Target driven IP Geolocation Algorithm

Zhanfeng Wang¹, Yan Niu²*, Hao Chen³, Guang Cheng⁴, Jiawei Cui⁵, Zhiyi Zhang⁵

¹School of Computer Science and Engineering, Southeast University, Nanjing, 211189, China
²China Academy of Industrial Internet, Beijing 100102, China
³China Telecom Jiangxi Branch Provincial Network Operation and Management Department, Nanchang 210003, China
⁴College of Cyberspace Security, Southeast University, Nanjing 211189, China
⁵The 54th Research Institute of CETC, Shijiazhuang 050081, China

*Corresponding author, E-mail: niuyan@china-aii.com

Abstract. IP Geolocation is an important technique to achieve network attack traceability, network security early warning, and improve the ability of cyberspace governance. Due to the limitation of collecting a large number of anchor nodes, it is impossible to accurately infer the IP location to block level or more in a large scale. Therefore, this paper proposes a target driven IP geolocation inference method. Firstly, the candidate anchor node is collected from the subnet or the adjacent subnet of the target IP, then the candidate anchor node is calibrated by their fingerprints, and finally the geolocation of target IP is comprehensively inference by its proximity to anchors. In order to verify the proposed method, 10,000 IPs in Beijing is selected randomly. The experimental results show that the method can effectively infer the geographical location of the target IP address.

1. Introduction

Cyberspace mapping technology is an hot issue of network security protection and network attack. Cyberspace mapping not only includes the detection and perception of various attributes of network entities in cyberspace, among them the the IP geolocation is very important one. IP geolocation is widely used in various fields, such as content positioning, targeted advertisements, identity authentication, network security, network performance optimization and so on. Because of its wide range of applications and important value, it has been highly concerned by the academic and industrial circles. At present, there are many commercial and academic IP location databases on the Internet, such as IP2location, Maxmind, Geobytes, Netacuity, Akami, Quova[1], AnchorX[2], IPIP.net, Chunzhen, Sina, Taobao, Baidu, etc.

IP geolocation rely on the routes, anchor nodes, bearer service information, whois data and other information[3][2]. These data come from the ISP or detection information, and its location correctness and accuracy correctness can not be guaranteed, which greatly affects the application of IP geolocation technique[4][5]. The original IP geolocation algorithms inference the geographical location of IP devices by querying the DNS server or mining the information hidden in the host name[4][11][12]. After that, some geolocation algorithms exploit the linear relationship between time delay and
geographical distance to estimate the host location, and reduce the location error through topological information[8]-[10]. In order to further improve the accuracy of IP geolocation, some methods based on delay estimation are proposed, which use nonlinear time delay geographic distance model [7][15][14]. On this basis, some integrated geolocation algorithms are proposed, which can infer the geographical location of IP to be located by comprehensively using the delay, network topology, anchor node and other information, and the highest accuracy can reach the block level, such as octant[5], TTG [10].

Although there are many IP location algorithms, it is still a very difficult to obtain high-precision IP geolocation information in the wild Internet. The main reasons are as follows: 1. Lack of a large number of anchor nodes, IP geolocation algorithms often infer the location by estimating the delay or topology distance between the target node and the anchor nodes. Therefore, the more the anchor nodes, the higher the accuracy. However, a large number of anchor nodes are difficult to obtain, even ISP cannot accurately tell the exact location of each IP. 2. Due to the huge scale of IP address, the network overhead for comprehensive location analysis is very large. At the same time, some IP addresses are dynamically allocated and need continuous iterative measurement.

Therefore, this paper proposes a target driven IP geolocation algorithm (TDIG). The idea of the algorithm is to locate the target IP by the dynamic collecting anchor nodes, and then inferring the location of the target IP address by comparing the path similarity. This method can avoid large-scale network space measurement and anchor node acquisition, so as to reduce the overhead of network space measurement and achieve high-precision IP address positioning. The positioning accuracy of TDIAL can reach the block level or even building level.

The rest of this paper is organized as follow. Section 2 summarizes the related work of IP geolocation, section 3 illustrates the basic idea of target driven IP address location technology, section 4 verifies the proposed method, and section 5 summarizes the full paper.

2. Related works
In cyberspace surveying and mapping, it is the key to infer the location of IP address as so to construct a map of the Internet. The basic principle of IP geolocation algorithm design is to reduce the measurement overhead as much as possible while ensuring the geolocation accuracy, and at the same time, it has good scalability. At present, all IP geolocation algorithms can be roughly divided into two types: client independent and client based.

The client based on geolocation algorithm has high accuracy, but it often needs the auxiliary of GPS, cellular base station, WiFi access point and other infrastructure. Recently, with the popularity of social networks, many applications publish the user's geographical location on the website. These user location data can also improve the geolocation accuracy. In addition, the large-scale deployment of IPv6 network and the proposal of new protocols such as locator / ID separation protocol (LISP) bring new opportunities and challenges to IP location technology.

According to its location principle, client independent geolocation algorithms can be divided into three categories: conjecture based geolocation algorithm, delay based geolocation algorithm and integrated geolocation algorithm. The location algorithm based on speculative information generally obtains the host name and street of the IP address by querying WHOIS database, or speculates the location of the IP device by the geographical location of the IP address segment. Typical algorithms include GeoTrack[8], Quova[6], MaxMind, EdgeScap, TraceWare, VisualRoute, IP2LL[11], NetGeo[12], GTrace[13], Neotrace, Software77, IPligence, etc. This kind of algorithms can locate IP by directly querying WHOIS database to infer the host location information, or inferring the host geographic location by measuring the host name and combining with the database information. Delay based location method estimates the location of the host by measuring the delay from the target host to the measurement point. In order to improve the location accuracy, it often combines the network topology information. Typical location algorithms and systems of this kind geolocation algorithm include Geoping[8], Shortest Ping[2], Constraint-based Geolocation (CBG)[9], Topology-based Geolocation (TBG)[2], NBIGA[7], Posit, Spotter, etc. Compared with the location algorithm based on
conjecture, this kind of location algorithm has a more perfect mathematical foundation. However, any of the above methods cannot locate the host accurately, so some researchers put forward comprehensive positioning methods, such as Octant[5] and TTG[10]. These algorithms use the above two or three algorithms to locate the host at the same time, and then carry out interactive verification to determine the location of the host.

3. Target driven IP geolocation algorithm
Reference[1] points out that the accuracy of IP geolocation depends on the number and distribution of validate anchor nodes, that is, the anchor nodes must be measurable and in the same network or adjacent network with the target IP. Anchor node is a kind of node which is active for a long time in the network, measurable and has a clear geographical location. For entire IPv4 space, if the location accuracy of the target is assumed to be block level, and there are at least 20 million anchor nodes in a 24 bit mask network. In order to maintain the available anchor node set, large amount of dynamic detection is needed, which will consumes a lot of network resources and causes disturbance to the network.

In order to reduce the overhead of network measurement while infer the geographic location of IP address accurately and effectively, a target driven geolocation algorithm is proposed. The basic idea is to dynamically collect the anchor nodes for locating the target IP address, find the effective anchor nodes, and then infer the location according to the geographical location of the anchor nodes. The algorithm is mainly divided into three stages: anchor node information collection, anchor node calibration and geographic location inference.

Anchor node information collection: in this phrase, all of the active IP address of the network segment where the target IP is located. If there is an active IP, the fingerprint characteristics of the active IP will be collected to determine whether it is a legal anchor node.

Anchor node calibration: it is mainly to determine whether each alternative anchor node is a legal anchor node. Because the anchor node is selected for a specific target, it only needs to have a clear geographical location, then it is judged to be valid, rather than long-term validity.

Geographical location inference: in this phrase, the geographical location is mainly inferred by measuring the similarity between the IP path to be located and the anchor node path. The Specific location of the target IP can be estimated by the centroid method or the geographical location of the adjacent anchor node.

3.1. Anchor node information collection
Due to the large amount of anchor nodes to be collected, the information that needs to be processed is very complex and cannot be fully automated. The target driven anchor node collection method can efficiently reduce the anchor nodes to collect. It consists of several steps as follow.

Firstly, the target IP address and all the other IPs which belong to the same subnet within a /24 netmask are scanned to collect ports, services, OS and the other fingerprint. Because many hosts will disable ICMP Protocol, besides Ping and traceroute, all the open ports are sniffered. Among the 65,535 ports opened by the host, they can be divided into three categories according to the port number: well-known ports, registered ports, dynamic and / or private ports. Among the three types of ports, dynamic and private ports are generally not used, so the 1-49151 ports are detected.

After detecting the active IP address, it is necessary to further detect the services provided by the IP to retrieve the possible geographical location of the IP. The IP’s possible location are usually hidden in the banner of some services, such as HTTP services, FTP services, email services, etc. For example, most of government, university and large companies will annotate their locations in the portal websites.

After the above steps, if there is no validate anchor node in subnet C where the target IP belongs to, the anchor node is searched from two adjacent subnets C-1 and C + 1 until sufficiently active anchor nodes are found.
In order to calibrate the anchor nodes, in addition to the previous detection information, but also to detect some additional information, such as path information, operating system, remote services (Windows remote desktop, SSH, telnet), so as to assist in determining the node types.

![Fig. 1 The procedure of anchor probing and information collection](image)

### 3.2. Anchor node calibration

After collecting enough candidate anchor nodes, we need to analyze and validate candidate anchor nodes whether their position are consistent with the actual positions, so as to determine whether the candidate anchor nodes are effective.

Before validating the candidate anchor nodes, some properties of the anchor nodes are illustrated to design the calibration algorithm.

1. The anchor node is located at the edge of the network, that is, the anchor node is not a relay node such as a router.
2. If the network topology distance of the end nodes is similar, the geographical locations of the hosts are adjacent, except for the host deployed in the cloud computing center.
3. If the host is independently deployed, its location is valid. The main hosts deployed independently are the websites of government, universities and large enterprises.

Based on the above properties, a classification decision tree of anchor nodes is designed. The basic idea is as follows: first, determine whether the node is a relay node according to the network topology. Any IP that is a relay node, if it is not in the last hop in any route, otherwise it is the end node. Furtherly, all of end nodes are classified into three types: NAT gateway, CDN nodes and server hosts. Among the three types of nodes, only the server hosts can be used as the anchor node.

In the above classification decision tree, whether it is an end node can be classified the position it is in the network topology. However, the classification of the node type depends on its fingerprints. Usually, the fingerprints include operating system, services/ports and other information. As we all know, in order to facilitate remote management, cloud computing nodes must open SSH (port 22), remote desktop service (Port 3389), VNC (port 5901) and other services. CDN nodes need to provide acceleration for many companies’ websites or services, so there are many open services. However, the above features often appear in different types of hosts. That means the host type cannot be determined by only one feature. Thus, a Bayesian decision network is introduced to implement classification.

Let $P_i$ denotes all the features of host $i$, $E_i$ is probability that the device host $i$ belongs to the the type $c_i$. 

\[
E_i = \text{arg max}_{c_i \in C} P(c_i \mid P_i)
\]  

(1)

Assuming that the open service or port of each host is independent, according to Bayesian theory, Formula (1) can be rewritten as.

\[
P(P_i|c_i) = P(p_1|c_i)P(p_2|c_i)\ldots P(p_n|c_i)
\]  

(2)

The probability $P(p_i|c)$ is estimated by one-dimensional kernel density, which has low computational complexity and high accuracy. The prior probability can be estimated by the IP allocation data, and it is illustrated in the section 4.1.

If a host is classified as a server host, its location will be estimated by if it is a cloud server or an independent host. With the large-scale application of cloud computing technology, most governments, high-tech, scientific research institutions and enterprises have established their own cloud computing centers to reduce the operation cost through centralized deployment. However, whether a node is deployed independently or not can be determined by the proximity of network topology and the discreteness of geographical location. The basic principle is as follows: if several nodes are
geographically far apart but topologically adjacent, it means that they are deployed in the cloud computing center, and their geographic coordinates cannot be denoted by the address in the website, but by the location of the cloud computing center. On the contrary, the geographical location of nodes can be denoted by the address in the website.

Let $H_i$ and $H_j$ denote two hosts, the distance between the two nodes is defined as the sum of the path length from the two anchor nodes to their common nearest common ancestors (LCA), as Formula (3). According to the similarity of network paths, the validate anchors nodes are clustered and divided into different groups.

$$\text{distance}(H_i, H_j) = \text{distance}(H_i, LCA_{ij}) + \text{distance}(H_j, LCA_{ij})$$ (3)

For the IPs of cluster $C_i$, the distance between the two farthest nodes in the cluster is selected as the geographic distance of the cluster. If the geographic distance of the cluster is greater than the radius of the city, it is a cloud node, otherwise it is an independent host. For simplicity, the paper assumes that the average radius of large, medium and small cities is 5km, 10km and 20km respectively.

$$\text{distance}(C_i) = \max \{\text{distance}(\text{loc}(IP_m), \text{loc}(IP_n))\}, \text{ where } IP_m, IP_n \in C_i$$ (4)

3.3. Geographical location inference

After the enough anchor node is collected, the geographic location of the target IP can be inferred by IP location technology. According to the response of target IP address to network measurement, it can be divided into two categories: responding IP and non-responding IP. The IP of the last hop in the route can be used to replace the non-responding IP.

For the responding node, it also needs to detect the fingerprint of the node to determine the target node type according to method in section 3.2. If the target node is a cloud computing node, the geographical location of the cloud service provider can be retrieved by calling Baidu interface. If it is an independent host, its location can be refereed according to the network topology proximity. The process is as follows: firstly, the network path to the target node is measured by traceroute, then the $N$ nearest IP to the node are selected by path matching method, and the geographic location of the target IP is determined by centroid method or nearest neighbor method.

Using the centroid method, the location of target IP is expressed as:

$$\left(\text{lat}_p, \text{long}_p\right) = \left(\frac{1}{N} \sum_{i=1}^{N} \text{lat}_i, \frac{1}{N} \sum_{i=1}^{N} \text{long}_i\right)$$ (5)

Where, $(\text{lat}_p, \text{long}_p)$ denotes the coordinates of the target node and $(\text{lat}_i, \text{long}_i)$ denotes the coordinates of node $i$.

Using the best neighbor rule, the location of target IP is expressed as:

$$\left(\text{lat}_p, \text{long}_p\right) \doteq \left(\text{lat}_{\text{nearest}_i}, \text{long}_{\text{nearest}_i}\right)$$ (6)

4. Experiment and verification

In order to verify the above proposed algorithm, we uses a cyber database AnchorX[2] to identify the key features of different type nodes, then compare the inference geolocation of IPs with the registration database. As we know, AnchorX is the most comprehensive Cyber database, which not only includes the geolocation data, but the web banners, open service/port, topology data, registration information, geocoding, CVE/CNVD/CNNVD and other related data.

4.1. The feature selection of different type nodes

In order to find the key features of different type nodes, the operating system and open services/ports of the node are detected. The CND nodes of TOP10 websites ranked by Alexs are selected as the representation of CND nodes, which includes Baidu, Tencent, Taobao, Sina, 360 and other head enterprises. By using the nationwide distributed detection capability provided by Anchorx, 242 acceleration nodes of these websites are found, and then all open ports and services are detected by Nmap, with a total of 533 active ports and average of 2.2 active ports. The two most open ports are
both HTTP service ports 80 and 443. In the same way, 54 virtual machine nodes are detected by measurement nodes of Anchorx, and a total of 300 active ports are found, with an average of 5.6 active ports. The most active port is SSH control port. For NAT nodes, 33 volunteers nodes are used to performance an analysis, a total of 365 active ports were found, with an average of 11 active ports.

| Node Type | Number | Port Number | High frequent port | Average active port |
|-----------|--------|-------------|--------------------|---------------------|
| CDN Serve | 200    | 533         | 443(98.0%), 80(98.0%) | 2.6                 |
| Cloud Serve | 54    | 300         | 22(75.0%)          | 5.6                 |
| NAT       | 33     | 365         | 593(72.0%), 4444(72.0%), 443(54.0%), 45(72.0%), 23(36.0%), 135(72.0%), 139(72.0%), 8080(57.0%), 1720(72.0%), 80(57.0%) | 11                  |

Through the above data, it can be found that the number of active ports and high frequency ports of different type nodes are quite different. Based on the above observation, the active port number and activeness of the high frequency ports are selected as key features to train Bayesian network. The parameters can be estimated by one dimensional kernel probability density.

4.2. Geolocation inference

To verifying our proposed algorithm, 10,000 IPs in AnchorX database are selected, distributed in 289 subnets. By measurement, there are 6,097 IPs which can find have effective anchors nodes. By the anchor node calibration algorithm, it’s find that most of them are university websites, government websites, enterprise websites. Fig.4 shows that cumulative distribution of the open service/port number of the anchor nodes. It can be find that the average port number of CND servers, cloud servers and independent-deploy servers are 2.1, 6.3, and 9.2 separately. It’s quite similar to the conclusion in the above section.

![Fig. 2 The cumulative distribution of the open ports/services number of the anchor nodes](image_url)

All the 10,000 IPs are measured, 6,097 IPs can find effective anchor node, distributed in 1460 subnets. In the experiment, if the location of target IP is same or more accurate than original record in the database, the inference results is accepted. By comparison, there are more than 5,651 IPs locations are more accurate than original records. The experiment results illustrates that TDIG can geolocate the target IP accurately, which can effectively avoid the need of large amount of anchor nodes.
Table 2 The verification of the algorithm

| Total IPs | Verified IPs | More Accurate | No Change |
|-----------|--------------|---------------|-----------|
| 10,000    | 6,097        | 5,651         | 446       |

5. Conclusion and Future work

As we all know the accuracy of IP geolocation relay on the available anchor nodes besides the target IP. However, it seems impossible that collecting and maintaining a large set of anchor nodes for any target IP. The experiment results show that Target driven IP Geolocation Algorithm can infer the geolocation accurately, effectively solve the problem of large-scale collection of anchor nodes. In the next step, the algorithm will be improved in a wider range, and more inferential features of nodes will be found, so as to make the algorithm more accurate and automated, and lay the foundation for cyberspace mapping and tracking.

Acknowledgments

This work was financially supported by the National Key Research and Development Program of China (No.2018YFB1800200 and No.2017YFB0801703), Foundation for Training Postdoctoral Innovative Talents in Southeast University (No.2242019R20024), Zhejiang public welfare Technology Application Research Project (No. LGG20F020014)

References

[1] WANG Zhan-Feng, FENG Jing, XING Chang-You, et al. Research on the IP Geolocation Technology[J]. Ruan Jian Xue Bao/ Journal of Software, 2014, 25(7): 1527-1540(in Chinese).
[2] Anchorx. [EB/OL]. https://www.anchorx.cn/, 2021.
[3] Ethan Katz-Bassett, John P. John, Arvind Krishnamurthy, et al. Towards IP Geolocation using Delay and Topology Measurements[C], In: Proc. of ACM IMC 2006, Brazil, 2006:71-84.
[4] Bamba Gueye, Steve Uhlig, S Fidia. Investigating the Imprecision of IP Block-Based Geolocation[C], In: Proc. of PAM 2007, LNCS, 2007(4427): 237-240.
[5] Bernard Wong, Ivan Stoyanov, Emin Gün Sirer. Octant: A Comprehensive Framework for the Geolocalization of Internet Hosts[C], In: Proc. of NSDI 2007, 2007.
[6] Mark Anderson, Ajay Bansal, Brad Doctor, et al. Method and apparatus for estimating a geographic location of a networked entity[P]. United States Patent 6,684,250, 2004.
[7] Brian Eriksson, Paul Barford, Joel Sommersy, et al. A Learning-based Approach for IP Geolocation[C], In: Proc. of PAM 2010, 2010.
[8] Venkata N. Padmanabhan, Lakshminarayanan Subramanian. An investigation of geographic mapping techniques for Internet hosts. In: Proc. of the ACM SIGCOMM Conference, 2001:173–185.
[9] Bamba Gueye, Artur Ziviani, Mark Crovella. Constraint-based geolocation of Internet hosts. ACM/IEEE Transactions on Networking, 2006,14(6):1219–1232.
[10] Yong Wang, Daniel Burgener, Marcel Flores, et al. Towards street-level client-independent IP geolocation. In: Proc. of the 8th USENIX, 2011.
[11] IP to Latitude/Longitude Server, University of Illinois, http://cello.cs.uiuc.edu/cgi-bin/slamm/ip2ll
[12] David Moore, Ram Periakaruppan, Jim Donohoe, et. al. Where is the World is netgeo.caida.org?, In: Proc of INET, 2000.
[13] Ram Periakaruppan, Evi Nemeth. GTrace-A Graphical Traceroute Tool. Usenix LISA, 1999.
[14] Inja Youn, Brian L. Mark, Dana Richards. Statistical Geolocation of Internet Hosts, In: Proc. of ICCCN, 2009.
[15] Sándor Laki, Peter Matray, Péter Hága, et al. Spotter: A Model Based Active Geolocation Service. In: Proc. of INFOCOM 2011, 2011.