Topology Mapping and Geolocating for China's Internet

Article *in* IEEE Transactions on Parallel and Distributed Systems · March 2012

DOI: 10.1109/INFCOM.2012.6195646

Ye Tian
University of Science and Technology of China

China's Internet: Topology Mapping and Geolocating

Ye Tian[‡], Ratan Dey[§], Yong Liu[§], Keith W. Ross[§]
[‡]Anhui Key Lab of High Performance Computing, Univ. of Science and Technology of China, Hefei, Anhui 230026, China
[§]Polytechnic Institute of New York University, Brooklyn, NY 11201, USA
Email: yetian@ustc.edu.cn, ratan@cis.poly.edu, yongliu@poly.edu, ross@poly.edu

Abstract—**We perform a large-scale topology mapping and geolocation study for China's Internet. To overcome the limited number of Chinese PlanetLab nodes and looking glass servers, we leverage several unique features in China's Internet, including the hierarchical structure of the major ISPs and the abundance of IDC datacenters. Using only 15 vantage points, we design a traceroute scheme that finds significantly more interfaces and links than iPlane with significantly fewer traceroute probes.**

We then consider the problem of geolocating router interfaces and end hosts in China. When examining three well-known Chinese geoIP databases, we observe frequent occurrences of null replies and erroneous entries, suggesting that there is significant room for improvement. We develop a heuristic for clustering the interface topology of a hierarchical ISP, and then apply the heuristic to the major Chinese ISPs. We show that the clustering heuristic can geolocate router interfaces with significantly more detail and accuracy than can the existing geoIP databases in isolation, and the resulting clusters expose the major ISPs' provincial structure. Finally, using the clustering heuristic, we propose a methodology for improving commercial geoIP databases.

I. Introduction

China[1] is the country with the largest number of Internet users and the second largest IP address space [1]. Nevertheless, China's Internet has received relatively little attention in the measurement community to date. This is perhaps because China's Internet lacks the infrastructure and resources that are essential for large-scale Internet measurement studies, such as those carried out in Rocketfuel [2] and iPlane [3]. For example, China has few PlanetLab nodes and looking glass servers, which are important infrastructure components for large-scale Internet measurement studies. Moreover, whereas many routers outside of China have names from which geolocation can be inferred, few router interfaces have names in China.

Nevertheless, China's Internet is complex and has its unique structural features, which makes it very different from the Internet in US and Europe. Unlike the US and Europe, China has a very simple AS-topology with few Chinese ASes [4]. Moreover, both of two major ISPs in China each have one giant AS that not only includes a national backbone network, but also includes regional networks in many provinces as well as residential networks. *Therefore, China's topology is largely shaped by the internal structure of its giant ASes rather than by its AS-topology.*

Of particular interest is geolocation services for China's Internet. More and more online businesses and services – including targeted advertising, spam filtering, and fraud prevention – are based on geolocation of IP addresses. Commercial geoIP databases for China and elsewhere typically incorporate multiple information sources, including information directly from ISPs, DNS reverse lookups, and end user inputs. As we will show in this paper, existing commercial geoIP databases for Chinese IP addresses have many incomplete and erroneous entries, particularly for router interfaces.

In this paper, we carry out a large-scale topology mapping and geolocation study for China's Internet. To overcome the small number of Chinese PlanetLab nodes, looking glass servers, and router interfaces with geographical names, we leverage several unique features in China's Internet, including the hierarchical structure of the major ISPs and the abundance of IDC datacenters. The contributions of this paper are as follows:

- We find that existing measurement practices do not adequately cover China's Internet. We develop two techniques, namely *nested IP block partitioning* and *collaborative tracerouting*, which allow us to perform a comprehensive and efficient traceroute measurement study of China's Internet using only 15 internal vantage points. In particular, our approach discovers significantly more interfaces and links than iPlane with significantly fewer traceroute probes.
- Using the IP addresses obtained from our traceroute measurements, we examine three well-known Chinese geoIP databases and MaxMind. We find that the three Chinese geoIP databases are only moderately accurate for end host geolocating, and substantially less accurate for router interfaces. In particular, we observe frequent occurrences of null replies and erroneous entries, suggesting that there is significant room for improvement.
- With the goal of accurately geolocating routers in China, we develop a heuristic for clustering the interface topology of a hierarchical ISP, so that each cluster is a connected component within a city. We then apply the heuristic to the major Chinese ISPs, leveraging the interface topologies derived from our traceroute measurements as well as the existing Chinese geolocation services. We show that this clustering heuristic can geolocate router interfaces with significantly more detailed location information than the existing geoIP databases in isolation.
- We analyze the clusters generated by our clustering heuristic. We show that they expose several characteristics

[1] By China we mean Mainland China.

of the Chinese Internet, including recent mergers of ISPs. We observe the provincial capital cities are not only government centers but are also hubs in the ISPs' networks.

- Using the geo-clustering heuristic, we propose a methodology for improving commercial geoIP databases. By evaluating with datacenter landmarks, we show that our approach is able to provide more detailed and accurate location information as compared with the original geoIP database. By improving on the best geoIP databases in China, we are currently providing the most accurate geolocation service for China's Internet.

II. Overview of China's Internet

Before presenting our methodologies for mapping and geolocating China's Internet, it is useful to briefly overview China's Internet. The two largest ISPs in China are China Telecom (a.k.a. ChinaNet and henceforth referred to as Telecom) and China Unicom (henceforth referred to as Unicom)[2]. Both Telecom and Unicom have high-performance national backbone networks, connecting regional and residential networks in China's provinces and major cities; and both also provide high-performance connections to the Internet outside of China [1]. Both Telecom and Unicom also have their own networks in many provinces and cities in China, and also provide access directly to end users. The other commercial ISPs in China are much smaller than Telecom and Unicom; they generally rely on Telecom/Unicom's backbone networks for accessing services connected to Telecom/Unicom, and for accessing services on the international Internet.

In addition to Telecom and Unicom, CERNET[3] is also a major ISP in China. As an academic network that connects the universities and research institutes all over China (analogous to Internet2 in the USA), CERNET is largely independent with its own national backbone and peers with many international commercial and academic networks.

III. Traceroute Measurement

Traceroute is one of the most fundamental measurement tools for studying the Internet. Unfortunately, existing large-scale traceroute measurement practices, such as iPlane [3] and CAIDA/Ark [5], do not satisfactorily cover China's Internet. These projects use very few vantage points within China: only two PlanetLab nodes from China are used in iPlane and only one Chinese monitor is used in Ark. As a result, these two projects use vantage points from outside China to collect most of their Chinese traceroute path segments. Moreover, it is well known that Telecom and Unicom have most of the international Internet connections in China [1]; therefore most of the traceroute probes originating from outside of China will enter China through a small number of ASes in Telecom and Unicom. *Thus, for traceroutes originating from outside of China, they are likely to follow similar paths when traversing China's Internet, thereby not revealing many diverse interfaces and links*. For comprehensively mapping China's Internet, we must therefore use vantage points located in China.

We face two challenges when attempting to map China's Internet with traceroute. The first is to identify a set of target IP addresses that is sufficiently, but not overly, dense within the Chinese Internet. Large-scale traceroute measurement studies (e.g., [3] and [5]) often use CIDR IP blocks from public BGP snapshots (e.g., from Oregon Routeviews [6] and RIPE RIS [7]); the blocks are used to partition the IP space, and then one address is selected from each block as the traceroute targets. However, there is no operational public BGP router in China's Internet [8]; therefore, we can only gather Chinese blocks from routers that are outside of China. Unfortunately, these blocks are generally too coarse for topology mapping, as they are likely to have been aggregated by the border routers in China's Internet. To establish this claim, we have downloaded eight BGP snapshots from different routers in Oregon Routeviews and RIPE RIS. (The routers are in USA, Europe, and Japan.) From these routing tables, we have observed many large blocks (e.g., blocks with prefix lengths smaller than 20, 18, and so on). For statistics on these Chinese blocks, see our technical report [9].

The other challenge is efficiency, that is, devising a traceroute strategy that sufficiently covers the Chinese Internet without overly burdening the traceroute sources (vantage points). iPlane and Ark spread their workload over hundreds of vantage points. In our traceroute measurements, we only use stable vantage points from within China, for which we have only identified 15 (7 PlanetLab nodes and 8 web-based traceroute servers). If we use iPlane's or Ark's probing strategy, we would overload our 15 vantage points with too many tasks. To address these two challenges, we devise two techniques, namely, *nested IP block partitioning* and *collaborative tracerouting*.

A. Nested IP Block Partitioning

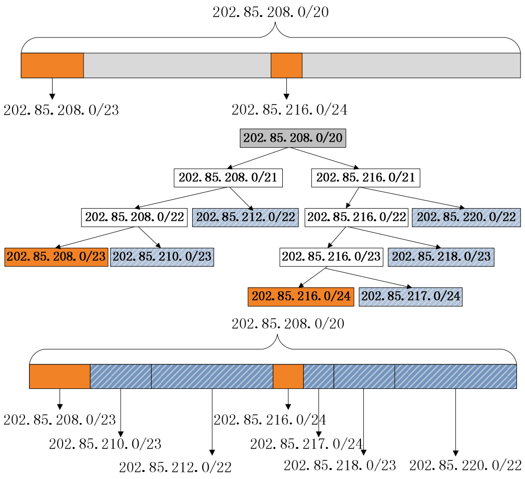

Fig. 1. Nested-block Partitioning

We need to partition the large Chinese IP address space, and then choose one IP address from each set in the partition

[2] Here we are referring to the current China Unicom, which merged with China Netcom (a.k.a. CNCGroup) in 2008.

[3] China Eduction and Research Network

as a traceroute target. For the partitioning, a simple approach is to evenly divide the large blocks obtained from the public BGP tables. However, taking a close look at these blocks, we find that *block nesting* [10], where a block from one BGP routing table entry resides in another block from a different entry, is very common; moreover, there are often several levels of nesting. An example of nested IP blocks is shown in the top graph in Fig. 1. In the graph, three blocks are obtained from BGP tables, i.e., 202.85.208.0/20, 202.85.208.0/23, and 202.85.216.0/24, where the latter two blocks are nested in the first one. Clearly, the smaller nested blocks suggest the existence of different subnets, as they appear as separate entries in the routing tables. If we set the granularity of the traceroute probing up to prefix /22, then for the block 202.85.208.0/20, we would obtain four equal-sized /22 blocks, but the smaller nested blocks would be masked. On the other hand, evenly dividing 202.85.208.0/20 into /24 blocks results in 16 blocks, which may overly increase the workload of the measurement.

We design a tree-based method to partition the Chinese IP address space with a minimal number of blocks while preserving the nested blocks obtained from the BGP tables. The blocks from the BGP tables are nodes in trees. We consider a block encompassing other blocks as the root of a binary tree, and all the nested blocks as leaves. With this tree the problem becomes: given the root node and a number of leaf nodes, construct a binary tree with the fewest leaves. After the tree is obtained, we use all the blocks corresponding to the leaf nodes (including the original nested blocks) to replace the root block. For example, in the case mentioned above, the corresponding binary tree is shown in the middle graph in Fig. 1, and we use seven blocks to replace the original /20 block, as shown in the bottom graph in Fig. 1. After the partitioning, we further evenly divide any blocks that are larger than our granularity, while reserving the smaller blocks for traceroute probing. For the example in Fig. 1, 7 blocks are probed instead of the 4 or 16 blocks that would be generated by evenly dividing. Thus, with nested-block partitioning, we can fully exploit the small nested blocks, suggesting different subnets, without naively dividing all the large blocks, which would geometrically increase the probing workload.

B. Collaborative Tracerouting

Ark and iPlane apply different strategies to reduce the workload (when probing the entire Internet): Ark groups its vantage points into teams, and each team only probes a subset of the targets; In iPlane, IP blocks from BGP snapshots with similar AS paths are further combined to reduce the workload [11]. However, we cannot apply either Ark's or iPlane's strategies for two reasons: (i) we have only 15 vantage points to spread the workload over; and (ii) we need to divide IP blocks from BGP snapshots rather than cluster them. Even after the nested-block partitioning, as described in Section III-A, there are still 223,714 CIDR blocks in China to be tracerouted. It is impractical to probe each block from each of the vantage points. A recent study [12] shows that there are many redundant probes in Ark and iPlane. We propose a mechanism for having the vantage points collaboratively and dynamically determine their traceroute targets, thereby avoiding redundant probes.

In our measurement, the IP blocks obtained in Section III-A (which partition the Chinese IP space) are the basic probe units. For each block, we always use its second IP address (i.e., a.b.c.1) as the traceroute target, as such addresses are usually used for gateways and are, thus, more likely to respond to a probe. In our collaborative tracerouting scheme, a vantage point actively uses the results of its previous probes and other vantage points' probes to avoid redundant probes. Specifically, each vantage point keeps a set, $reach\_set$, of all the addresses the vantage point has observed during its previous probes; and each IP block keeps a set, $source\_set$, containing all the IP addresses that lead to this block from previous traceroutes from all the vantage points. When a vantage point v encounters an IP block B it has not probed before, it examines v's $reach\_set$ and B's $source\_set$; if the two sets overlap, then an interface path can be found from v to the block B from previous traceroutes, so the vantage point v doesn't probe the block B.

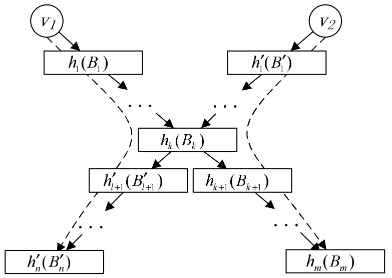

Fig. 2. An example of collaborative tracerouting

As an example, suppose a vantage point v_1 probes a target with the traceroute path

$$h_1, ..., h_k, h_{k+1}, ..., h_m$$

where the interfaces are in the blocks B_1, ..., B_k, B_{k+1}, ..., B_m, as shown in Fig. 2. v_1 inserts all the interface IP addresses it has reached, i.e., h_1, ..., h_m, into its $reach\_set$. For each interface IP, the corresponding block inserts all the IPs preceding it along the path into its $source\_set$. For example, h_1, ..., h_{m-1} are inserted into B_m's $source\_set$. Clearly, after this probing, v_1 can skip B_1, ..., B_{m-1} in future measurements, as v_1's $reach\_set$ overlaps with the $source\_set$ for each of these blocks. Moreover, suppose another vantage point v_2 has a traceroute path

$$h'_1, ..., h_k, h'_{l+1}, ..., h'_n$$

that traverses the blocks of B'_1, ..., B_k, B'_{l+1}, ..., B'_n. As a result of this probe, h_k will be included in $source_set$s of B'_{l+1}, ..., B'_n, which means that v_1 can skip these blocks as an interface path has already been found from v_1 to them via h_k, as shown in Fig. 2. (Shown as the dotted line in the left of the figure.) Similarly, v_2 can also skip the blocks of B_{k+1}, ..., B_m. (Shown as the dotted line in the right of the figure.)

C. Measurement Results

TABLE I
TRACEROUTE MEASUREMENT RESULTS

	iPlane (1 day)	iPlane (2 days)	cTrace	Both
Traceroutes	1,244,667	2,381,482	106,580	
Interfaces	17,308	17,761	71,047	10,023
Links	76,120	82,791	146,542	27,735

Using nested-block partitioning and collaborative tracerouting, we perform a traceroute measurement on China's Internet with 15 vantage points (from 9 different cities and in 4 ISPs) in China. We applied the nested-block partitioning algorithm on the IP blocks from 8 BGP snapshots and further divided them to prefix /22 blocks for obtaining the target addresses. The measurement was performed from Dec. 12, 2010 to Jan. 2, 2011. We also downloaded iPlane's traceroute data on Dec. 19 and Dec. 20, 2010 for comparison. For each path in iPlane, we extract the segment that is within China's Internet. We use a method similar to [4] to decide whether an address is in China by examining the AS it belongs to.

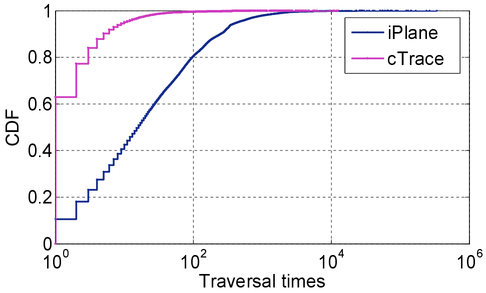

Fig. 3. Number of times the links are visited

Table I compares the iPlane data with our measurement results (referred to as *cTrace*). For iPlane, we present the results for both one and two days of measurement. As compared with iPlane, our approach employs only 5% of the number of traceroute probes but finds four times as many interfaces and twice as many interface links. This experiment therefore shows that using vantage points in China is much more efficient in exposing China's Internet, and collaborative tracerouting can effectively eliminate redundant probes. To further demonstrate our point, we plot the distributions of the number of times the links are visited in iPlane (over 2 days) and in cTrace in Fig. 3. In two days, iPlane visited some links thousands of times, even though most of its vantage points are far away from these Chinese links.

As iPlane contains interfaces and links that are located on the border of China's Internet, which cTrace may fail to discover by using vantage points within China, we therefore combine cTrace with the 2-day iPlane data, and use the combined data for further study in this paper.

In summary, we perform a traceroute measurement with as few as 15 vantage points on China's Internet. As compared to existing large-scale traceroute measurements, our scheme not only reveals a much larger number of Chinese links and interfaces, but also uses significantly fewer traceroute probes.

IV. GEOLOCATION SERVICES ON CHINA'S INTERNET

One goal of this paper is to develop a methodology for accurately geolocating Chinese IP addresses for both end hosts and router interfaces. In this Section, we briefly examine the geolocation services currently available for China's Internet. In the subsequent section, we will develop methodologies to improve these services.

We consider four geoIP databases in this study, namely, IP138 [13], QQWry [14], IPcn [15], and MaxMind [16]. The first three are Chinese databases well-known in the Chinese Internet community, whereas MaxMind is a leading global geolocation service provider. The locations returned by these databases generally have two levels: the province level and the city level. For the directly-controlled municipalities of Beijing, Shanghai, Tianjin and Chongqing, we consider them as both provinces and cities.

TABLE II
NULL REPLY RATIOS FOR ADDRESSES FROM TRACEROUTES AND FROM XUNLEI PEERS

	IP138	QQWry	IPcn	MaxMind
Province (traceroute)	0.105	0.074	0.108	0.186
City (traceroute)	0.240	0.212	0.280	0.227
Province (Xunlei)	0.011	0.004	0.021	0.161
City (Xunlei)	0.153	0.137	0.178	0.224

We consider null-reply ratios for each database. A database's null-reply ratio is defined as the fraction of the cases for which the database fails to provide location information [17]. We use the $78,229$ IP addresses from the combined traceroute data to examine the geoIP databases. The second and third rows in Table II show the null-reply ratios for the four databases at the province and city levels. We can see that each database frequently returns null replies, particularly for the city-level location information.

Two types of IP addresses are included in our traceroute data: router interface addresses and end host addresses. To gain further insight into the databases' performance for different types of addresses, we randomly selected 2,000 peer addresses from Xunlei DHT network [18] (a popular Chinese P2P download acceleration application) and fed these end host addresses to the geoIP databases. The fourth and fifth rows in Table II show the null-reply ratios on Xunlei peers. Comparing with the ratios for traceroute addresses, we can see that except for MaxMind, the three Chinese databases have fewer null replies for Xunlei peers, suggesting that they cover end host IP addresses better than router interface addresses.

In summary, we find that the three Chinese geoIP databases are moderately accurate for end host geolocating, and substantially less accurate for router interfaces. In particular, we observe frequent occurrences of null replies, suggesting that there is significant room for improvement.

V. Geolocating the Interface Topology

With the combined traceroute data obtained in Section III, we have obtained a separate interface topology for Telecom, Unicom and CERNET. Each of these interface topologies can be viewed as a directed graph: Each interface (IP address) forms a vertex, and each pair of successive interfaces from the traceroutes forms a directed edge. In this section, we seek to geolocate the interfaces in each of the three interface topologies. In many countries, router interfaces are often assigned names that indicate the interface's location. In such cases, the location of an interface can be determined by simply performing a reverse DNS lookup on the corresponding IP address. In China, however, very few router interfaces have names. We therefore must develop an alternative approach for geolocating the router interfaces. We develop a clustering approach, as described subsequently.

For a given interface topology T, we say a set of router interfaces S forms a *cluster* if (a) all the interfaces in S belong to the same city, and (b) the subgraph of T induced by S is weakly connected. We further say that a cluster S is a *maximal cluster* if it is not possible to create a larger cluster by adding more interfaces to it. Our goal is to determine the maximal clusters in each of three interface topologies. Note that a city could have more than one maximal cluster, for example, it could have two maximal clusters which do not have a direct link between them, but which have an indirect path between them via another city.

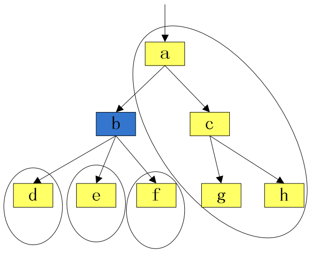

Fig. 4. Erroneous clusters example

A naive method to create the clusters is to directly use the city information provided by the geoIP databases on face value. However, this naive approach leads to a large number of small and disconnected erroneous clusters due to missing and erroneous entries in the geoIP databases. Fig. 4 provides an example, with boxes representing the interfaces and arrows representing the links. All the interfaces on the graph are at the same location, and should be included in one cluster. However, if interface b's location from geoIP database is wrong or missing, four instead of one cluster is formed, as shown in the figure. This example shows that a few errors in geoIP databases will cause many clusters to be erroneously formed. On the other hand, by combining the information in the geoIP databases with the topological information obtained from the traceroutes, it may be possible for us to identify the errors in geoIP databases and determine the interfaces' real locations. For example, for interface b in Fig. 4, as all the interfaces adjacent to it are at the same location, we can conclude that b's database location is likely incorrect and b is likely located at the same location as all the other interfaces on the graph. Inspired by this observation, we propose a heuristic for accurately determining the maximal clusters in each of the three interface topologies.

A. Geo-Clustering Heuristic

Geolocating an interface network using a partially accurate geoIP database is a challenging problem for an arbitrary interface topology. Fortunately, the major Chinese ISPs have a hierarchical structure, which makes the problem more tractable. We have developed a heuristic that could be used for any ISP with a hierarchical structure (not just Chinese ISPs). Due to space constraints, we only provide a summary of the heuristic here; for further details, please see [9].

For each of these ISPs, using the traceroute data, we first obtain an interface topology that expands from the ISP's backbone network to the traceroute targets in that ISP. For each of the resulting interface topologies, and for each of the databases, we infer the interfaces' city-level locations and cluster them through four steps. We refer to this four-step heuristic as the *geo-clustering* heuristic. Note that each ISP and database combination produces a different set of geo-clusters.

In Step 1, we select the interfaces that are at the edge of the interface topology, and form singleton clusters for each of them. We consider an interface to be at the edge of the topology if either (1) it has no out-linked interfaces, or (2) from each of its out-linked interfaces, there exists a path returning back to itself. For each singleton cluster formed at this step, we use the interface's DB location as the cluster's location. (The heuristic will possibly assign a different location to the interface in Step 4.)

Step 2 consists of a sequence of rounds. At the beginning of any round, some of the interfaces in the topology have been clustered, whereas the remaining remain to be clustered. At the beginning of each round, we select the unclustered interfaces that are one step closer to the backbone network as candidates for clustering. For each candidate, all it's out-linked interfaces use their DB or cluster locations to *vote* to decide the candidate's cluster location. If there is no majority winner from the voting, we use the candidate's DB location as its cluster location (see [9] for specific details) . After the candidate is assigned a cluster location, it merges all the clusters it links to that have the same location to form a new large cluster. After all the candidate interfaces are processed, the round is finished. We then continue with the next round by selecting new candidates. However, for a candidate interface, if more than one province appears in the voting, it is likely that this interface belongs to a backbone network. In this case, we abort the voting-based inference without forming or merging any clusters, and move on to the next candidate. The Step 2 heuristic stops when we can't form or merge any clusters during a round.

After Step 2 stops (without forming or merging any clusters in a round), Step 3 begins, merging interfaces that cannot be handled by Step 2 with new location inference rules. Step 3 in the heuristic works similarly as Step 2 by first selecting a set of candidate interfaces, inferring their cluster locations, and merging the clusters with the same cluster location. We use the same method as in Step 2 to select a candidate. However, unlike Step 2, where candidate interfaces are on routers in residential or provincial networks, in Step 3, nearly all the candidate interfaces are on backbone routers, which usually connect to many routers at different locations. Here we apply four different rules to infer an interface's cluster location by *combining the link delay* with the voting based approach [9].

After applying Steps 2 and 3, all the interfaces in the topology are clustered. Careful examination on the resulting clusters shows that for nearly all the cities, there are one or two large clusters containing most of the interfaces, as well as a number of singleton or small clusters. The objective of Step 4 is to merge these singleton and small clusters into larger ones. Here we categorize the clusters as mergeable *small clusters* and *large clusters* according to their sizes. For a small cluster, if it is only connected to one large cluster, then the location information given in the database for the small cluster is likely to be wrong; we therefore merge it into the large cluster, regardless of its original cluster location.

B. Geo-Clusters

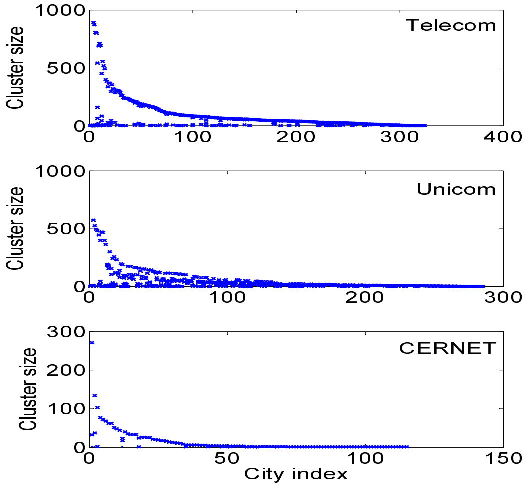

Fig. 5. Distribution of the sizes of geo-clusters across cities in three major ISPs

We applied the geo-clustering heuristic on the Telecom, Unicom, and CERNET interface topologies using each of three geoIP databases. By geo-clustering, we can group most of the interfaces on the interface topology into clusters with detailed city-level location information. We refer to a cluster with a city-level location as a *geo-cluster*. For example, for the Telecom's interface topology using the geoIP database of IP138, after four steps, 532 of the final geo-clusters containing 98.2% of the total interfaces have been formed. (The remaining clusters are singleton clusters for which the heuristic did not assign to a city since there was no clear majority winner in the voting.) Similar results were observed using the two other geoIP databases and for the two other ISPs. We omit them due to lack of space.

By examining the 532 geo-clusters obtained on Telecom's interface topology, we find they are located in 324 different cities, which are nearly all the cities in China. We show the sizes of the geo-clusters for each city for Telecom, Unicom, and CERNET in Fig. 5, where the x-axis is the city index, the y-axis is the cluster size, and each point on the figure corresponds to a geo-cluster. For each ISP, the cities are indexed according to the total number of IP addresses across all geo-clusters in the city. From Telecom and CERNET's figures, we can see that for many cities, there is only one geo-cluster. For a small fraction of the cities, multiple clusters are found, with one cluster containing the majority of the interfaces. There are two possible reasons for multiple clusters in a city: (i) the ISP has multiple networks serving different purposes in that city; and more likely (ii) some of the singleton and small clusters cannot be merged into large clusters in step four. Note that the Unicom's geo-cluster distribution is distinctly different from those of Telecom and CERNET. In particular, for Unicom in many cities there are two large geo-clusters of comparable size. Our heuristic is consistent with the fact that in 2008 Unicom merged with China Netcom, which used to be the second largest ISP in China. As a result, in many cities we can observe one large geo-cluster for the former Unicom network, and another large geo-cluster for the former Netcom network.

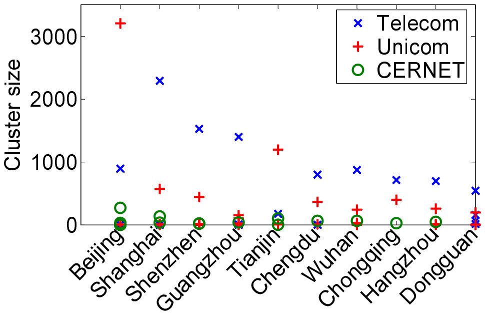

Fig. 6. Geo-clusters in the top-10 cities

Fig. 6 shows the geo-clusters of the top 10 cities. From the figure we an see that each top-10 city has only one major cluster per ISP (including Unicom for these cities); moreover, Unicom has much larger geo-clusters than Telecom in Beijing and Tianjin, located northern China, while Telecom has larger geo-clusters in other cities.

C. The Hierarchical Structure

TABLE III
STATISTICS OF INTER-CLUSTER LINKS

	Same province		Different province		
	Cap.	Other	2Cap.	Cap.	Other
Telecom	3,236	2,097	169	283	42
Unicom	1,504	1,281	199	25	0
CERNET	69	1	181	21	0

We now study the internal structure of each ISP. Table III categorizes inter-cluster links based on the locations of the two endpoints of the links. In this table we have removed the links with both endpoints on the backbone. The first and the second columns are for the intra-province links, where the first column is for links between the capital city and another non-capital city in that province, and the second column is for links between two non-capital cities. The third through fifth columns are for inter-province links: links between the capital cities of two different provinces (column 3), links with only one endpoint at a capital city (column 4), and links between two non-capital cities in different provinces (column 5). From the table we can see that for Telecom and Unicom, there are many intra-province links, and more than half of them are between capital and non-capital cities. There are relatively few inter-province links, and the majority of them connect to at least one capital city. *We can therefore conclude that the major Chinese ISPs are highly hierarchical following China's provincial organization, and that the provincial capital cities are not only government centers but also hubs in the ISPs' networks. This strikingly contrasts with flattening trends in the international Internet [19] [20].*

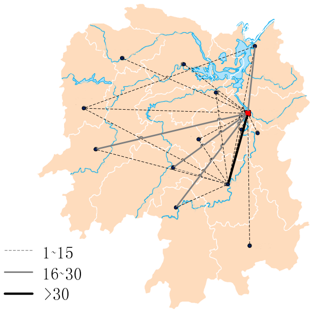

Fig. 7. Cluster topology for Hunan province

As an example, Fig. 7 shows the Telecom geo-cluster topology of all the cities in Hunan province, where the width of the edge between two cities represents the number of distinct interface links between geo-clusters located at the two cities. We can see that the topology is strongly centered around the capital city of Changsha, as shown by the red square on the graph.

D. Locating Interfaces with Null Replies

TABLE IV
NULL REPLY RATIOS

	DB province	Cluster province	DB city	Cluster city
IP138	7.7%	0.99%	21.7%	1.51%
QQWry	6.2%	1.00%	18.7%	1.64%
IPcn	8.0%	0.93%	26.4%	1.66%

After geo-clustering, each interface in an ISP's interface topology has two locations: the geoIP database location and its cluster location (with the clusters derived from the same database). In this section, we show that the cluster locations are significantly more complete and accurate.

We first examine the completeness by comparing the null reply ratios. In this comparison, all the IP addresses of the interfaces on Telecom, Unicom, and CERNET's interface topologies are included. Table IV shows the null reply ratios at the province and the city levels for both DB and cluster locations. Observe that the ratios for cluster locations are much smaller than those for the DB locations. The geoIP services give a high-level of null replies because many router addresses do not have city-level or province-level locations in the database. However, the cluster locations for many of these router interfaces have been inferred at the city level (by the voting in Steps 2 and 3 and by the merging in Step 4).

TABLE V
NUMBER OF THE INTERFACES THAT HAVE CONSISTENT LOCATIONS

	Telecom	Unicom	CERNET
Total	38,181	24,781	1,798
3DB identical	25,625 (67.1%)	15,794 (63.7%)	1,343 (74.7%)
3Cluster identical	35,376 (92.7%)	21,938 (88.5%)	1,602 (89.1%)

We now examine the accuracies of the DB and cluster locations. Unfortunately, given the lack of landmarks for router interfaces, it is not possible to say with 100% certainty whether a geoIP database location or a cluster location is correct. (However, we will be able to use landmarks in Section 6 when we study end host geolocation.) Instead, here we use cross validation to support our claim that clustering approach is substantially more accurate than the geoIP databases for router interfaces.

For an interface, if the locations from the three databases are the same, it is likely that the location is correct; if, however, all three databases do not give the same location, then we have a low level of confidence on the location information. Similarly, using the three sets of geo-clusters based on the three different geoIP databases, we can cross-validate the cluster locations. Table V shows for each of the three ISPs, the number of the addresses that have consistent locations for the two approaches. We see that the three geoIP databases agree only for 66.0% of the interfaces (average across the three ISPs), but after applying the geo-clustering heuristic, as many as 91.0% interfaces have the same cluster locations.

In summary, for a hierarchical interface topology, we propose a heuristic to geolocate the interfaces from traceroute measurements by forming geo-clusters. We apply the heuristic to China's Internet and provide evidence that resulting large geo-clusters are essentially the maximal clusters. The geo-clusters clearly expose China's hierarchical structure down to the city level. In addition, we show that our heuristic can geolocate router addresses with more detailed and accurate location information than can existing geoIP databases.

VI. IMPROVING GEOLOCATION SERVICES WITH GEO-CLUSTERS

In this section, we develop a methodology for accurately geolocating arbitrary Chinese IP addresses, including host interfaces. Our goal here is to provide a significant improvement over the existing Chinese geoIP databases.

A. Geolocating an Arbitrary IP Address

Our methodology relies on the geo-clustering heuristic described in Section V-A. For a given IP address p that we wish to geolocate, we first determine the ISP to which it belongs (e.g., by first determining the AS to which it belongs from BGP tables). This ISP has an interface topology, say T, which we obtained from our traceroute data.

To apply the geolocating algorithm in Section V to an arbitrary IP address p, we need to first augment T to reach p. This requires us to conduct additional traceroute probes. We choose a subset of existing vantage points, each of which keeps a queue of targets to be probed. For initialization, we put p into the target queue of each vantage point. Then vantage points conduct traceroute probes by working through their target queues: at each step, each vantage point dequeues a target t and performs a traceroute to t. Along the traceroute path, if there exists an interface i between T and t for which there is no anonymous router between T and i, we insert i into the target queues of all the vantage points (except for the one that just returned this path). This process continues until the queues of all the vantage points become empty.

We then use the new traceroutes to augment the topology T to create a new interface topology T'. Applying the geo-clustering heuristic to the new augmented topology T', we obtain a new set of geo-clusters. The location of p is then determined from these new geo-clusters using one of the following three cases:

- Case 1: p is in the topology T' and therefore is in one of the geo-clusters. In this case, we simply set p's location to the location of the cluster that encompasses it.
- Case 2: p can be reached by at least one traceroute path, but p is not in T' (due to the occurrence of anonymous routers in the traceroute paths). In this case, we find the geo-cluster that is closest to p among all the traceroute paths, which we refer to as the *last-hop geo-cluster*. If the distance between the last-hop geo-cluster and p is no larger than a threshold (2 hops in our evaluation), we set p's location to the location of the last-hop geo-cluster.
- Case 3: If we don't set p's location in Case 1 and 2, the location from the geoIP database is used.

B. Evaluation

1) Collecting Landmarks: We use a number of landmarks as the ground truth for evaluating the accuracies of the geoIP databases and of our methodology. For collecting landmarks, we leverage the numerous Internet datacenters (IDC) located in many cities in China. We skip our methodology of collecting landmarks here for space reason, interested readers can refer to our technical report [9]. We have successfully collected 305 landmarks – 199 on Telecom and 106 on Unicom – with their ground-truth locations detailed to the city level.

2) Evaluation Results: We use ten vantage points located in seven different cities to geolocate the 305 landmarks. Our methodology requires us to probe a few additional addresses for each landmark to extend the interface topology. For each landmark, 4 additional probes from each vantage point were required on average.

TABLE VI
EVALUATION USING TELECOM & UNICOM LANDMARKS

			Case 1	Case 2	Case 3	Total
Telecom	IP138	DB	105/115	11/15	56/69	172/199
		Impr.	110/115	15/15	56/69	181/199
	QQWry	DB	107/117	11/15	54/67	172/199
		Impr.	111/117	14/15	54/67	179/199
	IPcn	DB	102/117	11/15	57/67	170/199
		Impr.	111/117	14/15	57/67	182/199
	MaxMind	DB	N/A	N/A	N/A	85/199
Unicom	IP138	DB	46/55	9/10	34/41	89/106
		Impr.	52/55	9/10	34/41	95/106
	QQWry	DB	48/55	8/8	33/43	89/106
		Impr.	53/55	8/8	33/43	94/106
	IPcn	DB	44/55	8/10	28/41	80/106
		Impr.	52/55	9/10	28/41	89/106
	MaxMind	DB	N/A	N/A	N/A	57/106

For each landmark, we compare the location determined by our geo-clustering methodology and the location from the corresponding geoIP database with the landmark's ground truth location. The number of the landmarks that are accurately located by the different methods are shown in Table VI. We also evaluate the MaxMind database, and find that MaxMind is inaccurate comparing with the three Chinese databases in China's Internet.

From Table VI, we see that for both ISPs, our geo-clustering methodology can accurately geolocate more landmarks than can the geoIP databases. For the landmarks in Case 1 and Case 2, we are able to accurately geolocate over 7% more Telecom landmarks and over 10% more Unicom landmarks on average. In addition, more than 60% of the landmarks under evaluation fall into Case 1 and Case 2, suggesting that our methodology can improve the geolocation services for many IP addresses in the Chinese Internet.

In summary, we have designed a traceroute-based methodology for improving the Chinese geoIP databases. Our evaluation with ground-truth landmarks shows that the methodology provides more detailed and accurate location information. Finally, we point out that by improving the results from IP138, QQWry, and IPcn, which are currently considered as the best geoIP databases in China, we are indeed providing the (currently) best geolocation service for China's Internet.

VII. RELATED WORK

Rocketfuel [2] probes ISPs' networks and improves measurement efficiency by avoiding traceroutes through the same ingress and egress interfaces of the target ISP. The iPlane project [3] [11] also uses a public platform composed of PlanetLab nodes and traceroute servers; it reduces the probing workload by reducing the targets with BGP atoms [21]. On the other hand, CAIDA/Ark [5] works with dedicated monitors, dividing its monitors into teams for workload reduction. Several algorithms are proposed in recent years for improving the measurement efficiency on dedicated platforms [22] [12]. In this paper, we focus on the measurement efficiency on public platforms. However, unlike Rocketfuel and iPlane that use hundreds of vantage points, our proposed collaborative tracerouting scheme leverages the hierarchical structure of China's Internet to avoid redundant probes, probing the entire Chinese IP address space with only 15 vantage points.

For mapping the Internet, interfaces are typically clustered to routers and PoPs in order to reveal the Internet structure [2] [3] [5] [23]. However, router and PoP clusterings typically rely on having numerous vantage points and on the ability to reverse DNS router interface IPs, both of which are unavailable in China's Internet. In this paper we have developed a heuristic for a hierarchical topology, and have argued that the heuristic gives accurate results for China's Internet.

Many automatic IP address geolocation techniques based on landmarks and active delay measurement have been proposed in recent years [24] [25] [26]. However, Li *et al.* [27] show that the delay-distance correlation, which is a foundation for many delay measurement based geolocation techniques, is weak in China's Internet. Shavitt *et al.* [17] propose to use PoP-level topologies, which are derived from delay measurements [23], to compare and evaluate geoIP database services. Instead of developing a "pure" automatic geolocation technique, we combine the rich location information from commercial geoIP databases with the topological information, and show that our approach can better geolocate Chinese IP addresses than can the existing geoIP database in isolation.

There have only been a few studies focused on China's Internet. Xu *et al.* [4] analyze the Chinese Internet AS topology, we instead focus on the internal structure of the major Chinese ISPs and on geoIP location for China. Guo *et al.* propose Structon [28], which mines and extracts location information from web pages in order to provide a geolocation service within China. Our approach differs from Structon in that we use traceroutes, which directly reveal the underlying network structure, rather than the prefix partitioning rule, to infer addresses' locations; and instead of mining the error-prone web landmarks, we use commercial geoIP databases that provide richer and more accurate location information to drive our heuristic. Our approach outperforms the IPcn geoIP database, which Structon used as the ground truth, by geolocating considerably more router and end host addresses accurately.

VIII. Conclusion

China's Internet has received relatively little attention in the measurement community to date. In this paper, we carried out a large-scale topology mapping and geolocation study for China's Internet. We first developed two traceroute techniques, namely, nested-block partitioning and collaborative tracerouting, to comprehensively and efficiently probe China's Internet from a small number of vantage points inside China. Our approach is able to discover many more interfaces with significantly fewer traceroute probes than the existing traceroute schemes. By further exploiting the hierarchical structure of China's Internet, we proposed a geo-clustering heuristic that clusters interfaces within the same city. We show that the clustering heuristic can geolocate IP addresses with significantly more detail and accuracy than can the existing geoIP databases in isolation. The resulting clusters expose several characteristics of China's Internet. Finally, we demonstrate that the geo-clustering heuristic can be used to improve the accuracy of commercial geoIP databases for geolocating arbitrary IP addresses.

References

[1] China Internet Network Information Center, "Statistical report on Internet development in China," 2011.

[2] N. Spring, R. Mahajan, and D. Wetherall, "Measuring ISP topologies with rocketfuel," in *Proc. of SIGCOMM'02*, Pittsburgh, PA, 2002.

[3] "iPlane Project," http://iplane.cs.washington.edu/.

[4] X. Xu, Z. M. Mao, and J. A. Halderman, "Internet censorship in China: Where does the filtering occur?" in *Proc. of PAM'11*, Atlanta, GA, 2011.

[5] "CAIDA Archipelago Project," http://www.caida.org/projects/ark/.

[6] "Oregon Routeviews," http://www.routeviews.org/.

[7] "RIPE RIS," http://www.ripe.net/data-tools/stats/ris.

[8] "traceroute.org," http://www.traceroute.org/.

[9] Y. Tian, R. Dey, Y. Liu, and K. W. Ross, "China's Internet: Topology mapping and geolocating," University of Sci. & Tech. of China, Tech. Rep., 2011, http://staff.ustc.edu.cn/∼yetian/pub/ChinaInternet.pdf.

[10] Y. Zhu, J. Rexford, S. Sen, and A. Shaikh, "Impact of prefix-match changes on IP reachability," in *Proc. of IMC'09*, Chicago, IL, 2009.

[11] H. V. Madhyastha, T. Isdal, M. Piatek, C. Dixon, T. Anderson, A. Krishnamurthy, and A. Venkataramani, "iPlane: An information plane for distributed services," in *Proc. of OSDI'06*, Seattle, WA, 2006.

[12] R. Beverly, A. Berger, and G. G. Xie, "Primitives for active internet topology mapping: Toward high-frequency characterization," in *Proc. of IMC'10*, Melbourne, Australia, 2010.

[13] "IP138," http://www.ip138.com/.

[14] "QQWry," http://www.cz88.net/.

[15] "IPcn," http://www.ip.cn/.

[16] "MaxMind," http://www.maxmind.com/.

[17] Y. Shavitt and N. Zilberman, "A study of geolocation databases," 2010, preprint, arXiv:1005.5674v3 [cs.NI].

[18] P. Dhungel, K. W. Ross, M. Steiner, Y. Tian, and X. Hei, "Xunlei: Peer-assisted download acceleration on a massive scale," Polytechnic Institute of NYU, Tech. Rep., 2011.

[19] B. Augustin, B. Krishnamurthy, and W. Willinger, "IXPs: Mapped?" in *Proc. of IMC'09*, Chicago, IL, 2009.

[20] C. Labovitz, S. Iekel-Johnson, D. McPherson, J. Oberheide, and F. Jahanian, "Internet inter-domain traffic," in *Proc. of SIGCOMM'10*, New Delhi, India, 2010.

[21] Y. Afek, O. Ben-Shalom, and A. Bremler-Barr, "On the structure and application of BGP policy atoms," in *Proc. of the 2nd SIGCOMM Workshop on Internet Measurment*, Marseille, France, 2002.

[22] B. Donnet, P. Raoult, T. Friedman, and M. Crovella, "Efficient algorithms for large-scale topology discovery," in *Proc. of SIGMETRICS'05*, Banff, Alberta, Canada, 2005.

[23] Y. Shavitt and N. Zilberman, "A structural approach for PoP geo-location," in *Proc. of INFOCOM Workshop on Network Science for Communications (NetSciCom)*, San Diego, CA, 2010.

[24] B. Gueye, A. Ziviani, M. Crovella, and S. Fdida, "Constraint-based geolocation of Internet hosts," *IEEE/ACM Trans. Net.*, vol. 14, no. 6, pp. 1219 – 1232, 2006.

[25] E. Katz-Bassett, J. P. John, A. Krishnamurthy, D. Wetherall, T. Anderson, and Y. Chawathe, "Towards IP geolocation using delay and topology measurements," in *Proc. of IMC'06*, Rio de Janeiro, Brazil, Oct. 2006.

[26] Y. Wang, D. Burgener, M. Flores, A. Kuzmanovic, and C. Huang, "Towards street-level client-independent IP geolocation," in *Proc. of NSDI'11*, Boston, MA, 2011.

[27] D. Li, J. Chen, C. Guo, Y. Liu, J. Zhang, Z. Zhang, and Y. Zhang, "IP-geolocation mapping for involving moderately-connected Internet regions," Microsoft, Tech. Rep., 2009.

[28] C. Guo, Y. Liu, W. Shen, H. J. Wang, Q. Yu, and Y. Zhang, "Mining the web and the Internet for accurate IP address geolocations," in *Proc. of INFOCOM'09*, Rio de Janeiro, Brazil, 2009.