Jiang Q, Hu HY, Hu GM. Two-type information fusion based IP-TO-AS mapping table refining. JOURNAL OF COM-PUTER SCIENCE AND TECHNOLOGY 32(3): 571–584 May 2017. DOI 10.1007/s11390-017-1744-9

Two-Type Information Fusion Based IP-to-AS Mapping Table Refining

Qing Jiang, Hang-Yu Hu, Student Member, IEEE, and Guang-Min Hu*, Member, IEEE

School of Communication and Information Engineering, University of Electronic Science and Technology of China Chengdu 611731, China

E-mail: jq@std.uestc.edu.cn; huhangyuuestc@gmail.com; hgm@uestc.edu.cn

Received May 11, 2016; revised February 18, 2017.

Abstract The Internet topology at the autonomous system (AS) level is of great importance, and traceroute has been known to be a potential tool to obtain a complete AS topology. The original IP-to-AS mapping table maps the IP addresses in traceroute paths to their origin ASes, which may cause false AS links. The existing methods refine the original mapping table based on traceroute-BGP path pairs or alias resolution data. However, the information extracted from either of them is inaccurate and incomplete. In this paper, we present a two-type information fusion based method to refine the original mapping table. We extract four kinds of information from path pair and alias resolution data. Based on these information, we build a candidate AS set for each router. Then we choose the AS that is consistent with the existing information to be the owner AS of each router and map all of the IP addresses on the router to it. We validate the result with the ground truth from PeeringDB and Looking Glass severs. Compared with the existing methods, our method produces a more accurate mapping table. In addition, we discuss the coverage of our method and show that our method is convergent and more robust against the reduction of information or the increase of incorrect information.

Keywords network topology, Internet, routers, BGP, traceroute

1 Introduction

As the Internet technologies evolve to be mature, precisely understanding the topology of the Internet has been playing an increasingly important role in networking research community. Autonomous system (AS) topology, which is the top-level logical topology of the Internet, has drawn attention in the past decade. Network researchers have used the AS topology to understand the characteristics of the Internet topology and model it^[1], study the performance of routing protocol^[2], and infer AS relationships^[3], among others. A complete AS topology is needed for all of the studies. The existing methods for obtaining the AS topology can be divided into two main types: Border Gateway Protocol (BGP)-based passive measurement and traceroute-based active measurement. The former constructs the AS topology on the basis of the ASpath information extracted from BGP routing tables

and update messages. However, the AS topology obtained from BGP misses a large number of links, especially the peer links between lower-tier ASes, because of BGP export policies, route aggregation and best path selection. Considering these shortcomings, traceroutebased active measurement has a great advantage. The low cost of traceroute monitor allows the availability of more monitors in lower-tier ASes, thereby allowing traceroute to see many peer links that are not in the BGP-driven AS topology. Therefore, traceroute has a high potential to obtain a more complete AS topology.

Obviously, the key step of constructing the AS topology from traceroute paths is mapping each IP address in traceroute paths to the right AS, i.e., the owner AS of the router the IP address is on. The widely used IP-to-AS mapping method is looking up the original IP-to-AS mapping table extracted from BGP routing tables, in which each IP address is mapped to the origin AS of its longest matching prefix. However, in ac-

Regular Paper

This work was partially supported by the National Natural Science Foundation for Distinguish Young Scholars of China under Grant No. 61301274 and the National Natural Science Foundation of China under Grant No. 61471101.

^{*}Corresponding Author

 $[\]textcircled{O}2017$ Springer Science + Business Media, LLC & Science Press, China

tual network deployment, the origin AS of an IP address may not be the owner AS of the router that the IP address belongs to. Thus, using the original IP-to-AS mapping table directly may cause false AS links. An example is given and the topology of five ASes is shown in Fig.1. The letter on the link represents the origin AS of the prefix that the link uses. The prefix used in Internet Exchange Point (IXP) is announced by AS D. A traceroute path from the monitor in AS A to the destination in AS E is (IP1 \sim IP5), and IPx is on the interface of router Rx. Then we map IP addresses to their origin ASes to obtain the traceroute-AS path (ACDE), thereby discovering three AS links: $(A \sim C)$, $(C \sim D)$ and $(D \sim E)$. Compared with the actual topology, $(A \sim C)$ and $(D \sim E)$ are false links that do not exist because of the incorrect mapping of IP2 and IP4. Thus, we must refine the original IP-to-AS mapping table to map IP addresses to the owner ASes of their routers.



Fig.1. Example of the mistakes caused by the original mapping table.

Two methods are proposed to refine the original mapping table: path-pair-matching based method^[4-7] and alias-resolution based method^[8-9]. However, both path pair and alias resolution data are inaccurate, which means some information extracted from them is incorrect or deficient. For example, some path pairs are inconsistent and their matching connects certain IP addresses to the false ASes, and false negatives and false positives of the alias resolution data also influence the identification result directly. Moreover, as one type of data can provide limited information only, more than one AS may have the same probability occasionally, and the AS to which the IP address should be mapped cannot be determined.

In this paper, we attempt to fuse the information from traceroute-BGP path pair and alias resolution data to refine the original IP-to-AS mapping table with router granularity. Fusing different types of information can reduce the problem of information missing and inaccuracy in using only one type of information. There are two steps in our method. Firstly, we obtain the candidate ASes as many as possible, and then we validate the correctness of each candidate AS. In every step, we need the fusion of the two types of data, because it is possible that only one type of data can add the owner AS into the candidate set, and only another type of data has the capacity to identify it as the owner AS. In addition, the router granularity means that all the addresses on a router will be mapped to the same AS, which is in accordance with the goal of the IP-to-AS mapping table, i.e., mapping the IP address to the owner AS of its router. Compared with the existing methods, we can always produce a more accurate mapping table within a limited time.

The main contributions of this paper are summarized as follows.

• We make a detailed analysis of the shortcoming hidden in the information from traceroute-BGP path pair and alias resolution data. Then we define four kinds of information which can fully exploit the existing data.

• We propose a simple effective way of fusing the information from different types of data and design a method to refine the original mapping table with router granularity.

• We validate the accuracy of our method with two kinds of ground truth data. In addition, we design comparative experiments to show its robustness against the incomplete data and the data with a higher error rate. Moreover, we also analyze the coverage and the convergency of our method.

This paper is organized in the following manner. In Section 2, We describe briefly three kinds of methods which are used in refining the original IP-to-AS mapping table. In Section 3, we explain the four kinds of information extracted from the two types of data. In Section 4, we describe the steps of our method in detail. In Section 5, we introduce the datasets we used. In Section 6, we make experiments to evaluate our method and analyze the result, and then this paper is concluded in Section 7.

2 Related Work

The IP-to-AS mapping table is an essential component in all kinds of work that need extract AS-level information from IP addresses. Some of them use the AS-level information directly, such as [10] that utilizes traceroute to discover the existence of members and peering links in IXPs, and $Ark^{(1)}$ or $DIMES^{[11]}$ project that obtains the AS topology from traceroute paths. Some use the AS-level information indirectly, like [12] that focuses on mapping peering interconnections to the facilities where they occur, which need to map the IP addresses in interconnections to ASes to get the possible facilities. All these kinds of work require knowing the owner ASes of the IP addresses, i.e., the owner ASes of the routers that the IP addresses belong to. However, the original IP-to-AS mapping table is error-prone, and thus the result obtained by using it is inaccurate. There are three strategies used to solve this task.

The first method is based on the matching of the traceroute-BGP path pairs^[4-7]. It assumes that the AS routing path and the AS forwarding path are usually matched, and the only goal of refining the original mapping table is to maximize the number of matched path pairs. Mao *et al.*^[4-5] used the information from the</sup>optimal matching to modify the original mapping table with prefix granularity, which is not exact because IP addresses in the same prefix may belong to routers in different ASes. Thus, Zhang et al.^[6-7] proposed the IP granularity method (IGM), in which IP addresses in the same prefix can be mapped to different ASes. IGM can reach the highest accuracy in the current studies, but its coverage is limited because it cannot modify the mapping of the IP addresses that do not appear in the path pairs. In addition, [13] develops a systematic framework to quantify the potential errors of traceroute in AS topology inference.

The second method is based on alias resolution data^[8-9]. It identifies the owner AS of each router from the origin ASes of its IP interfaces, and then maps all of the IP addresses on the router to the owner AS. Huffaker *et al.*^[8] proposed five heuristics to choose the owner AS of the router. However, the false negatives of the alias resolution data may invalidate the method if all the IP interfaces that belong to the owner AS are

not discovered, and false positives may cause IP addresses on the routers of different ASes to appear as though they belong to the same router, thereby some IP addresses are mapped incorrectly no matter which AS is chosen. Pansiot *et al.*^[9] identified the owner AS in accordance with five rules. They used the dataset collected with mrinfo, a multicast tool that can discover all interfaces of a router. Nevertheless, only IPv4 multicast enabled routers reply to mrinfo.

The last method^[14] does not modify the original mapping table. It detects the real AS link directly from the traceroute-AS paths. First, it converts traceroute paths into AS-level paths based on the original mapping table. Afterwards, it uses IXP prefixes and AS sibling information to preprocess traceroute-AS paths and filter false links based on some heuristics. But, strict deletion rules may cause the method to obtain a result with some false negatives.

3 Two Types of Information

Due to the complexity and the noncooperativity of the network, the data we can obtain from it is incomplete and inaccurate. Therefore, it is impossible to extract the information that is absolutely right from the data. A way to solve the problem is extracting the useful information as much as possible and designing an information fusion method. In this section, we review the optimal matching of the pair path and introduce the four kinds of information that we extract from the path pair and alias resolution data.

3.1 Optimal Matching

The optimal matching of a path pair is proposed in [5], and we describe it here. MT is an IP-to-AS mapping table which includes a set P of IP prefixes, and each prefix P_x in P can be mapped to a nonempty set $MT(P_x)$ of ASes in MT. Every IP address is assumed to be able to find its longest matching prefix in P. To find the mapping AS set of an IP address I, firstly, we find the longest matching prefix P(I) of I in P, and then we map P(I) to its corresponding AS set MT(P(I)) in the mapping table MT. Therefore, MT(P(I)) is the mapping AS set of the IP address I. To simplify the notation in our paper, we use MT(I) instead of MT(P(I)) to represent the mapping AS set of IP address I. In most cases, the size of the set is 1. Given a path pair (p, q), in

⁽¹⁾http://www.caida.org/projects/ark, Feb. 2017.

which $p = (p_1, p_2, ..., p_n)$ is the traceroute path and $q = (q_1, q_2, ..., q_m)$ is the BGP-AS path, a matching of (p, q) is a function $a : \{1, 2, ..., n\} \rightarrow \{1, 2, ..., m\}$ and $a(i) \leq a(i+1)$. Every IP address p_n can be mapped to an AS $q_{a(n)}$ in a. The error of the matching a for (p, q) is

$$E_{MT}(a, p, q) = \left| \left\{ i \leqslant n : q_{a(i)} \notin MT(p_i) \right\} \right| + a(n) - \left| \{a(i) : 1 \leqslant i \leqslant n \} \right|.$$

The first part is the number of IP addresses that are not matched to the right AS, and the second part is the number of ASes that precede the last matched AS and are not matched to any IP address. Optimal matching of (p,q) is the mapping that minimizes $E_{MT}(a, p, q)$. If more than one matching satisfies the condition, we choose the one that maximizes $\sum_{i=1}^{n} a(i)$. The error of a path pair is the error of its optimal mapping. [5] proposes a dynamic programming algorithm for obtaining the optimal matching of a path pair and we use it in our method, since it is very effective in computing the optimal matching of a path pair. Given a path pair (p,q), it can find the optimal matching of the path pair in $O(mn^2)$ in which n is the number of the IP addresses in the traceroute path and m is the number of the ASes in the BGP-AS path.

For example, as shown in Fig.2, there is a path pair (p, q), in which p = (IP1, IP2, IP3, IP4) and q = (AS A, AS B, AS C, AS E, AS F). Thus, n = 4 and m = 5. At first, we map each IP address in the traceroute path to its mapping AS in the origin mapping table to get the traceroute-AS path: (AS A, AS C, AS C, AS D). Then we find the optimal matching of the path pair and the result is $\{a(1) = 1, a(2) = 3, a(3) = 3, a(4) = 4\}$ as shown in Fig.2. Therefore, the error of the matching a for (p,q) is 2, since AS C is not matched to any IP address and IP4 is matched to AS E that is not the mapping AS (AS D) of it. There are no any matchings that can make the error of the path pair less than 2.



Fig.2. Illustration of the optimal matching.

Same with [5], we think path pairs that have more than two errors are not good. They may be inconsistent path pairs caused by BGP multi-hop session^[15], abnormal routing^[13], etc. Of course some inconsistent path pairs have fewer than or equal to two errors; therefore they are unidentified and unavoidable. We will give an actual example of them in Subsection 5.2.

3.2 Information from Path Pair Data

As stated in the preceding paragraphs, the information from the optimal matching of path pairs is not absolutely right. One reason is the existence of inconsistent path pairs. Another reason is that optimal matching may provide incorrect information even in consistent path pairs. For example, Fig.3 shows the optimal matching of a path pair under the original mapping table. We can see that IP2 is matched to AS A instead of the owner AS of its router, AS B. Thus, the information of the optimal matching of a path pairs under the original mapping table is enough. In addition, in our method, AS B may be added in the candidate AS set of the router by other information (e.g., the origin ASes of the other IP addresses on the router). Then in order to be correctly identified as the owner AS, AS Bneeds to be recognized by the path pair, which means IP2 should be matched to AS B in the path pair in some condition.



Fig.3. Inaccuracy of the optimal matching.

To fully use the path pair data, we divide the information that the path pair data gives each IP address into two types.

Optimal Matching AS (OMAS). The OMAS of an IP address is its matching AS in the optimal matching under the current mapping table. Certainly, an IP address may have different OMASes in different path pairs. OMAS is the AS that the path pair thinks the router should belong to.

Feasible Matching AS (FMAS). We change the mapping AS of an IP address into AS A in the current mapping table. Then (1) is used to compute the

optimal matching of a path pair that has the IP address. If the IP address is matched to AS A in the optimal matching, we consider AS A as an FMAS of the IP address in the path pair. For example, in Fig.3, AS B is an FMAS of IP2 in the path pair.

$$E_{MT}(a, p, q) = 2 \left| \left\{ i \leqslant n : q_{a(i)} \notin MT(p_i) \right\} \right| + a(n) - \left| \left\{ a(i) : 1 \leqslant i \leqslant n \right\} \right|.$$
(1)

FMAS is the AS that an IP address can be matched in a path pair and (1) is used to identify the FMAS. Although the owner AS of an IP address may not be the OMAS of the IP address, it should be the FMAS of the IP address. In order to do this, when we change the corresponding AS of an IP address in the mapping table to its owner AS and use (2) to compute the optimal matching, we expect that this IP address can be matched to its owner AS (i.e., the owner AS can be identified as the FMAS). Therefore, we need increase the penalty of false matching. For example, as shown in Fig.4, the owner AS of R2 is AS D, but the OMAS of IP2 is AS C. When we change the corresponding AS of IP2 in the mapping table to AS D, and the penalty of false matching is higher than or equal to 2, then IP2 will be matched to AS D, i.e., AS D is an FMAS of IP2, which is the correct information that we extract from the path pair (if the penalty of false matching is 1, IP2 will be matched to AS B).

R1	R2
Traceroute: IP1	IP2
Traceroute-AS Path: (AS A)	(AS C)
BGP-AS Path: (AS A)	(AS B) (AS C) (AS D) (AS E)
	↓
Traceroute: IP1	IP2
Traceroute-AS Path: (AS A)	(AS D)
BGP-AS Path: (AS A)	(AS B) (AS C) (AS D) (AS E)

Fig.4. Illustration of the feasible matching AS.

However, if the penalty is higher than 2, for example, if it is 3, AS E will also become an FMAS of IP2. However, as discussed in Subsection 3.1, we think path pairs that have more than two errors are not good. If IP2 is matched to AS E, this path pair will have three errors (three ASes have no matching), which means this path pair may be inconsistent; thus the information (AS E is an FMAS of IP2) from the path pair is unreliable. In other words, if the cost of being FMAS is that the number of errors of the path pair is more than 2, the correctness of the information is doubtful. Therefore, setting the penalty of false matching to 2 can avoid introducing errors as far as possible.

3.3 Information from Alias Resolution Data

Alias resolution data lists the IP addresses on the interfaces of each router. The information extracted directly from alias resolution data is the origin ASes of these IP addresses. Obviously, the accuracy of this information is based on the accuracy of the alias resolution data. The alias resolution technique can now be classified into the measurement-based method (e.g., $MIDAR^{[17]}$ and iffinder^[18]) and the analysisbased method (e.g., kapar^[19]). The measurement-based method can obtain a result with high accuracy, but it includes many false negatives. The two methods can be used together to reduce the false negatives but at the same time increase the false positives. False negatives miss some IP interfaces, i.e., lose some information, which can be supplemented by extracting other information. However, the effect of the false positives is more disadvantageous than that of the false negatives. If IP addresses that belong to different ASes are identified to the same router, some of them will be mapped to the incorrect AS in any case. Thus we use the alias resolution data derived by the measurement-based method, i.e., MIDAR and iffinder, which is different from [8] that uses the result obtained through using the two methods together.

To make up for the lost information in false negatives, we define a new kind of router interface: AS interface. AS interface is an interface on a router and we do not know its IP address, but we can obtain the origin AS of its IP address. Fig.5 shows that a traceroute path passes through R1 to R2 and IP2 is from a prefix that belongs to AS B. In practice, the IP addresses on the same link are always from the same prefix. R1 should have an interface whose IP address is also from AS *B*. If we do not find any interface whose origin AS is AS B on R1, we will have a reason to believe that there is an interface lost on R1. This interface is called AS interface. In this example, we can know that R1 has an AS interface whose origin AS is AS B. Note that, by this definition, a router can have only one AS interface from the same AS.

There is a potential problem with this definition. The inference of the AS interface may be wrong when 1) the next router replies to the traceroute probe with the IP address on the outgoing interface used to send the response packet, 2) the outgoing interface is different from the incoming interface, and 3) the origin ASes of the two interfaces are different. However, we have two reasons to ignore the problem here. First, many researchers have conducted experiments and proved that the probability of this case is small. [16] concludes that most addresses in traceroute paths are configured on the in-bound interface of routers, and in the test of [4], most of the popular commercial routers use the address of the incoming interface as the source address of the response packet. Second, it is also possible that the previous router indeed uses the IP addresses from the origin AS of the outgoing interface of the next router. In this case, our inference is still right.



Fig.5. Illustration of an AS interface

4 Refining IP-to-AS Mapping Table

The method for refining the original IP-to-AS mapping table is described in this section. The most important part of our method is identifying the owner AS of each router. Two key steps are included: 1) the construction of the candidate AS set and 2) the score calculation of each candidate AS. We give the definition of candidate AS and the score of it in Subsection 4.1 and Subsection 4.2 respectively. The entire process of refining the original mapping table will be introduced in Subsection 4.3.

4.1 Candidate AS Set

The candidate AS set of a router includes all the ASes that are potentially the owner AS of the router. The candidate AS set is divided into two parts: the origin AS set and the optimal matching AS set (OMAS set), which are the information extracted from alias resolution data and path pairs data, respectively.

Origin AS Set. The origin ASes of all the IP interfaces and AS interfaces of a router comprise its origin AS set. A router must use at least one IP address announced by its owner AS; therefore all the origin ASes of the router have the potential to be its owner AS.

OMAS Set. The optimal matching AS set of a router is composed of the optimal matching ASes of all the IP addresses on the router. As a result of the limit of alias resolution data, the interfaces that use the IP addresses from the owner AS may not be discovered. Therefore the candidate AS needs to be collected from a new view. For a path pair that passes through the router, OMAS is the AS that the path pair thinks the router should belong to.

4.2 Score of Candidate ASes

We calculate the score of each candidate AS and choose the highest one to be the owner AS. The score represents the possibility of being the owner AS. The owner AS should satisfy two conditions to the greatest extent: 1) the owner AS should account for the largest proportion in origin AS set of the router, and 2) it is the feasible matching AS in all the path pairs that pass through the router. We assume that AS1 is one of the candidate ASes of a router. According to these conditions, the score of AS1 is defined as

$$Score_{AS1} = \frac{\sum_{i=1}^{n} x_i}{n} + \frac{\sum_{j=1}^{m} y_j}{m},$$
 (2)

in which n is the number of interfaces (i.e., the IP and AS interfaces) on the router, m is the number of path pairs that pass through the router, and x_i and y_i are both the Boolean variables. If the origin AS of the *i*-th interface is AS1, then $x_i = 1$; otherwise, $x_i = 0$. In addition, $y_i = 1$ if AS1 is an FMAS of the router in the *j*-th path pair; otherwise, $y_i = 0$. If the *i*-th interface has more than one origin AS and AS1 is one of them, then $x_i = 1$.

The two parts of (2) represent the approval of the alias resolution data and the path pair data on the candidate AS respectively. Since [6] and [8] use separately one type of data to refine the mapping table and obtain a relatively satisfactory result, here we consider that the two types of data have the same credibility, and thereby their weights are both 1.

4.3 Process of Refining Mapping Table

The entire process of refining the mapping table is an iterative process. Four steps are repeated in each iteration: 1) getting the optimal matching ASes of each IP address; 2) obtaining the candidate AS set of each router; 3) identifying the owner AS of each router and mapping all of its IP addresses to the owner AS; 4) verifying whether the mapping table is changed.

In order to describe our method in detail, we give the pseudo-code in Algorithm 1. We construct two dictionary trees, Trie1 and Trie2, to store the original IP-to-AS mapping table (OMT). The refining will be performed on Trie2, whereas Trie1 will not be changed. Then we filter the invalid path pairs. We define a path pair is invalid if the traceroute path in it has an IP loop, or an anonymous router, or a private IP address, or an IP address without the origin AS, or the BGP-AS path in it has an As-set object or a private AS. Afterwards, the iteration is initiated. In the first step (lines $6 \sim 9$), for each path pair, we use the dynamic programming (DP) algorithm to obtain the optimal matching of it and put the optimal matching AS (OMAS) of each IP address into the OMAS set of the IP address. After that, for each router, we identify its owner AS and refine Trie2. First, we get its candidate AS set (lines $12\sim17$). We merge the OMAS set of each IP address on the router into the OMAS set of the router, and we put the origin AS of each IP address into the origin AS set of the router. We also put the origin ASes of the next IP addresses behind the router in traceroute paths into the origin AS set, which means that in this step, we extract the AS interface information and add the origin ASes of the AS interfaces into the origin AS set. The OMAS and the origin AS sets are combined to

THEOREM AND THE THE COULD THE TOURS THE TOURS THE TOURS THE	Algorithm 1.	TIFM	(OMT,	IPPath.	ASPath.	Router
---	--------------	------	-------	---------	---------	--------

1:	$Trie1 \leftarrow OMT$
2:	$Trie2 \leftarrow OMT$
3:	$PathPair \leftarrow (IPPath, ASPath) //Get valid pathpair$
4:	
5:	while Trie2 is changed do
6:	foreach pathpair in PathPair
7:	DP(pathpair, Trie2)
8:	foreach IP in <i>pathpair</i>
9:	$OMASset[IP] \leftarrow OMAS[IP]$
10:	
11:	foreach router in Router
12:	foreach IP on router
13:	$OMASset \leftarrow OMASset[IP]$
14:	$OriginASset \leftarrow Trie1(IP)$
15:	foreach <i>pathpair</i> that has the IP
16:	$OriginASset \leftarrow Trie1(nextIP)$
17:	CandidateASset = OriginASset + OMASset
18:	foreach AS in $CandidateASset$
19:	getScore(AS)
20:	OwnerAS = AS that has the MaxScore
21:	foreach IP on router
22:	if $Trie2(IP) \neq OwnerAS$
23:	$Trie2 \leftarrow (IP, OwnerAS)$
24:	end while
25:	Output: Trie2 //Refined IP-to-AS mapping table

obtain the candidate AS set. Second, we calculate the score of each candidate AS and select the one with the highest score to be the owner AS (lines $18 \sim 20$). Trie1 is used to look up the origin AS in computing the first part of the score, and Trie2 is used to test the feasible matching AS in the second part. Then we map every IP address on the router to its owner AS (lines $21 \sim 23$). We check whether the mapping AS of the IP address in Trie2 is the owner AS. If not, we change the mapping by adding the IP address and the owner AS as a new node, or by altering the corresponding AS in the node of the IP address. After working through all the routers, we check whether Trie2 is changed in this iteration. If so, we will begin a new iteration; otherwise, we stop the program and output Trie2, i.e., the refined IP-to-AS mapping table.

Note that we refine the mapping table immediately after identifying the owner AS of a router. Therefore, the latest mapping table can be used to identify the owner AS of the next router, which can ensure the convergency of our method. Originally, we update the IPto-AS mapping table after the end of an iteration. But in this way, we find that our program cannot be convergent, and the owner ASes of some routers are changed in every iteration. Here we give an example to explain this phenomenon. As shown in Fig.6, router R3 and router R4 appear in three path pairs in our dataset (the second path pair appears twice in our dataset). There are three IP interfaces on R3 (including IP3), and the origin ASes of them are AS1(IP3), AS2, and AS3 respectively. There are two IP interfaces on R4 (including IP4), and the origin ASes of them are AS1, and AS2(IP4) respectively. The owner AS of these two routers is AS2. The missing of AS2 in the first BGP-AS path may be due to the BGP route aggregation. We get the traceroute-AS path based on the original mapping table as shown in Fig.6.

In the first iterative, the candidate AS set of R3 is {AS1, AS2, AS3} and the scores of them are 2/3(1/3+1/3), 1 (1/3+2/3), 1/3 (1/3+0), respectively. Thus we identify AS2 as the owner AS of R3. Then for R4, note that the mapping AS of IP3 is still AS1 in the mapping table, thereby R4 does not know that IP3 has been correctly refined. Thus, the candidate AS set of R4 is AS1, AS2 and the scores of them are 3/2 (1/2+1), 7/6 (1/2+2/3), respectively. Therefore we identify AS1 as the owner AS of R4. After the first iterative, the refining result is shown in Fig.4, the mapping AS of IP3 is changed to AS2 and the mapping AS of IP4 is changed to AS1. Then we start the sec-

			R1	R2	R3	R4 →●
		Traceroute:	IP1	IP2	IP3	IP4
1	×	Traceroute-AS Path:	(AS3)	(AS1)	(AS1)	(AS2)
		BGP-AS Path:	(AS3)	(AS1)		
			R5	R6	R3	R4
		Traceroute:	IP5	IP6	IP3	IP4
2	×	Traceroute-AS Path:	(AS1)	(AS2)	(AS1)	(AS2)
		BGP-AS Path:	(AS1)	(AS2)		
		Iteration 1	AS1	AS2	AS2	AS1
		Iteration 2	AS1	AS2	AS1	AS2
		:	:	÷	:	:
		Iteration $n-1$	AS1	AS2	AS1	AS2
		Iteration n	AS1	AS2	AS2	AS1
		÷	:	÷	÷	:

Fig.6. Illustration of the convergency of TIFM.

ond iterative, the candidate AS set of R3 is still {AS1, AS2, AS3} and the scores of them are 4/3 (1/3+1), 1 (1/3+2/3), 1/3 (1/3+0), respectively. Therefore, we identify AS1 as the owner AS of R3. For R4, the candidate AS set of R4 is still AS1, AS2 and the scores of them are 5/6 (1/2+1/3), 7/6 (1/2+2/3), respectively. Therefore, AS2 is identified as the owner AS. Obviously, this is an endless loop. In the third iterative, the mapping AS of IP3 will be changed to AS2 and the mapping AS of IP4 will be changed to AS1 again.

In order to avoid this endless loop and make our program convergent, we refine the mapping table immediately after identifying the owner AS of a router. In this example, in the first iterative, if the mapping AS of IP3 is modified to AS2 after the owner AS of R3 is identified, then for R4, since AS1 will not be the FMAS in the second path pair, the score of AS1 will be 5/6 (1/2+1/3) which is lower than the score of AS2 (7/6 (1/2+2/3)). Therefore, AS2 will be correctly identified as the owner AS of R4 and the endless loop will not appear.

5 Data Collection

Three kinds of data are needed in our method: 1) original mapping table, 2) BGP-AS path pairs, and 3)

alias resolution data. We describe the way of collecting them in this section.

5.1 Original IP-to-AS Mapping Table

We collected all of the BGP routing tables in Routeviews⁽²⁾ and RIPE RIS⁽³⁾ at 8:00 a.m., 2012-6-30. Then we extracted the IP prefixes and their corresponding origin AS to construct the original IP-to-AS mapping table. 484k IP prefixes are included in the original mapping table.

5.2 Traceroute-BGP Path Pairs

To obtain the traceroute-BGP path pairs, the traceroute monitors and BGP feeders need to be located in the same ASes. There are 13 traceroute monitors that meet the requirement in the Ark project of CAIDA⁽⁴⁾. Table 1 shows the monitors and feeders. We collected a cycle of data of the 13 monitors from 2012-6-30 to 2012-7-3 as our traceroute path data. To avoid the influence of the changes of routing, we used the BGP routing tables at 0:00 a.m. everyday to match the traceroute paths in the day.

For each traceroute path, we matched it with the AS path of its destination in the corresponding BGP rout-

²http://www.routeviews.org, Feb. 2017.

³https://www.ripe.net/data-tools/stats/ris, Feb. 2017.

⁽⁴⁾http://www.caida.org/home, Feb. 2017.

579

ing table. The number of the path pairs is listed in the fourth column in Table 1. Some monitors, such as cdgfr, have few path pairs, because of the incompleteness of the BGP table provided by the BGP feeders. In the last column, we list the ratios of path pairs that have more than two errors. The last four monitors have relatively higher ratios. The reason is that some probing packets left the located AS through a different border router, instead of the one that we obtain BGP table from, which caused the path pairs to be inconsistent.

 Table 1. Information of Path Pairs

Monitor	Feeder	Collector	Pair (k)	Not Good (%)
arn-se	AS 2603	ris-rrc07	4.0	1.13
sql-us	AS 1280	ris-rrc14	490.0	2.21
sjc2-us	AS 6393	rv-rv2	541.0	2.37
syd-au	AS 7575	ris-rrc14	499.0	2.66
nrt-jp	AS 7660	rv-rv2	452.0	2.68
lax-us	AS 2152	rv-rv2	460.0	4.34
bjc-us	AS 3356	rv-rv2	70.0	6.27
hkg-cn	AS 22548	ris-rrc00	533.0	9.58
cdg-fr	AS 30781	ris-rrc04	0.3	12.89
zrh2-ch	AS 34288	ris-rrc12	493.0	20.15
sao-br	AS 22548	ris-rrc15	469.0	27.79
ams-nl	AS 1103	ris-rrc03	515.0	29.35
lej-de	AS 680	ris-rrc12	489.0	37.14

Here we give an example of inconsistent path pairs which are unavoidable. We found 31k path pairs in nrtjp which are caused by HKIX⁽⁵⁾, an IXP in Hong Kong that uses a route server to offer better service. In an IXP with a router server, the IXP members keep BGP sessions with the route server to obtain routing information, but they exchange data directly between each other. Thus, compared with the traceroute-AS path, the BGP-AS path that passes through this IXP will have an extra AS (i.e., the AS of the IXP). As a result, these path pairs are inconsistent, but contain one error only.

5.3 Alias Resolution Data

As discussed in Subsection 3.3, we collected the alias resolution data derived by MIDAR^[17] and iffinder^[18] from CAIDA in 2012-7-31. It includes 34 935k routers, with the 125k having more than one interface.

6 Experiments and Validation

In this section, we show the accuracy, robustness and convergence of our method. We use three different sets of path pairs to evaluate the robustness against the reduction of path pairs and the increase of inconsistent path pairs. Both evaluations are necessary, because the lack and the inaccuracy of pair paths are unavoidable. Compared with the existing methods^[6,8], we have a better result and can always get the result within a limited time.

6.1 Ground Truth Datasets

PeeringDB⁽⁶⁾ is a website that provides many of IXP's IP addresses and the owner ASes of the routers they are assigned to. Since PeeringDB is a non-profit organization, all the information on it is maintained by peering networks themselves. This information may contain some inaccuracies because no one can ensure it is updated timely. However, the ground truth of IP-AS pairs is very difficult to obtain, and the previous work (e.g., [6] and [20]) has used the PeeringDB data as their ground truth. Therefore, we also consider the IP-AS pairs from PeeringDB as one kind of our ground truth dataset. We collected 8444 IP-AS pairs from PeeringDB as one of the ground truth datasets. In addition, we also collected ground truth from Looking Glass (LG) severs. Running the show ip bgp summary command on a router, we can see the BGP sessions established with the router and the owner AS and IP address of its peering BGP router for each session. Since we obtain these IP-AS pairs from the routers directly, we consider they are correct. First, we selected 44 LG severs that support the show ip bgp summary command from LG severs $list^{\bigcirc}$. We ran the command on the routers of these LG severs and collected 9521 IP-AS pairs. Then we filtered out the pairs whose AS number is private. Moreover we want the second kind of datasets can represent more other IP addresses instead of IXP's IP addresses. Therefore we built an IXP prefixes list from PeeringDB and PCH⁽⁸⁾ and filtered out the pairs whose IP addresses are in the IXP prefixes list. At last, we obtained 1816 pairs as the second ground truth dataset.

⁵http://www.hkix.net, Feb. 2017.

⁽⁶⁾https://www.peeringdb.com, Feb. 2017.

⁽⁷⁾http://traceroute.org, Feb. 2017.

⁽⁸⁾https://prefix.pch.net/applications/ixpdir, Feb. 2017.

580

6.2 Experiments

We conducted experiments on three different path pair datasets. In all datasets the path pairs that have more than two errors are filtered. The first dataset is 13-monitor, in which the path pairs of 13 monitors are included. To show the robustness against the smaller dataset, the second dataset is 6-monitor, which includes only the path pairs of the first six monitors in Table 1. At last, because nrt-jp has many inconsistent and unavoidable path pairs, we used it as the third dataset to show the robustness against the increase of incorrect information.

We ran our method on the three path pair datasets to refine the original mapping table. All experiments used the same alias resolution data obtained in Section 5. To compare with the other two methods, we implemented IGM^[6] and ran it on the same datasets. Moreover we collected the node-AS file from CAIDA on 2012-9-14. Each router is assigned to an AS in the node-AS file. Then we mapped the IP addresses of routers to their corresponding ASes to generate the IP-to-AS mapping table as the result of the alias resolution based method (ARBM).

6.3 Validation

By using the ground truth, we validated the above refined IP-to-AS mapping tables and list the result in Table 2 and Table 3. The first and the second columns are the path pair datasets and the methods used in the experiments, (TIFM is two-type information fusion method), respectively. For TIFM and IGM, we validated the IP addresses that are included both in the ground truth and on the routers that appear in path pairs. For ARBM, we validated the IP addresses that are included both in the ground truth and in the result of ARBM. The total number of the verifiable IP addresses is shown in the "Total" column. The percentage of the IP addresses that are mapped correctly is shown in the "Correct Rate" column. In addition, for TIFM and IGM, we classified these verifiable IP addresses into three types, as shown in Fig.7. Type3 is the IP address that appears in path pairs and on the router that has more than one IP interface. Type2 meets the first condition but fails to meet the second one, while Type1 is contrary to Type2. We also showed the number and correct rate of each type.

Different types of IP addresses will obtain different information and thus influence the accuracy of the result. Both Type1 and Type3 can obtain the information from the origin ASes of other IP addresses on the routers and the optimal matching of path pairs that include the routers. The difference is that the IP addresses in Type1 do not appear in the path pairs, thereby they cannot obtain their own optimal matching information. In addition, the IP addresses in Type2 can

Table 2.	Validation	Result	Based	on	PeeringDB
----------	------------	--------	-------	----	-----------

Path Pair	Method	Туре	el	Туре	e2	Туре	e3	Total	Correct
		Number of IP	Percentage	Number of IP	Percentage	Number of IP	Percentage	-	Rate $(\%)$
		Addresses		Addresses		Addresses			
13-monitor	TIFM	324	86.4	919	94.1	639	94.2	1882	92.8
	IGM		2.5		89.5		89.2		74.4
6-monitor	TIFM	446	87.9	581	93.5	437	94.1	1464	91.9
	IGM		1.8		86.6		86.3		60.7
nrt-jp	TIFM	485	88.3	58	37.9	66	87.9	609	83.4
	IGM		2.5		36.2		25.8		8.2
-	ARBM	_	_	-	_	—	-	3123	55.9

Table 3.	Validation	Result	Based	on	Looking	Glass
----------	------------	--------	-------	----	---------	-------

Path Pair	Method	Тур	e1	Туре	e2	Туре	e3	Total	Correct
		Number of IP	Percentage	Number of IP	Percentage	Number of IP	Percentage		Rate $(\%)$
		Addresses		Addresses		Addresses			
13-monitor	TIFM	51	90.2	146	78.8	194	86.6	391	84.1
	IGM		66.7		23.3		28.9		31.7
6-monitor	TIFM	63	93.1	71	71.8	139	84.2	273	81.3
	IGM		69.8		12.7		12.2		25.6
_	ARBM	_	_	_	—	-	_	485	62.3

just obtain the information from the optimal matching and the origin ASes of themselves, thereby the accuracy of Type2 is always the lowest one. Moreover, in different ground truth data, the value of these information is different, which can be shown in the validation of the accuracy.



Fig.7. Examples of different types of verifiable IP addresses. The red points are the interfaces that own the verifiable IP address: IP1 is Type1, IP2 is Type2, and IP3 is Type3.

Accuracy. Table 2 reports the validation result on PeeringDB. In 13-monitor, 92.8% of the IP addresses can be mapped to the right AS with our method, while IGM can reach only 74.4%. Of course, one reason is that IGM cannot cover the IP addresses in Type1 that do not appear in path pairs. It is the problem of the coverage of IGM and we discuss it later. But since these IP-AS pairs are also in the refined mapping table of IGM, we listed their correct rate here. Thus, the accuracy of IGM is 89.4% by combining Type2 and Type3, which is also lower than that of TIFM because it uses only one type of information. TIFM has the lowest accuracy in Type1 as compared with Type2 and Type3. The reason is that the information that IXP's IP addresses obtain from path pairs is usually correct, but Type1 cannot use it. Moreover, the IGM did not refine the mapping of IP addresses in Type1. This means that the accuracy in Type1 is the accuracy of the original mapping table. We can see it is relatively low (2.5%)because the IXP prefixes are usually announced by only one participant of the IXPs. Thus, other participants are assigned the IP addresses whose original mappings are not themselves.

We then look at the result in Table 3. The accuracy of IGM (23.3% and 28.9%) decreases significantly from PeeringDB, while the accuracy of TIFM (84.1%) is still acceptable. This result is due to the LG dataset containing many non-IXP's IP addresses that acquire less correct information from path pairs than the IXP's IP addresses in PeeringDB, and IGM uses only the information from path pairs. This result fully demonstrated the deficiency of using only one kind of information and the necessity of two-type information fusion. Then we focused on the difference between these types. In Algorithm 1, Type2 has the lowest accuracy because it cannot obtain the information from other IP interfaces on the router. Meanwhile, Type1 has a better result than Type3. We note that the mean value of the IP interface number of the routers in Type1 and Type 3 is 17 and 4.98, respectively. We made a further analysis on this phenomenon. First, the IP addresses in the LG dataset are used in BGP sessions, thereby the probe packets that go through the routers in Type3 are more likely to enter an AS, while for Type1, they are more likely to leave an AS. Second, we found that 9 monitors in Table 1 are in the top 1000 of the AS $\operatorname{rank}^{\textcircled{9}}$ (51171 ASes are present in the rank), which means that most of the probing paths are from top to bottom. In other words, probes always leave a provider AS and enter its customer AS, and the provider AS is generally larger than the customer AS. A conclusion is provided based on these two points. The routers in Type1 mostly belong to the larger AS compared with those in Type3. In addition, the border routers in the larger AS have more interfaces and are more inclined to using the IP addresses from the owner AS. Thus, the routers in Type1 are easier to map correctly. In IGM, Type1 is higher than Type2 and Type3, which means that for the non-IXP's IP addresses, the accuracy of the original mapping is better than the refined mapping of IGM.

At last, we can see that ARBM has a quite low accuracy but covers more IP addresses. For the total number of correct IP addresses, it is almost equal to TIFM. We found that many IP addresses that belong to different ASes in the PeeringDB data are identified on the same router in the alias resolution data used by ARBM. This situation is obviously due to the false positive of the alias resolution method on the routers of IXP. Note that since ARBM is only based on the alias resolution data, in order to avoid the missing of information it uses the alias resolution data driven by the combination of measurement-based method and analysis-based method^[8], and this kind of alias resolution data has less false negatives. But this kind of alias resolution data has more false positives and this problem is more obvious on the routers of IXP. Thus, the accuracy of ARBM in PeeringDB is lower than the accuracy in LG. It is no doubt that ARBM can obtain a better result if the alias

⁽⁹⁾http://as-rank.caida.org, Feb. 2017.

resolution data is more accurate and complete, and of course, this will also improve the result of TIFM.

Coverage. We can see intuitively from the validation result, in both the PeeringDB dataset and LG dataset, ARBM has refined the largest number of IP-AS pairs, and TIFM has refined more IP-AS pairs than IGM. For example, in PeeringDB dataset, ARBM has refined 3123 pairs, while TIFM and IGM have refined 1882 pairs and 1558 pairs respectively on the 13-monitor. For the difference between ARBM and TIFM, the reason is that ARBM can refine all the IP addresses that appear on the routers and the TIFM can only refine the IP addresses that appear on the routers which are included in the path pairs. It is no doubt that TIFM will cover more IP addresses, if we use more path pairs. But note that the vantage points that can be used to provide the path pairs are part of all the vantage points from which the alias resolution data is obtained. Therefore, due to the incompleteness of alias resolution data, some IP addresses cannot be correctly identified on the same router with the IP addresses that appear in the path pairs; thus they still cannot be covered by TIFM, even if we increase the number of the path pairs. In addition, the low coverage of IGM is because it can just refine the IP addresses that appear in the path pairs. If we use more path pairs, the number of the covered IP addresses will be increased in both IGM and TIFM, and TIFM can always cover more IP addresses than IGM due to the combination with the alias resolution data.

Robustness. We validated the robustness against the reduction of information and the increase of incorrect information. Firstly, for the path pair data, compared with 13-monitor, the accuracy of TIFM of Type2 and Type3 falls by 0.9% and 2.8% respectively in 6-monitor, while IGM falls by 3% and 14.1% respectively (combining the accuracy of Type2 and Type3). For TIFM, the reduction of the path pairs means the reduction of the number of the covered routers, and thus the number of covered IP addresses is decreased as shown in Table 2 and Table 3. And for the routers that are still covered, the reduction of the path pairs means the reduction of the information from the path pairs. However, since TIFM is based on the fusion of more than one type of information, it can make up for the missing information with another type of information, and then the reduction of the information from the path pairs has less effect on TIFM than IGM. Then in nrt-jp, TIFM falls by 9.4%, while IGM falls by 62.2%. These results showed that TIFM has a great robustness and is better

than IGM. In nrt-jp, the accuracy of TIFM in Type3 (87.9%) is obviously higher than that of IGM (25.8%), thereby indicating clearly the advantage of the fusion of more than one type of information.

For robustness against the reduction of the information from the alias resolution data, we used the alias resolution data derived by MIDAR, iffinder and kapar which is the same with ARMB in our method. This data is more complete because it has less false negatives, but at the same time, it has more false positives. There are a lot of false positives in this alias resolution on router of IXP. In order to reduce the effect of the false positive, we just evaluated the robustness on 13-monitor in LG dataset. At the same time, we used the alias resolution data derived by MIDAR and iffinder which is the same with our method in ARBM. In this situation, the accuracy of TIFM and ARBM is 84.9% and 58% respectively, which means the accuracy of TIFM falls only by 0.8% while ARBM falls by 4.3%. This is also because TIFM uses more than one type of information. Thus the reduction of one type of information will have little influence on the result of TIFM. These results also indicated the robustness of TIFM against the reduction of the information from the alias resolution data is better than ARBM.

Convergency. In all experiments, we could always get the results in a finite number of iterations. Each iteration has the running time complexity of $O(mn^2)$, where n is the number of path pairs and m is the number of routers that appear in path pairs. For 13monitor, TIFM finished in seven iterations, and every iteration takes about 2 hours. For 6-monitor, TIFM used eight iterations, and every iteration takes about 30 minutes.

6.4 Contribution

Our method is based on the fusion of two types of data. For each type of data, we extracted two kinds of information. The two kinds of information extracted from alias resolution data are the original ASes of IP interfaces and AS interfaces on routers. For path pairs data, the two kinds of information extracted from it are the optimal matching ASes and the feasible matching ASes of the IP addresses in path pairs. To simplify our notation in what follows, we used "Type-1" and "Type-2" to denote the information from alias resolution data and path pairs data, respectively. In this subsection, we evaluated the contribution of each type of information and moreover we also evaluated the importance of their contribution.

Firstly, we divided the contribution into two classes: adding the owner AS into the candidate AS set (class-1) and identifying the owner AS correctly from the candidate AS set (class-2). Both of the two types of information have made contribution in these two aspects. Then for the routers that are identified correctly, we divided them into four classes. The first two classes are: 1) the routers whose owner ASes appear in the origin AS sets and 2) the routers whose owner ASes appear in the OMAS sets. The proportion of these two classes indicates the contribution in terms of adding the owner AS into the candidate AS set. The next two classes are: 1) the routers whose owner ASes get the highest scores in the alias resolution data and 2) the routers whose owner ASes get the highest scores in the path pairs. The proportion of these two classes shows the contribution in aspect of identifying the owner AS correctly from the candidate AS set. The proportion of these four kinds of routers is shown as the first ratio value in Table 4 and Table 5. For example, the routers whose owner ASes appear in their origin AS sets account for 61.9% of the routers that are identified correctly in PeeringDB.

 Table 4. Contribution of Each Type of Information in PeeringDB

	Class-1 Contribution	Class-2 Contribution
Type-1	61.9%/27.1%	52.2%/4.7%
Type-2	72.9%/38.1%	85.6%/47.8%

 Table 5. Contribution of Each Type of

 Information in Looking Glass

	Class-1 Contribution	Class-2 Contribution
Type-1	86.0%/68.7%	79.3%/2.0%
Type-2	31.3%/14.0%	90.0%/12.7%

In addition, we also evaluated the importance of the contribution of each type of information. We defined four types of routers: 1) the routers whose owner ASes appear only in the origin AS sets, 2) the routers whose owner ASes appear only in the OMAS sets, 3) the routers whose owner ASes only obtain the highest scores in the alias resolution data, and 4) the routers whose owner ASes only obtain the highest scores in the path pairs. The higher the proportion of these data, the more important the contribution of the corresponding type of information. The proportion of these four kinds of routers is shown as the second ratio value in Table 4 and Table 5. For example, the routers whose owner ASes appear only in the OMAS sets account for 38.1% of the routers that are identified correctly in PeeringDB.

AS shown in Table 4, we can find that the information from path pair data makes more contribution on both the two aspects, which is because the router in IXP can always obtain the correct information from the path pair data. For the result in Table 5, in terms of adding the owner AS into the candidate AS set, the contribution of alias resolution data is more than that of path pair data. It is because that the owner AS of a BGP router usually appears in its origin AS set. Instead, the path pair data makes more contribution on identifying the owner AS correctly from the candidate AS set, because the number of the IP interfaces on a BGP router is not enough to correctly identify the owner AS in most cases. In addition, we note that in the class-1 contribution of Table 4 and Table 5, the sum of the first number in one type of information and the second number in another type of information is equal to 100%. For example, in Table 4, the sum of 61.9%and 38.1% is equal to 100%. This is because all the routers in these tables are correctly identified in TIFM, which means the owner ASes of these routers must be added in the candidate set by one type of information. Moreover, the same sum in class-2 contribution is less than 100%, because it is possible that the owner AS does not get the highest score in both the two types of data, but the sum of the two kinds of score is the highest in the candidate set. This situation also shows the advantage of two-type information fusion.

7 Conclusions

Mapping IP addresses to AS numbers is always a crucial step when traceroute-AS mapping is required. In this paper, we fused the information from path pair and alias resolution data to refine the original IP-to-AS mapping table. Compared with the existing methods, our method can reach the highest accuracy and can always obtain the result within acceptable time. In terms of the robustness, our method could maintain a high accuracy even with a smaller dataset or more incorrect information.

In future work, we plan to explore two directions: 1) running more experiments to measure the impact of these four kinds of information on the accuracy of the result; 2) generating an AS topology with traceroute and the refined mapping table, and comparing it with the AS topology driven by the original mapping table, which will show the advantage and necessity of refining the IP-to-AS mapping table.

J. Comput. Sci. & Technol., May 2017, Vol.32, No.3

References

- Boguna M, Papadopoulos F, Krioukov D. Sustaining the Internet with hyperbolic mapping. *Nature Communications*, 2010, 1(6): Article No. 62.
- [2] Papadopoulos F, Krioukov D, Bogua M, Vahdat A. Greedy forwarding in dynamic scale-free networks embedded in hyperbolic metric spaces. In *Proc. the 29th IEEE INFOCOM*, March 2010, pp.2973-2981.
- [3] Gao L. On inferring autonomous system relationships in the Internet. *IEEE/ACM Trans. Networking*, 2001, 9(6): 733-745.
- [4] Mao Z M, Rexford J, Wang J, Katz R H. Towards an accurate AS-level traceroute tool. ACM SIGCOMM Computer Communication Review, 2003, 33(4): 365-378.
- [5] Mao Z M, Johnson D, Rexford J, Wang J, Katz R. Scalable and accurate identification of AS-level forwarding paths. In *Proc. the 23rd IEEE INFOCOM*, Mar. 2004, pp.1605-1615.
- [6] Zhang B, Bi J, Wang Y, Zhang Y, Wu J. Refining IP-to-AS mappings for AS-level traceroute. In Proc. the 22nd IEEE Computer Communications and Networks (ICCCN), July 30-Aug. 2, 2013.
- [7] Zhang B, Bi J, Wang Y, Wang Y, Zhang Y, Wu J. Revisiting IP-to-AS mapping for AS-level traceroute. In Proc. ACM CoNEXT Student Workshop, Dec. 2011, pp.900-902.
- [8] Huffaker B, Dhamdhere A, Fomenkov M, Claffy K. Toward topology dualism: Improving the accuracy of AS annotations for routers. In Proc. the 11th International Conference on Passive and Active Measurement (PAM), Apr. 2010, pp.101-110.
- [9] Pansiot J J, Merindol P, Donnet B, Bonaventure O. Extracting intra-domain topology from mrinfo probing. In Proc. the 11th International Conference on Passive and Active Measurement (PAM), Apr. 2010, pp.81-90.
- [10] He Y, Siganos G, Faloutsos M, Krishnamurthy S. Lord of the links: A framework for discovering missing links in the Internet topology. *IEEE/ACM Trans. Networking*, 2009, 17(2): 391-404.
- [11] Shavitt Y, Shir E. DIMES: Let the Internet measure itself. ACM Computer Communication Review (CCR), 2005, 35(5): 71-74.
- [12] Giotsas V, Smaragdakis G, Huffaker B et al. Mapping peering interconnections to a facility. In Proc. ACM International Conference on Emerging Networking Experiments and Technologies (CoNEXT), Dec. 2015, pp.37:1-37:13.
- [13] Zhang Y, Oliveira R, Wang Y, Su S, Zhang B, Bi J, Zhang H, Zhang L. A framework to quantify the pitfalls of using traceroute in AS-level topology measurement. *IEEE Journal on Selected Areas in Communications*, 2011, 29(9): 1822-1836.
- [14] Chen K, Choffnes D R, Potharaju R, Chen Y, Bustamante F E, Pei D, Zhao Y. Where the sidewalk ends: Extending the Internet as graph using traceroutes from P2P users. In Proc. the 5th International Conference on Emerging Networking Experiments and Technologies, Dec. 2009, pp.217-228.
- [15] Amini L, Shaikh A, Schulzrinne H. Issues with inferring Internet topological attributes. *Computer Communication*, 2004, 27(6): 557-567.

- [16] Luckie M, Claffy K. A second look at detecting third-party addresses in traceroute traces with the IP timestamp option. In *Proc. the 15th PAM*, Mar. 2014, pp.46-55.
- [17] Keys K, Hyun Y, Luckie M, Claffy K. Internet-scale IPv4 alias resolution with MIDAR. *IEEE/ACM Trans. Networking*, 2013, 21(2): 383-399.
- [18] Pansiot J J, Grad D. On routes and multicast trees in the Internet. ACM SIGCOMM Computer Communication Review, 1998, 28(1): 41-50.
- [19] Keys K. Internet-scale IP alias resolution techniques. ACM SIGCOMM Computer Communication Review, 2010, 40(1): 50-55.
- [20] Durairajan R, Sommers J, Barford P. Layer 1-informed Internet topology measurement. In *Proc. ACM IMC*, Nov. 2014, pp.381-394.



Qing Jiang is a Ph.D. candidate in School of Communication and Information Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu. Her research interests include Internet topology discover and complex network. In 2013, she received her B.S. degree in

communication engineering from UESTC, Chengdu.



Hang-Yu Hu is a Ph.D. candidate in School of Communication and Information Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu. His research interests include network security, anomaly detection and behavior profiling. In 2011, he received his

B.S. degree in communication engineering from UESTC, Chengdu.



Guang-Min Hu received his B.E. degree in computer sience from Nanjing University, Nanjing, in 1986, and his M.S. and Ph.D. degrees in geophysical prospecting from Chengdu University of Technology, Chengdu, in 1992 and 2000, respectively. From 2000 to 2003, he was a post-doctor in the School of

Communication and Information Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu. From 2002 to 2003, he was a visiting scholar at the Hong Kong Polytechnic University. He is now a full professor of the School of Communication and Information Engineering, UESTC, Chengdu. His current research interests include computer network and signal processing. He is a member of IEEE.

584