201	129	283	361
/30	180	49	211	125	37	146	191	148	131	245
/31	144	168	215	0	57	178	183	141	153	257

4.3.3 The Help for SLG

A sufficient number of landmarks is the premise and foundation of the IP geolocation algorithm that based on the connections between landmarks, the intermediate routers and the target. Whether suitable landmarks can be found in target area will determine whether target IP can be successfully geolocated by these algorithms. As mentioned above, SLG [24] is one of the most representative algorithms. Through three-layer geolocation, SLG continuously analyzes and narrows the geographic range of the target IP, and finally obtains the location. The general idea of the three-layer geolocation is: at the first layer, three-point geolocation is used to restrict the target IP to a wide area; at the second layer, by finding and probing landmarks in the area, the landmarks that have the closest common router with the target are obtained, and the distance constraints between landmarks and the target IP are used to narrow the target area; at the third layer, a number of landmarks are obtained and probed continually, and the landmark with the smallest relative delay to the target IP are selected, and its location is used as the target's location. The number of landmarks has a great influence on the success rate of SLG. Therefore, in order to further verify the effectiveness of the proposed algorithm in this manuscript, as well as the help for the existing geolocation methods, SLG is used for IP geolocation.

1000 IPs with known locations are selected as targets to be geolocated respectively from 10 cities, including Beijing, Shanghai, Guangzhou, New York, Chicago, and Washington, etc. At different times, the original and expanded landmarks of each city are used to geolocate the targets with SLG (its 2th and 3th layer geolocation are mainly used). Links between VPs, target IPs and landmarks are different in different periods, so the test is conducted twice. In the 2th and 3th layers of SLG, when landmarks can be found to narrow the area to which a target belongs and estimate its location, geolocation is considered to be successful, otherwise failed. When geolocating, a smaller homogeneous block containing each original landmark is preferentially used to ensure that the geographic distance between the expanded landmarks and the original landmark is not too large. Fig. 5 shows using different types of landmarks, the comparison of geolocation results of SLG.

It can be seen that whether it is the 2th or the 3th layer of SLG, with a large number of expanded landmarks, the success rates of geolocation are high, and compared with using original landmarks, the success rates of the two layers are increased by about 113.3% and 242.9%. Results show that with homogenous blocks, the success rate of SLG can be significantly improved.

The above experimental results indicate the good effect of the proposed algorithm on city-level homogeneous blocks identification, as well as the help for landmark expansion and IP geolocation.

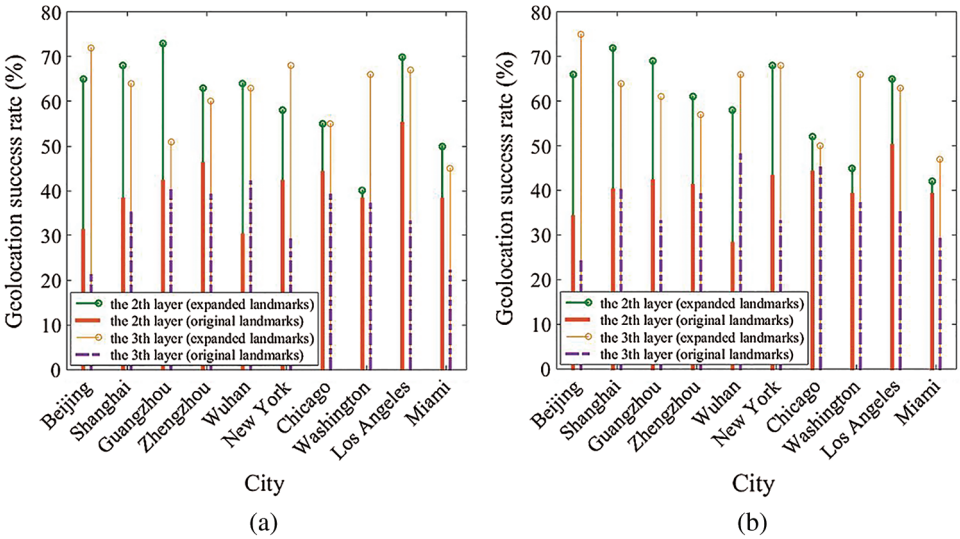

Figure 5: Geolocation results of SLG. (a) The first time. (b) The second time

5 Conclusion

There are few studies on homogenous block identification. Based on the definition of homogenous block in the existing studies and combined with the statistical analysis results of the probe data in the actual Internet environment, this manuscript proposes a city-level homogenous block identification algorithm for IP geolocation. A large number of city-level homogenous blocks can be obtained by the proposed algorithm to better help geolocate sensitive targets and maintain network security. Experiments are carried out in different regions and results show that the effect of the proposed algorithm is good, and the locations of the identified homogeneous blocks are accurate. Using the identified blocks for landmark expansion, the number of landmarks can be greatly increased and the success rate of existing geolocation method can be significantly improved. Although the proposed algorithm has achieved good results, it relies on existing landmarks and cannot handle blocks that have fewer alive IPs or that do not respond to probes. This will be further studied in the future.

Funding Statement: This work was supported by the National Natural Science Foundation of China (Nos. U1804263, U1636219), and the Science and Technology Innovation Talent Project of Henan Province (No. 184200510018).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] P. Gill, Y. Ganjali and B. Wong, "Dude, where's that IP?: Circumventing measurement-based IP geolocation," in *Usenix Security Sym.*, Washington, DC, USA, pp. 1–16, 2010.

[2] Z. Dong, R. D. W. Perera, R. Chandramouli and K. P. Subbalakshmi, "Network measurement based modeling and optimization for IP geolocation," *Computer Networks*, vol. 56, no. 2, pp. 85–98, 2012.

[3] W. Han, Z. Tian, Z. Huang, L. Zhong and Y. Jia, "System architecture and key technologies of network security situation awareness system YHSAS," *Computers, Materials & Continua*, vol. 59, no. 1, pp. 167–180, 2019.

[4] H. Zhang, Y. Yi and J. Wang, "Network security situation awareness framework based on threat intelligence," *Computers, Materials & Continua*, vol. 56, no. 3, pp. 381–399, 2018.

[5] D. Feldman, Y. Shavitt and N. Zilberman, "A structural approach for PoP geo-location," *Computer Networks*, vol. 56, no. 3, pp. 1029–1040, 2012.

[6] Y. Tian, R. Dey, Y. Liu and K. W. Ross, "Topology mapping and geolocating for China's internet," *IEEE Transactions on Parallel and Distributed Systems*, vol. 24, no. 9, pp. 1908–1917, 2013.

[7] I. Poese, S. Uhlig, M. A. Kaafar, B. Donnet and B. Gueye, "IP geolocation databases: Unreliable?," *ACM SIGCOMM Computer Communication Review*, vol. 41, no. 2, pp. 53–56, 2011.

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

[9] X. Cai and J. Heidemann, "Understanding block-level address usage in the visible internet," in *ACM Special Interest Group on Data Communication*, *New Delhi, India*, pp. 99–110, 2010.

[10] L. Quan, J. Heidemann and Y. Pradkin, "When the internet sleeps: Correlating diurnal networks with external factors," in *Internet Measurement Conf.*, Vancouver, BC, Canada, pp. 87–100, 2014.

[11] J. L. Zhang, W. Z. Wang, X. W. Wang and Z. H. Xia, "Enhancing security of FPGA-based embedded systems with combinational logic binding," *Journal of Computer Science and Technology*, vol. 32, no. 2, pp. 329–339, 2017.

[12] S. Zhou, M. Ke and P. Luo, "Multi-camera transfer GAN for person re-identification," *Journal of Visual Communication and Image Representation*, vol. 59, no. 1, pp. 393–400, 2019.

[13] F. Peng, J. Yang, Z. X. Lin and M. Long, "Source identification of 3D printed objects based on inherent equipment distortion," *Computers & Security*, vol. 82, pp. 173–183, 2019.

[14] Z. Zhang, Y. B. Li, C. Wang, M. Y. Wang and Y. Tu, "An ensemble learning method for wireless multimedia device identification," *Security and Communication Networks*, vol. 2018, 5264526, 2018.

[15] F. Chen, R. K. Sitaraman and M. Torres, "End-user mapping: Next generation request routing for content delivery," *ACM SIGCOMM Computer Communication Review*, vol. 45, no. 4, pp. 167–181, 2015.

[16] L. Quan, J. Heidemann and Y. Pradkin, "Trinocular: Understanding internet reliability through adaptive probing," *ACM SIGCOMM Computer Communication Review*, vol. 43, no. 4, pp. 255–266, 2013.

[17] Z. Hu, J. Heidemann and Y. Pradkin, "Towards geolocation of millions of IP addresses," in *Internet Measurement Conf.*, Boston, MA, USA, pp. 123–130, 2012.

[18] M. Gharaibeh, H. Zhang and C. Papadopoulos, "Assessing co-locality of IP blocks," in *IEEE Conf. on Computer Communications Workshops*, San Francisco, CA, USA, pp. 503–508, 2016.

[19] M. Calder, R. Gao and M. Schroder, "Odin: Microsoft's scalable fault-tolerant CDN measurement system," in *Networked Systems Design and Implementation*, Renton, WA, USA, pp. 501–517, 2018.

[20] M. Calder, X. Fan and L. Zhu, "A cloud provider's view of EDNS client-subnet adoption," in *Network Traffic Measurement and Analysis Conf.*, Vienna, Austria, pp. 129–136, 2019.

[21] P. Sermpezis, V. Kotronis, A. Dainotti and X. Dimitropoulos, "A survey among network operators on BGP prefix hijacking," *ACM SIGCOMM Computer Communication Review*, vol. 48, no. 1, pp. 64–69, 2018.

[22] B. Augustin, T. Friedman and R. Teixeira, "Measuring multipath routing in the internet," *IEEE/ACM Transactions on Networking*, vol. 19, no. 3, pp. 830–840, 2010.

[23] B. Donnet, M. Luckie, P. Mérindol and J. J. Pansiot, "Revealing MPLS tunnels obscured from traceroute," *ACM SIGCOMM Computer Communication Review*, vol. 42, no. 2, pp. 87–93, 2012.

[24] Y. Wang, D. Burgener and M. Flores, "Towards street-level client-independent IP geolocation," in *Sym. on Network System Design and Implementation*, Boston, MA, USA, pp. 365–379, 2011.

[25] J. Chen, F. Liu and Y. Shi, "Towards IP location estimation using the nearest common router," *Journal of Internet Technology*, vol. 19, no. 7, pp. 2097–2110, 2018.

[26] R. Li, Y. Liu, Y. Qiao, T. Ma, B. Wang *et al.,* "Street-level landmarks acquisition based on SVM classifiers," *Computers, Materials & Continua*, vol. 59, no. 2, pp. 591–606, 2019.

[27] K. Keys, Y. Hyun, M. Luckie and K. Claffy, "Internet-scale IPv4 *alias* resolution with MIDAR," *IEEE/ACM Transactions on Networking*, vol. 21, no. 2, pp. 383–399, 2013.

[28] J. F. Grailet and B. Donnet, "Towards a renewed *alias* resolution with space search reduction and IP fingerprinting," in *Network Traffic Measurement and Analysis Conf.*, Dublin, Ireland, pp. 1–9, 2017.