electronics

MDPI

Article

Root Mirror Sites Identification and Service Area Analysis

Jiachen Wang [1], Zhiping Li [2,*], Zhaoxin Zhang [1,*], Jian Chen [2], Chao Li [1] and Yanan Cheng [1]

[1] Faculty of Computing, Harbin Institute of Technology, Harbin 150001, China
[2] China Academy of Information and Communications, Beijing 100191, China
* Correspondence: lizhiping@caict.ac.cn (Z.L.); zhangzhaoxin@hit.edu.cn (Z.Z.)

Abstract: The operation of today's Internet can only be achieved with the domain name system (DNS), and the essential part of the DNS is the root servers. Adding anycast mirrors has been used to maintain the security of root servers, but many problems accompany this technique. In this paper, we used 36198 probe points deployed worldwide to probe 1160 root mirror sites and analyzed the data with root mirrors' identification and localization (RMIL). RMIL is a method to identify and locate root mirrors. It contains probing and analyzing the network services ID (NSID) and traceroute data to identify and locate root mirror sites. Using this method, 821 (70.78% of the total) sites were accurately identified and located, and city-level localization was achieved for 281 other sites. Finally, the identification results were used in the service area analysis. The analysis contained multiple dimensions: locations, autonomous system numbers (ASN), internet service providers (ISP), and IPV4 prefixes. As such, we helped identify and locate root mirror sites more precisely and discover which ones have a greater service area in different dimensions.

Keywords: IP anycast; root name servers; anycast identification; service area analysis

Citation: Wang, J.; Li, Z.; Zhang, Z.; Chen, J.; Li, C.; Cheng, Y. Root Mirror Sites Identification and Service Area Analysis. *Electronics* **2023**, *12*, 1737. https://doi.org/10.3390/electronics12071737

Academic Editors: Leandros Maglaras, Helge Janicke and Mohamed Amine Ferrag

Received: 18 February 2023
Revised: 1 April 2023
Accepted: 3 April 2023
Published: 5 April 2023

1. Introduction

The domain name system (DNS) is a hierarchical and decentralized system that converts domain names into internet protocol (IP) addresses. The process of resolving a domain name to an IP address involves several steps:

(a) A user types a domain name into a web browser;
(b) The browser sends a request to a local DNS resolver which is usually provided by the internet service provider (ISP) to resolve the domain name;
(c) If the local resolver has the IP address for the domain name cached, it returns the IP address to the browser. If not, it sends a request to a root server;
(d) The root server responds to the local resolver by referring to the appropriate top-level domain (TLD) server;
(e) The local resolver sends a request to the TLD server;
(f) The TLD server responds by referring to the authoritative DNS server for the domain name;
(g) The local resolver sends a request to the authoritative DNS server;
(h) The authoritative DNS server responds to the local resolver with the IP address of the domain name.

This process happens in milliseconds, allowing users to access websites using easy-to-remember domain names instead of numerical IP addresses. The DNS system plays a crucial role in the functioning of the Internet, and it is essential for providing reliable and efficient access to web resources [1–3]. In this system, the root servers are responsible for resolving the IP addresses of 1592 top-level domain name servers, which control the IP address resolving process of 349.9 million domain names [4]. Without the root servers, any website accessed by domain names will not be accessible.

Even if there is only one root server that goes wrong, it means that a large number of clients will face the risk of high network latency [5], network unavailability [6,7], and

even hijacking [8,9], which is why the root servers are so vital. Many threats can have a significant impact, especially distributed denial-of-service (DDoS) attacks [10,11]. To reduce such threats, various methods, especially adding anycast mirrors, were adopted to improve the availability of root servers [12]. Adding anycast mirrors ensures that the root servers can resist not only DDoS flooding [13–15] but also DNS manipulation [16], prefix hijacking [17,18], and many other attacks.

Nevertheless, these mirrors lead to some new problems. For example, ICANN predicts that the increase of root zone files will lead to root zone file synchronization delay, and the mirrors will aggravate this situation. [19]. Another fundamental issue is that the current anycast root mirrors belonging to the same root have the same IP, making it challenging to accurately identify which mirror was accessed during probe works.

So, this paper started the identification and localization research on the root mirror sites, which mainly includes using many probe methods to obtain multiple sets of data with different dimensions, extending the data dimensions through some publicly available application programming interface (API) and other information, and compared the results by several sets of algorithms to see which set has the best identification and localization accuracy. The contributions of this research paper are as follows:

- Root Mirrors' Identification and Localization (RMIL): a complete set of root mirrors' identification and localization methods and algorithms is proposed. Based on the probe results, this method can identify root mirrors' naming rules and geographic locations. Furthermore, it also helps to analyze the service area of each mirror, continuously identify and locate root mirrors and update their information in time. The localization process of this method uses multiple dimensions of data to verify each other, which significantly improves the reliability of localization. At the same time, the process compares several different algorithms. Finally, it arrives at a set of relatively suitable algorithms for handling this problem
- Naming Rules of Network Services ID (NSID): using RMIL, a complete list of naming rules for the current root mirrors is derived. Summarizing the root mirror sites' naming rules for each root can be used to improve the efficiency of subsequent identification and localization. It can also be utilized to pre-determine whether the responses come from a real root mirror, reduce other studies on NSID noise, or quickly cut out geographic information from NSID for root mirror localization.
- Root Servers' Service Areas Analysis: an accurate service area analysis of the root mirrors is performed. Analyzing the service area of the root mirrors can reveal fundamental problems, such as: which regions' servers are facing higher network loads? Which server will affect a wider area when network attacks happen? Which routers' configurations are unreasonable, resulting in the root domain name resolution not getting the optimal solution?

The remainder of this paper is organized as follows. Section 2 analyzes related studies, and Section 3 provides a detailed description of the RMIL method and its methodology. The results of the study are presented in Section 4. Finally, the discussion is presented in Section 5, and Section 6 provides a summary and conclusion.

2. Related Works

There were some efforts to identify and locate the root mirrors and analyze their service areas.

2.1. Root Mirrors' Identification and Location

The administration of the root servers considered how different root mirrors should be distinguished at the beginning of adding anycast mirrors. As early as 2007, with the increased use of DNS anycast, multiple root mirrors were used to share a single IP address for load balancing and other mechanisms; so, it is sometimes difficult to determine which anycast mirror responded to a query. To identify different DNS servers, the administration published the material on the DNS NSID option [20] with corresponding technical

support to provide a way to identify the sources of the responses. Chao L., Yanan C., and Hao M. et al. [8] researched the DNS root anycast nodes' performance based on the active measurement. Their work integrated various data types to identify root anycast instances, such as string matching of the anycast instance name, domain resolving, traceroute pathing, and so on, instead of relying just on the source data. Zhang, F., Lu, C., and Liu, B. [21] used China's DNS censorship mechanism to determine whether the accessed root mirrors were located in China. Sarat, S., Pappas, V., and Terzis, A. [22] analyzed the access patterns of some anycast nodes by leveraging the similarity of network topology and geographical locations. They found that 37% to 80% of probing points' queries accessed the geographically closest server, which could be utilized in server identification and localization. Jones B, Feamster N, and Paxson v et al. [23] used various methods to identify different root mirrors while studying the root mirrors' manipulation. Their probing points were deployed with related probes such as HOSTNAME.BIND and traceroute to obtain information about the NSID and the root mirror server's Internet neighbor nodes for identification and localization. Fan X., Heidemann J., and Govindan R. [24] focused on identifying and characterizing anycast in the DNS. They obtained TXT records of CHAOS class and traceroute information for locating anycast server nodes and revealed the limitations of each method in identifying anycast nodes.

However, since root mirror identification was not the focus of their works, they only conducted identification studies for root mirrors of several roots (e.g., B, F, H, L, and M) or root mirrors of some regions (e.g., China). However, our work contains identifying and localizing all roots, and we provide a complete model for probing and processing data.

2.2. Root Mirror Server Service Area Analysis

Zhang F., Lu C., and Liu B. et al. [21] presented a study on the impact of DNS root server instances in China, measured their service areas, and evaluated the effect of deploying new instances on query performance. Recommendations were made to improve the performance of the DNS root servers. Moura G., Schmidt R. d. O., Heidemann J. et al. [25] examined a large number of attacks on root servers in 2015, evaluated the behavior of some root zone administrators and root zone users, and gave some recommendations on root service state evaluation and server-side and client-side configuration. The work of Fejrskov M., Pedersen J., and Vasilomanolakis E. [26], on the other hand, focused on DNS manipulation and the impact of DNS prefix hijacking. The approach used actively probed data, filtered at various levels to obtain valid data, and then, dimensionally expanded and clustered using machine learning and other methods to simultaneously analyze the attacked servers and predict the attack's impact.

Their study measured the service area of root name servers from several perspectives, but they lacked an appropriate method to identify those root mirrors. That is why they can only study the service areas of different roots but not different root mirrors.

In summary, the identification and localization of root mirror servers are fundamental to many studies on root mirrors. However, existing localization studies only cover a part of the roots, and the actual situation is complex, making it difficult to achieve precise localization. In addition, analyzing the service areas of root mirrors provides insight into the potential risks and direct impact areas in the event of an attack, but the existing analysis needs more granularity. To address these issues, this paper proposes the RMIL method, which utilizes cyclic probing data to identify and locate root mirrors and enable service area analysis at a more precise level.

3. Methodology and Method

As mentioned earlier, the root servers in the DNS system are extremely important. One of the biggest challenges in studying the security of root servers is the difficulty in identifying root mirrors. Therefore, this paper uses the root mirrors' identification and localization (RMIL) method to address this issue. The methodology and final structure of the RMIL method are explained in this section.

The RMIL method includes several steps: obtaining data, handling data, comparisons, and obtaining results. The specific process and logical relationships are shown in Figure 1.

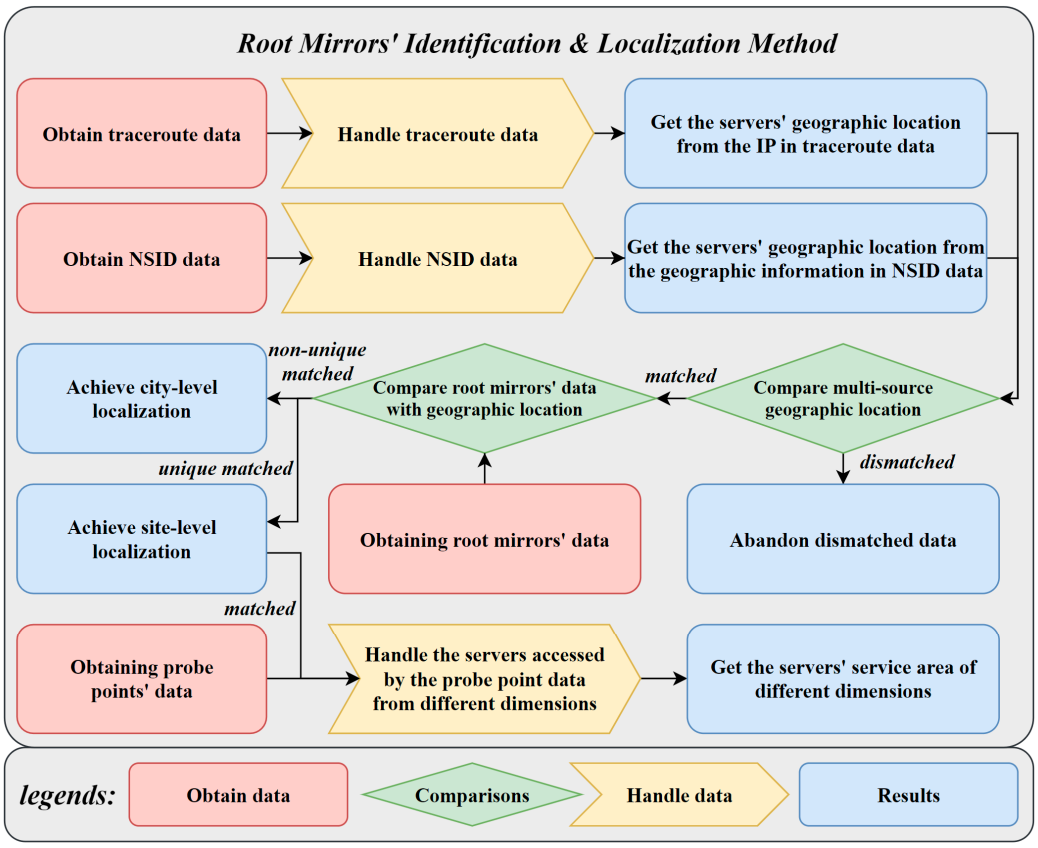

Figure 1. Root Mirrors' Identification and Localization Method.

3.1. Obtaining Data

3.1.1. Obtaining Active Data

NSID: The NSID of the root name servers is well-suited for identifying works. Additionally, there are various ways to obtain the NSID. One of them is to request the HOSTNAME.BIND in the TXT record of the CHAOS class, but this method can only work with DNS servers that use the BIND software; the other one is to request the ID.SEVER information in the TXT record of the CHAOS class, but G root will only return invalid identifiers when it replies to this type of request. Theoretically, turning on the NSID option in Extended-DNS (EDNS) can obtain the NSID of most root name servers. However, difficulties may be encountered when packets pass the routing firewalls after using EDNS because some firewalls assume that the maximum length of DNS messages is 512 bytes and will not forward extra-long DNS packets. Therefore, this paper's probe process uses different methods for different root mirrors. For the G root, we query its HOSTNAME.BIND information. For the rest of the root, we query the ID.SEVER information. All dataset comes from Ripe atlas' probe points [27].

Traceroute: Traceroute data is relatively simple to obtain, and although there are many existing traceroute programs based on different protocol versions, the result is similar. Therefore, deploying traceroute probes for all thirteen root server IPs on all probing points is sufficient.

3.1.2. Obtaining Public Data

Root mirrors' data: As the administration of the A and J roots, Verisign also maintains root-server.org [4] and its public data, which contains public data for all thirteen roots. The publicly available data provides geographic location information for all mirrors of each

root. However, the NSID of each mirror is only available for four roots (roots B, H, L, and M). Unfortunately, the remaining roots have incomplete data, making it necessary to extract information from actively probed data and match it with the publicly available data to obtain a complete portrait. This data mining process allows us to compensate for the lack of NSID information and ensure our analysis is as comprehensive as possible.

The public data maintained from Verisign's root-server.org is relatively simple to process. It includes sites for four roots: B, H, L, and M, and the corresponding identifiers of each site. However, the identifiers of the L root do not correspond to the NSID of the L root in our actual probing. However, they look like domain names in public data and can be processed into the NSID after cropping and splicing, thus facilitating the subsequent service area research.

Probe points' data: As of December 2022, the total number of root mirror sites deployed worldwide is 1160, and the total number of instances is 1603. It is challenging to probe data from all or most of such a large number of root mirrors simultaneously. In reality, it is even more difficult because the service type of some root mirrors restricts them to serving only local-type probe points, meaning that only probe points that meet multiple requirements can probe their data. Although there were some papers on discovering anycast mirrors [11], the cost of discovering and probing many nodes using iterative deployment is still very demanding. In this paper, we adopted probe data from the Ripe Atlas, but since not all probes belong to Ripe Atlas itself, some of the probes came from unsolicited contributions from the public, we could not just download the probe points' data from its website. There was some missing information about ipv4 addresses, operators, ipv4 prefixes, and geographic locations of these probes. To complete the probe points' information, we used the offline database of AIWEN TECH [28] to expand the dimensionality of the missing information of probe points with the ipv4 address by using their IP. Meanwhile, the Ripe Atlas platform collects the relatively accurate latitude and longitude information of all detection points, and so, we used the information from Baidu's geocoding service [29] to extend the dimensionality of the probe points with missing geographic location information and ipv4 address.

3.2. Identification and Localization

The network topology structure may only sometimes align with the physical geographic layout; so, a more precise method than simply using the geographic distance between the probe point and the root mirrors is required, making locating root mirrors challenging. However, since root mirrors are the most authoritative anycast mirrors, their NSID naming rules contain valuable geographic location information. The localization process involves several parts, and we conducted all our tests on the 12th Gen Intel(R) Core (TM) i7-12700H, CentOS-7-x86_64, Python 3.8 platform.

3.2.1. Handling Traceroute Data

Traceroute data describes the routers which the probe package accessed. Since these root mirrors are all anycast mirrors, it is insignificant to geolocate them based on the IP address of the last hop. This paper chose to geolocate based on the penultimate hop's IP, and the localization source was the IP location database mentioned above [28]. To improve the accuracy of the localization process, it was essential to consider the network topology structure of different root mirrors. However, the probe process was more complicated due to the variable and complex network topology. It cannot be ensured that a probe point will always traceroute and request NSID from the same mirror, leading to various routing-NSID correspondences, as depicted in Figure 2. Therefore, it was necessary to carefully analyze and process the probe data to identify and locate the root mirrors accurately.

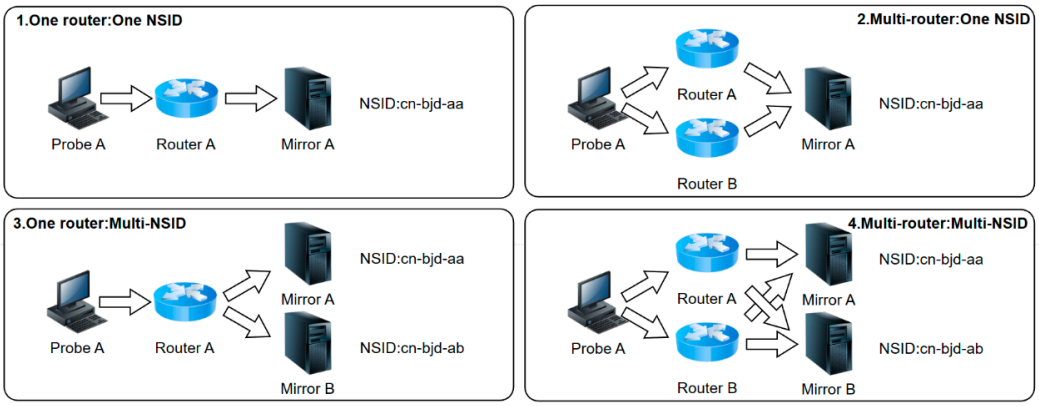

Figure 2. The correspondence between the router and NSID.

For the first three cases shown in Figure 2, we can rely on the IP address of the penultimate hop router in traceroute to locate that mirror. At the same time, following the similarity between the end-to-end distance in the network and the real world, we can approximate the root mirror's location by replacing it with the penultimate hop router's location, thus achieving a city-level root mirror localization.

3.2.2. Handling NSID Data

The geographic location information implied in NSID is vital to achieving root mirror identification and location. It is more difficult to filter out this part of the information and compare it with existing data. Figure 3 shows the processing of analyzing NSID:

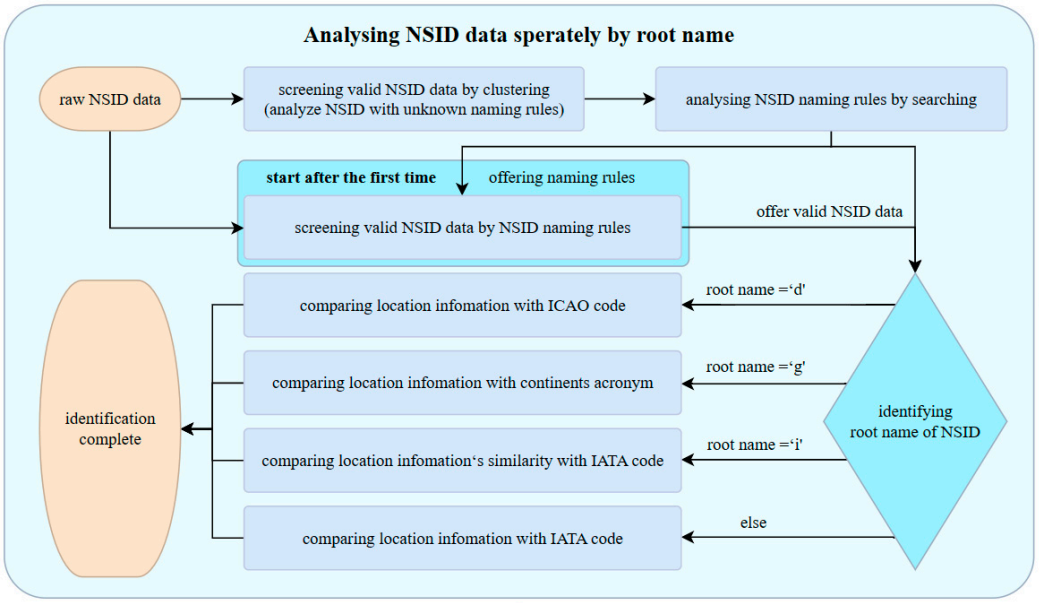

Figure 3. NSID analyzing flow chart.

Noise identification: Because the data from the Ripe Atlas platform are actual probe results from probe points obtained in a complex and variable network environment, there was much noise in the obtained data. They may have come from the network administrator's testing, the network administrator built a private root mirror for accelerating DNS resolution, or other root mirrors in the network are not authorized, etc. To eliminate the noise and retain valid data for subsequent studies, the team used various screening methods and evaluated them.

Our team initially attempted to filter out responses from different servers that contained identical invalid substrings and remove them. Since the naming rules for root

mirrors vary among organizations maintaining the roots (except for A and J roots maintained by Verisign), different roots should have distinct mirrors that do not return responses with identical substrings. Conversely, malicious mirrors often return responses with similar substrings. However, the accuracy of this filtering method may be lower than other methods during the screening process.

While analyzing the data of root mirrors, our team discovered that the identifiers of genuine root mirrors usually have a fixed length. Therefore, a length-limiting approach can be used for filtering. However, the length of NSID varies among different roots, so different parameters must be applied to separate roots during the filtering process. Moreover, the naming rules of some roots are complex and diverse, which results in multiple correct lengths of NSID. The wide range of legal NSID lengths makes it challenging to filter out malicious mirrors effectively. This method is more appropriate for roots with a relatively simple NSID naming rule.

We also experimented with several clustering algorithms to solve the clustering problem of legitimate data and noise, including k-means algorithms and hierarchical clustering (HC). To reasonably apply these algorithms, it was necessary to give a reasonable measure of the distance between the NSID strings. We tried the jellyfish (JF) algorithm and the edit distance (ED) algorithm to measure the difference between the strings. Finally, hierarchical clustering and edit distance pairing can most effectively complete the screening of the active data under suitable parameter constraints.

The public data of the four roots B, H, L, and M mentioned earlier contained the NSIDs of all their root mirror sites, and so, we used this part of the public data to compare the screening results accuracy of different screening methods. The screening accuracy of various methods is referenced in Figure 4.

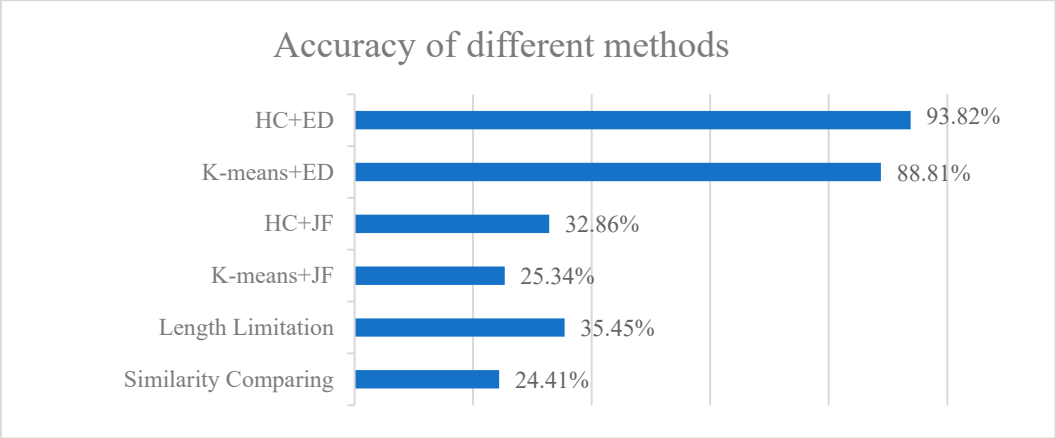

Figure 4. The proportion of noise screened by different methods to the total noise.

Where the accuracy of the methods shown in Figure 4 is calculated by Equation (1)

$$Accuracy = \frac{Number\ of\ noise\ NSID\ correctly\ screened}{Actual\ number\ of\ \ noise\ NSID} \quad (1)$$

According to Figure 4, two simple methods, relying on similarity comparison and length limit, had a limited effect. The JF algorithm, on the other hand, could not reflect the difference between strings at an appropriate range; so, when it cooperates with the two clustering algorithms, they judged a deficient proportion of noise. The ED algorithm, on the other hand, is good at judging the distance between legal identifiers as small and the distance between noise and legal identifiers as significant, which is conducive to clustering. The k-means algorithm relies on K's value and the selection of initial vectors for each set, so there is a particular random factor in the clustering results. The final performance requires several experiments to determine the proper K value and consistently yields poor results.

In contrast, hierarchical clustering can distinguish valid NSIDs from noise well as long as the threshold value of clustering distance is set reasonably, together with the distance measure of ED.

Clustering by region: the use of hierarchical clustering and edit distance algorithm in screening noise can achieve better results, but in the distinction of the NSID from different root mirror sites, it cannot reach the ideal consequence by still following this method, which is mainly due to two reasons:

- Hierarchical clustering is relatively simple, and the misjudgment cannot be re-clustered;
- The ED algorithm considers the impact of letter substitutions occurring in different positions of two strings as the same. For instance, when analyzing NSIDs, the geographic location information difference and the difference at the end of two NSIDs are treated the same. However, the two differences should be distinct, with the geographic location information difference being more significant and the difference at the end of the two NSIDs being smaller.

To solve the above problems, this paper modified the original ED algorithm. Because the NSIDs of the same root generally have the same naming rules, only the geographic location information part will be inconsistent. Accordingly, the distance judgment of the ED algorithm was modified as follows:

- Substitute occurs in three consecutive English letters (THEL); add 5 to the distance;
- Other than the above one, follow the original algorithm.

Then, the modified ED algorithm is:

For any two strings A and B, the length of them is L_A and L_B, respectively, there is a corresponding matrix M of row L_A and column L_B as follow Equation (2):

$$M = \begin{bmatrix} D_{0,0}\ D_{0,1} & \cdots\cdots & D_{0,L_B-1}\ D_{0,L_B} \\ D_{1,0}\ D_{1,1} & \cdots\cdots & D_{1,L_B-1}\ D_{1,L_B} \\ \cdots\cdots & & \cdots\cdots \\ D_{L_A-1,0}\ D_{L_A-1,1} & \cdots\cdots & D_{L_A-1,L_B-1}\ D_{L_A-1,L_B} \\ D_{L_A,0}\ D_{L_A,1} & \cdots\cdots & D_{L_A,L_B-1}\ D_{L_A,L_B} \end{bmatrix} \tag{2}$$

The $D_{i,j}$ ($0 \le \mathrm{i} \le \mathrm{L_A}, 0 \le \mathrm{j} \le \mathrm{L_B}$) in M means the modified Edit Distance between the first i characters of A and the first j characters of B. Each $D_{i,j}$ is calculated as Equation (3), where $A[i]$ represents ith characters of A and $B[j]$ represents jth characters of B:

$$D_{i,j} = \begin{cases} 0; i = 0 \ or \ j = 0 \\ min \begin{cases} D_{i,j-1} + 1; \\ D_{i-1,j} + 1; \\ D_{i-1,j-1}; A[i] = B[j] \\ D_{i-1,j-1} + 1;\ A[i] \neq B[j] \end{cases} \\ D_{i-1,j-1} + 5;\ A[i] \neq B[j] \ and \ they \ all \ in \ \textbf{\textit{THEL}} \end{cases} \tag{3}$$

Additionally, the $D_{LA,\,LB}$ is the modified Edit Distance between A and B.

The above rules can be applied to distinguish some similar NSIDs very well, as shown in Table 1.

Table 1. Two pairs of NSID (F root) measured by different algorithms.

NSID Type	Examples	Edit Distance	Modified Edit Distance
NSID from different clusters	KIV.cf.f.root-servers.org KIX.cf.f.root-servers.org	1	5
NSID from same clusters	KIX.cf.f.root-servers.org KIX1f.f.root-servers.org	2	2

The above modifications can ensure that the new algorithm can accurately distinguish NSIDs with different geographic location information. Applying the algorithm to the distance judgment of clustering avoids the occurrence of misjudgment, which fills the defect that hierarchical clustering cannot re-judge after misjudgment.

3.2.3. Comparing Geographic Location Information

Comparing geolocation information consists of two main parts: comparing geolocation information in NSID and comparing geolocation information in traceroute data.

NSID: The geographic location information in NSID usually includes several types: country code, area code, and continent code. Most root NSIDs use the IATA code [30] as the geographic location information (except D and G root). IATA code is a kind of code that uniquely represents a regional airport with three letters, while the D root uses the four-letter code ICAO [31], which is also the identification of certain regional airports; the G root uses the abbreviation of the continent as the geographic location information of different root mirror sites. It is nice that G-root has fewer deployed root mirror sites, so the identification of root mirror sites can be made only by continent information and service area data. In addition, this paper also used the depth-first search method to search the NSID naming rules of each root's current root mirrors after clustering and accurately cuts out the geographic location information from the NSIDs.

Since multiple airports exist in some regions, the IATA codes in the NSIDs may differ for the same area. Therefore, in this paper, based on the geographic location data of the root mirrors on the root-server.org [4], all airport codes of the regions where the root mirrors were distributed were queried and stored in the database. After that, we queried the database to determine whether there was a unique root mirror site in each region based on the geographic location information obtained from active probes and public information and store the query results as location results.

Traceroute: As mentioned before, we relied on the IP address of the penultimate hop router in traceroute to locate that router to localization root mirrors. We utilized the positioning IP method offered by AIWEN TECH [28] to obtain the root mirrors' geographical location.

3.3. Service Area Analysis

The localization and identification of the root mirror sites are fundamental to the service area analysis. This paper achieved a fine-grained service area analysis of the root mirror sites from several dimensions, due to previous research. In the geolocation-based root mirror portrait, the location of the root mirrors can be determined according to the latitude and longitude, and the geographic service area distribution pattern of the root mirrors can be visualized on the map. Additionally, the ASN service area of some root mirrors, the ISP service area of mirrors, and the ipv4 prefix service area of mirrors can be obtained. In Section 4, we analyzed and revealed some identified and located root mirrors with extra extensive service areas.

4. Results

4.1. Root Mirror Sites Identification and Localization

This paper presented a relatively reliable RMIL method, and the implemented method and process were elaborated. 821 root mirrors were detected and accurately identified by geographical location, accounting for 70.78% of the total servers. These identification results are indispensable in the subsequent service area analysis and other possible studies. The identification percentage for each root is shown in Figure 5.

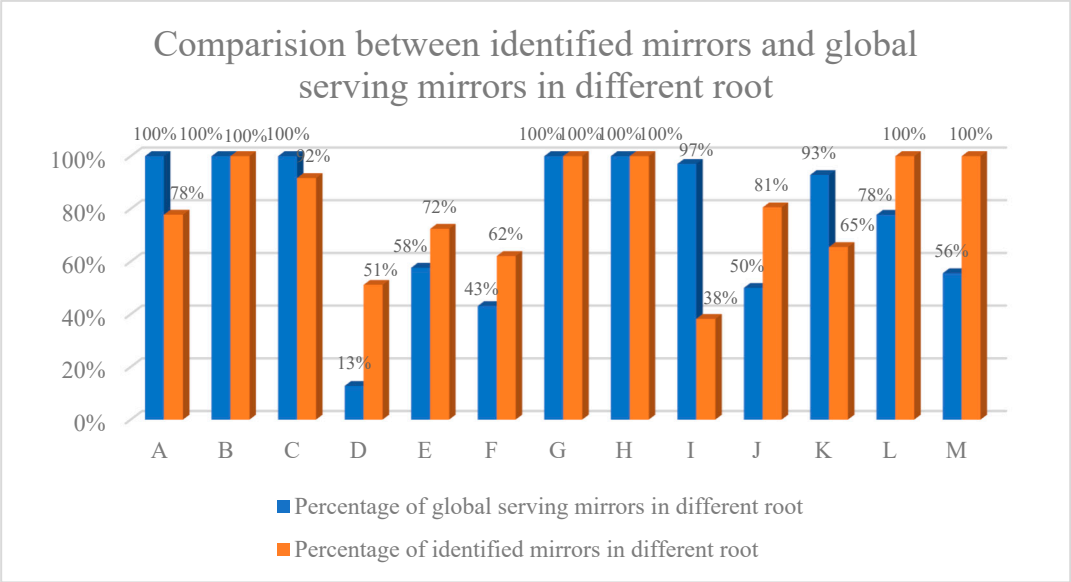

Figure 5. The percentage for each root's identified mirrors and global serving mirrors. The global serving mirrors are easier to be probed and identify.

Figure 5 shows the proportion of root mirror sites with global serving and the proportion of identified root mirror sites in each root. It is easy to see that the RMIL method identified all root mirror sites with global serving in most roots. At the same time, the technique performed even better in city-level identification and localization. It completed the city-level localization of 1102 root mirrors, accounting for 95.00% of the total root mirrors.

4.2. Identify the Root Mirrors' NSID Naming Rules

This paper gave a set of methods to recursively and continuously identify the naming rules of the root mirror's NSID. Including obtaining data from the detection, then using clustering and other ways to assist in screening out the valid data, and finally, by using clustering and depth-first search, we can obtain the NSID naming rules. At the same time, our team derived all thirteen roots currently available NSID naming rules based on existing data, and the results are listed in Table 2.

Table 2. Naming rules of different root's NSIDs.

Root	Naming Rules of NSIDs
A	nnn1-###% rootns-&&###-%* rootns-###% rootns-el###%
B	b%-###
C	###%*.c.root-servers.org
D	####%.droot.maxgigapop.net
E	*%%. ###.eroot
F	###%*.f.root-servers.org ###.cf.f.root-servers.org
G	groot-###%-%
H	%%%. ###.h.root-servers.org
I	s1.###

Table 2. *Cont.*

Root	Naming Rules of NSIDs
J	nnn1-###% rootns-&&###-%* rootns-###% rootns-el###%
K	ns%. &&-###.k.ripe.net
L	&&-###-**
M	m-###-% m-nrt-####-% m-not-#####-%

Symbol replaced contents are as follows: #: Location identifier (IATA or ICAO or else). &: Country code *: Alphabet serial code %: Order number.

4.3. Root Mirrors Service Area Analyzing

In addition to the previous contributions and studies, this paper relied on the identification and localization results to analyze the root mirror service area. Specifically, it included four dimensions: server geographic location, service ISP, service ASN, and service ipv4 prefix. Figure 6 shows the heat map of the geographic distribution of the number of instances of root mirror sites located globally, where the size and color of the circles represent the number of instances distributed in the geographic location. The colors from red to green indicate the decreasing density of the root mirror instances distribution in the region.

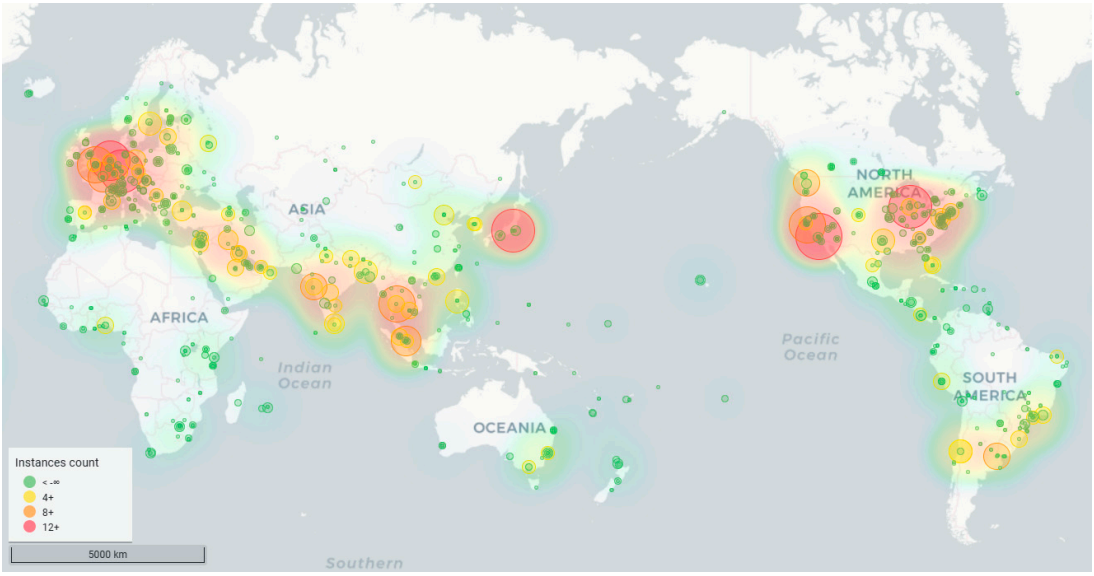

Figure 6. The global distribution map of DNS root mirror instances.

Our team also analyzed the service area of the root mirrors in terms of geographic area, ASN, ISP, and IPv4 prefixes. The results of the top 10 root mirror NSIDs with the highest count of services in four dimensions are shown in Table 3.

The geographic location information of these NSIDs mainly focused on FRA, AMS, LCY, and ORY, which corresponded to the four regions of Frankfurt, Germany; Amsterdam, Netherlands; London, UK; and Paris, France. The most significant request counts of root mirrors were concentrated in these European countries. More root mirrors will likely need to be deployed in these regions for load balancing and to reduce the hazard when a single mirror is unavailable.

Table 3. Top 10 root mirror NSIDs for the count of services in four different dimensions.

Dimension	NSID	Count
NSID order by serving location	001.fra.h.root-servers.org	1452
	rootns-fra5	1099
	nnn1-fra5	1097
	b3-ams	1012
	b4-ams	998
	b1-ams	976
	fra1a.c.root-servers.org	843
	fra1b.c.root-servers.org	821
	001.lcy.h.root-servers.org	786
	M-ORY-2	724
NSID order by serving ASN	001.fra.h.root-servers.org	1814
	rootns-fra5	1489
	nnn1-fra5	1484
	b3-ams	1060
	b4-ams	1025
	b1-ams	984
	fra1a.c.root-servers.org	940
	fra1b.c.root-servers.org	879
	M-ORY-2	854
	M-ORY-1	843
NSID order by serving ISP	001.fra.h.root-servers.org	1698
	rootns-fra5	1380
	nnn1-fra5	1379
	b3-ams	995
	b4-ams	985
	b1-ams	937
	fra1a.c.root-servers.org	866
	fra1b.c.root-servers.org	829
	M-ORY-2	806
	M-ORY-1	793
NSID order by serving ipv4 prefix	001.fra.h.root-servers.org	3297
	rootns-fra5	2558
	nnn1-fra5	2550
	b4-ams	1946
	b3-ams	1938
	b1-ams	1856
	fra1a.c.root-servers.org	1689
	fra1b.c.root-servers.org	1650
	s1.dex	1503
	001.lcy.h.root-servers.org	1473

5. Discussion

This paper described how the RMIL method can effectively identify and locate root mirror sites with multiple naming rules in different roots. It includes how to obtain the NSID and traceroute data of each root mirror site, how to use the public data for data dimension extension, how to process the geographic location information of all the data obtained for comparison and verification, and so on. Compared with previous studies, our work had many advantages:

- Identified and located root mirrors. The RMIL method covers a more comprehensive scope, including all thirteen root mirrors distributed worldwide. The previous studies only focused on roots B, F, H, L, and M. Simultaneously, our method achieved more precise localization. Site-level localization was reached for 70.78% of root mirror sites, which means these sites will be identifiable and localizable in subsequent studies.
- Provided the trace basis for the root mirror abnormality. As mentioned in Section 1, various root mirror anomalies were difficult to trace due to the lack of a reliable method

to identify and locate root mirrors. However, the RMIL method offers a new process to deal with such problems. It means we can analyze which root mirror has root zone file synchronization delay, network unavailability, DDoS attack, and even hijacking.

- Obtained the naming rules of all roots. Applying the RMIL method, this paper identified root mirror NSID naming rules for all thirteen roots. This result can be used in many fields, such as root server NSID legitimacy pre-determination, fast slicing NSID geographic location information, etc. They can help with more detection tasks and analysis tasks for root mirror servers to be launched with less cost.
- Analyzed the root servers' service areas. An accurate service area analysis of the root mirrors was performed. It showed which root mirror served a broader area in different dimensions, and it also means that these root mirrors are facing more requests and loads.

The RMIL method had some limitations, too. First, its identification and localization ability need to be improved, and many roots still need to improve their identification proportion. Second, more comparison cases should be carried out to verify the effectiveness of our method; we only compared several algorithm combinations, and there may be more suitable methods waiting for us to try. Finally, the service area analysis may also be used for the root mirrors' identification and localization, which is worth trying in the future.

6. Summary

Our work was mainly about the root mirrors' identification and localization. We proposed the RMIL method, which includes the process of probing active data of root mirrors' NSID and traceroute, obtaining public data from publicly available API, analyzing NSID data by hierarchical clustering and modified edit distance algorithm, analyzing traceroute data by IP positioning, and analyzing the service area of each root mirror based on the identification result. Using this method, we obtained the root mirrors' identification and localization result, the naming rules of all roots, and the service area of each root mirror. These results provide a reliable identification method for subsequent research and help with more detection tasks and analysis tasks for root mirrors to launch with less cost.

Our future work will rely on the existing identification and location method and information. We will continue strengthening the current method, and analyzing and tracing the root mirror anomaly based on the existing identification results. There will be much to study from this approach.

Author Contributions: Conceptualization, J.W., C.L. and Z.Z.; methodology, J.W. and C.L.; software, J.W.; validation, Z.L., Y.C. and C.L.; formal analysis, J.C., C.L. and Y.C.; investigation, Z.L.; resources, Y.C.; data curation, C.L.; writing—original draft preparation, J.W.; writing—review and editing, J.W., C.L. and Y.C.; visualization, J.W.; supervision, J.C.; project administration, Z.Z.; funding acquisition, Z.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the 2020 Industrial Internet Innovation and Development Project, Network Identifier Construction Project, the Natural Science Foundation of Shandong Province [GrantNo.ZR2020KF009], and the Young Teacher Development Fund of Harbin Institute of Technology [Grant No.IDGA10002081].

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

Domain Name System	DNS
Root Mirrors' Identification and Localization	RMIL
Network Services ID	NSID
Autonomous System Numbers	ASN
Internet Service Providers	ISP
Internet Protocol	IP

Top-Level Domain	TLD
Distributed Denial-of-Service	DDoS
Application Programming Interface	API
Extended-DNS	EDNS
Hierarchical Clustering	HC
Jellyfish	JF
Edit Distance	ED
Three Consecutive English Letters	THEL

References

1. Zhauniarovich, Y.; Khalil, I.; Yu, T.; Dacier, M. A survey on malicious domains detection through DNS data analysis. *ACM Comput. Surv. (CSUR)* **2018**, *51*, 1–36. [CrossRef]
2. Alieyan, K.; Almomani, A.; Manasrah, A.; Kadhum, M.M. A survey of botnet detection based on DNS. *Neural Comput. Appl.* **2017**, *28*, 1541–1558. [CrossRef]
3. Wang, Y.; Zhou, A.; Liao, S.; Zheng, R.; Hu, R.; Zhang, L. A comprehensive survey on DNS tunnel detection. *Comput. Netw.* **2021**, *197*, 108322. [CrossRef]
4. Root Server Technical Operations Association. Available online: https://root-servers.org/ (accessed on 20 December 2022).
5. Saridou, B.; Shiaeles, S.; Papadopoulos, B. DDoS attack mitigation through Root-DNS Server: A case study. In Proceedings of the 2019 IEEE World Congress on Services, SERVICES 2019, Milan, Italy, 8–13 July 2019; pp. 60–65.
6. Fachkha, C.; Bou-Harb, E.; Debbabi, M. Fingerprinting Internet DNS Amplification DDoS Activities. In Proceedings of the 6th International Conference on New Technologies, Mobility and Security (NTMS), Dubai, United Arab Emirates, 30 March–2 April 2014; pp. 1–5.
7. Alieyan, K.; Kadhum, M.M.; Anbar, M.; Rehman, S.U.; Alajmi, N.K.A. An overview of DDoS attacks based on DNS. In Proceedings of the International Conference on Information and Communication Technology Convergence (ICTC), Jeju, Republic of Korea, 19–21 October 2016; pp. 276–280.
8. Li, C.; Cheng, Y.; Men, H.; Zhang, Z.; Li, N. Performance Analysis of Root Anycast Nodes Based on Active Measurement. *Electronics* **2022**, *11*, 1194. [CrossRef]
9. Houser, R.; Hao, S.; Li, Z.; Liu, D.; Cotton, C.; Wang, H. A Comprehensive Measurement-based Investigation of DNS Hijacking. In Proceedings of the 40th International Symposium on Reliable Distributed Systems (SRDS), Online, 20–23 September 2021; pp. 210–221.
10. Events of 2015-11-30. Available online: https://root-servers.org/media/news/events-of-20151130.txt (accessed on 20 December 2022).
11. Events of 2016-06-25. Available online: https://root-servers.org/media/news/events-of-20160625.txt (accessed on 20 December 2022).
12. Fan, X.; Heidemann, J.; Govindan, R. Evaluating anycast in the domain name system. In Proceedings of the IEEE INFOCOM, Section V, Turin, Italy, 14–19 April 2013; pp. 1681–1689.
13. Nazario, J. DDoS attack evolution. *Netw. Secur.* **2008**, *2008*, 7–10. [CrossRef]
14. Zargar, S.T.; Joshi, J.; Tipper, D. A survey of defence mechanisms against distributed denial of service (DDoS) flooding attacks. *IEEE Commun. Surv. Tutor.* **2013**, *15*, 2046–2069. [CrossRef]
15. Yazdani, R.; Hilton, A.; van der Ham, J.; van Rijswijk-Deij, R.; Deccio, C.; Sperotto, A.; Jonker, M. Mirrors in the Sky: On the Potential of Clouds in DNS Reflection-based Denial-of-Service Attacks. In Proceedings of the 25th International Symposium on Research in Attacks, Intrusions and Defenses, Limassol, Cyprus, 26–28 October 2022; pp. 263–275.
16. Weaver, N.; Kreibich, C.; Paxson, V. Redirecting DNS for Ads and Profit. *FOCI* **2011**, *2*, 2–3.
17. Gersch, J.; Massey, D. Rover: Route origin verification using DNS. In Proceedings of the 22nd International Conference on Computer Communication and Networks (ICCCN), Nassau, Bahamas, 30 July–2 August 2013; pp. 1–9.
18. Ballani, H.; Francis, P.; Zhang, X. A study of prefix hijacking and interception in the Internet. *ACM SIGCOMM Comput. Commun. Rev.* **2007**, *37*, 265–276. [CrossRef]
19. Akkerhuis, J.; Chapin, L.; Fältström, P.; Kowack, G.; Liman, L.; Manning, W.J. Scaling the Root Report on the Impact on the DNS Root System of Increasing the Size and Volatility of the Root Zone; Prepared by the Root Scaling Study Team for the Root Scaling Steering Group. [PDF] Scaling the Root Report on the Impact on the DNS Root System of Increasing the Size and Volatility of the Root Zone Prepared by the Root Scaling Study Team for the Root Scaling Steering Group | Semantic Scholar. 2009. Available online: https://www.semanticscholar.org/paper/Scaling-the-Root-Report-on-the-Impact-on-the-DNS-of-Akkerhuis-Chapin/bc4a738d4a068c78dbfc12467475dd41b717ea41 (accessed on 20 December 2022).
20. RFC 5001: DNS Name Server Identifier (NSID) Option. Available online: https://www.rfc-editor.org/rfc/rfc5001 (accessed on 20 December 2022).
21. Zhang, F.; Lu, C.; Liu, B. Measuring the Practical Effect of DNS Root Server Instances: A China-Wide Case Study. In *International Conference on Passive and Active Network Measurement*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 247–263.
22. Sarat, S.; Pappas, V.; Terzis, A. On the use of anycast in DNS. In Proceedings of the IEEE International Conference on Computer Communications and Networks (ICCCN), Arlington, VA, USA, 9–11 October 2006; pp. 71–78.

23. Jones, B.; Feamster, N.; Paxson, V.; Weaver, N.; Allman, M. Detecting DNS root manipulation. In Proceedings of the Passive and Active Measurement: 17th International Conference, PAM 2016, Heraklion, Greece, 31 March–1 April 2016; pp. 276–288.
24. Fan, X.; Heidemann, J.; Govindan, R. *Identifying and characterizing anycast in the domain name system*; USC/ISI Technical Report ISI-TR-671; USC/Information Sciences Institute: Marina del Rey, CA, USA, 2011; pp. 1–13.
25. Moura, G.C.; Schmidt, R.D.O.; Heidemann, J.; de Vries, W.B.; Muller, M.; Wei, L.; Hesselman, C. Anycast vs DDoS: Evaluating the November 2015 root DNS event. In Proceedings of the 2016 Internet Measurement Conference, Santa Monica, CA, USA, 14–16 November 2016; pp. 255–270.
26. Fejrskov, M.; Pedersen, J.M.; Vasilomanolakis, E. Detecting DNS hijacking by using NetFlow data. In Proceedings of the IEEE Conference on Communications and Network Security (CNS), Austin, TX, USA, 3–5 October 2022; pp. 273–280.
27. RIPE Atlas. Measurements. Available online: https://atlas.ripe.net/measurements/ (accessed on 20 December 2022).
28. More Accurate Global IP Address Positioning Platform_IP Asking—AIWEN Technology. Available online: https://ipplus360.com/ (accessed on 20 December 2022).
29. Inverse Geocoding Rgc Anti-Geo Search | Baidu Map API SDK. Available online: https://lbsyun.baidu.com/index.php?title=webapi/guide/webservice-geocoding-abroad (accessed on 20 December 2022).
30. IATA. Codes—Airline and Location Codes Search. Available online: https://www.iata.org/en/publications/directories/code-search/ (accessed on 20 December 2022).
31. Location Indicators. Available online: https://www.icao.int/safety/OPS/OPS-Tools/Pages/location-indicator.aspx (accessed on 20 December 2022).