Towards Region-level IP Geolocation Based on the Path Feature

Jingning Chen[1, 2], Fenlin Liu[1, 2], Tianpeng Wang[1,2], Xiangyang Luo[1,2], Fan Zhao[1,2], Guang Zhu[1,2]

[1]State Key Laboratory of Mathematical Engineering and Advanced Computing
[2]Zhengzhou Science and Technology Institute
jingning_chen@sina.com, fenlinliu@vip.sina.com, wangtp_mail@sina.com, luoxy_ieu@sina.com
zhaofan_123@yeah.net, mail_guang@foxmail.com

***Abstract*—While the delay-distance correlation among nodes is weak, calculating the distance constraints according to delay may fail to geolocate the target into a right region. Aiming at this problem, a new geolocation method is proposed in this paper, and according to extract path features for all possible regions in use of a large number of landmarks, this method can estimate the region for a target IP. The experiment results shown that, the proposed method can give the geolocation region with high accuracy.**

***Keywords*— IP geolocation, path feature, landmark, region-level geolocation, tracroute**

I. INTRODUCTION

IP geolocation refers to use the corresponding IP address to determine the location of a network entity in some level of granularity [1, 2]. According to the evolution of location based service (LBS), IP geolocation based applications are more and more popular, such as, the targeted advertising according to the users' location, adjusting the language on the site by ISP automatically according to the clients' regions, tracing cyber fraud and attacks and extracting the system logs for computer forensics, developing the deployment strategy of the network infrastructure and discovering fault nodes. Further, specifying the geographic region of a cloud service is increasingly common, and geographic region options are provided to help customers achieve a variety of objectives, including performance, continuity and regulatory compliance [3,4]. Therefore, IP geolocation is widely applied in Internet commerce and security, as well as the cloud service.

IP geolocation was openly discussed since 2001 [1], and with a decade of development, there are a variety of geolocation methods, including GeoTrack [1], CBG [5] (Constraint- Based Geolocation), Shortest Ping [6], TBG [6] (Topology Based Geolocation), Octant [7], SLG [8] (Street-Level Geolocation), LBG [9], etc. There are also a number of researches on the related key technologies [10~12]. Generally, Region estimation is the base of IP geolocation with high precision, and this means that, an accurate result (such as a single point denoted by longitude and latitude) can be got only for the network entity with known regional location. However, while geolocating a target IP, the corresponding region is usually unknown, and if the delay-distance correlation among nodes is weak, calculating the distance constraints according to delay will fail to get effective geolocation region for the target. Therefore, according to analysis the paths to different regions, and extract path feature for each region, the region estimation method based on path feature is proposed in this paper.

II. GEOLOCATION PRINCIPLE OF REGION-LEVEL GEOLOCATION

Different with the delay which changes dynamically, the path between two nodes is much more stable, and a large number of paths detection show that, interface IPs of intermedia routers on the paths from a probe point to different destination IPs located in the same region are similar, while the interface IPs are different for different regions. Taking the above conclusion as premise condition, a region-level geolocation method is proposed. For a certain probe point (P) and destination region (D), the path feature (PF) in this paper referred to a series of IPs corresponding to intermediate routers along with the paths from P to D. Obviously, for one D, different P will result in different PF. More specifically, PF consisted of the triples with hop, IP and probability, and denoted as <hop, IP, probability >.

The geolocation principle of proposed method is shown in Figure 1. Firstly, combined with prior knowledge of the target IP, find out the all possible regions in which target IP may be located, and those regions referred to destination regions. Then, select landmarks within destination regions from

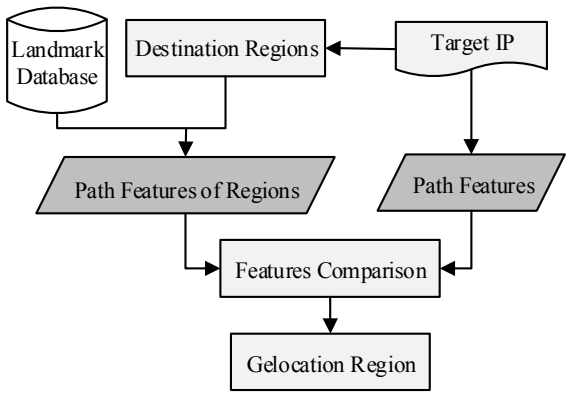

Figure 1. The diagramatic flow of proposed method

landmark database and obtain the paths from the probe to the selected landmarks; calculating the path feature for each destination region according to the paths to corresponding region. Finally, calculate a path feature for the target IP, and select a region from all destination regions according to path features of those regions.

Knowing from Figure 1 that, there are two key parts of region-level geolocation.

A. Extracting path features of the Region

The path features of a region consisted of multiple <hop, IP, probability>. For a certain P and D, and <5, 180.149.140.*, 0.243> is a component of PF, this means that, among the selected landmarks within D, on the paths from P to those landmarks, there are about 24.3% of landmarks whose the fifth intermedia router is identified with 180.149.140.*. In order to balance the load, and avoiding network fault caused by failure of single router, the key nodes on the communication path between two network entities are deployed multiple routers, and usually, same class-C IP addresses are assumed in the same city, therefore, while extracting path features, only the first three bytes of IP address related to router interfaces will be recorded.

The path features extraction of a region can be divided into the candidate feature extraction and probability calculating. For a given P and D, the main steps of candidate feature extraction for a region D are as follows:

1）Selecting landmarks: Find out part of landmarks within D from database, and make sure that those landmarks to cover all segments of IP address corresponding to D as much as possible.

2）Obtain paths: Probing all communication paths form P to selected landmarks using traceroute, and record the intermedia router interface and hops on each path.

3）Removing useless router interfaces: Delate interfaces near to P and selected landmarks from the above record.

4) Modifying IP addresses of interfaces: replace IP address of intermediate routers on paths with the corresponding class-C IP address, IP1.IP2.IP3.IP4 will be replaced by IP1.IP2.IP3.*.

5) Counting the same interfaces: for intermediate routers of each hop on all paths, statistics the number of each IP address.

After the extraction process, a set of candidate features of each region will be obtained, and a candidate feature is a triple with hop, IP and count, and denote as <hop, IP, count>. Since the intermedia routers may appear on the paths to different regions, we need to calculate the probability of the router on the paths related to a Specific region. The probability calculation of each candidate feature is as follows: 1) for <hop_i, IP_i, count_i>, find out all <hop, IP, count> if hop is equal to hop_i and IP as same as IP_i, sum each "count" and denoted by "Base_i"; 2) count_i divide by Base_i and take this as probability,update each <hop_i, IP_i, count_i> by the corresponding probability, and the <hop_i, IP_i, probability _i> is the path feature.

B. Geolocating based on path feature

For a certain probe point, the path features to network entities within same region are similar to each other, and different for regions, so path feature can be used for region-level geolocation. The main steps of geolocating target IP are as follows:

1）extracting path feature of target IP: Obtain the path from probe point to target IP, and extract path feature as region feature extraction, and then remove the useless feature and modifying the IPs of interfaces.

2) Select the most possible region: Compare target' feature with path features of all possible region, calculate the probability of target IP located in each region and choose the region with maximum probability as the geolocating result. This process is shown in figure 2.

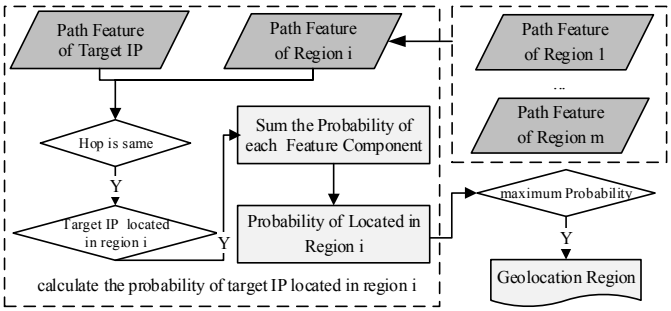

Figure 2. Geolocating based on path feature

III. EXPERIMENTAL RESULTS

The geolocation accuracy of proposed method depends on granularity of region feature. While the path region extracted in province-level, our method can estimate the target' province, and city-level feature can provide city-level geolocation result. In order to verify the effectiveness of the proposed geolocation method, we extract path feature for 3 cities（Shanghai, Shenzhen and Xi' an）, and geolocation accuracy of the city-level are given. In experimental, the single probe point is located in Beijing, and the distribution of probe point and 3 possible regions is shown in Figure 3.All paragraphs must be indented. All paragraphs must be justified, i.e. both left-justified and right-justified.

Figure 3. Distribution of probe point and possible regions

TABLE 1. THE ACCOUNTS OF TESTED LANDMARKS AND TARGETS

IPs	Total	Shanghai	Shenzhen	Xi'an
Landmarks	247	115	72	60
Targets	70	30	20	20

The accounts of landmarks and targets for 3 city are shown in Table 1. While extracting path features from landmarks for 3 cities, we select the intermedia interfaces from fifth to thirteen hops as candidate feature. Ignoring the fourth byte of IP addresses of router interface, the count of IP addresses of router interface and the corresponding hops are shown in Figure 4, while the Base_i of each hop are shown in Table 2.

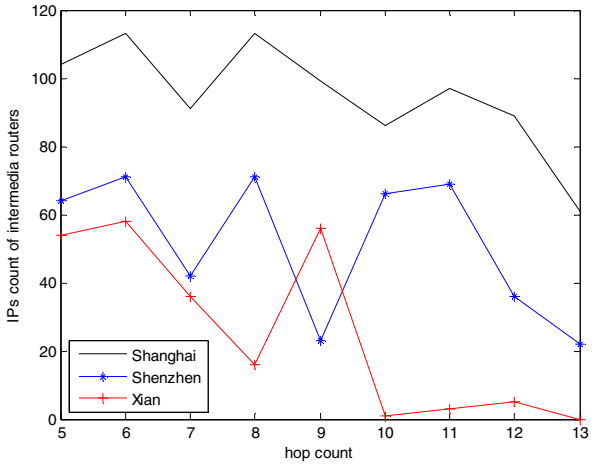

Figure 4. The count of IP addresses and corresponding hops

TABLE 2. THE BASE_I OF EACH HOP

Hop	5	6	7	8	9	10	11	12	13
Base_i	222	242	169	200	178	153	169	130	83

According to the Base_i and candidate features, calculate the probability of candidate comments, and then path feature can be extracted. The path feature of possible region (Xi'an) is shown in Table 3.

TABLE 3. THE PATH FEATURE OF XI'AN

Hop	IP	probability
5	180.149.140.*	0.243
6	180.149.128.*	0.24
7	180.149.128.*	0.172
7	220.181.177.*	0.041
8	202.97.65.*	0.03
8	202.97.80.*	0.05
9	117.36.240.*	0.107
9	218.30.19.*	0.084
9	218.30.69.*	0.124
10	117.36.240.*	0.007
11	117.36.120.*	0.006
11	117.36.123.*	0.006
11	61.150.6.*	0.006
12	124.115.221.*	0.008
12	125.76.192.*	0.008
12	125.76.242.*	0.015
12	61.150.6.*	0.008

For the 72 target IPs, obtain the path form probe point and record the corresponding path feature, then compare this feature with the path regions of 3 possible cities, choose the city with maximum probability the target located in, the geolocation accuracy of city-level is shown in Table 4.

TABLE 4. THE GEOLOCATION RESULT

IPs	Shanghai Y/N	Shenzhen Y/N	Xi'an Y/N
Target Region	30/0	20/0	20/0
Geolocation Region	30/2	20/5	13/0

The geolocation result in Table 4 shows that most of the targets can be geolocated into the right region. For total 30 targets located at Shanghai, the geolocation regions given by proposed method are Shanghai. For total 20 targets located at Shenzhen, the geolocation regions given by proposed method are Shenzhen. For total 20 targets located at Xi'an, 13 geolocation regions given by proposed method are Xi'an, while the another 7 are geolocated into the wrong cities (Two targets are supposed to at Shanghai, and five at Shenzhen).

For 218.30.15.156 and 218.30.66.101,the two targets supposed to at Shanghai, none of the intermediate routers on paths from probe point to this two IPs is on the paths to Xi'an, and then the probability (calculated by the geolocation method) of two IP located at Xi'an is 0. The reason of geolocation error of target located at Xi'an is larger than other two cities is that, the number of landmarks used to extract path features of Xi'an is not enough and uniform, and this can be improved by adding more landmarks to geolocation method.

IV. CONCLUSIONS

When fixing the probe point and destination region, the IP addresses of intermedia routers on paths to the region can be used as a path feature, and then the corresponding region can get a complete set of path features using landmarks with region-level location. For target IP geolocation, calculating the probability of each region with which the target located at, and the target will be geolocated into the most possible region. The experiment results shown that, the proposed method can give the geolocation region with high accuracy. .

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (No. 61379151, 61272489, 61302159 and 61401512), the Excellent Youth Foundation of Henan Province of China (No. 144100510001), the Innovation Scientist and Technicians Troop Construction Project of Zhengzhou City (No. 10LJRC182), and Foundation of Science

and Technology on Information Assurance Laboratory (No. KJ-14-108).

REFERENCES

[1] Padmanabhan, V. N., Subramanian, L., An investigation of geographic mapping techniques for Internet hosts, ACM SIGCOMM Computer Communication Review, vol. 31, no. 4, 2001, pp. 173-185.

[2] J. A. Muir and P. C. V. Oorschot, Internet geolocation: evasion and counterevasion, ACM Computing Surveys, vol. 42, no. 1, pp.1-22, 2009.

[3] Gondree, M., Peterson, Z. N., Geolocation of data in the cloud, Proceedings of the third ACM conference on Data and application security and privacy, ACM, 2013, pp. 25-36.

[4] Peterson, Z. N., Gondree, M., Beverly, R., A Position Paper on Data Sovereignty: The Importance of Geolocating Data in the Cloud, Proceedings of the 8th USENIX conference on networked systems design and implementation, 2011.

[5] Gueye, B., Ziviani, A., Crovella, M., Fdida, S, Constraint-Based Geolocation of Internet Hosts, IEEE/ACM Transactions on Networking, vol. 14, no. 6, 2006, pp.1219-1232.

[6] Katz-Bassett, E., John, J. P., Krishnamurthy, A., Wetherall, D., Anderson, T., Chawathe, Y., Towards IP geolocation using delay and topology measurements, Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement, 2006, pp. 71-84.

[7] Wong, B., Stoyanov, I., Sirer, E. G., Octant: A comprehensive framework for the geolocation of internet hosts, Proceedings of USENIX NSDI Conference, 2007, pp. 23-36.

[8] Wang, Y., Burgener, D., Flores, M., Kuzmanovic, A., Huang, C., Towards street-level client-independent IP geolocation, Proceedings of the 8th USENIX Conference on Networked Systems Design and Implementation, 2011, pp. 27-36.

[9] Eriksson, B., Barford, P., Sommers, J., Nowak, R., A Learning-based Approach for IP Geolocation, Proceedings of Passive and Active Measurements Conference, 2010, pp.171-180.

[10] Ping Guo, Jin Wang, Bing Li, Sungyoung Lee, A Variable Threshold-value Authentication Architecture for Wireless Mesh Networks, Journal of Internet Technology, vol. 15, no. 6, pp. 929-936, 2014.

[11] Xie Shengdong, Wang Yuxiang, Construction of Tree Network with Limited Delivery Latency in Homogeneous Wireless Sensor Networks, Wireless Personal Communications, vol. 78, no. 1, pp. 231-246, 2014.

[12] Jian Shen, Haowen Tan, Jin Wang, Jinwei Wang, Sungyoung Lee, A Novel Routing Protocol Providing Good Transmission Reliability in Underwater Sensor Networks, Journal of Internet Technology, vol. 16, no. 1, pp. 171-178, 2015.

Jing-ning Chen was born in 1985, China. She received the B.S. degree and the M.S. degree from Zhengzhou Information Science and Technology Institute, Zhengzhou, China, in 2008 and 2011, respectively. She is currently a Ph.D. candidate in the State Key Laboratory of Mathematical Engineering and Advanced Computing at Zhengzhou Information Science and Technology Institute. Her primary interest is in Network entity geolocation.

Fen-lin Liu was born in 1964, China. He received his B.S. from Zhengzhou Institute of Science and Technology in 1986, M.S. from Harbin Institute of Technology in 1992, and Ph.D. from the East North University in 1998. Now, he is a professor of Zhengzhou Institute of Information Science and Technology. His research interests include network and information security.

Tian-peng Wang was born in in 1979, China. Wang received the B.S. degree from Zhengzhou Science and Technology Institute, Zhengzhou, China, in 2002. He is currently an instructor in The State Key Laboratory of Mathematical Engineering and Advanced Computing at Zhengzhou Information Science and Technology Institute. His research interest includes landmark mining of network entities, IP geolocation and information Security.

Xiang-yang Luo was born in 1978, China. Luo received the B.S. degree, the M.S. degree and the Ph.D. degree from Zhengzhou Science and Technology Institute, Zhengzhou, China, in 2001, 2004 and 2010, respectively. He is currently an associate professor of Zhengzhou Information Science and Technology Institute. His research interest includes information security, image steganography and steganalysis.

Fan Zhao was born in 1989, China. Zhao received the B.S. degree from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2012. He is currently a M. S. candidate. His research interests focus on network security.

Guang Zhu was born in in 1990, China. ZHU got the B.S. degree from Henan University of Technology, Zhengzhou, China, in 2013. He is currently a M.S. candidate in The State Key Laboratory of Mathematical Engineering and Advanced Computing at Zhengzhou Information Science and Technology Institute. His research interest includes landmark mining of network entities, IP geolocation and information Security.