Sept. 2005, Vol.20, No.5, pp.689–701 J. Comput. Sci. & Technol.

Techniques for Determining the Geographic Location of IP Addresses in ISP Topology Measurement

Yu Jiang[1] (姜 誉), Bin-Xing Fang[1,2] (方滨兴), Ming-Zeng Hu[1] (胡铭曾), and Xiang Cui[2] (崔 翔)

[1] *School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, P.R. China*

[2] *National Computer Network Emergency Response Technical Team/Coordination Center, Beijing 100029, P.R. China*

E-mail: jiangyu@hit.edu.cn; bxfang@cert.org.cn; mzhu@hit.edu.cn; cuix@cert.org.cn

Received August 25, 2003; revised January 20, 2005.

Abstract A brief survey on the state-of-the-art research of determining geographic location of IP addresses is presented. The problem of determining the geographic location of routers in Internet Service Provider (ISP) topology measurement is discussed when there is inadequate information such as domain names that could be used. Nine empirical inference rules are provided, and they are respectively (1) rule of mutual inference, (2) rule of locality, (3) rule of ping-pong assignment, (4) rule of bounding from both sides, (5) rule of preferential exit deny, (6) rule of unreachable/timeout, (7) rule of relay hop assignment, (8) rule of following majority, and (9) rule of validity checking based on interface-finding. In totally 2,563 discovered router interfaces of a national ISP topology, only 6.4% of them can be located by their corresponding domain names. In contrast, after exercising these nine empirical inference rules, 38% of them have been located. Two methods have mainly been employed to evaluate the effectiveness of these inference rules. One is to compare the measured topology graph with the graph published by the corresponding ISP. The other is to contact the administrator of the corresponding ISP for the verification of IP address locations of some key routers. The conformity between the locations inferred by the rules and those determined by domain names as well as those determined by *whois* information is also examined. Experimental results show that these empirical inference rules play an important role in determining the geographic location of routers in ISP topology measurement.

Keywords network topology, Internet topology measurement, geographic location, network deployment structure, routing, domain name, whois information

1 Introduction

In recent years, the Internet has experienced rapid expansion in topology, and studies concerning Internet connectivity have attracted considerable attention in research communities. During the course of this development, some techniques for determining geographic locations of routers and end-hosts have been exploited and have played significant roles in various aspects, such as Internet routing topology measurement[1−5], characterization of Internet routing properties[6,7], location-aware applications servicing (e.g., targeted advertising)[8,9], Web personalized searching[10,11], Web geo-spatial navigating[12], and so on.

Currently, there are mainly two research efforts aiming at measuring an Internet Service Provider (ISP) routing topology from multiple vantage points. The researchers at the University of Washington employ multiple public *traceroute* servers to collect routing paths of an ISP network[1] and 94.2% of probing sources are outside a target ISP. In parallel with that study, we have developed a distributed architecture that uses multiple probing engines to collect routing paths of national ISPs in China[2]. In this research, all of the probing sources are equivalent to be inside a target ISP[13]. We have collected a set of routing paths from three different active sources for the China Education and Research Network (CERNET). CERNET is a nationwide ISP and plays an important role in Internet connectivity in China. Presently, CERNET has a four-layer hierarchy, i.e., the nationwide backbone, regional networks, provincial networks and campus networks, which provides a connected infrastructure for colleges, universities and other higher educational institutions around the mainland China. The domain names belonging to the *.edu.cn* or *.cernet.net* domain are registered within CERNET.

An important problem in ISP topology measurement is how to map an IP address (especially for router interface) to its real geographic location. Provided that routers can be located, it is possible 1) to form an intuitionistic geographical layout for routing topology presentation and further analysis[1,2,4], 2) to explore the properties of routing dynamics (for example, finding routing pathologies such as circuitous routing, connectivity altered mid-stream, rapidly variable routing, i.e., fluttering, etc.)[6,7,14], 3) to generate a synthesized topology with an underlying geographic model[15], or 4) to investigate the spreading scope and behavior of worm viruses, etc. Mapping an IP address to its location is an interesting and challenging task because an IP address does not inherently contain information about its geographic location.

To date, to tackle this mapping problem, researchers have developed various location-mapping tools by using some methods and techniques, such as *VisualRoute*①, *IP2Geo*[16], *GeoPing*[8], *GTrace*[17], *NetGeo* (now li-

Regular Paper

Supported by the National Natural Science Foundation of China under Grant Nos. 60203021 and 60403033.

①Visual Traceroute Utility, http://www.visualware.com/visualroute/

censed to Ixia's *IxMapping*)[18], and those in [19]. However, these techniques may not be very efficient for geographically locating routers in some ISPs due to insufficient *whois* information, inadequate or inaccurate DNS (Domain Name System) naming information (i.e., domain name), or other various reasons, especially for some backbone routers.

To deal with the ineffective locating problem of above methods, in this paper we present nine empirical inference rules for inferring the geographic location of routers: 1) rule of mutual inference, 2) rule of locality, 3) rule of ping-pong assignment, 4) rule of bounding from both sides, 5) rule of preferential exit deny, 6) rule of unreachable/timeout, 7) rule of relay hop assignment, 8) rule of following majority, and 9) rule of validity checking based on interface-finding. These empirical inference rules are very effective when there is insufficient *whois* information, or there are inadequate domain names, or the assigned domain names are inaccurate or lack of location information due to naming convention for determining locations of routers. Usually, the factors of Internet routing principles, network deployment structures, and economic constraints can be exploited in inferring routers' geographic location, and these nine empirical inference rules are obtained from taking these factors into account in determining geographic locations of CERNET routers.

Our experiments show that applying these rules to the collected routing paths has gained impressive results. In the totally 2,563 discovered CERNET router interfaces, only 163 (6.4%) router interfaces can be located by their corresponding domain names. After exercising these nine empirical inference rules, 974 (38%) interfaces have been located. In terms of the collected routing paths, two different measured geographic topology graphs generated before and after exercising these nine empirical inference rules are shown in Figs.1 and 2 respectively.

In order to evaluate the correctness of inference, on one hand, we have compared the measured topology graph with the graph published by CERNET; on the other hand, we have contacted the administrator of CERNET for the verification of IP address locations of some key routers. The evaluation results show that these empirical inference rules are informative and effective in locating routers.

Additionally, we have also examined the conformity between the locations inferred by our rules and those determined by DNS naming information as well as those determined by *whois* information. Our experiments reveal that these rules are able to detect some inaccurate locations determined by DNS naming information or *whois* information. Refer to Section 5 for the detail.

The rest of this paper is organized as follows. In Section 2, we present a brief survey on the state-of-the-art techniques used for determining locations of IP addresses and the limitation of these techniques. In Section 3, we provide some preliminary knowledge needed in the rest of the paper. In Section 4, we describe the nine empirical inference rules aiming at determining geographic location of routers in ISP topology measurement. We evaluate these rules through practical measurements and draw conclusions in Sections 5 and 6 respectively.

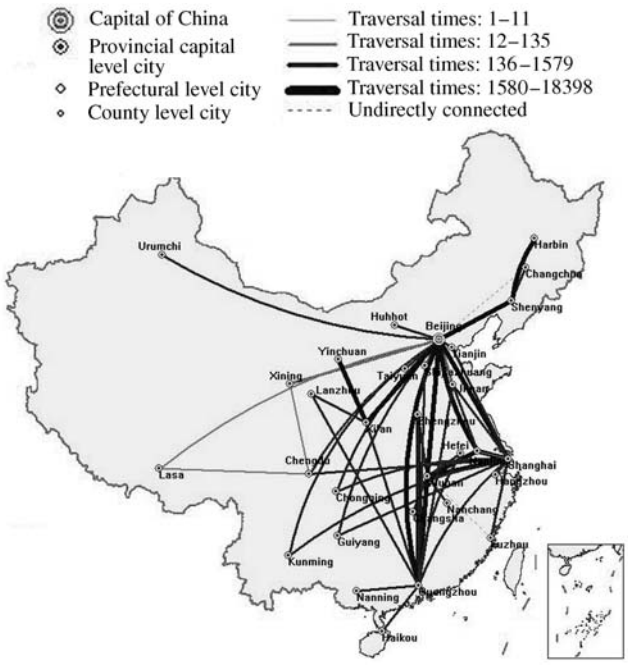

Fig.1. Measured topology graph at the 1st presentation level in which locations are determined by *whois* information and DNS naming information.

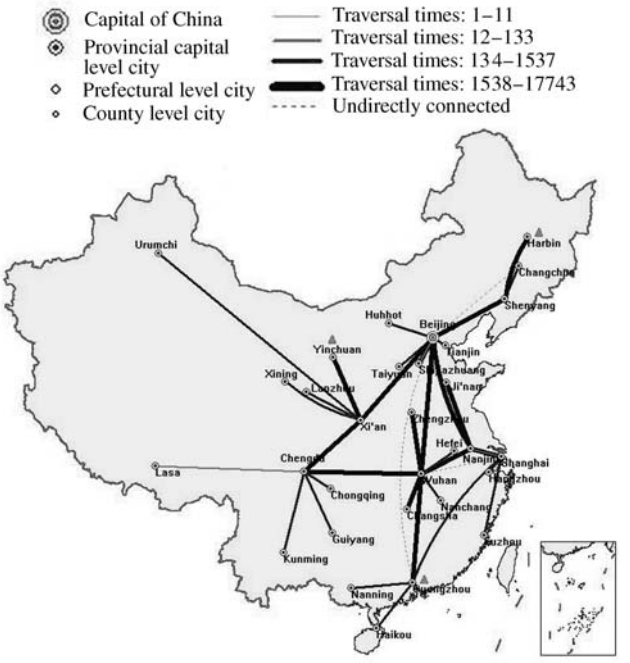

Fig.2. Measured topology graph at the 1st presentation level after exercising our empirical inference rules.

2 General Methods and Their Limitations

In this section, we will briefly summarize the general methods of determining the geographic location of IP addresses[8,9,18–21].

A number of techniques have been developed for the purpose of mapping an IP address to its geographic location. Each of them has varied strengths and weaknesses. The quantitative evaluation of the tools corresponding to the techniques is beyond the scope of the paper, and thus we will only list typical tools that use the relevant techniques. Notice that a tool often employs more than one technique.

1) *Using DNS location record for mapping an IP address to its geographic location.* In RFC (Request for Comments) 1876 (Jan. 1996), a new DNS Resource Record format was defined for associating host locations with host names within a given domain. The latitude and longitude information is registered in the resource record so as to locate the geographic location of the corresponding host, server, or router. A number of applications, such as *GMT*②, Cooperative Association for Internet Data Analysis′ (CAIDA) Mapnet project ③, *VisualRoute*, and *NetGeo*, use DNS location record for geographic display of packet routes.

The main limitation of this DNS location record-based approach lies in that the latitude and longitude information related to an IP address is registered mainly for end-hosts or Web servers, but rarely for routers due to various reasons. Meanwhile, the record deployment requires a modification of the record structure of the DNS records. This also burdens administrators with the task of entering the location records and thus there are very few DNS location records available at present. Moreover, there is no easy way to verify the accuracy of the entered location data[9].

2) *Using the domain name for mapping an IP address to its geographic location.* Some domain names contain geographic location information. For example, the domain name *p-0-3-r2-c-jsnj-1.cn.net* may imply that the corresponding IP address "resides" in Nanjing, Jiangsu Province of China. Some domain names contain organization information. For example, *www.tsinghua.edu.cn* is the domain name for the Tsinghua University of China, which is in Beijing. From the explicit or implicit geographic hints in a domain name, a corresponding IP address can be mapped to its geographic location. Examples that have employed this approach in their implementation are *Rocketfuel*[1], *VisualRoute*, and *GeoTrack*[8,9].

The main limitation of this approach is revealed in two respects. On one hand, while *Rocketfuel* researchers observed very few routers with no DNS naming information for the ten ISPs they measured[1], for some ISPs, there may be only a fraction of IP addresses that have corresponding domain names, and the problem of lacking domain names is even more striking for router interfaces. For example, merely about 6.4% of IP addresses associated with interfaces on the routers of CERNET have corresponding domain names.

On the other hand, it is challenging for this approach to identify the embedded location information for the following reasons[9]. First, there is no standard naming convention used by all ISPs, therefore, we have to compile different parsing lists for different ISPs accordingly. Second, since domain name registrants do not insist on keeping the database accurate or current, the data might be incorrect or out-of-date.

3) *Using whois information for mapping an IP address to its geographic location.* In terms of RFC 954 (Oct. 1985), each entry in a *whois* database corresponds to an IP address block that is allocated to an organization. There may be multiple entries related to one and the same organization due to Regional Internet Registries' "no guarantee of contiguous allocations" policy. Often, there is some geographic information corresponding to the IP address block in an entry. This information can be utilized to map IP addresses to their locations. *IP2LL*④, *VisualRoute*, and *NetGeo* are *whois*-based tools to infer the geographic location of an IP address.

For mapping a backbone router to its location, the fatal weakness of this approach is that the registered *whois* information for backbone routers is often at a coarse granularity. That is, in a *whois* information entry, there is usually no detailed geographic location information for each IP address associated with an interface on a router. For example, a *whois* information entry shows the whole IP address block *202.112.40.0/21* "resides" in Shanghai, China. But actually this block includes a sub-block with prefix/24s which is used by CERNET nationwide backbone routers.

However, the entries in *whois* database concerning customer organizations are usually of much value for mapping the corresponding IP address blocks to geographic locations.

4) *Obtaining partial IP-to-location mappings from HTTP cookies of Web-hosting sites, or from users' registration records either of free email services or of online TV program guide Web-servers.* The technique of *GeoCluster*[8,9] combines this kind of partial information of a few hosts with address prefix information derived from Border Gateway Protocol (BGP) routing tables to build a location mapping list for a large subset of IP address space.

The main limitations of this approach are the inaccuracy of the users' registration records out of the consideration of privacy protection, and its focus of locating hosts, not routers. We argue that obtaining partial IP-to-location mapping from user committed order forms in relevant logs at E-Commerce web-sites such as *www.china-pub.com* and *www.amazon.com* will benefit the accuracy of IP-to-location mapping, since on delivering goods, either directly to subscribers or via post offices, geographical addresses are guaranteed to be true

②The Generic Mapping Tools, http://gmt.soest.hawaii.edu/.

③Mapnet—Macroscopic Internet Visualization and Measurement, http://www.caida.org/tools/visualization/mapnet/

④Host name to Latitude/Longitude, http://cello.cs.uiuc.edu/cgi-bin/slamm/ip2ll/

by the subscribers. What we are concerned about here is only the information of the subscribers' locations and their corresponding IP addresses, yet any other private information is not involved.

5) *Estimating location from delay measurements.* In [8], the authors made an attempt to estimate geographic distance from network delay measurements, and developed *GeoPing* technique to exploit the relationship between latency and distance for determining the geographic location of a host. This method is appropriate to the environment with approximately the same bandwidth capacities and time delays in the network.

However, the problems of congestion, the Internet Control Message Protocol (ICMP) traffic density and "the last one mile" are the main disadvantages for the accuracy of estimation.

6) *Making an exhaustive tabulation of IP addresses and their corresponding locations.* This method may produce a very accurate IP-to-location mapping list, but it demands too much effort. The main limitation of this method lies in its difficult and time-consuming process as well as continuous maintenance. Two of such location-mapping services are Akamai's *EdgeScape*⑤ and Digital Island's *TraceWare*⑥[8].

From the above analyses, we learn that these techniques have played important roles in the relevant research fields, especially for the use of domain name and *whois* information. Most of these techniques employ various kinds of static information for mapping an IP address to its location, which is of much use for locating non-router IP addresses, but of little avail for locating backbone routers with no DNS naming information for some ISPs. Therefore, in Section 4 we will put forward some efficient empirical inference rules for determining geographic location of a router (interface) after some preliminaries in Section 3.

In the rest of the paper, for simplicity, a router and an IP address associated with an interface on the router are used interchangeable, and so are an IP address and a host when needed. An IP address, without particular explanation, mainly refers to an interface on a router, not a host.

3 Preliminaries

Before we present the empirical inference rules for determining geographic locations of Internet routers, this section provides the classification of IP addresses and cities and a few definitions that will later be used in the paper.

3.1 Classification of IP Addresses and Cities

In most cases, on tracing an end-host with *traceroute*-like tools, when the last packet has reached the destination, all of the IP addresses except the last one in the output list are IP addresses representing certain interfaces on corresponding routers; when the trace process terminates before reaching the destination, all of the IP addresses in the output list are IP addresses representing certain interfaces on corresponding routers, and so is the case for tracing a non-end-host destination.

Since the inclusion of end-hosts in the measured routing topology would make the size of graph much larger and it is neither good for visual effects nor good for checking the connections between routers, the end-hosts are usually not included in the measured routing topology[2,4]. Therefore, in the rest of the paper, we will not consider the end-host IP addresses when discussing the IP addresses in the *traceroute*-like outputs.

We classify the IP addresses used for routers into two sets. One set contains the IP addresses in the custody of customer communities (not ISPs), which are contained in the allocated IP address blocks for the customers. The other set contains the IP addresses in the custody of providers (ISPs), which are used only for routers (not including unassigned or unallocated IP address blocks). Hence, all of the IP addresses in routing paths belong to either the former set or the latter set.

We assume that each customer community (not ISP) in a city obtains its IP addresses from Local Internet Registries. The case of portable IP address blocks will not be considered here. That is, each IP address block registered in one entry of *whois* information, which is used by a customer community, is located in only one city. Therefore, the geographic information contained in an entry of *whois* information of a customer community can be used to infer the geographic location of routers within those blocks. This assumption is reasonable since allocating IP address blocks in this way is beneficial for routing aggregation policy.

Therefore, the main task left now is to determine the geographic location of IP addresses that are in the custody of the target ISP and used for (backbone) routers. In other words, we are concentrating on determining the locations of intermediate hops in routing paths, not including the last hop, and determining the locations of interfaces on routers is equivalent to determining the locations of routers.

In order to accomplish the task of determining geographic locations of CERNET nationwide routers from the data collected by us[2], we refer to a full list of names and codes of cities in China published by the National Bureau of Statistics of China⑦ and classify these cities into four categories: 1) *municipal cities under the Central Government*, namely, Beijing, Shanghai, Tianjin, and Chongqing, short termed as *prov-cities*, 2) *cities of provincial capitals*, such as Nanjing (the capital of Jiangsu province), short termed as *prov-capital-cities*, 3) *cities on prefecture level*, such as Suzhou (in Jiangsu province), short termed as *pref-cities*, and 4)

⑤Akamai Inc. http://www.akamai.com/

⑥Digital Island Inc. http://www.digitalisland.com/

⑦http://www.stats.gov.cn/tjbz/index.htm

cities at county level, such as A'Cheng (in Heilongjiang province), short termed as *county-cities*. We use these categories for our determination of location and they are used as the granularity of the geographic location in our topology presentations.

There are totally four presentation levels. The first presentation level contains all cities of the first two levels. The second presentation level contains all cities of the first three levels. The third presentation level contains all the cities confined within a single province, namely, the provincial capital city and all of the *pref-cities* and *county-cities* within that province. The layout for the first two levels comprises one map each. The layout for the third level comprises multiple maps, each corresponding to a province. All of the links between routers within the same city are not displayed in these three levels of city-coverage presentation. The fourth presentation level comprises multiple logical layouts, each mainly containing routers (interfaces) confined within a specific city. This logical level has no relevant geographic information and hence we do not explain much about it here.

3.2 Related Definitions

In order to present the empirical inference rules in a concise way, in this section we give some definitions that are related to the collected routing paths.

Definition 1. *A target is a final destination IP address to which a series of probing packets with different Time-To-Live values are sent. A probing source is an IP address from which probing packets are sent to various targets.*

Definition 2. *A path is a hop series from a probing source to a target, consisting of IP addresses or asterisks. A targeted path is a path ending with the target IP address. A customer-targeted path is a targeted path in which the targeted IP address is allocated to a customer. A customer-targeting path is a non-targeted path in which the target IP address is also allocated to a customer.*

Definition 3. *A downstream sequence is a sequence of consecutive IP address hops from a probing source to a target in a path. An upstream sequence is a reversal sequence of a corresponding downstream sequence.*

Definition 4. *A downstream pair (IP_1, IP_2) is a pair of two consecutive IP address hops in a downstream sequence, short termed as down-pair. Here, IP_1 is called as a starting point, and IP_2 is called as a successive point.*

Definition 5. *A relay hop is a hop that only connects two other hops.*

In Fig.3, only nodes B and D are relay hops.

Definition 6. *A ping-pong state refers to an occurrence of the following short-period-loop phenomenon in a path: $\cdots \to A \to B \to A \to B \to A \to B \to \cdots$. We call such loop paths as ping-pong paths. An inter-city ping-pong path is a ping-pong path in which the link between the ping-pong routers is an inter-city link. An intra-city ping-pong path is defined similarly.*

A *ping-pong* state usually means that packets are stuck and cannot proceed to reach the target, and this state may occur in any position of mid-stream.

Definition 7. *A petiole refers to the shared part in which all probed paths from one probing source must traverse. It consists of a sequence of hops except the farthest one away from the probing source in the shared part*[4].

The petiole part is not presented in our layout of measured routing topology[4]. In Fig.4, S is a probing source, and edges (S, A) and (A, B) make up a *petiole*. These two edges will be removed from the generated graph except for node B.

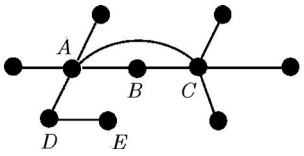

Fig.3. Example of relay hops (B and D).

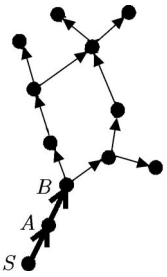

Fig.4. Example of the *petiole*, (S, A) and (A, B).

3.3 Categories of Probed Paths

The collected routing paths must be normalized before further processing, so we normalize and classify the probed paths by their characteristic features.

1) The case of *targeted*. In the course of sending ICMP or UDP probing packets to an IP address, when receiving an ICMP "echo reply" packet (in an ICMP probing way) or receiving an ICMP "protocol unreachable" or "port unreachable" error message (in a UDP probing way), it is confirmed that the probing packet arrives at the destination (target), and we record the path to the IP address that is prior to the destination IP address. In the meantime we record the last two IP addresses (including the target) into another list.

If the destination happens to be a backbone router's interface, it will likely present itself in other paths. Otherwise, it would be a leaf in the topology graph and could be deleted without any side effect. Thus, recording the targeted path in this way does not affect displaying the completeness of the routing connectivity.

2) The case of *ping-pong*. If a *ping-pong* state $\cdots \to A \to B \to A \to B \to \cdots$ occurs, we terminate this probing process and record the path to the first B.

3) The case of *max-hop*. When the hop reaches the maximum (30 hops by default), if it happens to arrive at the target, we classify this case to the first categorization; if it still does not arrive at the target, we record the whole path. The case of *max-hop* may be further divided into two sub-cases. The first sub-case is that there does not exist any repeated segment of an IP address sequence in the path. The second sub-case is the *long-period-loop* paths in which there exists a repeated

segment of an IP address sequence. This is also a kind of routing loop.

4) The case of *timeout*. If three asterisk hops appear consecutively, when the third asterisk occurs, we terminate this probing process and only record the path to the IP address prior to the first asterisk. We assume that the subsequent hops are also timeout hops in this probing period.

5) The case of *unreachable*. When receiving a "network unreachable" or "host unreachable" message, or a "communication administratively prohibited" message, or any other unreachable message with codes defined in RFC1812 (June 1995), except CODE 2 and CODE 3, we record the whole path.

Consequently, we classify paths into five categories: *targeted*, *ping-pong*, *max-hop*, *timeout*, and *unreachable*. There is no overlap between these five cases in terms of the corresponding normalized recording principles.

Hop	IP address
⋮	⋮
3	202.112.19.193
4	202.112.1.77
5	202.112.46.69
6	202.112.46.182
7	202.112.53.74
8	192.168.0.2
9	202.112.14.2
10	202.112.15.122
11	210.41.116.1
12	210.41.116.1
(Trace complete.)	

Fig.5. Example of end replica phenomenon in probed paths (omit the first two hops and RTT values here).

In the collected routing paths, some of them exhibit the phenomenon of end-replica, as illustrated in Fig.5. End-replica means that the last few IP addresses in the path are identical, and the known maximum number of end-replica hops in our experiment is four. Except the case of ping-pong, the end-replica phenomenon exists in all the other four cases mentioned above. Usually, this may be caused by the administrative or firewall configuration for preventing the outside from probing the structure of the local network. Since the replicated hops other than the last one are not the real interfaces of routers, they must be refined from paths.

In the next section, we will provide and explain the empirical inference rules.

4 Empirical Inference Rules Based on Routing Principles, Network Deployment Structure, and Economic Constraints

In this section, we present some more effective empirical inference rules for determining the geographic location of routers.

4.1 Motivation

For simplicity, we say an IP address is *located* if we are able to map its geographic location; otherwise, it is *un-located*.

In general, the mapping problem solved by the previous mapping methods stated in Section 2 is with limitation. For example, we employ some general methods for mapping CERNET routers:

1) Using the reversal domain name resolution technique to obtain the possible domain name(s) for each IP address. Then, using the regular expression matching method, we make most of the geographic hints in each domain name to map the IP address to its location. For the rest of *un-located* IP addresses, we resort to the next approach.

2) Using the technique of querying *whois* database to obtain the possible geographic information registered in the corresponding *whois* information entry. If available, the target ISP's *whois* information server, such as *whois.edu.cn* of CERNET, is referred to first. Then for still *un-located* IP addresses, the Asia Pacific Network Information Center's (APNIC) *whois* information server (*whois.apnic.net*) is referred to [8]. That is, querying the *whois* information servers from particular to general according to the regions involved.

3) Using our mapping list manually compiled from other resources to rectify possible inaccurate mapping.

Experiments show that in the totally 2,563 discovered interfaces of CERNET routers, only about 6.4% of interfaces can be located by DNS naming information, and the remaining interfaces cannot be located without relevant *whois* information. However, the anticipated accuracy of relevant information in *whois* database is far from good. An experiment of sampling shows that 45.6% of the sampled router interface IP addresses' *whois* information is inaccurate. Therefore, some "rule of thumb" approaches must be explored to improve the accuracy of locating IP addresses.

4.2 Empirical Inference Rules

Generally, a nationwide ISP network may comprise one or more autonomous systems. An autonomous system (AS) usually employs the routing protocols that utilize the principle of selecting the shortest path within the autonomous system, and inter-domain routing protocols such as BGP-4 (version 4) are usually employed between ASes[22]. A national ISP network such as CERNET usually employs a hierarchical structure in its deployment.

An ISP typically employs "cold-potato" routing where it carries the packets on its own network as far as possible before handing off to the next ISP, or employs "hot-potato" routing where it hands off packets to the next ISP as soon as possible[7]. An ISP is likely to use "cold-potato" routing policy to deliver packets whose destination IP addresses are in the custody of itself, while it is likely to use "hot-potato" routing policy

[8] In the portable assignment case, other *whois* databases may be necessary. However, we do not consider this case in the paper.

to deliver packets whose destination IP addresses are in the custody of other ISPs. Moreover, Internet employs a kind of destination-oriented routing and provides best-effort service for packets.

Out of economic consideration, it is unlikely for an ISP to build more than three inter-city links with one and the same router between a pair of cities, though it is common that routers in different cities link with one and the same router in another city for an ISP, since it is usually costly to build inter-city links.

The basic network deployment structures, routing principles, and economic constraints provide a ground for the following empirical inference rules, and these empirical inference rules are derived from the practice of determining geographic locations of CERNET routers.

Rule 1 (Rule of Mutual Inference). *In a customer-targeted path, the last two IP addresses in the path are assumed to be in the same city.*

According to the principle of constructing and linking to Internet, there would be two possible scenarios for a customer's network in a city. One is that there exists a router to connect with its provider in the customer's local area network (LAN). The other is that no router exists in the customer's LAN and the customer's LAN employs a layer-two device such as a switch to connect with its provider.

In the former scenario, the router in the customer's LAN is in the same city as the hosts of that LAN, and Rule 1 is applicable to this case. That is, as long as we know where one of the last two IP addresses is located, we will be able to infer the location of the other. This is very useful in determining locations of some routers near hosts since we are usually able to know hosts' locations such as by *whois* information. For example, suppose that IP_1, IP_2 form a downstream subsequence of the last two IP addresses in a customer-targeted path and IP_2 is a host in city C_1, then we can infer IP_1 is also in C_1.

In the latter scenario, what we assume is that the destination IP address and its access router are in the same city with a high probability. If wrong inter-city link(s) is (are) introduced, we may rely on Rule 8 or Rule 9, or evaluation methods to rectify wrong inter-city link(s).

Rule 2 (Rule of Locality). *In a customer-targeted path, if the destination and the probing source are at the same location, then all of the IP addresses in this path are assumed to be at the same location as the probing source, which is already known.*

From the perspective of users, only under the condition that the probing source and the destination hosts' IP addresses are in the custody of the same ISP can this rule be applied, since in this case, the scope of the routing path is guaranteed by the cold-potato routing and the shortest path routing principles. For instance, if we measure the topology of CERNET in Guangzhou, and the IP addresses of probing source and destination hosts are both in Guangzhou as well as in the custody of CERNET, then the locations of all the IP addresses in all the probed paths from this source to the destination hosts can also be inferred to be Guangzhou.

This rule may not be appropriate for the paths with routing pathologies such as circuitous routing or loop routing, especially the case of *long-period-loop*. Hence, this rule is not workable for non-targeted paths.

Even the precondition is satisfied, Rule 2 may also introduce wrong inter-city link(s) in the measured topology of a target ISP, when a path exists from a source to a destination in the same city crossing over the border of the city. However, if this scenario is the result of mis-configuration on some router that leads to circuitous routing, then this rule may tend to help identify the existence of mis-configuration when validating the measured graph. In our experiments no such case is encountered.

For Rule 1 and Rule 2, it is also important that the targeted IP address is allocated to a customer. Otherwise, if the targeted IP address (randomly selected) happens to "belong to" a backbone ISP and is used on a router, we would not know its precise location beforehand and could not use it as a prerequisite to determine other locations.

Rule 3 (Rule of Ping-Pong Assignment). *For customer-targeting ping-pong paths, we assume that the ping-pong pair is at the prov-capital-city of the province in which the target is located, or at the same location as the target ISP's regional network center that is the nearest to the targets, depending on the locations of the targets and the preceding hop of the ping-pong pair.*

This is a reasonable assumption since Internet is a kind of IP network that provides best-effort services, and in principle, destination-oriented routing based on network addresses would make the phenomenon of *ping-pong* not far away from the city of destination. Note that there are four levels in the deployment structure of CERNET, a *prov-capital-city* is at the third level of this deployment structure, and the provincial networks' administration center is usually located in *prov-capital-city*.

If the preceding hop of the *ping-pong* pair and the starting point of the downstream *ping-pong* pair are inferred to be at different locations by this or other rules, then we rectify the location of the successive point of the downstream *ping-pong* pair to be at the same location as the preceding hop. This rectification will not introduce wrong inter-city links since when a *ping-pong* phenomenon happens (as Fig.6 illustrates), the successive point of the downstream *ping-pong* pair (B in Fig.6) and the preceding hop of the *ping-pong* pair (P in Fig.6) must be two interfaces on the same router that generates the ICMP time exceeding packets sent either from the ingress port or from the egress port.

Rule 4 (Rule of Bounding from Both Sides). *If there is an IP address sequence of consecutive hops* $\cdots A \to B_1 \to \cdots \to B_k \to C \cdots$ *in a path (here* k *is a positive integer), A and C are varied addresses yet the locations of A and C are the same, then the locations of* $B_1, \ldots, B_k$ *are assumed to be the same as that of A (or C).*

Here k is a threshold value and is usually not greater than 2. In this paper, we set $1 \leqslant k \leqslant 2$. When k equals

1, this rule is equivalent to the approach employed in [6]. Note that this rule is not fit for a larger k due to possible circuitous routing. It is a prerequisite for this rule that A and C are different IP addresses at the same location.

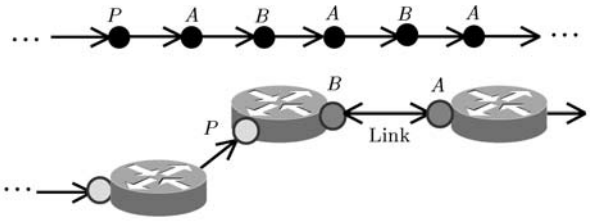

Fig.6. Illustration of a *ping-pong* phenomenon.

Rule 5 (Rule of Preferential Exit Deny). *If an IP address, say IP_1, in a valid path has already entered the area of the target location, then the probability that all of the other downstream IP addresses after IP_1 in this path and the target are at the same location is very high, and thus it can be preferentially accepted.*

This rule differs from Rule 4 in two points. One is the number of IP addresses between the two reference IP addresses. For Rule 4, k is definite ($1 \leqslant k \leqslant 2$), while for Rule 5, it is not necessarily the case. The other difference between Rule 4 and Rule 5 is that for Rule 4, the IP address sequence of consecutive hops $\cdots A \to B_1 \to \cdots \to B_k \to C \cdots$ may be any segment in the path, while for Rule 5, IP_1 must be in the same city as the target location.

The exceptional cases to this rule are usually the *long-period-loop* case and *ping-pong* case. In these two cases, locations of IP addresses after IP_1 in the path may have been out of the target location. However, we believe that the probability for this situation in very low.

Like Rule 2, another low probability case is that Rule 4 may also introduce wrong inter-city link(s) in the measured topology of a target ISP when there actually exists a path between two IP addresses in the same city crossing over the border of the city. That is, it first comes into the city, then runs outside the city for some hops, and then comes back again to the city and gets to the destination in the city. This is a kind of circuitous routing that might make this rule inapplicable. However, if it is found that this rule has made wrong inter-city link inference, it is possible that this wrong inter-city link should be attributed to mis-configuration (resulting in a sub-optimal route) on some routers.

Rule 6 (Rule of Unreachable/Timeout). *For still un-located IP addresses of the last one or two hops in an unreachable or timeout path, we assume that the last one or two hops are at the same location with its preceding hop.*

In IP networks, probing packets are capable of traversing the furthest according to the best-effort service and the destination-oriented routing. Thus, the last hop of the unreachable or timeout path is likely to "stop" at a location near the location of its preceding hop along the path. Exercising this rule may decrease the number of possible inter-city links, but it would not introduce wrong inter-city links.

Rule 7 (Rule of Relay Hop Assignment). *We believe the probability that a relay hop is at the same location as either its preceding hop or its subsequent hop is very high. In this paper,* (1) *if the preceding hop happens to be an IP address that is not in the custody of the target ISP, and we do not know its location, then the relay hop is assumed to be at the location of its subsequent hop;* (2) *if the subsequent hop happens to be an IP address that is not in the custody of the target ISP, and we do not know its location, then the relay hop is assumed to be at the location of its preceding hop;* (3) *if the preceding hop is at the location of a regional center and the subsequent hop is at a prov-city or prov-capital-city, then the relay hop is assumed to be at the location of a regional center nearest to the prov-capital-city;* (4) *if the preceding hop is at the location of a regional center and the subsequent hop is at a pref- or county-city, then the relay hop is assumed to be at the prov-capital city of the province in which the pref-city or county-city is located;* (5) *for other cases, the relay hop is assumed to be at the location of its preceding hop.*

Fig.7 illustrates the way of exercising the Rule 7(1) and Rule 7(2). In this figure, R is the relay hop and inferred node, while P is R's preceding hop, and S is R's subsequent hop. These two inferences are secure since they would not introduce wrong inter-city links. It is a prerequisite to this rule that the location of the subsequent hop in Rule 7(1), the location of the preceding hop in Rule 7(2) and Rule 7(5), and the locations of both the preceding hop and subsequent hop in Rule 7(3) are already known.

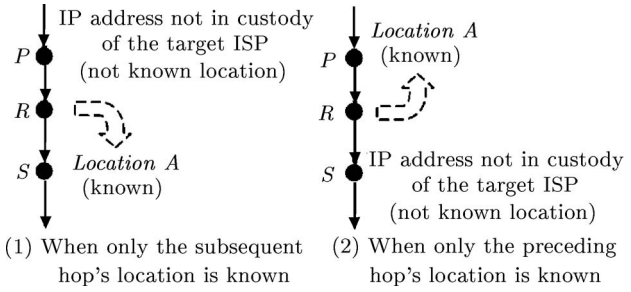

Fig.7. Illustrations of exercising Rule 7(1) and Rule 7(2).

If the two nodes adjacent to the relay hop are at two different locations and there indeed exists a direct link between them, then the assumption does not affect the accuracy of city-coverage layout of routing topology, since this method will not change the pattern of links between the two cities and links within the same city are not displayed at this presentation level. In Fig.3 node B is an example of such case. Certainly, if there is no direct link between these two nodes adjacent to the relay hop, that is, the relay hop is at the third city, the assumption would lead to an error. In Fig.3 node D is an example of such case. Therefore, we must be careful to exercise this rule. This rule cannot be applied to an individual path. Instead, it can only be applied to united paths.

It is much safer to apply this rule to the union of probed paths collected from multiple probing sources than to the paths collected only from a single source, since a relay hop in routing paths collected from a single source may no longer be a relay hop in united routing paths collected from multiple sources due to possible cross-links or branch-links. Moreover, Rule 7(3) does not fit the situation that the inter-city link is a satellite channel, even for the united paths collected from multiple probing sources.

Rule 8 (Rule of Following Majority). *Within an ISP,* (1) *if there are no less than k* (*k is a positive integer*) *down-pairs with the same starting point, but the successive points are of different IP addresses yet at the same known location, then the location of the starting point is assumed to be the same as the k successive points*; (2) *if there are no less than three down-pairs with the same starting point, but the successive points are at different pref-cities or county-cities within the same province, then the location of the starting point is assumed to be at the prov-capital city of the province*; (3) *if there are no less than three down-pairs with the same starting point, but the successive points are at different prov-cities or prov-capital-cities near the same regional network center and are not at probing sources, then the location of the starting point is assumed to be at the location of that regional network center.*

Here k is a threshold value and is usually no less than 3. Otherwise, it would lead to errors. We set it to 3 in this paper (for larger ISP, the threshold may need to be set to 4 or even larger). Fig.8 illustrates the way of exercising Rule 8(1), where D is the inferred node.

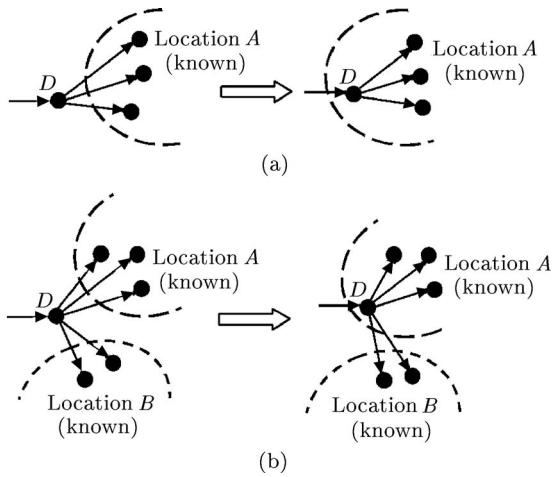

Fig.8. Illustrations of exercising Rule 8(1). (a) All the successive points are at the same location. (b) Majority of successive points are at the same location.

This rule is deduced from the fact that it is a regular case for an ISP to build no more than two inter-city links with one and the same router between a pair of cities. It is unlikely for an ISP to build more than three inter-city links with one and the same router between a pair of cities, since it is usually costly to build inter-city links. That is, if for an ISP there are no less than three links with one and the same router, these routers are likely to be in the same city.

It is a common phenomenon that routers in different cities link with one and the same router in another city for an ISP. It is also a common phenomenon for a pair of cities to have more than three links that belong to different ISPs.

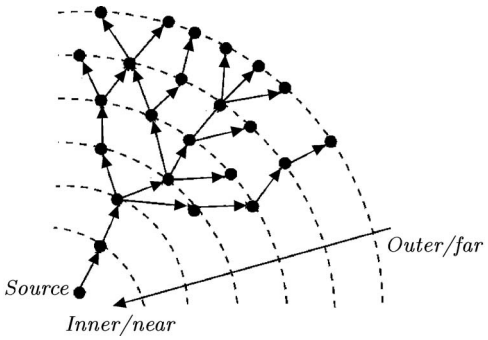

Fig.9. Illustration of how to observe the collected routing paths from a probing source.

If we view the graph constructed from paths collected from a probing source in the following way: taking the probing source as the center, and assuming each downstream hop with the same hop number in every path on one and the same curve (as shown in Fig.9), then Rule 8 (rule of following majority) must be applied in the way from outer curves to inner ones consecutively. Otherwise, it would introduce errors. Like Rule 7, this rule can only be applied to united paths instead of individual paths.

We learn that a router has at least two interfaces and each interface is associated with a distinct IP address belonging to different subnets labeled by subnet masks. If we know the geographic location corresponding to one of these IP addresses, we then know the router's location and also the locations of other interfaces' IP addresses on the same router. With the help of interface finding mechanism (for example, *iffinder* mechanism in [3]), we are able to validate, verify, and rectify the locations of some routers. This constitutes our inference Rule 9.

Rule 9 (Rule of Validity Checking Based on Interface Finding). (1) *If there are k interfaces found on the same router and more than half of them are at the same location, then the location of the other interfaces on this router is also assumed to be that location.* (2) *If there are two IP addresses found on the same router and one of them is in the custody of a customer, then the location of the other address is assumed to be at the location of that customer.* (3) *If there are two IP addresses, for example, IP_1 and IP_2, found on the same router, IP_1 is located by its domain name and its location is not modified by other inference rules, IP_2 is in the custody of a provider, then the location of IP_2 is assumed to be at the same location as IP_1.*

Rule 9(1) is also a kind of following majority principle, and Rule 9(2) has its root in the fact that, for a customer, the ISP to which the customer will access usually provides an IP address for the access-link to the

customer. Rule 9 functions in two aspects. One is to validate the above inferences, and the other is to provide some more inferences not made by other previous rules. This rule should be applied after all of the previous eight rules, since earlier application of this rule may lead to inaccurate inferences due to the possibly inaccurate locations and then may lead to wrong conclusions.

4.3 Exercising Order and the Roles

The empirical inference rules presented above have the corresponding preconditions and preferential exercising order. From the analysis we learn that Rule 1, Rule 2, Rule 3, and Rule 6 are applicable to different categories of individual routing paths, while Rule 4 and Rule 5 are applicable to various categories of individual routing paths provided that they satisfy the conditions required by these two rules. Exercising Rule 1, Rule 2, and Rule 3 will obtain more geographic locations besides those provided by possible domain names. In addition, considering the preconditions of exercising Rule 4 and Rule 5, we should exercise Rule 4 and Rule 5 after Rule 1, Rule 2, and Rule 3, and exercise these rules in numerical order. Then, Rule 6 is applied.

Differing from previous rules, Rule 7, Rule 8, and Rule 9 can only be applied to united paths instead of individual paths, as we have pointed out above. While Rule 7 and Rule 9 may be applicable to united routing paths collected from a single probing source, they function much better when applied to the overall united routing paths collected from multiple vantage points. Moreover, inferences made by Rule 1 through Rule 7 are the bases of exercising Rule 8 and Rule 9. Therefore, Rule 8 and Rule 9 should be applied after the first seven rules and in numerical order as well. In fact, the basic guideline for the order of exercising these empirical inference rules is from simple to complex, from individual paths to united paths/graph.

After exercising Rule 8, if there is any location rectification we re-exercise rules from Rule 6, then Rule 7, and then Rule 8, until there are no more rectifications, since after exercising Rule 8 some locations may be changed and thus will have impacts on the locations of hops related to Rule 6, Rule 7, and then Rule 8. Finally, Rule 9 is exercised.

It is important that Rule 1 does not fit the targeted paths with end replica, timeout at the penultimate hop, and non-customer as destination. It is only fit for the case of customer-targeted paths; Rule 2 is inapplicable to both non-targeted and non-customer targeted paths; Rule 4 is not fit for *max-hop* paths with *long-period-loop*; Rule 5 does not fit both *long-period-loop* case and *ping-pong* case. All of these rules are also inapplicable to the IP addresses that are not in the custody of the target ISP and to the private addresses that fall in blocks *10.0.0.0/8*, *172.16.0.0/12*, or *192.168.0.0/16* (RFC1918, Feb. 1996) appearing in the paths (see hop 8 in Fig.5).

Experiments show that after exercising Rule 1 through Rule 7, the presentation layout of the measured topology is modified to some extent, but the visual effect is still similar to that of the graph in Fig.1, and that exercising Rule 8 fundamentally changes the visual effect such that the clarity of the presentation layout of measured topology is greatly improved, much similar to that of the graph in Fig.2. Then, exercising Rule 9 gets the final result. However, Rules 1 through 7 are not unnecessary because they constitute the base of and are a prerequisite to exercising Rule 8 and Rule 9.

In addition, we admit that some existing domain names have also played an important role in determining the locations of CERNET routers. Without them, the conditions for exercising some of these rules might not exist and these empirical inference rules might not work successfully, though Rule 1, Rule 2, Rule 3, Rule 5 and Rule 8 might still manage to make some inferences.

In the next section, we will present the experiment and evaluation results of applying these empirical inference rules to the probed paths.

5 Experiments and Discussions

In this section, we demonstrate the inspiring results obtained from applying these inference rules to the collected routing paths of CERNET topology.

5.1 Experiments

Using the data collected in [2], we conducted the analysis and inference work of determining the locations of IP addresses in the probed paths. The data were collected from three different probing sources, Guangzhou, Yinchuan, and Harbin (indicated by three solid triangles on the map from bottom to top in Fig.2), in 2003, for mapping the routing topology of CERNET.

When we intend to provide a geographic layout of the measured routing topology graph from the collected routing paths, we first resort to the general methods for mapping an IP address to its geographic location. After accomplishing steps in Subsection 4.1, we find that the results shown in Fig.1 do not come up to the desired effects, since in the totally 2,563 discovered router interfaces of CERNET there are merely about 6.4% of interfaces on routers with domain names, and the remaining 93.6% of interfaces cannot be located without relevant *whois* information. Meanwhile, CERNET national topology graph published in the corresponding period on the web page[23] is shown in Fig.10.

Comparing Fig.1 with Fig.10, we learn that merely depending on the limited DNS naming information and *whois* information in determining locations of routers of CERNET does not work well. We are convinced that there must be some router interfaces that are inaccurately mapped. Therefore, we must resort to other methods.

As stated in Subsection 4.3, before exercising each of these empirical inference rules, we must exclude the IP addresses and paths to which the rule is inapplicable. In the collected paths, there are 308 private addresses and 303 IP addresses that are not in the custody of CERNET. The total number of discovered interface addresses

on CERNET routers is 2,563 IP addresses that should be located. Moreover, the following five kinds of routing paths, *targeted* paths with end replica, *targeted* paths with a non-customer address to be destination, *targeted* paths with a timeout at penultimate hop, *targeted* paths with a private address at penultimate hop, and *max-hop* paths with *long-period-loop*, are excluded before exercising the corresponding empirical inference rules.

Fig.10. CERNET national topology graph[23].

We apply these inference rules to the refined paths collected from multiple probing sources in the following way: applying Rule 1 to refined routing paths from Guangzhou, then Yinchuan, and then Harbin, and then applying Rule 2 to the paths from each of the three sources in the same order, and then applying Rules 3–6 successively in the same way. After that, Rule 7 is applied to the overall united graph derived from the above three sets of the collected routing paths, and Rule 8 is applied to the individual united graph derived from the routing paths collected at each single source. After Rule 8 is applied and there is no more location rectification, we apply Rule 9 to the overall united graph. For Rule 9, we first employ the technique presented in [3, 5] for alias resolution and then check the location of each interface of a router. For still *un-located* interfaces, we check the related paths to find the mainstream and make a non-conflicting inference.

In this way, the total number of distinct interfaces to which the empirical inference rules are applied is less than the sum of interfaces in different categories related to individual rules, as shown in Table 1, since for some interfaces the condition for exercising an empirical inference rule may also be feasible for exercising some following rules, and thus they would be counted more than once if more than one rule is applicable.

The classified numbers of interfaces to which empirical inference rules are applied are shown in Table 1 (the counted interface addresses may overlap due to different probing sources and the rectification of inferred locations.) From this table we may see that, in the totally 2,563 discovered CERNET router interfaces, the locations of 33.1% of interfaces have been inferred by the first eight rules, which includes the number of some interfaces with domain names to which empirical inference rules have also been applied.

The effectiveness of exercising these empirical inference rules is shown in Table 2. From this table we may see that before exercising these rules, in the totally 2,563 discovered CERNET router interfaces, merely about 6.4% of router interfaces can be located by their corresponding DNS information, and the remaining 93.6% of interfaces are unable to be located without relevant *whois* information. However, after exercising these nine empirical inference rules, 38% of all the discovered router interfaces have been located, almost six times as many locations as inferred merely by DNS naming information directly. Locating other interfaces still rely on the *whois* information.

Furthermore, as we have stated above that Rule 9 plays two roles, one is to validate the above inferences, and the other is to provide some more inferences not made by other previous rules. After alias resolution, the total aliases/interfaces involve 396 different IP addresses in routing paths, without including 8 pairs of alias of IP addresses that are all in the custody of providers and 11 IP addresses not in routing paths. For these 396 addresses, the number of locations validated by Rule 9 is 373 (94.2%) and the number of locations rectified by Rule 9 is 23 (5.8%), including 6 locations rectified from those located by *whois* information and 17 locations rectified from those inferred by Rule 1 through Rule 8. All of these 23 rectified mappings of IP addresses belong to the case of relocating their locations from one end of a direct inter-city link to the other. Thus the visual effect of the measured graph is not influenced much by the rectifications.

Additionally, in the totally 848 locations inferred by the first eight rules, although only 29 locations are inferred by Rule 8 (refer to Table 1), the locations of these 29 interfaces have much impact on the visual effect of the measured graph, as we stated in Subsection 4.3.

Table 1. Number of Interface Addresses to Which Empirical Inference Rules Are Applied

Probing source	Totally inferred	Rule 1	Rule 2	Rule 3	Rule 4		Rule 5	Rule 6	Rule 7	Rule 8
					$k=1$	$k=2$				
Union (distinct)	848	181	41	418	64	20	138	64	45	29

Table 2. Number of Interface Addresses That Are Located before and after Exercising Empirical Inference Rules

Probing source	Totally determined	Before exercising inference rules		After exercising inference rules	
		By domain names	By *whois*	By domain names and inference rules	By *whois*
Union (distinct)	2,563	163	2,400	974	1,589

The final visual layout of our measured topology is shown in Fig.2, and comparing Fig.1, Fig.2, and Fig.10 we may see that these empirical inference rules work effectively.

5.2 Validity of Empirical Inference Rules

One good way to check the effect of exercising these rules is to inspect the visual presentation of the measured topology. For the nationwide measured graph, we use two different presentation levels to check the exercising effect. The first presentation level includes all *prov-cities* and all *prov-capital-cities*, and the second presentation level further includes all *pref-cities*, as categorized in Subsection 3.1. Compared with the officially published graph, this second presentation level is in a fine granularity.

After exercising these empirical inference rules, the first presentation level of our measured CERNET topology is shown in Fig.2 and the second presentation level is shown in Fig.11. Comparing Fig.1 (before exercising empirical inference rules), Fig.2, Fig.10 (published by CERNET), and Fig.11, we are convinced that these empirical inference rules work very efficiently since most of backbone inter-city links in measured topology graphs are in conformity with those in Fig.10.

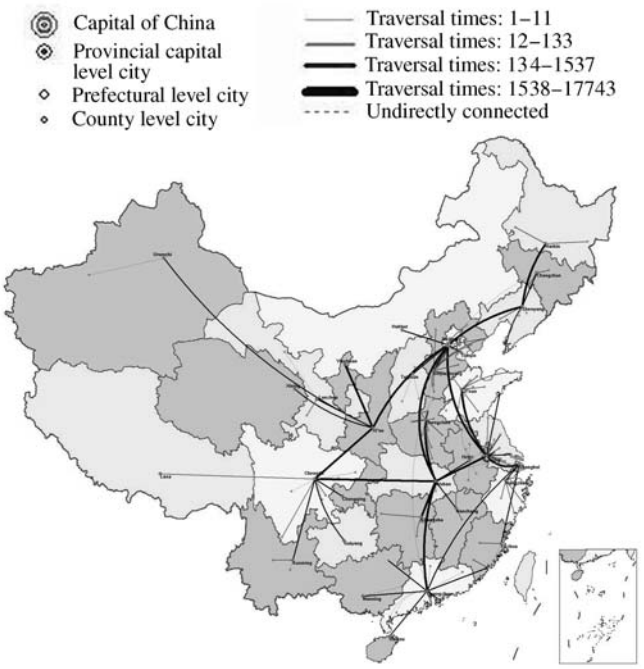

Fig.11. Measured topology graph at the 2nd presentation level after exercising our empirical inference rules.

Moreover, we select twelve inferred backbone router interfaces' IP addresses and request the operator of CERNET National Center for verifying the accuracy of their locations. The check result shows that 5 interfaces are inaccurately located by their DNS naming information, and in contrast our rules have made accurate mapping.

For those interfaces with domain names, we have also examined the conformity between the locations determined by DNS naming information and those inferred by our empirical inference rules. In the 163 interfaces with DNS naming information, 39.3% of them are also located by the first eight rules, and 33.7% of them are verified by Rule 9. It validates these empirical inference rules.

Furthermore, for those interfaces with no domain names, we have also checked the conformity between the locations inferred by these empirical inference rules and those determined by *whois* information. Comparing Fig.1 with Fig.10 we are convinced that there must have been some locations inaccurately determined by *whois* information. After we exercise these empirical inference rules, 811 locations formerly determined by *whois* information have been relocated or verified by these nine empirical inference rules. Among these 811 locations, 45.6% of them are different from those located by *whois* information, caused by lacking detailed location information in *whois* entries, and 54.4% of them are the same as those located by *whois* information. At the granularity of geographic city, these empirical inference rules behave inspiringly well.

6 Conclusion

Determining the geographic locations of IP addresses is not only important to the study of some routing problems and Internet topology measurement, but also important to the study of worm virus spreading scope and behavior, the service of location-aware applications, Web geo-spatial navigation, and so on. However, it is difficult to obtain accurate information of geographic location of an IP address.

This paper focuses on the study of determining the geographic location of routers and presents nine empirical inference rules. Experiments of applying these inference rules to the collected routing paths of CERNET network have shown the effectiveness of these inference rules in determining the geographic location of routers in ISP topology measurement. The geographic locations of routers may be located from paths collected from multiple vantage points with the help of cross-inference, yet they usually cannot be completely determined from merely single traced path when there is inadequate locating information that could be used, and these rules are based on the network deployment structure, routing principles, and economic constraints. In this regard, this paper opens up a new line of research in determining the geographic location of routers.

Different ISPs may have different deployment policies for their infrastructure topologies. Therefore, the type of inferred cities in Rule 3, Rule 7, and Rule 8(3) may be adjusted according to the corresponding deployment policies, and the relevant empirical inference rules for the corresponding ISPs will be obtained.

Acknowledgements We would like to thank Prof. Xing Li at the CERNET National Network Center for his help in verifying some IP addresses' geographic loca-

tions of backbone routers of CERNET, and the operator at the Network Information Center of Nanchang University for the help of verifying the relevant inter-city link. The authors are grateful to Associate Professor/Dr. Li-Xin Gao at the University of Massachusetts, Dr. Zhen-Ying Liu at the University of Houston, and the area editor and anonymous reviewers for providing detailed comments and constructive suggestions that helped improve the presentation of this paper.

References

[1] Spring N, Mahajan R, Wetherall D. Measuring ISP topologies with Rocketfuel. *ACM SIGCOMM Comp. Comm. Rev.* (*CCR*), 2002, 32(4): 133–145.

[2] Jiang Y, Fang B X, Hu M Z *et al.* A distributed architecture for Internet router level topology discovering systems. In *Proc. 4th Int. Conf. Parallel and Distributed Computing, Applications and Technologies* (*PDCAT'03*), Chengdu, China, Aug. 27–29, 2003, IEEE Press, pp.47–51.

[3] Huffaker B, Plummer D, Moore D *et al.* Topology discovery by active probing. In *Proc. 2002 Symp. Applications and the Internet Workshop* (*SAINT'02w*), Nara City, Japan, Jan. 28–Feb. 1, 2002, IEEE Press, pp.90–96.

[4] Jiang Y, Hu M Z, Fang B X *et al.* An Internet router level topology automatically discovering system. *Journal of China Institute of Communications*, Dec. 2002, 23(12): 54–62.

[5] Govindan R, Tangmunarunkit H. Heuristics for Internet map discovery. In *Proc. INFOCOM 2000,* Tel-Aviv, Israel, March 26–30, 2000, pp.1371–1380.

[6] Paxson V. Measurements and analysis of end-to-end Internet dynamics [Dissertation]. Lawrence Berkeley National Laboratory, UC, Berkeley, April 1997.

[7] Subramanian L, Padmanabhan V N, Katz R H. Geographic properties of Internet routing. In *Proc. the USENIX Ann. Technical Conf.,* Monterey, CA, 2002, pp.243–259.

[8] Padmanabhan V N, Subramanian L. An investigation of geographic mapping techniques for Internet hosts. *ACM SIGCOMM CCR,* 2001, 31(4): 173–185.

[9] Subramanian L. On inferring the geographic properties of the Internet [Thesis]. UC, Berkeley, 2002.

[10] Buyukkokten O, Cho J, Garcia-Molina H *et al.* Exploiting geographical location information of Web pages. In *Proc. ACM SIGMOD Workshop on the Web and Databases* (*WebDB'99*), Philadelphia, June 3–4, 1999, pp.91–96.

[11] Ding J, Gravano L, Shivakumar N. Computing geographical scopes of Web resources. In *Proc. VLDB 2000,* Cairo, Egypt, Sept. 10–14, 2000, pp.545–556.

[12] McCurley K S. Geo-spatial mapping and navigation of the Web. In *Proc. 10th Int. WWW Conf.,* Hong Kong, May 1–5, 2001, pp.221–229.

[13] Jiang Y, Fang B X, Hu M Z. Techniques in mapping router-level Internet topology from multiple vantage points. In *Lecture Notes in Computer Science 3320,* Liew K M, Shen H, See S *et al.* (eds.), Springer-Verlag, 2004, pp.410–415.

[14] Wang F, Gao L X. On Inferring and characterizing Internet routing policies. In *Proc. 2003 ACM SIGCOMM Internet Measurement Conf.* (*IMC*), Florida, 2003, pp.15–26.

[15] Lakhina A, Byers J W, Crovella M, Matta I. On the geographic location of Internet resources (Abstract). In *Proc. 2002 ACM SIGCOMM Internet Measurement Workshop* (*IMW*), Marseille, France, Nov. 6–8, 2002, pp.249–250.

[16] Padmanabhan V N, Subramanian L. Determining the geographic location of Internet hosts. *ACM SIGMETRICS Performance Evaluation Review*, June 2001, 29(1): 324–325.

[17] Periakaruppan R, Nemeth E. *GTrace* – A graphical traceroute tool. In *Proc. 13th LISA Systems Administration Conf.* (*USENIX LISA'99*), Seattle, Nov.7–12, 1999, pp.69–78.

[18] Moore D, Periakaruppan R, Donohoe J *et al.* Where in the world is netgeo.caida.org? In *Proc. 10th Ann. Internet Society Conf.* (*INET 2000*), Yokohama, Japan, 2000. (poster) http://www.caida.org/outreach/papers/2000/inet_netgeo/

[19] Raz U. Finding a host's geographical location. http://www.private.org.il/IP2geo.html

[20] Davis C. DNS LOC: Geo-enabling the domain name system. http://www.ckdhr.com/dns-loc/

[21] Cyberspace directory: Mapping the Internet. http://www.cybergeography.org/mapping.html

[22] Zhao Y X, Yin X, Wu J P. Problems in the information dissemination of the Internet routing. *J. Comput. Sci. & Technol.*, Mar. 2003, 18(2): 139–152.

[23] CERNET National Topology Graph. http://www.edu.cn/20010101/21585.shtml.

Yu Jiang received the B.S. degree in computer software from the Heilongjiang University, Harbin, China, in 1990 and the M.S. degree in computer software and theory from the Harbin Institute of Technology (HIT), Harbin, China, in 1999. From August 1990 to August 1999, he worked on computer applications for the Statistics Bureau of Heilongjiang Provincial Government. At present, he is with the professional title of senior electronic engineer, and he is currently working toward the Ph.D. degree at the HIT. His current research interests are in the areas of Internet measurement and distributed network computing.

Bin-Xing Fang received the B.S. degree in computer applications from the HIT, Harbin, China, in 1981 and the M.S. degree in computer architecture from Tsinghua University, Beijing, China, in 1984 and the Ph.D. degree in computer architecture from the HIT, Harbin, China, in 1989. He conducted post-doctoral research work at the National University of Defense Technology, Changsha, China. He is a part time professor in the School of Computer Science and Technology at the HIT, Harbin, China, and the head of the National Computer Network Emergency Response Technical Team/Coordination Center, Beijing, China. His research interests are in the areas of computer network and information security, distributed network computing, and high performance computer architecture.

Ming-Zeng Hu graduated from the HIT, Harbin, China, in 1958. He is a professor in the School of Computer Science and Technology at the HIT, Harbin, China. His research interests are in the areas of high performance computer architecture, computer information security, and parallel computing.

Xiang Cui received the B.S. degree and the M.S. degree in computer architecture from the HIT, Harbin, China, in 2001 and in 2003, respectively. He works presently for the National Computer Network Emergency Response Technical Team/ Coordination Center, Beijing, China. His current research interests are in the areas of computer virus inspection and computer information security.