Anonymous node location based on clustering

Jihai Luan¹, Yong Xia¹, Yuancheng Xie¹, Dong Zhao¹, Zhaoxin Zhang¹,²* and Ning Li¹,²

¹ School of Computer Science, Harbin Institute of Technology (HIT), China
² Universidad Politécnica de Madrid, Spain

*Corresponding author. Email: heart@hit.edu.cn

Abstract. With the rapid development of network technology, the network's scale is increasing and becoming more complex. It is imperative to understand the structure of the network for network management and network security. Therefore, the construction of the routing topology has become a top priority. Routing topology analysis is already a relatively mature technology, but anonymous nodes' positioning has not received enough attention. Nowadays, the primary solutions to anonymous nodes are graph analysis and tomographic tree construction. Based on these two methods, this paper proposes a rough location of anonymous nodes based on the method of three-level clustering. They are the K-means clustering of data before anonymous identification, hierarchical clustering of feature values in anonymous identification, and central point calculation based on clustering ideas during anonymous positioning. Experiments have proved that performing the first-level K-means clustering before anonymous node identification can effectively improve anonymous positioning accuracy. And the three-layer clustering idea proposed in this paper can roughly get the location of anonymous nodes based on reducing the complexity of the original method.

1. Introduction

With the development of Internet, the progress of the third industrial Revolution has swept the world. People are stepping into the information age. This further promotes the process of globalization. The Internet's development is gradually accelerating. The number of users and scale of the Internet have also become very large. The network structure is becoming increasingly complex. Under such a vast network scale, how to quickly recognize the Internet's composition is necessary. And an efficiently control and manage for internet has become vital. The key to solving the problem mentioned above is to get an accurate network topology. It's also necessary to measure and analyze the critical network parameters. Therefore, it is indispensable to construct the topology for the IP level.

Anonymous nodes are precisely a problem that cannot be ignored in IP-level topology measurement. The most intuitive display of anonymous nodes is that there will be'*' when using traceroute command.

The reasons for Anonymous node here may be as follows. 1)Private address. 2)Network congestion. 3)Misconfiguration of the gateway node. 4)Some routers set the sending and receiving rate to prevent DDOS and other attacks. 5)The network administrator sets the router not to respond to ICMP packets. In either case, the path finally detected will be different from the real network environment.

Anonymous routing node's real meaning is that the exact IP address cannot be obtained directly. Due to the invisibility of the IP address, the identity of the point cannot be marked. When measuring multiple destination nodes, anonymous routers under different paths may be the same route. However,
due to unrecognizable reasons, they will be mistakenly identified as multiple nodes. Figure 1 shows the interference of anonymous nodes on real links.

![Figure 1. (a) is the network structure we analysed using traceroute command, and (b) is the real link.](image)

In terms of topology construction, the main problem is obtaining as accurate IP node information as possible. The main problem is how to solve the anonymity and alias problems. For the problem of anonymous nodes, the mainstream solutions are mainly divided into two. The first is the traditional way. This method can be divided into the following categories. 1)The direct connection method [1]. 2)Combination method [2]. 3)The only existence method [3]. 4)Graph induction [4].

The second way is based on the level of construction. Ratnasamy[5] who is the researcher of Berkeley University, and the researchers from the University of Electronic Science and Technology of China, all of them have studied in this area for a long time. The main difference of these methods are how to construct the chromatographic tree.

Comparing the two methods, the first method only follows a particular graph rule. It will cause the erroneous deletion of nodes and links. That will bring errors to topology recognition. The second method's advantage is that it combines the real network environment. However, the method based on tomographic construction needs to send "sandwich packets" to all points at the same time. The purpose is to obtain the real point-to-point network environment. Then analyze the similarity between links. But it cannot send a large number of detection packets to all points in a short time. Too much time difference will affect the fusion result. Therefore, this study divides the entire topology environment into several intervals. The purpose is to perform detection and fusion within the class simultaneously without destroying the topological integrity. And reduce the pressure and increase the accuracy of simultaneous detection on all nodes.

Based on the problems mentioned earlier, this paper proposes a method for rough positioning of anonymous nodes based on the idea of three-level clustering.

- The first layer is to rely on K-means clustering to select the cluster center as the source point. Then topologically construct within the class. So, it only needs to send the probe packets to the points in the class at the same time. The function of K-means clustering can also solve the problems of the fusion of multiple sources to a certain extent.

- The second layer is the hierarchical clustering within the class to construct the tomographic tree. Because the nodes within the class have a certain degree of similarity in the network environment, the interference will be reduced when obtaining the delay curve's feature vector.

- The third layer uses the Euclidean distance to calculate the position coordinates of the anonymous node. Based on the idea of clustering. This study uses the anonymous nodes' previous-hop and next-hop to calculate the anonymous router's location. Because the anonymous node's geographic location cannot be directly located by known means. So, the information we calculated can only be used to construct rough geographic information of anonymous nodes to facilitate topology display.

The main contributions of this article are as follows:

- Simplify the traditional tomographic construction method. By using the ping command multiple times to obtain the transmission delay for a specific destination node. Then construct the delay curve. Finally use the wavelet basis decomposition to get the feature vector.
• Other people's articles have analyzed a lot of how to deal with multi-source node fusion methods. Still, they have not reached a very effective way. So, this article proposes based on K-means’ multi-source anonymous recognition. Clustering the data and classification according to geographic location. Each source point only detects the destination node in the class. That can effectively improve the positioning accuracy of anonymous nodes.

• In terms of anonymous node fusion, this article uses graph analysis and merging method. Unlike the traditional integer programming model, this study only processes two paths at a time. Rules are formulated according to the possible conditions in the topological graph. Then, according to the hierarchical clustering results and the two traceroute paths, a set of anonymous nodes will be obtained by fusion analysis.

This article will introduce related technical research in Chapter 2. Chapter 3 will introduce the detailed methods of this research. Chapter 4 will show the results and summarize the conclusions. Chapter 5 will be a summary of the full text and experimental results.

2. Related works
For the two types of methods summarized in Chapter 1, many universities and research institutions conduct related research and all gains.

In the first method, the more commonly used method is graph analysis. Through the process of studying anonymous routers, M.Gunes et al. found that most of the complex structures formed by anonymous routers have the law of graphs to some extent. They also give a simple alternative graph to characterize. By traversing the entire topological structure to find a graph that conforms to this graph, the entire network's topological structure can be simplified. Furthermore, in 2015, through the analysis of historical data, M.Gunes further proposed a combination of minimizing graph theory and data mining to solve the problem of individual anonymous nodes [7]. In the "PoP network topology-based city-level IP geolocation algorithm" proposed by S. Zu et al [8], but they only analyze the graph without additional analysis of the real network status. So, it will cause the accidental deletion of nodes and links. Besides, in sub-graph extraction, the amount of calculation is tremendous and cannot be resolved. The occurrence of consecutive anonymous nodes brings errors to topology recognition.

The second method is constructing a tomographic tree. This method can merge anonymous routers in the same path into an anonymous route. And it extends to an anonymous route identification algorithm based on network tomography. This method aims to use a specific algorithm to cluster all anonymous routers in the network. The purpose of clustering is to aggregate the routes with a close logical relationship to achieve anonymous identification. Furthermore, many scholars have studied the method of the tomographic tree. In 2000, Castro et al. proposed to use unicast detection packets to estimate the packet loss rate and use the ALT algorithm to estimate the network topology. At this time, the traceroute method has not been combined [9-10]. In 2004, Duffield et al. proposed a BDT method that uses the delay variance to estimate the tree topology [11].

The team of Professor Hu of the University of Electronic Science and Technology of China proposed the multi-source anonymous node location method. They combine the network tomography construction and the integer programming model [12]. Moreover, in 2019, the University of Electronic Science and Technology of China team proposed an "anonymous route identification method based on network structural characteristics" [13]. They use the method of graph embedding to identify the structural characteristics of the network. And obtains the continuous characteristics of the network's anonymous routing nodes as a basis. Aims to calculate the vector distance of anonymous routes, and cluster the anonymous router.

3. Materials and methods
The overall framework proposed in this paper mainly includes six processing stages: 1) Select and collect spare data for the detection source. 2) Perform K-means clustering on all points. 3) Construction of the chromatographic tree within the class. 4) Anonymous node topology integration. 5) Rough positioning of anonymous nodes 6) Error analysis.
First, obtain the domestic IP address in China, store the surviving IP, obtain the urban development level, expand the data according to the weight. And then perform the K-means algorithm on the data (k=3 is selected in this experiment). After that, servers are deployed in cities near the cluster center to obtain the delay curve and traceroute path. Then constructs the tomographic tree and merges the paths. Figure 2 summarizes the overall flow chart.

3.1. Node data pre-processing
The original data only contains the cities and their IP addresses. First, this point's weight is set according to the city where the IP node is located. The weight is calculated by combing the city development level and the number of destination IPs in the city. In clustering, the weight will be processed according to the degree of economic development and the node's location. The flowchart is shown in Figure 3.

3.2. Weighted K-means clustering
The construction of the tomographic tree needs to be based on data that can truly reflect the network environment. The delay of the network may vary greatly every second, and it is impossible to obtain the delay curve for a large number of points at the same time during detection. Therefore, this study proposes a method based on K-means for intra-class detection. On the one hand, it alleviates the
problem of inaccurate data for detecting all points simultaneously. On the other hand, since the network environment of points belonging to the same class will not be too far apart, the experimental results will also be more accurate. This study's partitioning method is a triplet, including IP address, corresponding weight, and location information. The use of the weight in this experiment is to expand the data accordingly. So, the first step is to expand the number of points based on the weight of the k-means result. The second step of clustering will generate K cluster centers, corresponding to k clusters. Each cluster represents a clustering interval. The number of k is artificially designated. The specific k clusters also need to meet the following conditions: each cluster must contain at least one IP address, and each IP address can only belong to one cluster. In this paper, the standard for the definition of similarity is based on Euclidean distance to group the latitude and longitude into a cluster. The different distances are as far as possible. Figure 4 shows the distribution of nodes after clustering (in this study, let K=3).

3.3. Construction of the chromatographic tree within the class
The way of network tomography is accepted by most people now. the method of constructing a tomographic tree in this experiment can be combined with graph inference and the real state of the network. Its advantage is that there is no need to pay too much attention to nodes other than the source and destination nodes. This experiment uses the ping command to acquire the delay.

In the actual network environment, the time delay of the network is a curve that changes with time. In the network, different destination nodes' delay change waves are similar on the shared path. So that this experiment can use the ping command. From each source point S={s_1,s_2,s_3}(this experiment uses K-means clustering results to select three cities in China as the source point finally), use the ping command multiple times (send about 30-50 times) to the destination nodes and record the delay each time. Draw the delay curve. Use the wavelet basis decomposition to obtain the delay curve's eigenvalues to form the eigenvectors through DBN wavelet decomposition. Wavelet decomposition can finely decompose the high-frequency part of the signal. And this decomposition is without redundancy and omission[14]. Then perform hierarchical clustering on the feature vector D. Draw a tomographic tree based on the degree of similarity. According to the clustering results, the servers are deployed in cities near the cluster center. Then perform hierarchical clustering to obtain a tree diagram.

3.4. Anonymous node topology fusion
When performing anonymous fusion, this study adopts the merge method. First, the continuous anonymous nodes are grouped into one node. Anonymous fusion combining tomographic tree and graph analysis. The anonymous node fusion has the following points:

1. In two different paths, the parent node and child node of the anonymous nodes are agreed nodes. Then the two anonymous nodes are the same node. This situation is shown in Figure 5(a).
2. In two different paths, anonymous nodes with different parent nodes but the same child nodes can also be classified into one category, as shown in Figure 5(b).
The paths obtained by traceroute and the result obtained by hierarchical clustering are used for fusion analysis. The fusion of anonymous nodes is performed according to the following rules. Two paths are compared each time, and each fusion is performed from the destination node to the source node. Finally, for each group of anonymous node sets \((a_i)\), the set of the previous hop \((u_i)\) and the next-hop \((f_i)\) is constructed, which is denoted as the adjacency set \(C_i\) here. First, according to the hierarchical clustering results, the two paths with the greatest similarity are selected from the database each time. If the previous nodes are always the same, find the first different and the last same node \((u_i, u_j, f_i, f_j)\). And store the four nodes into the adjacency set \(C_i\) (corresponding to the first case, search from back to front the location of the anonymous node). After finding the first different node, the first two different nodes \(f_i\) and \(u_j\) are stored in the adjacency set \(C_i\) (Corresponding to the second case). The pseudo code is as follows.

**Algorithm** Part of Anonymous node fusion

| Input: tracer1, tracer2 |
|-------------------------|
| Output: ans: The set of up hop and next hop of anonymous node |

1: start ← 0, end ← 0, i ← 0, res ← [], L ← min(len(tracer1), len(tracer2))
2: while i < L do:
3:   if tracer1[i] == tracer2[i]:
4:     if * not in tracer[i]:
5:       start ← i
6:     elif tracer1[i] != tracer2[i] and * not in tracer1[i]:
7:       end ← i
8:       break
9:   i ← i + 1
10: if start and end and start < end do:
11:   for i ∈ (start, end + 1) do:
12:     if * in tracer1[i] and * in tracer2[i] do:
13:       res.append((tracer1[i],tracer2[i]),(tracer1[start],tracer1[end],tracer2[start],tracer2[end]))
14:     break
15: if start do:
16:   for i ∈ (0,start) do:
17:     if * in tracer1[i] do:
18:       m1 ← tracer1[i]
19:     if * in tracer2[i] do:
20:       m2 ← tracer2[i]
21:     if m1 and m2 do:
22:       res.append((m1,m2),(tracer1[i-1],tracer1[i+1],tracer2[i-1],tracer2[i+1]))
23: if start and not end and start < L - 1 do:
24:   if * in tracer1[start+1] and * in tracer2[start+1] do:
25:     res.append((tracer1[start+1],tracer2[start+1]),(tracer1[start]))
26: return res

3.5. Anonymous node rough positioning

Next is the rough positioning of anonymous nodes. The anonymous node's adjacency set is \(C\) (the up and down jump sets of anonymous nodes). Since these points are known IP, the geographic information of these points is also known. In this experiment, this set is regarded as a category, and the center point of
the set is taken because it can be converted into geographic coordinates. The anonymous node position is within the range of these points. So, the approximate geographic coordinates of anonymous nodes can be located.

3.6. Error analysis module
Set the threshold to view the variance of the distance between the anonymous node that is coarsely located and its up and down nodes. If the distance is too far, it can indicate the accuracy of anonymous node positioning to a certain extent to determine the accuracy of positioning.

4. Result

4.1. Result display
In the process of traceroute, a tiny number of nodes will jump to foreign countries. These nodes are due to the lease of the server. And these points will be ignored in subsequent processing.

To prove the effectiveness of data preprocessing, a comparative test is set up. There are three groups which source points are selected near the cluster center. And the comparative test group is the result of detection from City 3 (Weihai) to all destination nodes. It is used to compare with the last three groups. Aims to see whether it is valuable to perform K-means clustering for intra-class detection, and whether improve the accuracy of anonymous nodes. Detect the corresponding intra-class nodes from City 1, City 2, and City 3 to obtain the transmission delay curve and perform traceroute detection. The final result of chromatographic tree construction is shown in the figure 6. And Table 1 shows part of the database data.

Figure 6. (a)(b)(c) is the tomographic tree display of detection and fusion from the cluster center to the points in the cluster, (d) is the tomographic tree display of the detection and fusion from one point to all points.

Table 1. Dsip is the destination IP address, Tznumpy is the eigenvalue decomposed by wavelet base, Trace is the original path obtained by the traceroute command.

| Id | Desip      | Tznumpy             | Trace                      |
|----|------------|---------------------|----------------------------|
| 1  | 1.180.141.230 | 0.11183987, ......, 0.16581239 | 100.120.18.13 *13 ... 1.180.153.19 |
| 2  | 1.180.153.190 | 0.03535305, ......, 0.10606017  | 100.120.18.13 *7 ... 1.180.154.22  |
| 3  | 1.180.128.82  | 0.04999999, ......, 0.08660540  | 100.120.183.13 *3 ... 1.180.164.9 |
After anonymous topology fusion, the last-hop and next-hop set $C_i$ of anonymous nodes will be obtained. The data is shown in the following table 2. When there is only one node in the anonymous node-set, it means that the anonymous node is not fused.

Table 2. Anonymous node and its up and down set.

| Anonymous node collection ($a_i$) | Anonymous node last hop and next-hop set ($C_i$) |
|----------------------------------|-----------------------------------------------|
| ('248)                           | [221.131.221.226, '111.1.68.20']              |
| ('417, '247)                     | ['112.6.224.1', '120.222.48.157', '211.137.196.173'] |
| ('275, '397)                     | [221.183.48.69, '221.183.57.250', '221.183.57.234'] |

Set multiple thresholds. When it is greater than this threshold, the located anonymous node deviates too far from the top and bottom. Prove that there is a large error. Check the error comparison between the clustering results of multiple source points and the comparison group under different thresholds. The following figure 7 shows the number of nodes within the threshold for each detection source point, and $x$ in the table represents the value of variance.

5. Conclusion

It can be seen from the above information, when $x=0$ (when $x$ is 0, the anonymous node is not fused) it can be seen that, in general, the results of multi-source point detection are mostly in $(0, 0.5]$ In this interval, most of the comparison test groups are in the interval greater than 1. Therefore, from the perspective of the data's accuracy. The variance from the anonymous node detected by the multi-source point to the up and down node is lower than the result of single point detection. So, the accuracy of the data has been improved. Therefore, this experiment fully proves that the addition of the k-means method can effectively improve the results' accuracy before anonymous node positioning. Furthermore, the rough location of anonymous nodes can be realized through the idea of three-layer clustering.

Acknowledgement

This work is supported by National Key R&D Program of China under Grant No.2018YFB1800202, No.2017YFB0803001, and the National Science Foundation of China under grant No.61571144, No.61972089, No61672156.

References

[1] Bilir, S, Sarac, K., & Korkmaz (2005). Intersection characteristics of end-to-end Internet paths and trees. Network Protocols, 2005. ICNP 2005. 13th IEEE International Conference on.
IEEE.

[2] Cheswick, B., Burch, H., & Branigan, S. (2000, June). Mapping and visualizing the Internet. In USENIX Annual Technical Conference, General Track (pp. 1-12).

[3] Broido, A., & Claffy, K. (2002). Internet topology: connectivity of ip graphs. Proceedings of SPIE the International Society for Optical Engineering, 4526.

[4] Gunes, M. H., & Sarac, K. (2008, April). Resolving anonymous routers in Internet topology measurement studies. In IEEE INFOCOM 2008-The 27th Conference on Computer Communications (pp. 1076-1084). IEEE.

[5] Ratnasamy, S., & McCanne, S. (1999, March). Inference of multicast routing trees and bottleneck bandwidths using end-to-end measurements. In IEEE INFOCOM'99. Conference on Computer Communications. Proceedings. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies. The Future is Now (Cat. No. 99CH36320) (Vol. 1, pp. 353-360). IEEE.

[6] Gunes, M. H., & Sarac, K. (2008, April). Resolving anonymous routers in Internet topology measurement studies. In IEEE INFOCOM 2008-The 27th Conference on Computer Communications (pp. 1076-1084). IEEE.

[7] Kardes, H., Gunes, M. H., & Sarac, K. (2015). Graph Based Induction of unresponsive routers in Internet topologies. Computer Networks, 81, 178-200.

[8] S. Zu, X. Luo, S. Liu, Y. Liu and F. Liu.City-Level IP Geolocation Algorithm Based on PoP Network Topology," in IEEE Access, vol. 6, pp. 64867-64875, (2018).

[9] Y. Vardi. Journal of the American Statistical Association. Network tomography: Estimating source-destination traffic intensities from link data. Journal of the American Statistical Association. (1996).

[10] Liang G, Yu B. Maximum pseudo likelihood estimation in network tomography[J]. IEEE Transactions on Signal Processing, (2003, 51(8): 2043-2053).

[11] N. G. Duffield, Francesco Lo Presti. Network tomography from measured end-to-end delay covariance.(2004, 12(6):978-992).

[12] Li Nan. Research on Network Topology Optimization Detection and Identification Method [D]. University of Electronic Science and Technology. (2018).

[13] Gaolei, Fei, et al. Anonymous Router Identification Method Based on Network Structure Characteristics. (2019).

[14] Wang Xin. Research on Non-stationary Network Topology Estimation Method Based on Wavelet Transform. Diss. (2019)