Topology Mapping and Geolocating for China's Internet

Ye Tian, *Member, IEEE,* Ratan Dey, *Student Member, IEEE,* Yong Liu, *Member, IEEE,* Keith W. Ross, *Fellow, IEEE*

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 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 consistency than can the existing geoIP databases in isolation. We show that the resulting clusters expose several characteristics of the Chinese Internet, including the major ISPs' provincial structure and the centralized inter-connections among the ISPs. Finally, using the clustering heuristic, we propose a methodology for improving commercial geoIP databases and evaluate using IDC datacenter landmarks.

Index Terms—Network topology, Measurement techniques.

1 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. China has a very simple AS-topology with few Chinese ASes [4]. However, 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. As China's Internet is dominated by few major ISPs, it is therefore largely shaped by the internal structure of these giant ASes rather than the AS-topology. Interested readers can refer to our brief overview of China's Internet in Section 1 of the supplementary file, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety/XXXXX, and we also present a survey on the previous Internet measurement studies and methodologies in Section 6 of the supplementary file.

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, whois databases, DNS reverse lookups, and end user inputs. However, location information from these sources may be stale or inaccurate, which lead to errors in geoIP databases. 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 insufficient number of Chinese PlanetLab nodes, looking glass servers, and router interfaces with geographical names, we leverage unique

- *Y. Tian is with the Anhui Key Laboratory on High Performance Computing, School of Computer Science and Technology, University of Science and Technology of China, Hefei, Anhui 230026, China. E-mail: yetian@ustc.edu.cn*
- *R. Dey and K. W. Ross are with the Department of Computer Science and Engineering, Polytechnic Institute of NYU, Brooklyn, NY 11201, USA. E-mail: ratan@cis.poly.edu, ross@poly.edu*
- *Y. Liu is with the Department of Electronic and Computer Engineering, Polytechnic Institute of NYU, Brooklyn, NY 11201, USA. E-mail: yongliu@poly.edu*

1. By China we mean Mainland China.

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 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, and inter-ISP connections are concentrated to a few routers across China.
- 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, and the methodology can also differentiate the results from geoIP databases with different confidence levels. By improving on the best geoIP databases in China, we are currently providing the most accurate geolocation service for China's Internet.

This paper substantial extends the earlier conference paper [5] in both methodologies and insightful observations in topological mapping and geolocating for China's Internet.

2 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 [6], 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. As it is well known that Telecom and Unicom have most of the international Internet connections in China [1], 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.

TABLE 1
Distribution of IP block prefix length

$\leq$18	19	20	21	22	23	24	$\geq$25
2,793	1,197	1,987	1,210	1,175	1,133	4,404	104

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 [6]) often use CIDR IP blocks from public BGP snapshots (e.g., from Oregon Routeviews [7] and RIPE RIS [8]); the blocks are used to partition the IP space, and then one address is selected from each block in the partition as the traceroute targets. However, there is no operational public BGP router in China's Internet [9]; therefore, we can only gather Chinese blocks from routers that are outside of China. Because these blocks are likely to have been aggregated by the border routers in China's Internet, they are generally too coarse for topology mapping. To establish this claim, we have downloaded eight BGP snapshots from different routers in Oregon Routeviews and RIPE RIS. (The routers are located in USA, Europe, and Japan.) Table 1 lists the numbers of distinct IP blocks in China with their prefix lengths. We can see that there are many large blocks (e.g., blocks with prefix lengths smaller than 20, 18, and so on).

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); furthermore, all the PlanetLab nodes are in CERNET. If we use iPlane's or Ark's probing strategy, we would overload our 15 vantage points with too many tasks. For example, although iPlane uses the PlanetLab nodes to exhaustively probe target addresses (typically 140,000∼150,000 addresses per day), it only schedules a few tens of targets to looking glass servers. As we rely on a few looking glass servers to effectively measure majority of China's Internet (i.e., Telecom & Unicom), we can not expect them to probe as many targets as the PlanetLab nodes in iPlane. To address the two challenges, we devise two techniques, namely, *nested IP block partitioning* and *collaborative tracerouting*.

2.1 Nested IP Block Partitioning

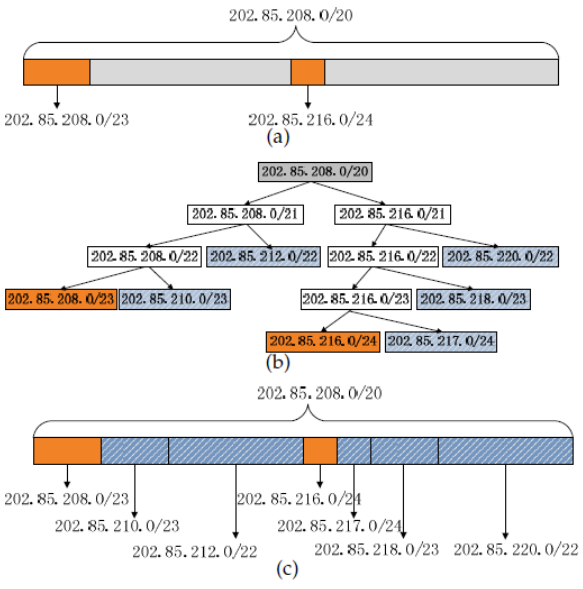

Fig. 1. Nested-block Partitioning

We need to partition the large Chinese IP address space, and then choose one IP address from each set in the partition 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: among the $14,003$ Chinese blocks collected from the BGP snapshots, $11,091$ of them are nested in larger blocks. Moreover, there are often several levels of nesting. An example of nested IP blocks is shown in Fig. 1(a). In the figure, 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, for the case in Fig. 1(a), the block binary tree is shown in Fig. 1(b), and we use seven blocks to replace the original large block 202.85.208.0/20, as shown in Fig. 1(c). After partitioning the nested IP blocks, 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, if the granularity is prefix /22, then 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 into smaller ones, which would geometrically increase the probing workload.

2.2 Collaborative Tracerouting

Ark and iPlane apply different strategies to reduce the workload (when probing the entire Internet). In Ark, each /24 block is probed in one measurement round, but Ark groups its vantage points into teams, with each team having a geographically distributed set of members. Each team only probes a subset of the targets. Although a target is only probed by one team, the number and the geographical distribution of the team's vantage points ensure the diversity of the traceroutes. In iPlane, IP blocks from BGP snapshots with similar AS paths are further combined to reduce the workload [11].

We, however, cannot apply either Ark's or iPlane's strategies for two reasons: (i) we have only 15 vantage points (8 are looking glass servers with low probing rate) 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 2.1, 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 2.1 (which partition the Chinese IP space) are the basic probe units. When a vantage point probes a block, we always use the second IP address of that block (i.e., a.b.c.1) as the target, as such addresses are usually used for gateways and are, thus, more likely to respond to a probe than other addresses. 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. An example demonstrating the collaborative tracerouting scheme can be found in Section 2.1 of the supplementary file.

2.3 Measurement Results

TABLE 2
Traceroute measurement results

	iPlane (one day)	iPlane (two 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 different 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 12 December 2010 to 2 January 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 on 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.

Table 2 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. In the fifth column of the table, we list the numbers of interfaces and links that appear both in cTrace and in iPlane (2 days). 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 compare the link traversal frequencies in iPlane and cTrace, and present the results in Section 2.2 of the supplementary file.

Finally, from Table 2 we note that although cTrace contains many more interfaces and links than iPlane, there are still 7,738 interfaces and 55,056 links in iPlane that are not discovered by cTrace. Furthermore, if we only consider the traceroutes that are from the vantage points outside China by removing the only two PlanetLab nodes located in China from iPlane, we can still find 6,754 interfaces and 53,035 links that are missing in cTrace. We believe these interfaces and links are located on the border of China's Internet that connects to the international Internet. These links are thus unlikely to be traversed by cTrace, which uses vantage points within China. For this reason, we 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.

3 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 that are 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. For cases when bogus locations are returned (e.g., a non-exist location name), we consider the corresponding level location information as null.

TABLE 3
Null reply ratios for the addresses from traceroute

	IP138	QQWry	IPcn	MaxMind
Province (all)	0.105	0.074	0.108	0.186
Province (router)	0.185	0.143	0.184	0.167
City (all)	0.240	0.212	0.280	0.227
City (router)	0.283	0.271	0.290	0.225

We first 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 all the $78,229$ IP addresses from the combined traceroute data to examine the geoIP databases. The second and forth rows of Table 3 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 examine the null reply ratios for all the IP addresses that do not appear at the last hop of the traceroutes. These addresses are bound on router interfaces. A total number of 31,920 addresses are examined, and the null reply ratios for the four databases are listed in the third and fifth rows of Table 3. From the table we can see that except for MaxMind, the three Chinese databases have more null replies for router interfaces, suggesting that the three Chinese databases cover better end host IP addresses than router interface addresses. In Section 3 of the supplementary file, we also geolocate Xunlei peers' IP addresses, and compare with the traceroute addresses to support our claim.

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 and erroneous entries, suggesting that there is significant room for improvement.

4 Geolocating the Interface Topology

With the combined traceroute data obtained in Section 2, 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 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.

A naive method to create the clusters is to simply use the city information provided by the geoIP databases on face value. However, we show in Section 4.1 of the supplementary file that it will lead to a large number of small and disconnected erroneous clusters. In the following subsections, we propose a heuristic for accurately determining the maximal clusters in each of the three interface topologies.

4.1 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. The heuristic we present here could be used for any ISP with a hierarchical structure (not just Chinese ISPs).

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. After obtaining the interface topology, the clustering algorithm starts from the interfaces at the edge of the topology, then gradually moves towards the backbone interfaces located at the core. The heuristic algorithm consists of four steps. In the first step, we form singleton clusters using the interfaces at the edge of the interface topology. In the second step, we repeatedly select the interfaces that are one step closer to the backbone network and, based on their inferred locations, group them into existing clusters. In this step, we cluster router interfaces in the residential and provincial networks. In the third step, we cluster the router interfaces in the backbone network using

a method similar to step 2, but we apply different rules for inferring the interfaces' locations. Finally, in step 4, we merge the singletons and small clusters that remain after step 3 to create the maximal clusters. In the following we describe each of the steps in detail.

4.1.1 Step 0: Preprocessing

Before clustering the interfaces, we first filter out the influence of the vantage points and anonymous routers in the interface topologies. For a typical traceroute path traversing Telecom, Unicom, or CERNET, it contains three subpaths: the subpath from the vantage point to the first backbone router, the subpath inside the backbone network, and the subpath from the last backbone router to the target. We only include the last two subpaths in our interface topology. By filtering out all the "up backbone" links in the first subpath, the resulting interface topology can be viewed as expanding from the backbone network to the traceroute targets. In addition, if an anonymous router is found in the third subpath, all the interfaces after the anonymous interface are removed.

In this preprocessing step, we need to identify the backbone routers. Although most routers in China are nameless, and thus cannot be reverse DNS queried, the three Chinese databases do return backbone network information, indicating whether an address being queried belongs to the Telecom, Unicom, or CERNET backbone. By filtering we remove only a small fraction of the addresses from the traceroute data. For example, only 2.5%, 5.9%, and 15.3% of the interfaces on the interface topologies of Telecom, Unicom, and CERNET are removed using the geoIP database of IP138.

4.1.2 Step 1: Startup Clusters from the Edge

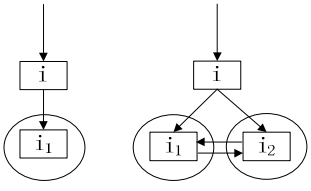

Fig. 2. Example of step 1 clustering

In the first step, we select the addresses at the edge of the interface topology to form startup singleton clusters. Specifically, for an interface in the topology, say i_1, if it has no outgoing links, or for each interface i_2 it links to, there exists a path from i_2 to i_1 in the interface topology, a singleton cluster containing only i_1 is formed. We use i_1's location from the geoIP database (referred to as i_1's *DB location*) as the cluster's location. A simple example is shown in Fig. 2. In the left figure, i_1 has no outgoing links, and a singleton cluster is formed. In the right figure, i_1 and i_2 link to each other, so there is a return path for both of them; thus two singleton clusters, one containing i_1 and the other containing i_2, are formed.

4.1.3 Step 2: Clustering Residential and Provincial Networks

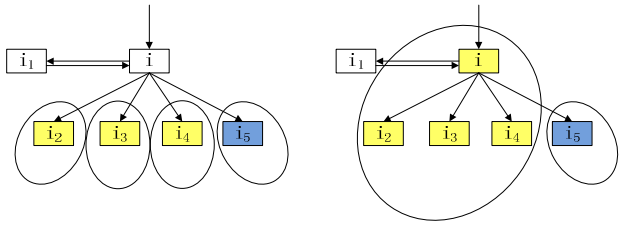

Fig. 3. Example of step 2 clustering

After obtaining the startup singleton clusters, we continue to cluster more router interfaces. The heuristic works in rounds. In each round, we select some of the unclustered addresses in the interface topology as candidates for clustering. An unclustered interface i is selected as a candidate if each of i's out-linked interfaces is (a) either clustered (as are interfaces i_2, i_3, i_4, and i_5 in Fig. 3) or (b) there exists a path in the interface topology from the out-linked interface back to i (as is i_1 in Fig. 3).

For each candidate interface, we use its out-linked interfaces to infer its location. Suppose a candidate interface i links to interfaces i_1, i_2, ..., i_n, which belong to clusters c_1, c_2, ..., c_m. We then have i_1, i_2, ..., i_n *vote* to infer i's location. For each out-linked interface, say i_k, if its DB location contains a city-level location, it uses its DB location to vote; otherwise it uses the location of the cluster it belongs to (referred to as i_k's *cluster location*). After voting, if there exists a city-level location x that wins the voting by exceeding a threshold of c_vote, x is assigned as i's cluster location (but i still keeps its DB location). We further merge all clusters among c_1, c_2, ..., c_m that have cluster location x into one larger cluster, and also put i into this newly merged cluster. When there is no winner from the voting, if i's DB location has a city-level location, say y, we merge all clusters among c_1, c_2, ..., c_m that have cluster location y into one larger cluster, and put interface i into this newly merged cluster. Otherwise, i cannot be put into any cluster and will form a singleton cluster containing only itself. A simple example of voting and cluster merging is shown in Fig. 3: in the left graph, the candidate interface i's DB entry does not include city-level information; however, its cluster location is inferred by the voting among its out-linked interfaces i_1, ..., i_5; after the voting, i is assigned the same location as the clusters of i_2, i_3 and i_4, and they join i to form a larger cluster, as shown in the right graph.

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. After all the candidate interfaces are processed, the heuristic finishes a round and selects new candidates for the next round. Step 2 stops when we can't form or merge any clusters during a round.

4.1.4 Step 3: Clustering Interfaces in the Backbone

Step 3 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 many routers at different locations. In addition, the links that connect backbone interfaces are usually traversed many times during the traceroute measurement. This makes it possible to accurately estimate delays on those links. For a candidate interface i, we sequentially apply three rules based on delay and majority voting to infer its cluster location. Please refer to Section 4.2 of the supplementary file for the details of the three rules.

4.1.5 Step 4: Merging Singleton and Small Clusters

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 a large one. We consider the clusters containing less than c_size interfaces as mergeable *small clusters*, and the others are regarded as *large clusters*. 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. By repeatedly merging small clusters, we can eliminate most of them.

We refer to this four-step heuristic as the *geo-clustering* heuristic on the interface topology.

4.2 Geo-Clusters

We applied the geo-clustering heuristic on the Telecom, Unicom, and CERNET interface topologies using each of three geoIP databases. For the parameters, we use $c\_vote = 0.5$, $l\_delay = 1ms$, $t\_times = 5$, $p\_vote = 0.5$ and $c\_size = 5$. For example, for the Telecom's interface topology using the geoIP database of IP138, there are $38,181$ interfaces. After Steps 1, 2, 3, and 4, we get $26,518$, $7,326$, $7,467$, and $1,125$ clusters, respectively. 532 of the final clusters contain $37,488$ interfaces and have been assigned city level locations. (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.) The observation indicates that by geo-clustering, we can group most of the interfaces into clusters with detailed city-level location information. We refer to a cluster with a city-level location as a *geo-cluster*. Similar results are observed using the two other geoIP databases and for the two other backbone 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. 4, 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, as shown in Fig. 4(b). 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 where the network of former Unicom and the former Netcom's network do not connect to each other, we observe one large geo-cluster for the former Unicom network, and another large geo-cluster for the former Netcom network.

Fig. 5 shows the geo-clusters of the top 10 cities. For clarity we remove the singleton clusters. 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 in southern and western China.

4.3 Intra- and Inter-ISP Structures

With the geo-clusters, we now study the internal structure of each ISP, as well as the inter-connectivity among the major ISPs in China. We only list our main results here, and interested readers can refer to Section 4.3 and Section 4.4 of the supplementary file for more details.

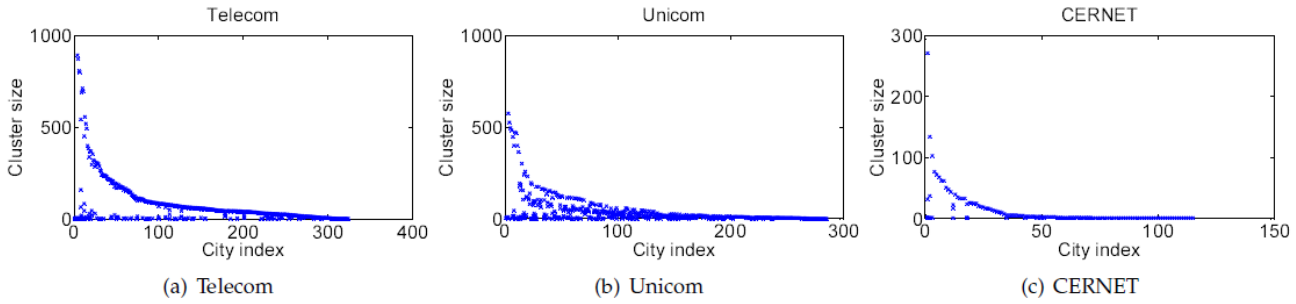

Fig. 4. Distribution of the sizes of geo-clusters across cities in three major ISPs

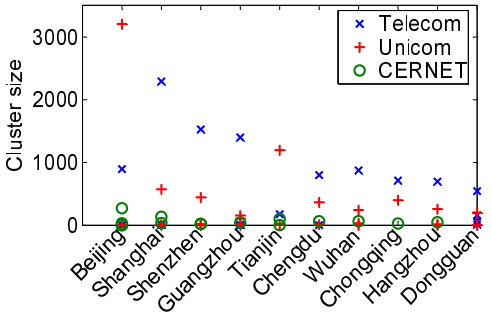

Fig. 5. Geo-clusters in the top-10 cities

First, we find that 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 [18] [19].

Second, we observe routes from all over China within a major ISP will concentrate to a few routers for accessing other ISPs' networks, potentially making these routers bottlenecks for inter-ISP traffic in China.

4.4 Locating Interfaces with Null Replies

TABLE 4
Null reply ratios

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

Each interface in an ISP's interface topology has now been assigned 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 4 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 5
Number of the interfaces that have consistent locations

	Total	3DB identical	3Cluster identical
Telecom	38,181	25,625 (67.1%)	35,376 (92.7%)
Unicom	24,781	15,794 (63.7%)	21,938 (88.5%)
CERNET	1,798	1,343 (74.7%)	1,602 (89.1%)
Total	64,760	42,762 (66.0%)	58,916 (91.0%)

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 by showing that clustering approach leads to substantially more consistent results 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 5 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 tracer-

oute 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. We also observe a concentration of inter-ISP connections at a relatively small number of interfaces. In addition, we show that our heuristic can geolocate router interface addresses with more detailed location information than can existing geoIP databases, and the consistency of the location information suggest better accuracy of the clustering heuristic.

5 Improving Geolocation Services with Geo-Clusters

In the previous section, we showed how our methodology can geolocate router interface addresses that have null or erroneous entries in the geoIP databases. In this section, we develop a methodology for accurately geolocating arbitrary Chinese IP addresses. Our goal here is to provide a significant improvement over the existing Chinese geoIP databases.

5.1 Geolocating an Arbitrary IP Address

Our methodology relies on the geo-clustering heuristic described in Section 4.1. 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 4 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' (using Step 0 in the heuristic, as described in Section 4.1). 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 included 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. However, if there are multiple last-hop geo-clusters with different cluster locations for p, then Step 2 is inconclusive, and we proceed to Step 3.
- Case 3: If we don't set p's location in Case 1 and 2, the location from the geoIP database is used.

5.2 Evaluation

5.2.1 Collecting Landmarks

We use 199 landmarks on Telecom and 106 landmarks on Unicom as the ground truth for evaluating the accuracies of the geoIP databases and our geolocating approach. In Section 5.1 of the supplementary file, we describe how to collect the landmarks from the IDC datacenters.

5.2.2 Evaluation Results

TABLE 6
Evaluation using Telecom landmarks

		Case 1	Case 2	Case 3	Total
IP138	DB	105/115	11/15	56/69	172/199
	Improve	110/115	15/15	56/69	181/199
QQWry	DB	107/117	11/15	54/67	172/199
	Improve	111/117	14/15	54/67	179/199
IPcn	DB	102/117	11/15	57/67	170/199
	Improve	111/117	14/15	57/67	182/199
MaxMind	DB	N/A	N/A	N/A	85/199

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.

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 numbers of the Telecom landmarks that are accurately located by the different methods are shown in Table 6, and we also present the evaluation results on Unicom landmarks in Table 4 of the supplementary file. We further classify the landmarks into three cases based on how their locations are determined by our methodology. As an example, consider the case of Telecom

and the IP138 geoIP database. Of these 199 Telecom landmarks, 115 fall into Case 1. Of these 115 landmarks, the IP138 database correctly located 105; whereas our methodology (using the same location database) correctly located 110. We also evaluate the MaxMind database, and find that MaxMind is inaccurate comparing with the three Chinese databases.

From the landmark geolocating results, 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. Although these improvements for locating end host IPs are not as dramatic as our results for locating router interface IPs, we believe that the improvements are nevertheless significant and useful.

In addition to improved accuracy, a less obvious benefit of our methodology is that it provides a means for users to assess the quality of the results returned from geoIP databases. Specifically, each geoIP database is significantly more accurate for targets falling into Case 1 or 2 than those falling into Case 3. Thus, when using a geoIP database, if the target falls into Case 1 or 2, the user can be relatively confident about the result, but less confident when the target falls into Case 3. Furthermore, we find that when a geoIP database gives an accurate result, our methodology always provides the same result, with only one exception of a Unicom landmark using the databases of IP138 and QQWry.

For the landmarks belonging to Case 3, by examining the traceroute paths to them, we find that the distances between their last-hop geo-clusters and the traceroute targets are larger than 2 hops, and many paths never reach the landmarks. We remark that for an IP address that is unreachable with traceroute, or is far behind anonymous devices, it becomes difficult to geolocate for any traceroute-based mechanism.

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, and also allows users to assign levels of confidence to the results returned from the geoIP databases. 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.

6 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.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 61202405 and 61103228) and the Fundamental Research Funds for the Central Universities of China (Grant No. WK0110000007 and WK0110000024).

References

[1] China Internet Network Information Center, "Statistical report on Internet development in China," Jan. 2011.
[2] N. Spring, R. Mahajan, and D. Wetherall, "Measuring ISP topologies with rocketfuel," in *Proc. of SIGCOMM'02*, Pittsburgh, PA, USA, Aug. 2002.
[3] "iPlane: An information plane for distributed services," http://iplane.cs.washington.edu/.
[4] H. Yin, H. Chang, F. Liu, T. Zhan, Y. Zhang, and B. Li, "A complementary and contrast view of the Chinese Internet topology," in *Proc. of IEEE International Conference on Ubiquitous Computing and Communications*, Liverpool, UK, Jun. 2012.
[5] Y. Tian, R. Dey, Y. Liu, and K. Ross, "China's Internet: Topology mapping and geolocating," in *Proc. of IEEE INFOCOM'12 Mini-conference*, Orlando, FL, USA, Mar. 2012.
[6] "Archipelago measurement infrastructure," http://www.caida.org/projects/ark/.
[7] "University of Oregon route views project," http://www.routeviews.org/.
[8] "Routing information service," http://www.ripe.net/datatools/stats/ris/routing-information-service.
[9] "traceroute.org," http://www.traceroute.org/.
[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, USA, Nov. 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, USA, Nov. 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, Nov. 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 geolocation databases study," *IEEE J. on Selected Areas in Communications*, vol. 29, no. 10, pp. 2044 – 2056, 2011.
[18] B. Augustin, B. Krishnamurthy, and W. Willinger, "IXPs: Mapped?" in *Proc. of IMC'09*, Chicago, IL, USA, Nov. 2009.
[19] C. Labovitz, S. Iekel-Johnson, D. McPherson, J. Oberheide, and F. Jahanian, "Internet inter-domain traffic," in *Proc. of SIGCOMM'10*, New Delhi, India, Aug. 2010.

Ye Tian is an associate professor at the School of Computer Science and Technology, University of Science and Technology of China (USTC). He joined USTC in August 2008. Ye Tian received his Ph.D. degree from the Department of Computer Science and Engineering at The Chinese University of Hong Kong (CUHK) in December 2007. He received his Bachelor of Engineering degree in electronic engineering and Master of Engineering degree in computer science from the University of Science and Technology of China, in July 2001 and 2004, respectively. His research interests include Internet and network measurement, Peer-to-Peer networks, overlay networks, online social networks, and multimedia networks. He is a member of IEEE and ACM, and a senior member of China Computer Federation (CCF).

Ratan Dey is a Ph.D student at Polytechnic Institute of New York University (NYU - Poly) in the Department of Computer Science and Engineering. His research adviser is Prof. Keith W. Ross. He received B.Sc. degree in Computer Science and Engineering from Bangladesh University of Engineering and Technology (BUET) and M.Sc. degree in Computer Science from Polytechnic Institute of New York University (NYU - Poly). His research focus is in the areas of Privacy, Social Networks, and Networking Measurement.

Yong Liu is an associate professor at the Electrical and Computer Engineering department of the Polytechnic Institute of New York University (NYU-Poly). He joined NYU-Poly as an assistant professor in March, 2005. He received his Ph.D. degree from Electrical and Computer Engineering department at the University of Massachusetts, Amherst, in May 2002. He received his master and bachelor degrees in the field of automatic control from the University of Science and Technology of China, in July 1997 and 1994 respectively. His general research interests lie in modeling, design and analysis of communication networks. His current research directions include Peer-to-Peer systems, overlay networks, network measurement, online social networks, and recommender systems. He is the winner of the IEEE Conference on Computer and Communications (INFOCOM) Best Paper Award in 2009, and the IEEE Communications Society Best Paper Award in Multimedia Communications in 2008. He is a member of IEEE and ACM. He is currently serving as an associate editor for IEEE/ACM Transactions on Networking, and Elsevier Computer Networks Journal.

Keith W. Ross is the Leonard J. Shustek Chair Professor in Computer Science at Polytechnic Institute of NYU. He is also the Head of the Computer Science and Engineering Department. Before joining NYU-Poly in 2003, he was a professor at University of Pennsylvania (13 years) and a professor at Eurecom Institute (5 years) in France. He received a Ph.D. in Computer and Control Engineering from The University of Michigan. Keith Ross has worked in security and privacy, peer-to-peer networking, Internet measurement, video streaming, among other areas. He is an IEEE Fellow and recipient of three major recent best paper awards. His work on privacy has been featured in the New York Times, NPR, Bloomberg Television, Huffington Post, Fast Company, Ars Technia, and the New Scientist. Keith Ross is co-author of the popular textbook, Computer Networking: A Top-Down Approach Featuring the Internet, published by Addison-Wesley. It is the most popular textbook on computer networking, both nationally and internationally, and has been translated into fourteen languages. In 1999-2001, Professor Ross took a leave of absence to found and lead Wimba, which develops voice and video applications for online learning. He was the Wimba CEO and CTO during this period. Wimba was acquired by Blackboard in 2010.