A City-Level Boundary Nodes Identification Algorithm Based on Bidirectional Approaching

Zhiyuan Tao[1,2], Fenlin Liu[1,2], Yan Liu[1,2*] and Xiangyang Luo[1,2]
[1] State Key Laboratory of Mathematical Engineering and Advanced Computing
Zhengzhou, Henan, 450001 China
[2] Key Laboratory of Cyberspace Situation Awareness
Zhengzhou, Henan, 450001 China
[e-mail: tzy_idle@sina.com, liufenlin@vip.sina.com, ms.liuyan@foxmail.com, luoxy_ieu@sina.com]
[*]Corresponding author: Yan Liu

Received February 6, 2021; revised April 4, 2021; revised May 17, 2021; accepted June 7, 2021; published August 31, 2021

Abstract

Existing city-level boundary nodes identification methods need to locate all IP addresses on the path to differentiate which IP is the boundary node. However, these methods are susceptible to time-delay, the accuracy of location information and other factors, and the resource consumption of locating all IPes is tremendous. To improve the recognition rate and reduce the locating cost, this paper proposes an algorithm for city-level boundary node identification based on bidirectional approaching. Different from the existing methods based on time-delay information and location results, the proposed algorithm uses topological analysis to construct a set of candidate boundary nodes and then identifies the boundary nodes. The proposed algorithm can identify the boundary of the target city network without high-precision location information and dramatically reduces resource consumption compared with the traditional algorithm. Meanwhile, it can label some errors in the existing IP address database. Based on 45,182,326 measurement results from Zhengzhou, Chengdu and Hangzhou in China and New York, Los Angeles and Dallas in the United States, the experimental results show that: The algorithm can accurately identify the city boundary nodes using only 20.33% location resources, and more than 80.29% of the boundary nodes can be mined with a precision of more than 70.73%.

Keywords: Internet measurement, Network topology, Boundary node identification, Bidirectional approaching.

This work was supported by the National Natural Science Foundation of China (No. U1804263, U1736214), the Zhongyuan Science and Technology Innovation Leading Talent Project (No. 214200510019).

http://doi.org/10.3837/tiis.2021.08.004 **ISSN : 1976-7277**

1. Introduction

City-level boundary nodes of a network usually refer to nodes that perform the task of cross-city data transmission in communication between different cities. The set of inter-city boundary nodes constitutes the network boundary of a city [1]. Identifying city boundary nodes is very important for preventing external attacks, deploying targeted security protection measures [2, 3, 4] and defining electronic taxation based on city.

The method of boundary node identification firstly originated from the boundary research of AS (Autonomous System). For example, Ref.[5] developed an AS-level boundary identification system called bdrmap, and based on the measurement results, constructed a router level network topology combined with the topology constraints inferred from BGP (Border Gateway Protocol) data [6] to narrow the link set and associated IP addresses of the boundary between networks, thus inferring the boundary of AS. Ref. [7] introduces the method of MAP-IT, which combines the data of AS switch to infer the boundary interface of AS from traceroute data. Giotsas [8] iteratively improved the inferred possible peer interconnection facility by using the inter-AS links derived from the router level diagram constructed by Midar AS input to the constraint facility. The Ref. [9] combines the content of Ref. [5, 7] and adds the voting mechanism to identify the boundary of AS. The granularity of boundary identification of AS is relatively coarse. The mapping relationship between IP and AS is studied more, rather than the boundary nodes between different regions in reality.

With the continuous expansion of the scale of the Internet, the maintenance needs of network boundaries have promoted the study of city boundary nodes on the basis of AS boundary identification. Unlike boundary routers of AS, city boundary routers have no apparent protocol features, so researchers need to mine boundary nodes from other perspectives. There are still few researches on city-level boundary node identification, which are mainly divided into two categories: one is based on time-delay characteristics in the network, and the other is based on location methods.

The first kind of methods is based on communication delay. Considering that the delay between routers that are close to each other should be small, the delay between routers within the same cities is small, and the delay between routers between different cities is large. For example, Ref. [10] probes the single-hop delay in the path and finds that the delay presents a "low-high-low" distribution. According to this distribution feature, the boundary IP is found, and the path is divided. When the single-hop delay meets the above characteristics, the boundary IP can be obtained by this method. Ref. [11] identifies the boundary IP based on the difference in single-hop delay of the path and the difference in router hostname of different cities. If the difference in single-hop delay is noticeable, the boundary IP can be identified by comparing each single-hop delay with the target city delay threshold. For paths where the differences between single-hop delays are not significant, the composition of IP hostnames per hop in the path is analyzed, and the boundary IP is further identified based on the differences in hostname strings. However, this kind of method is difficult to distinguish the boundary effectively when the delay information has no obvious distribution characteristics, and the corresponding host information is lacking.

The second kind of method locates each hop in the detection path based on IP location algorithm and divides the path into inner and outer parts of the city according to the location results, so as to obtain the boundary nodes of the city. Such as Ref. [12] using the statistical information of probing landmarks to identify the border IP, probing some of the city's landmarks, for each landmark, the method extracts each IP from the path, and check the IP whether it is in the same city landmarks; if belongs to the same city, will the IP as the boundary

of the corresponding city IP, and continue the above analysis on the next-hop of the path. In addition, other IP location algorithms such as SLG [13], Lencr [14] and GEO-RMP [15], as well as existing IP location databases, can also be used for this kind of boundary node identification method. The above methods need to locate each IP address on the path, for medium-sized cities, in such a way to identify the boundary requires higher probing cost and computing resources, is not suitable for the actual research, and cannot be carried out in the absence of landmark data and high-precision location information.

In view of the problems of the above methods, this paper intends to propose a city-level boundary node identification algorithm based on bidirectional approaching. In this algorithm, the router-city mapping set is established by two-way sampling measurement, and the communication path between the external city and the target city is obtained. Based on this, candidate boundary nodes are obtained. After obtaining the high-precision location data, the algorithm can further verify the candidate boundary nodes, so as to obtain the accurate boundary nodes. The main contributions of this paper are as follows:

- Propose an iterative measurement method based on target sampling to measure the nodes of the target city iteratively, construct the router set of target city, and obtain relatively complete topological information of target city;
- Adopt the path intersection point selection strategy of bidirectional approaching to aggregate the results of internal and external vantage points to find intersection points between cities, so as to reduce the size of nodes to be identified;
- Design a lightweight boundary node determination algorithm based on IP address database to improve the precision of boundary recognition and to label some errors in the existing IP address database.

The rest of this paper is organized as follows: In Section 2, describe the problems studied in this paper, elaborates the problems studied in this paper and gives an explanation of the symbols used in this paper. In Section 3, the main steps and principle analysis of the proposed algorithm are given. In Section 4, the effect of boundary node recognition is verified experimentally, and its recognition rate, precision and performance are analyzed. In Section 5, the paper is summarized and prospected.

2. Problem Formulation

To ease the understanding of the proposed algorithm, in this section, we first define the key concepts used in this paper. The problem formulation in the process of city-level boundary node identification is also presented here.

Probe paths. $p_{i\to j}=\{v_i,\mathrm{K},v_B,\mathrm{K}\ v_j\}$ represents the probe path from node v_i to node v_j and consists of each hop's IP address on the communication path.

Network topology. $G(\mathbf{V},\mathbf{E})$ represents the network topology composed of the distribution of computers and their connection relations. $\mathbf{V}$ represents a collection of nodes in the topology. $\mathbf{E}$ represents the collection of edges in the topology. $e_{i,j}$ represents the edge from node v_i to node v_j. $\mathrm{C}(v_i)$ represents the city of the node $v_i\in\mathbf{V}$. $\mathrm{IP}(v_i)$ represents the IP address of the node $v_i\in\mathbf{V}$.

Boundary routing nodes. $\mathbf{V}_B(X,Y)$ represents the set of boundary routing nodes between city X and city Y. Boundary routing node v_B refers to the intermediate router connecting

two cities. Assume that node v_x is located in city X and node v_y is located in city Y, $p_{y\to x}=\{v_y,\mathrm{K},v_z,v_B,\mathrm{K}\ v_x\}$ is the probe path from node v_y to node v_x, v_B is a boundary node between city X and city Y, then $\mathrm{C}(v_z)\neq X$ and $\mathrm{C}(v_B)=X$. That is, v_B is the first hop into city X.

External path of the city. $p_{\text{external}}=\{v_y,\mathrm{K},v_B\}$ represents the path from the IP node v_y located outside the target city to the intermediate router v_B.

Internal path of the city. $p_{\text{internal}}=\{v_B,\mathrm{K},v_X\}$ represents the path between the intermediate router v_B and the IP node v_x inside City *X*. All $v_i\in p_{\text{internal}}$ satisfy $\mathrm{C}(v_i)=X$.

The problem studied in this paper is: given all IP address blocks of target city X, to find the boundary nodes of the city. In this paper, we plan to divide the probe path $p\in\mathbf{P}_{Y\to X}$ from city X to city Y into two parts: internal path of the city p_{internal} and external path of the city p_{external}. The intersection points v_B of these two paths is found to be the city boundary nodes, and the set of such nodes on all probe paths is the city boundary nodes set $\mathbf{V}_B(X,Y)$.

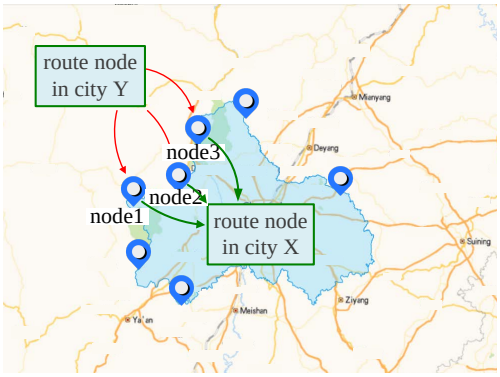

Fig. 1. Schematic diagram of intercity communication

To ease the understanding, take **Fig. 1** as an example to illustrate. The communication path from the routing node in city Y to the routing node in city X is divided into two parts. The red line represents the external path of the city, and the green line represents the internal path of the city. The nodes transmission information between the two parts, such as node1, node2 and node3, are the boundary nodes between city X and city Y.

The symbols used in this paper are described in **Table 1**.

Table 1. List of notations

Notations	Description
$p_{i\to j}$	The probe path from node v_i to node v_j
$G(\mathbf{V},\mathbf{E})$	Network topology
$\mathbf{V}$, $\mathbf{E}$	Node set; Edge set
$\mathrm{C}(v_i)$, $\mathrm{IP}(v_i)$	The city of the node $v_i\in\mathbf{V}$; The IP address of the node $v_i\in\mathbf{V}$
$\mathbf{V}_B$	Collection of boundary nodes of the target city
$\mathbf{V}_B(X,Y)$	Collection of boundary nodes between city *X* and *Y*
p_{external}, p_{internal}	External path of the city; Internal path of the city

X	Target city
$\mathbf{P}_{Y\to X}$	Collection of probe paths from city X to city Y
$\mathbf{V}_V^I$, $\mathbf{V}_V^O$	Vantage points located inside / outside the target city
n_I , n_O	Number of the inside / outside vantage points
$\mathbf{V}_V$, $\mathbf{V}_T$	Vantage points set; Target IP set
D	IP address database
t_{p}	Probing cycle
$\mathbf{V}_N$	Collection of new routing nodes
M_{RL}	Router-city map
$\mathbf{V}_C$	Collection of candidate boundary nodes

3. Algorithm of City-Level Boundary Node Identification Based on Bidirectional Approaching

To solve the problem that the existing boundary node identification algorithms' location resource cost is high and cannot work in the environment of lacking high-precision location information. This paper adopts a measurement algorithm of internal and external approaching, establish a route-city mapping set to select candidate boundary nodes in the absence of high-precision location data. After obtaining the high-precision location data, the accurate results can be obtained only by verifying the location information of one hop before and after the candidate boundary nodes. The schematic diagram of the algorithm in this paper is shown in Fig. 2.

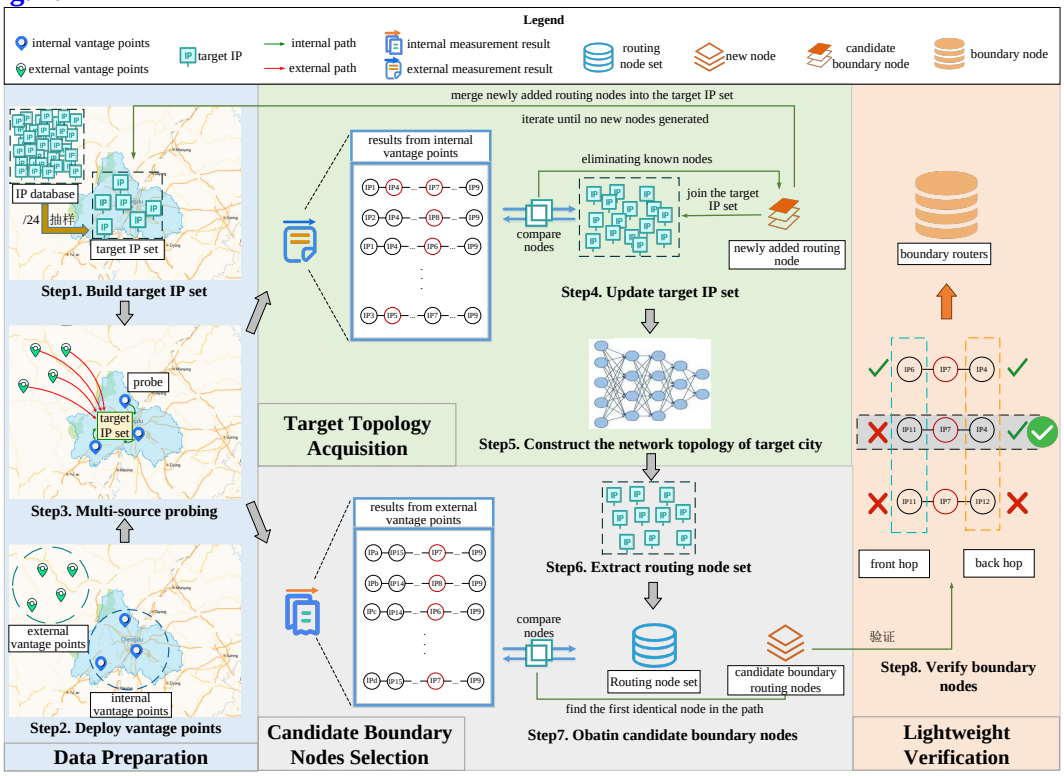

Fig. 2. Schematic diagram of city-level boundary node identification algorithm

The detailed steps of the city-level boundary node identification algorithm based on bidirectional approaching are as follows:

Step 1: Select the vantage points. n_I vantage points $\mathbf{V}_V^I$ located inside the target city X, and n_O vantage points $\mathbf{V}_V^O$, located outside the target city are selected to form the vantage points set $\mathbf{V}_V$.

Step 2: Sampling target IP. Get the IP address blocks assigned to the target city A from the IP address database $D=\{D_1, D_2, D_3, D_4, D_5, D_6\}$. In each IP address blocks, the representative IP of the block is selected to form the target IP set $\mathbf{V}_T$.

Step 3: Internet measurement. The vantage point set $\mathbf{V}_V$ is used to probe the target IP set $\mathbf{V}_T$ with the probing cycle t_{p}.

Step 4: Update the target IP set $\mathbf{V}_T$. By comparing the probing results obtained in the current round and the last round, the new added routing node $\mathbf{V}_N$ is taken as the target IP, and the target IP set $\mathbf{V}_T$ is updated.

Step 5: Iterate step 3-4 until the number of routing nodes in the target area stops increasing to build the topology G of the target city.

Step 6: Build the router-city map list M_{RL}. According to the constructed target city topology G, the routing nodes in the city are added to the router-city list.

Step 7: Obtain the candidate boundary nodes $\mathbf{V}_C$. According to the probing results returned by the vantage points distributed inside and outside the city, the measurement results are approaching bidirectional. The routing nodes in the path from external vantage points are compared with the router-city mapping list M_{RL} to find the first identical node in the path, and serve as the candidate boundary node $\mathbf{V}_C$ in the target city.

Step 8: Verify the candidate boundary nodes. The hop before and after the candidate boundary routing nodes $\mathbf{V}_C$ are verified by combining existing high-precision IP location databases D. If the hop before the candidate node is outside the target city and the hop after is inside the target city, the node will be added to the boundary routing node set $\mathbf{V}_B$.

Step 1 and 2 are the data preparation stage, corresponding to the blue module in the figure. Step 3-5 is the topology acquisition part, corresponding to the green module in the figure. Step 6-7 are the part of selecting candidate boundary nodes, corresponding to the gray modules in the figure. Step 8 is the part of verifying boundary nodes, corresponding to the orange module in the figure. Since the last three parts are the core steps of the algorithm in this paper, these three modules are mainly introduced in the following sections.

3.1 Iterative City Topology Acquisition Based on Target Sampling

The resource consumption of probing all IP addresses in Medium-sized city and above areas is vast, and the cycle of probing is long, while the target IP is likely to be in different network environments (for example, the network may be congested), each target IP can only be measured once during the probing cycle, which may result in accidental results [16]. Simultaneously, existing research shows that under the same network segment, a group of IP addresses assigned to the same organization often have the same or similar characteristics, for example, the export routers they use for external communication tend to be the same [17, 18]. Based on this network feature, this paper obtains the topology of the target city. The process diagram of this step is shown in **Fig. 3**.

1) Sampling target IP.

The target IPes of a specific city is screened. On the basis of ensuring that each /24 prefix network retains at least one IP, one or more IP addresses are selected for each network segment according to the IP blocks divided in the IP address database. Select the representative IP of this network segment from each IP block from the IP address database $D=\{D_1,D_2,D_3,D_4,D_5,D_6\}$ to form the target IP set $\mathbf{V}_T$, and then only probe the IP addresses screened out, which can shorten the cycle and reduce resource consumption.

2) Measure the newly added nodes to supplement the gap iteratively.

n_I vantage points located inside the target city X are selected to probe $\mathbf{V}_T$ with the probing cycle t_p. In each round of probing, new routing nodes that do not belong to the initial IP set $\mathbf{V}_T$ are acquired in the path, which are added to $\mathbf{V}_T$ and a new round of probing is carried out. The completeness of the topology obtained by probing is ensured through iteratively adding routing nodes in the city.

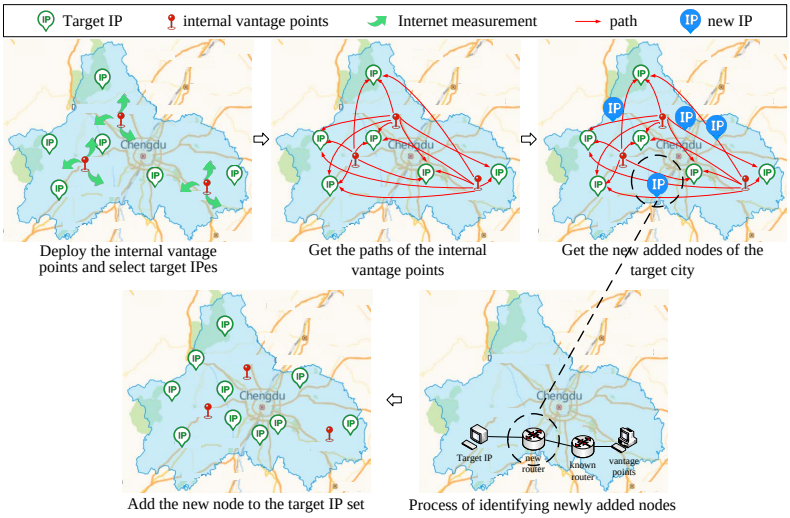

Fig. 3. Schematic diagram of iterative measurement

The detailed city topology acquisition process is shown in Algorithm1.

Algorithm 1: Algorithm of Iterative City Topology Acquisition Based on Target Sampling

Input: Target city X, target IP set $\mathbf{V}_T$,

Output: network topology of target city G

1: deploy internal vantage points $\mathbf{V}_V^I$

2: initial node set of X: $\mathbf{V}_X=\phi$

3: initial edge set of X: $\mathbf{E}_X=\phi$

4: **while** $\text{card}(\mathbf{V}_N)\neq 0$ **do**

5: obtain the result of internal network measurement: $\mathbf{P}_{\text{internal}-X}\leftarrow \text{Measure}(\mathbf{V}_V^I,\mathbf{V}_T)$

6: initial new node set: $\mathbf{V}_N=\phi$

7: **for** every path $p\in\mathbf{P}_{\text{internal}-X}$ **do**

8: **for** every node $v_i\in p$ **do**

9: **if** $v_i\notin\mathbf{V}_T$ **then**

10: add new node v_i to $\mathbf{V}_N$
11: update target IP set: add v_i to $\mathbf{V}_T$
12: **end if**
13 **end for**
14: **end for**
15: **end while**
16: build network topology of X : $G \leftarrow \text{BuildGraph}(\mathbf{V}_X, \mathbf{E}_X)$

3.2 Selection of Intersection Points between Bidirectional Paths Based on Location Recommendation

The identification of boundary nodes can be transformed into the division of internal and external paths between cities. In this section, the boundary nodes are approached by external measurement, and the candidate boundary nodes are mined by looking for the intersection point between the path of the external and the internal topology. This paper assumes that every node in the probing path from the vantage points inside the city to the target IP inside the city should be located in the city, and constructs the router-city mapping list. This list contains the internal topology information obtained by the above procedure to find the intersection points of the bidirectional paths by comparing the nodes in the list with the external probe results.

The process of this step is shown in **Fig. 4**.

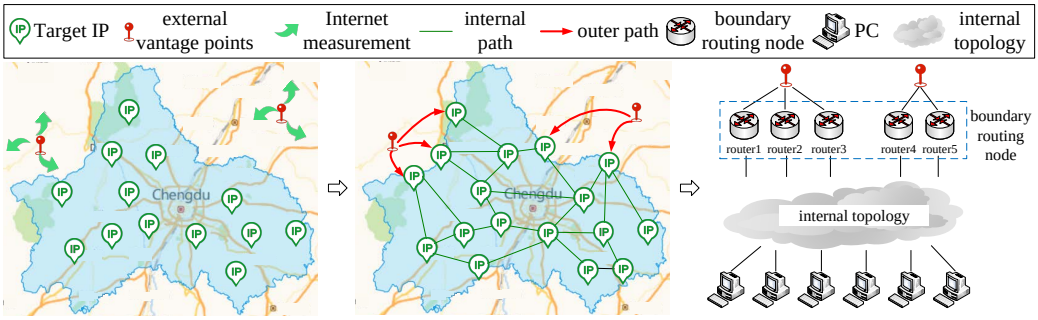

Fig. 4. Schematic diagram of intersection points selection

The routing-city mapping list M_{RL} is constructed using the internal routing nodes obtained in the above steps. n_o vantage points are deployed outside the target city to probe the target IP set $\mathbf{V}_T$. The route node $\mathbf{V}_{\text{external}-X} = \{v \in p \mid p \in \mathbf{P}_{\text{external}-X}\}$ is compared with the router-city mapping list M_{RL} based on the returned probe result $\mathbf{P}_{\text{external}-X}$, the candidate boundary routing nodes set $\mathbf{V}_C$ of the target city is obtained.

Assuming that the vantage point outside the target city is v_y and the target node in the target city X is v_x, the probing path can be expressed as:

$$p_{y \to x} = \{v_y, \text{K}, v_k, \text{K}\ v_x\} \tag{1}$$

Then start from v_y to traverse each node in $p_{y \to x}$, find the first node that meets the following conditions, and stop traversing:

$$v_k \in M_{\mathrm{RL}}(A) \tag{2}$$

v_k is the intersection point in the topology information obtained by bidirectional probing. The above process is performed for each result, and the nodes found are added to the candidate boundary routing node set $\mathbf{V}_C$.

The candidate boundary routing nodes are selected on the basis of the crossover characteristics of actual bidirectional measurement paths, and have low dependence on the accuracy of location data. In the absence of high-precision location data, the candidate boundary routing nodes can still be used to find the city boundary nodes.

The detailed candidate boundary node identification process is shown in Algorithm 2.

Algorithm 2: Algorithm of Selection of Intersection Points between Bidirectional Paths Based on Location Recommendation

Input: Target city X, target IP set $\mathbf{V}_T$, network topology of target city G

Output: candidate boundary node set $\mathbf{V}_C$

1: build the map between routing nodes and cities: $M_{\mathrm{RL}} \leftarrow \mathrm{Map}(G, X)$
2: deploy external vantage points $\mathbf{V}_V^O$
3: obtain the result of internet network measurement: $\mathbf{P}_{\mathrm{external}-X} \leftarrow \mathrm{Measure}(\mathbf{V}_V^O, \mathbf{V}_T)$
4: initial candidate boundary node set: $\mathbf{V}_C = \phi$
5: **for** every path $p \in \mathbf{P}_{\mathrm{external}-X}$ **do**
6: **for** every node $v_i \in p$ **do**
7: **if** $M_{\mathrm{RL}}(v_i) = X$ **then**
8: add v_i to $\mathbf{V}_C$
9: break
10: **end if**
11: **end for**
12: **end for**

3.3 Lightweight Boundary Node Verification Based on Location Database

The boundary node identification based on IP location methods need to locate the IP of each hop on the probing path; this kind of algorithm costs a lot. This section only verifies the front and back hop of the candidate boundary nodes, which can reduce the location scale. The verification method is shown in Fig. 5.

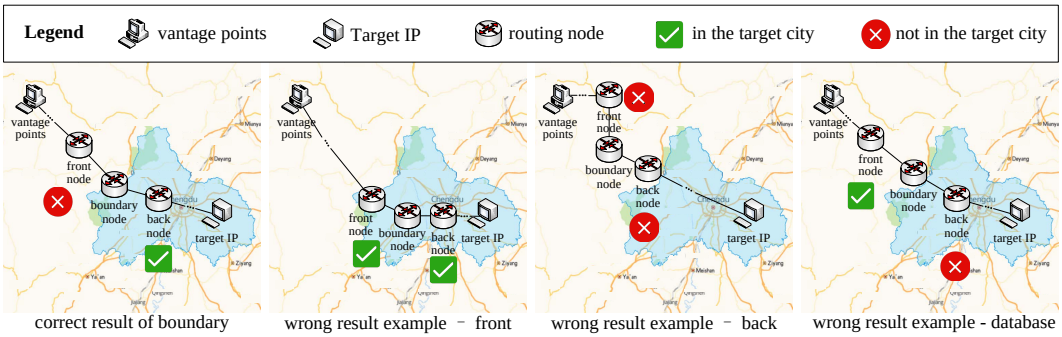

Fig. 5. Schematic diagram of boundary node verification

On the basis of obtaining the candidate boundary routing node $\mathbf{V}_C$, combined with the existing IP address database and location algorithm, the front and back hop's location of the candidate boundary routing node is verified.

For each candidate node $v_c \in \mathbf{V}_C$, the probe path in (1) can be expressed as follows:

$$\text{path}_{y\to x} = \left\{ v_y, \mathrm{K}\ , v_{c-1}, v_c, v_{c+1}, \mathrm{K}\ v_x \right\} \tag{3}$$

where, v_{c-1} is the hop before reaching the candidate boundary node v_c from the external vantage points, and v_{c+1} is the hop after reaching v_c. The verification conditions for candidate boundary nodes are as follows:

$$\text{result} = \left\{ \begin{array}{l} \text{TRUE}, \mid \mathrm{C}\left(v_{c-1}\right) \neq X \text{ and } \mathrm{C}\left(v_{c+1}\right) = X \\ \text{FALSE}, \mid \text{else} \end{array} \right\} \tag{4}$$

If the front hop v_{c-1} of the candidate node is located outside the target city and the back hop v_{c+1} is located inside the target city, the node v_c is determined to be a boundary routing node.

If the node does not meet the above conditions, it is judged to be a non-boundary routing node; at this condition, the path is traced back, each hop on the path is located. Judge the error of candidate nodes is due to the wrong of initial data or the wrong of mining method is determined. For the wrong initial data, the annotation is given in the original location database.

The detailed boundary node verification process is shown in Algorithm 3.

Algorithm 3: Algorithm of Lightweight Boundary Node Verification Based on Location Database

Input: Target city X, IP address database D, candidate boundary node set $\mathbf{V}_C$, result of external Internet measurement $\mathbf{P}_{\text{external}-X}$

Output: boundary node set $\mathbf{V}_B$

1: initial boundary node set $\mathbf{V}_B = \phi$

2: **for** every node $v_i \in \mathbf{V}_C$ **do**

3: find the front IP and back IP of v_i in corresponding path $p \in \mathbf{P}_{\text{external}-X}$:

$$\mathrm{IP}\left(v_i.\text{front\_IP}\right), \mathrm{IP}\left(v_i.\text{back\_IP}\right) \leftarrow \text{GetNode}\left(p\right)$$

4: **if** $M_{\mathrm{RL}}\left(v_i.\text{front\_IP}\right) \neq X$ and $M_{\mathrm{RL}}\left(v_i.\text{back\_IP}\right) = X$ **then**

5: add v_i to $\mathbf{V}_B$

6: **else then**

7: **for** every node $v_i \in p$ **do**

8: get the city of the node: $\mathrm{C}\left(v_i\right) \leftarrow \mathrm{Loc}\left(D, v_i\right)$

9: **end for**

10: annotation error information in D

11: **end if**

12: **end for**

4. Experiments

In order to verify the feasibility and effectiveness of the proposed algorithm, experiments on boundary identification are carried out in this section. It includes four parts: experimental setup, topology integrity analysis, boundary recognition rate analysis, identification precision analysis and algorithm performance analysis.

4.1 Experimental Setup

1) Datasets

This paper uses Scamper [19] developed by CAIDA for probing. The IP address blocks of the three target cities (Zhengzhou, Hangzhou and Chengdu) were selected from 6 IP address databases released in November 2019: IPIP[1], Whois[2], IPPlus[3], IP2location[4], Maxmind[5] and IPcn[6]. There were 6,174 IP blocks, including 12,748,117 IP addresses. Combined with the IP selection method adopted in this paper, the target IP set constructed for Zhengzhou, Hangzhou, and Chengdu contains 60,337 target IP addresses in total. The number of IP blocks, full IP and target IP of the three cities are shown in **Table 2**.

Table 2. Statistics of the number of IP addresses in the target city

Target City	# IP blocks	# Full IP	# Target IP
Zhengzhou	1,702	2,725,327	11,598
Hangzhou	2,221	7,501,838	30,694
Chengdu	2,251	4,2747,36	18,045

The probing period t_p is 2 hours, 12 rounds of probing are carried out every day, 360 rounds of probing are carried out, and a total of 65,163,960 results are obtained.

Due to the limitation of probing resources, the experimental data of New York, Dallas and Los Angeles used the measurement results provided by IPIP and CAIDA in 2020, containing a total of 2,609,529 results.

2) Evaluation Metrics

The effectiveness and feasibility of the algorithm are evaluated by using the recognition rate, precision and cost commonly used in previous researches.

- Recognition Rate

The recognition rate is the proportion of probing results that can find the boundary node among all results. The calculation formula is as follows:

$$R = \frac{\text{card}\left(\mathbf{P}_X^S\right)}{\text{card}\left(\mathbf{P}_X\right)} \tag{5}$$

[1] http://www.ipip.net/
[2] http://www.whois.com/
[3] https://www.ipplus360.com/
[4] http://www.ip2location.com/
[5] http://www.maxmind.com/
[6] http://www.ip.cn/

where, $\mathbf{P}_X$ is the set of all the paths from the external vantage points to the target city X, and $\mathbf{P}_X^S$ is the path set that can find the boundary node.

- Precision

Precision is the proportion of candidate nodes that meet the characteristics of boundary nodes, and the calculation formula is as follows:

$$\text{Precision} = \frac{\text{card}\left(\mathbf{P}_X^{\text{TRUE}}\right)}{\text{card}\left(\mathbf{P}_X^S\right)} \tag{6}$$

where, $\mathbf{P}_X^{\text{TRUE}}$ is the path set of the candidate nodes that meet the boundary node characteristics.

- Cost

The cost is the ratio of the number of IP addresses required to be located by the algorithm in this paper in the process of city boundary node identification to the number of IP addresses required to be located by the boundary node identification method relying on locating. The calculation formula is as follows:

$$\text{cost} = \frac{N_{\text{loc}}^{'}}{N_{\text{loc}}} \tag{7}$$

where, $N_{\text{loc}}^{'}$ is the number of IP required to be located by the algorithm in this paper, and N_{loc} is the number of IP required to be located by the boundary node identification methods based on the locating.

3) Experiment Settings

The relevant experiment settings in this paper is shown in **Table 3**.

Table 3. Experiment settings

Parameter	Setup
X	ZZ (Zhengzhou), HZ (Hangzhou), CD (Chengdu), LA (Los Angeles), NYC (New York), Denver, MIA (Miami), Dallas
D	Maxmind, IP2location, Whois, IPIP, IPPlus, IPcn
$\mathbf{V}_V$	211.149.219.168, 47.110.233.88, 122.114.14.202, 155.94.254.7, 104.219.168.124, 107.161.88.35, 155.94.222.154, 104.131.176.211
t_{P}	2 hours

In **Table 3**, X represents the target cities, D represents the IP database adopted, $\mathbf{V}_V$ represents vantage points, and t_{P} represents the probing cycle.

4) Baseline Methods

In this paper, the method proposed by Zhao et al. [10] and the method proposed by Liu et al. [12] is used as baseline methods.

4.2 Analysis of Topology Integrity

Taking the probing results in China as an example, topological integrity was analyzed. By probing the target IPes for 30 days (360 rounds in total), the new nodes in each round of data acquired by vantage points inside cities were statistically analyzed. The results are shown in **Fig. 6.**

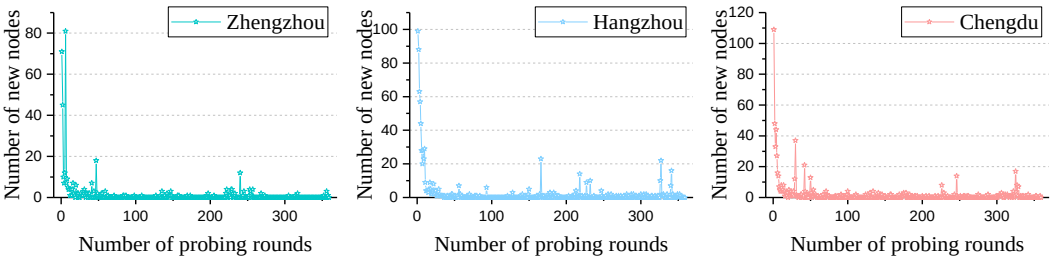

Fig. 6. Statistical of added new nodes each round

It can be seen that as the probing continues, the new nodes approaching zero, indicating that the obtained network tends to be complete. Meanwhile, the data obtained by the method in this paper is compared with CAIDA and IPIP, and the results are shown in **Table 4**.

Table 4. Comparison of measurement results between different data sets

City	Probing All IPes		Probing Target IPes		CAIDA		IPIP	
	$N_{/24}$	N_R	$N_{/24}$	N_R	$N_{/24}$	N_R	$N_{/24}$	N_R
Zhengzhou	5,519	9,264	5,466	8,809	173	556	3,633	4,187
Hangzhou	16,641	23,354	16,443	22,220	325	1,785	15,057	16,159
Chengdu	7,526	13,383	7,462	10,956	224	1,104	5,816	6,715
Total	29,686	46,001	29,371	41,985	722	3,445	24,506	27,061

In **Table 4**, $N_{/24}$ is the number of /24 prefix IP blocks covered by measurement results, and N_R is the number of routing nodes probed.

It can be seen from **Table 4**, the coverage of /24 prefixed IP blocks and nodes in the network using the method proposed in this paper reaches 98% and 91% of probing all IPes, respectively; while the data CAIDA provided can only cover less than 3% of the /24 prefix IP blocks and 8% of nodes; the coverage rate of IPIP's data reached 82%, 59% respectively. This result proves the rationality of probing method and IP sampling method of the algorithm proposed in this paper; it can obtain the basic topology of the target region to carry out the next step of analysis.

4.3 Analysis of Recognition Rate

The recognition rate of boundary identification obtained using the algorithm in this paper is calculated and compared with the algorithm in Ref. [10] and Ref. [12]. The results in three Chinese cities and three American cities are shown in **Fig. 7** and **Fig. 8**, respectively.

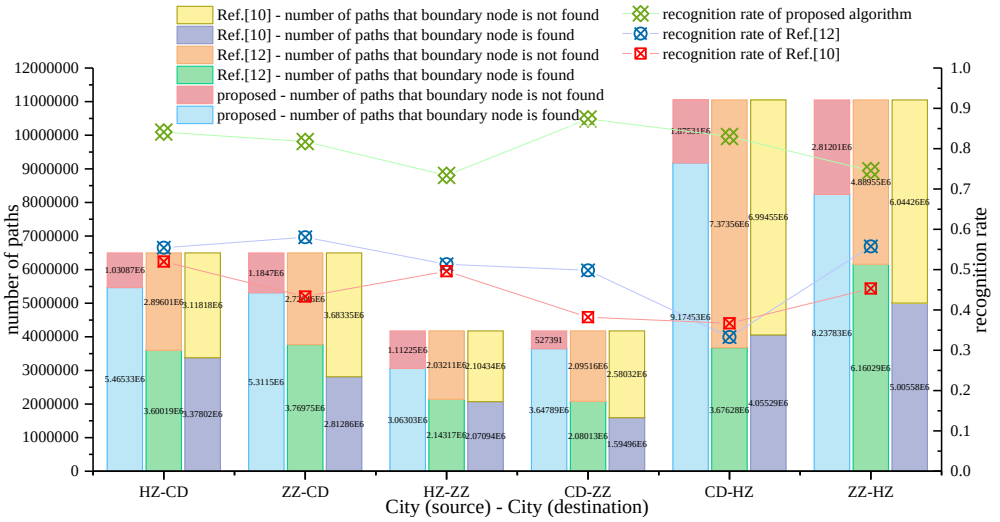

Fig. 7. Statistics of recognition rate (China)

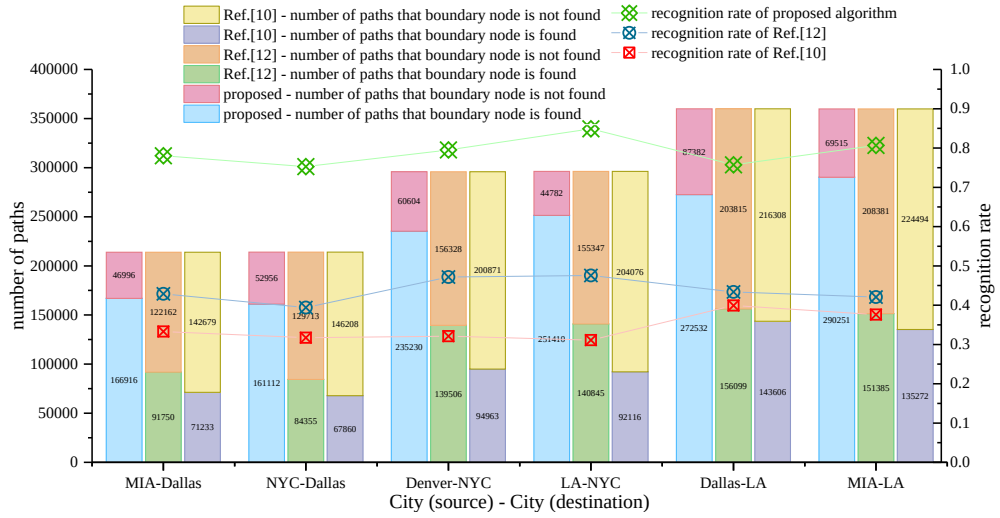

Fig. 8. Statistics of recognition rate (American)

In the figure, the abscissa represents the cities where the vantage points and target IPes are located; the left ordinate represents the number of probing results, and the right ordinate represents the proportion of paths that can find the boundary.

It can be seen from the figure that, in the results of six cities in China and the United States, the algorithm in this paper can find more boundaries than the algorithm in Ref. [10] and Ref. [12]. The specific data are shown in **Table 5**.

Table 5. Statistical results of recognition rate

City (VPs) - City (Target)	Ref. [10]	Ref. [12]	Proposed
Hangzhou - Chengdu	52.00%	55.42%	**84.13%**
Zhengzhou - Chengdu	43.30%	58.03%	**81.76%**
Hangzhou - Zhengzhou	49.60%	51.33%	**73.36%**
Chengdu - Zhengzhou	38.20%	49.82%	**87.37%**
Chengdu - Hangzhou	36.70%	33.27%	**83.03%**
Zhengzhou - Hangzhou	45.30%	55.75%	**74.55%**
China (average)	43.55%	49.33%	**80.34%**
Miami - Dallas	33.30%	42.89%	**78.03%**
New York - Dallas	31.70%	39.41%	**75.26%**
Denver - New York	32.10%	47.16%	**79.51%**
Los Angeles - New York	31.10%	47.55%	**84.88%**
Dallas - Los Angeles	39.90%	43.37%	**75.72%**
Miami - Los Angeles	37.60%	42.08%	**80.68%**
American (average)	34.78%	43.91%	**79.18%**

It can be seen from **Table 5** that the recognition rate of this algorithm in the detection results of six cities in China and the United States is higher than that of Ref. [10] and Ref. [12], and the average recognition rate of this algorithm in the two countries is 80.34% and 79.18% respectively. In contrast, only 43.55%, 34.78% and 49.33%, 43.91%, were found in Ref. [10] and Ref. [12]. This shows that the proposed algorithm can still identify the boundary without the support of high-precision location and landmark data, and can guarantee a high recognition rate.

4.4 Analysis of Identification Precision

The boundary nodes identification precision of the three methods is analyzed and compared, as shown in **Fig. 9** and **Fig. 10.**

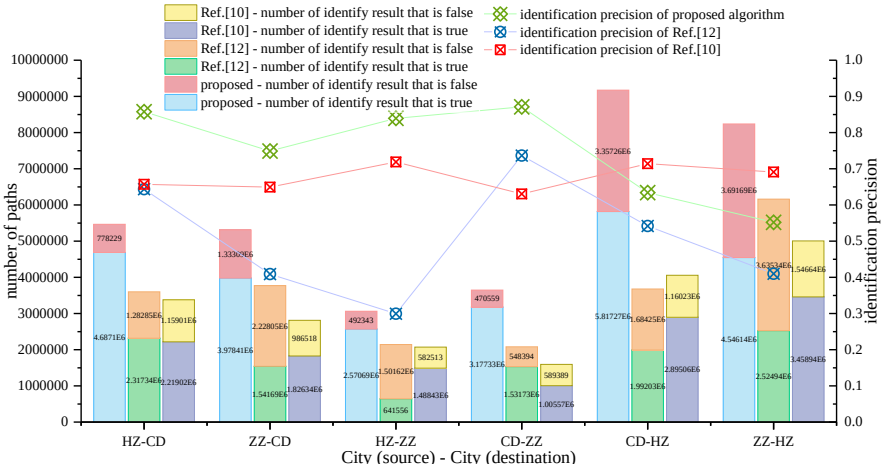

Fig. 9. Identification precision analysis (China)

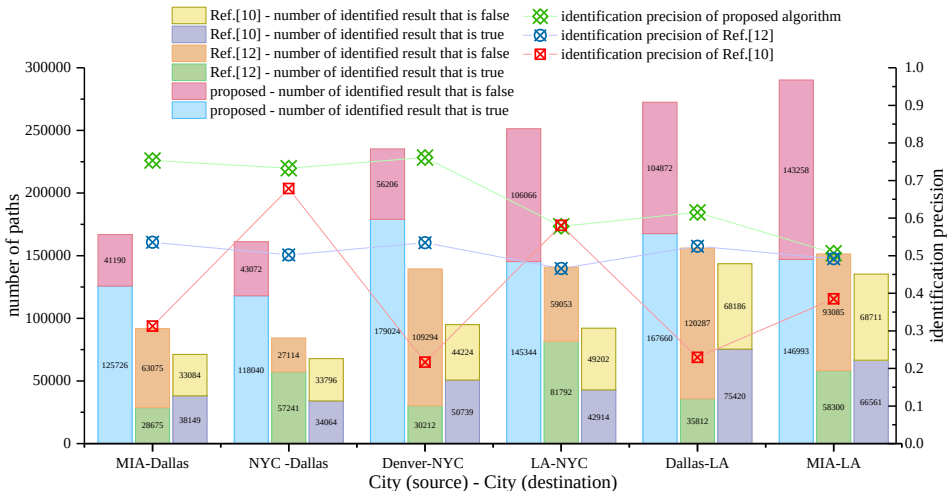

Fig. 10. Identification precision analysis (American)

As shown in the figure above, among 12 groups of experimental data in six cities, the algorithm in this paper performed better than the other two methods in 9 groups of experiments. The specific data are shown in **Table 6**.

Table 6. Statistical results of identification precision

City (VPs) - City (Target)	Ref. [10]	Ref. [12]	Proposed
Hangzhou - Chengdu	65.69%	64.37%	**85.76%**
Zhengzhou - Chengdu	64.93%	40.90%	**74.90%**
Hangzhou - Zhengzhou	71.87%	29.93%	**83.93%**
Chengdu - Zhengzhou	63.05%	73.64%	**87.10%**
Chengdu - Hangzhou	**71.39%**	54.19%	63.41%
Zhengzhou - Hangzhou	**69.10%**	40.99%	55.19%
China(average)	67.88%	49.23%	**70.99%**
Miami - Dallas	53.56%	31.25%	**75.32%**
New York - Dallas	50.20%	67.86%	**73.27%**
Denver - New York	53.43%	21.66%	**76.11%**
Los Angeles - New York	46.59%	**58.07%**	57.81%
Dallas - Los Angeles	52.52%	22.94%	**61.52%**
Miami - Los Angeles	49.21%	38.51%	**50.64%**
American (average)	50.88%	38.23%	**64.09%**

It can be seen from **Table 6** that the average precision of the algorithm in this paper in the two countries is 70.99% and 64.09%, respectively, while that Ref. [10] and Ref. [12] is only 67.88%, 50.88% and 49.23%, 38.23%. This indicates that the proposed algorithm can obtain more accurate results than the existing algorithms on the premise of guaranteeing a higher recognition rate.

4.5 Analysis of Algorithm Performance

The following is an example based on location algorithm SLG and probing results of three cities in China for performance analysis. The path lengths from vantage points in three cities to three different cities are shown in **Fig. 11**.

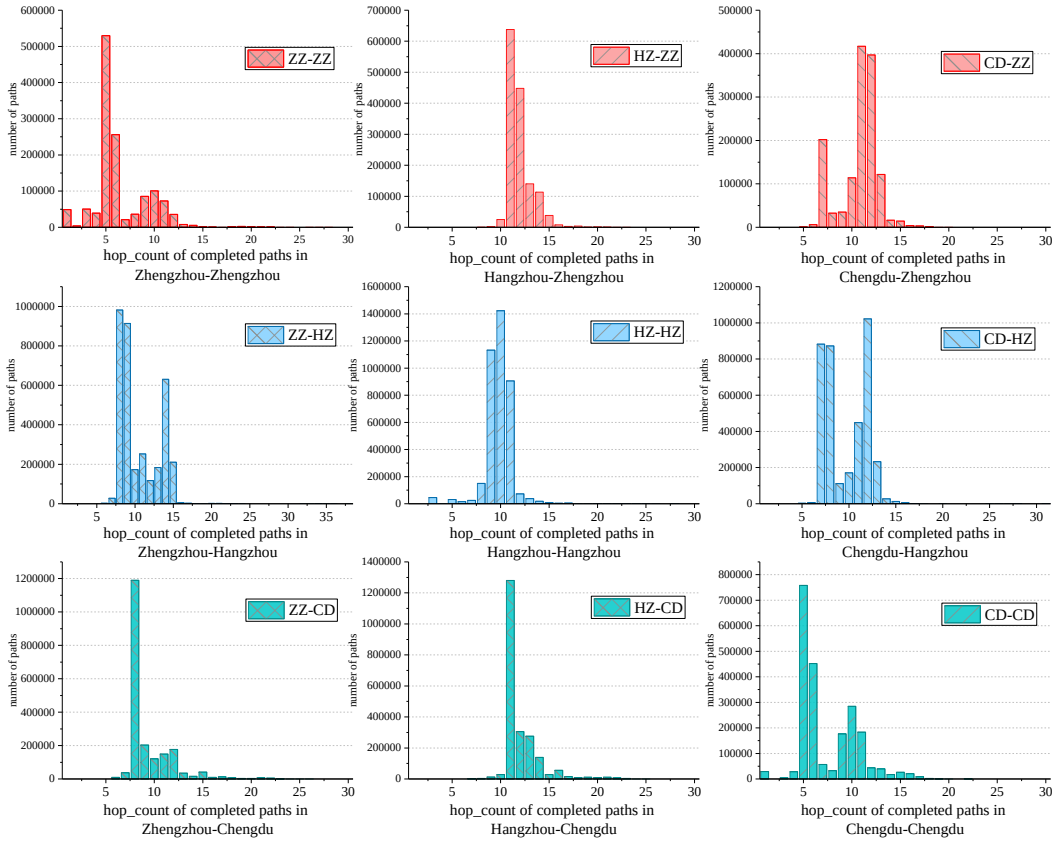

Fig. 11. Statistics of path length

As shown in **Fig. 11**, the average length of paths to Zhengzhou, Hangzhou and Chengdu is 9.78, 10.01 and 9.57, respectively; If the SLG location algorithm is used for boundary identification, each hop in the path needs to be located, while the algorithm proposed only needs to locate one hop before and one hop after the candidate nodes. The location resources consumed in the three cities were only 20.45%, 19.98% and 20.89% of SLG, with an average of 20.33%. It greatly reduces resource consumption and improves the efficiency of boundary node identification.

5. Conclusion

Considering the existing algorithm is vulnerable to time-delay, location accuracy, and has high consumption, this paper proposes a city-level boundary node identification algorithm based on bidirectional approaching. The proposed algorithm can accurately identify the boundary nodes on the basis of consuming only 20.33% of the location resources of existing algorithms. Compared with the existing algorithms, the proposed algorithm can identify city-level boundary nodes in the absence of high-precision location databases, and the recognition rate and precision can reach more than 80.29% and 70.73%, respectively. At the same time, the algorithm in this paper can also annotate some errors in the existing IP address database. Of course, the algorithm still has some shortcomings; the precision and recognition rate of the algorithm will still be disturbed by the accuracy of the initial IP blocks. In future work, we

will study how to further improve the precision and recognition rate of the proposed algorithm and how to evaluate the accuracy of candidate boundary nodes without locating information.

References

[1] G. Ciavarrini, M. S. Greco, and A. Vecchio, "Geolocation of internet hosts: accuracy limits through Cramér-Rao lower bound," *Computer Networks*, vol. 135, pp. 70-80, Apr, 2018. Article (CrossRef Link)

[2] J. Chen, Y. Luo and R. Du, "The impact of privacy seal on users' perception in network transactions," *Computer Systems Science and Engineering*, vol. 35, no.3, pp. 199-206, May, 2020. Article (CrossRef Link)

[3] S. Kaur and V. K. Joshi, "Hybrid soft computing technique based trust evaluation protocol for wireless sensor networks," *Intelligent Automation & Soft Computing*, vol. 26, no.2, pp. 217-226, Jan, 2020. Article (CrossRef Link)

[4] G. Swathi, "A frame work for categorise the innumerable vulnerable nodes in mobile adhoc network," *Computer Systems Science and Engineering*, vol. 35, no.5, pp. 335-345, Jan, 2020. Article (CrossRef Link)

[5] L. Matthew, D. Amogh, H. Bradley, C. David, and C. Kc, "Bdrmap: inference of borders between IP networks," in *Proc. of Internet Measurement Conference*, Santa Monica, CA, USA, pp.381-396, 2016. Article (CrossRef Link)

[6] M. Luckie, B. Huffaker, A. Dhamdhere, V. Giotsas, and K. Claffy, "AS relationships, customer cones, and validation," in *Proc. of the ACM SIGCOMM Internet Measurement Conference*, Barcelona, Spain, pp. 243-256, 2013. Article (CrossRef Link)

[7] A. Marder, and J. M. Smith, "MAP-IT: multipass accurate passive inferences from traceroute," in *Proc. of the ACM SIGCOMM Internet Measurement Conference*, Santa Monica, CA, USA, pp. 397-411, 2016. Article (CrossRef Link)

[8] V. Giotsas, G. Smaragdakis, B. Huffaker, M. J. Luckie, and K. C. Claffy, "Mapping peering interconnections to a facility," in *Proc. of the ACM Conference on Emerging Networking Experiments and Technologies*, Heidelberg, Germany, pp. 1-13, 2015. Article (CrossRef Link)

[9] A. Marder, M. Luckie A. Dhamdhere, B. Huffaker, and J. M. Smith, "Pushing the Boundaries with bdrmapIT: Mapping router ownership at Internet scale," in *Proc. of the ACM SIGCOMM Internet Measurement Conference*, Boston, MA, USA, pp. 56-69, 2018. Article (CrossRef Link)

[10] S. Q. Liu, F. L. Liu, F. Zhao, L. X. Chai, and X. Y. Luo, "IP city-level geolocation based on the pop-level network topology analysis," in *Proc. of International Conference on Information Communication and Management*, Hatfield, UK, pp. 109-114, 2016. Article (CrossRef Link)

[11] F. X. Yuan, F. L. Liu, R. Xu, Y. Liu, and X. Y. Luo, "Network topology boundary routing IP identification for IP geolocation," in *Proc. of International Conference on Artificial Intelligence and Security*, Hohhot, China, pp. 534-544, 2020. Article (CrossRef Link)

[12] F. Zhao, X. Y. Luo, Y. Gan, X. D. Zu, J. N. Chen, and F. L. Liu, "IP geolocation based on identification routers and local delay distribution similarity," *Concurrency Computation*, vol. 31, no. 22, pp. 1-15, Nov., 2018. Article (CrossRef Link)

[13] Y. W, D. Burgener, F. Marcel, K. Aleksandar; and C. Huang, "Towards street-level client-independent IP geolocation," in *Proc. of USENIX Symposium on Networked Systems Design and Implementation*, Boston, MA, USA, pp. 365-379, 2011. Article (CrossRef Link)

[14] J. N. Chen, F. L. Liu, Y. F. Shi, and X. Y. Luo, "Towards IP location estimation using the nearest common router," *Journal of Internet Technology*, vol. 19, no. 7, pp. 2097-2110, 2018. Article (CrossRef Link)

[15] F. Zhao, R. Xu, R. X. Li, M. Zhu, and X. Y. Luo, "Street-level geolocation based on router multilevel partitioning," *IEEE Access*, vol. 7, pp. 59237-59248, 2019. Article (CrossRef Link)

[16] J. P. Liu, X. C. Kang, C. Dong, and F. H. Zhang, "Simulation of real-time path planning for large-scale transportation network using parallel computation," *Intelligent Automation & Soft Computing*, vol. 25, no.1, pp. 65-77, Jan., 2019. Article (CrossRef Link)

[17] B. Donnet, P. Raoult, T. Friedman, and M. Crovella, "Deployment of an algorithm for large-scale topology discovery," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 12, pp. 2210-2220, Dec, 2006. Article (CrossRef Link)
[18] Y. Tian, R. Dey, Y. Liu, and K. W. Ross, "Topology mapping and geolocating for China's Internet," *IEEE Transactions on Parallel and Distributed Systems*, vol. 24, pp. 1908-1917, Sept. 2013. Article (CrossRef Link)
[19] M. Luckie, "Scamper: a scalable and extensible packet prober for active measurement of the internet," in *Proc. of ACM SIGCOMM conference on Internet measurement*, Melbourne, Australia, pp. 239-245, 2010. Article (CrossRef Link)

Zhiyuan Tao is a postgraduate at the State Key Laboratory of Mathematical Engineering and Advanced Computing. His research interests include network security, network topology and network data analysis.

Fenlin Liu is currently a Professor with the Zhengzhou Science and Technology Institute. He has authored or co-authored more than 90 refereed international journal and conference papers. His research interests include network topology and network geolocation.

Yan Liu is currently an Associate Professor. She has authored or co-authored more than 50 refereed international journal and conference papers. Her research interests include network topology discovery and network data analysis.

Xiangyang Luo is currently a Professor at Zhengzhou Science and Technology Institute. His research interests lie in multimedia security and cyberspace surveying and mapping. He is the author or co-author of more than 100 refereed international journal and conference papers.