Supplementary File of the TPDS Manuscript

Ye Tian, Ratan Dey, Yong Liu, Keith W. Ross
E-mail: yetian@ustc.edu.cn, ratan@cis.poly.edu, yongliu@poly.edu, ross@poly.edu

Abstract—This supplementary file contains the supporting materials of the TPDS manuscript – "Topology Mapping and Geolocating for China's Internet". It improves the solidity and completeness of the TPDS manuscript.

1 Overview of China's Internet

We briefly overview China's Internet in this Section. 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)[1]. 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. Unicom holds the dominant market share in the northern provinces, whereas Telecom dominates southern China [2]. 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[2] 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.

Since Telecom and Unicom are government-owned companies, it is not surprising that their topological structure is hierarchical and mimics the provincial organization of the Chinese government. Using Telecom as an example, Telecom Corporation is in charge of Telecom's national backbone network. For each province, Telecom has a provincial subsidiary that manages its provincial network; and for most cities in the province, there is a city company under the provincial company that is responsible for constructing and maintaining residential networks and providing Internet services to end users in that city. In addition, Telecom assigns blocks of IP addresses along this three-level structure, that is, from its address space, it allocates blocks to each of the provincial subsidiaries, and each provincial subsidiary allocates blocks of addresses to cities. In fact, many Chinese geoIP databases exploit this hierarchical IP address assignment to provide basic geolocation services.

Internet Datacenters (IDCs) are widely used in China. An IDC datacenter can provide many services to enterprises and individual customers, including server hosting, server leasing, and virtual servers. Most of the datacenters connect directly to either the Telecom or Unicom backbones, and some connect to both (and directly to other Chinese ISPs). Because the datacenters have high-speed connections directly to the backbones, they are attractive for hosting websites and other online services for various customers. In fact, many Chinese university web sites are not hosted on the university campuses, but instead in IDC datacenters; similarly, many government web sites are hosted in IDC datacenters. Furthermore, many popular web sites in China are mirrored across several datacenters.

1. Here we are referring to the current China Unicom, which merged with China Netcom (a.k.a. CNCGroup) in 2008.
2. China Education and Research Network

2 Traceroute Measurement

2.1 Collaborative Tracerouting Example

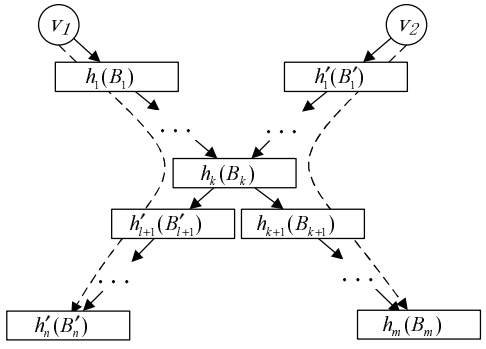

Fig. 1. An example of collaborative tracerouting

We present an example to demonstrate the collaborative tracerouting scheme in this Section. 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. 1. 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. 1. (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.)

2.2 Comparison of iPlane and cTrace

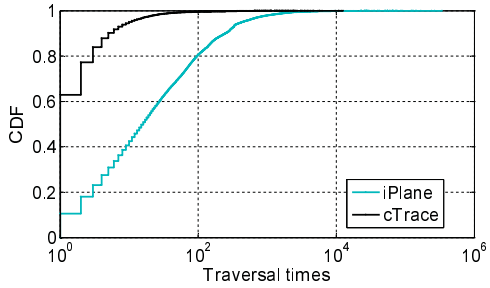

Fig. 2. Number of times the links are visited

To further demonstrate that cTrace is more effective in exploring China's Internet, we plot the distributions of the number of times that links are traversed in iPlane (over 2 days) and in cTrace in Fig. 2. In two days, iPlane visited some links thousands of times, even though most of its vantage points are outside of China and, thus, are far away from these links.

3 Evaluation of GeoIP databases using Xunlei Peers

TABLE 1
Null reply ratios for Xunlei peers

	IP138	QQWry	IPcn	MaxMind
Province	0.011	0.004	0.021	0.161
City	0.153	0.137	0.178	0.224

To gain further insight into the databases' performance for different types of addresses (i.e., router addresses and end host addresses), we randomly selected 2,000 IP addresses from peers collected by crawling the Xunlei DHT [3] (a popular P2P download acceleration application in China) and fed these end host addresses to the geoIP databases. Table 1 shows the null-reply ratios on Xunlei peers. Comparing with Table 3 in the manuscript main file, we can see that except for MaxMind, the three Chinese databases have fewer null replies for Xunlei peers, suggesting that the three Chinese databases cover better end host IP addresses than router interface addresses.

4 Geolocating the Interface Topology

4.1 Discussion on the Naive Clustering Method

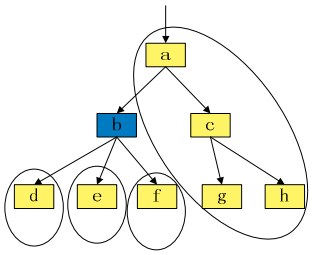

Fig. 3. Erroneous clusters example

A naive method to create the clusters is to simply 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. 3 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. 3, 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.

4.2 Rules for Inferring Interface Location in Step Three

In Step 3 of the geo-clustering heuristic, we sequentially apply three rules to infer a candidate interface i's cluster location.

- We first examine each link incident from i to a clustered interface, say i_1, that has been traversed more than t_times times. For such a link, we use the median of the delays estimated from different vantage points, as proposed in [4], as its delay. If the delay is smaller than the threshold of l_delay, we assign i_1's cluster location x to i, and merge i with all the clusters that connect to i with location x. However, if more than one link is observed, suggesting different cluster locations, we apply the next rule.
- We use the same voting-based method as in Step 2 to infer candidate i's city-level location. But here we don't apply the stop condition when more than one province appears. We use the location that wins the voting as i's cluster location and merge its out-linked clusters if possible. If there is no winner from the voting, apply the next rule.
- We conduct a province-level voting using a method similar to Step 2. For each of i's out-linked interfaces, only its province-level cluster location information is used. We use a threshold of p_vote to find the winner. After the voting, if a province-level location wins, we assign the capital city of that province as the candidate's city-level location. This is because a provincial network usually accesses the backbone network at the capital city of that province (see Section 4.3). We then perform cluster merging with the assigned city-level location if possible.

Finally, if none of above rules can be applied to i, a singleton cluster is formed for it.

4.3 The Hierarchical Structure

TABLE 2
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

Using our clustering heuristic, we now study the internal structure of each ISP. Table 2 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 [5] [6].* Finally, we observe relatively few intra-province links in CERNET, as CERNET only reaches cities that have many universities and research institutes, and such cities are usually the provincial capital cities in China.

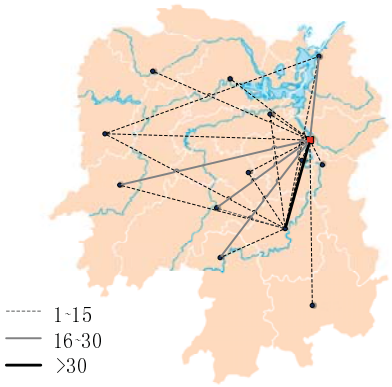

Fig. 4. Cluster topology for Hunan province

As an example, Fig. 4 shows the topology of the Telecom geo-clusters 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.

4.4 Inter-Connectivity among Major ISPs

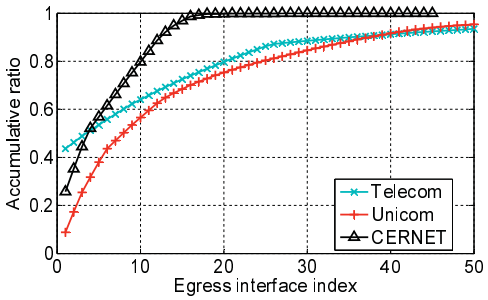

Fig. 5. Ratio distribution for paths traversing egress interfaces

We investigate the inter-connectivity among the major Chinese ISPs in this section. For a major ISP, we focus on its backbone interfaces through which

traceroute paths exit the ISP to enter another major ISP's backbone network. We refer to such an interface as the ISP's *egress interface*. From the combined traceroute data, we collected 150, 98, and 45 distinct egress interfaces for Telecom, Unicom, and CERNET, respectively.

We rank egress interfaces by the number of inter-ISP traceroute paths traversing them. Fig. 5 shows the accumulative ratios for the three ISPs. For each of the three ISPs, the paths are distributed unevenly among the egress interfaces: 99.9%, 86.2% and 80.3% of the paths departing CERNET's, Telecom's and Unicom's backbone, respectively, are through the first 25 egress interfaces.

TABLE 3
Shared interfaces and traceroutes across distant vantage points in same ISP

	Vantage point	Interface	Traceroute
Telecom	Chengdu	76.1%	53.6%
	Xiamen	67.3%	50.1%
Unicom	Tianjin	76.2%	95.9%
	Shenzhen	84.3%	92.7%
CERNET	Beijing	78.6%	64.7%
	Changsha	68.8%	79.7%

We now investigate how the egress interfaces are traversed. For each major ISP, we select two vantage points on its network that are geographically far away from each other. For each vantage point, we examine the egress interfaces traversed by its traceroutes, identifying the egress interfaces that are shared by the two vantage points in a same ISP. Table 3 lists the percentages of the shared egress interfaces among all the egress interfaces of each vantage point, as well as the percentages of the paths through these shared egress interfaces among all the inter-ISP traceroute paths from each vantage point. From the table we can see the two distant vantage points in the same ISP share an astonishing percentage of interfaces when accessing other ISPs' backbone networks. *Although we only use two vantage points for each ISP, their geographical locations suggest that 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.*

5 Improving Geolocation Services with Geo-Clusters

5.1 Collecting Landmarks

We use a number of landmarks as the ground truth for evaluating the accuracies of the geoIP databases and of our methodology. Generally it is difficult to obtain ground truth landmarks, particularly for China's Internet, as many websites are hosted by IDCs, including university and government websites (which are often used as landmarks in other studies [7] [8]). In this paper, we leverage the numerous IDC datacenters located in many cities in China, for collecting the landmarks. For a datacenter, we find one or more of its IP addresses, and associate the datacenter's location with the IP addresses to get landmarks.

We combine the IDC datacenter IP addresses we have found from the websites listed on *www.IDCquan.net* and the mirror servers of the popular software downloading site *onlinedown.net* to collect IDC datacenter landmarks. We have successfully collected 305 landmarks – 199 on Telecom and 106 on Unicom – with their ground-truth locations detailed to the city level.

5.2 Evaluation with Unicom Landmarks

TABLE 4
Evaluation using Unicom landmarks

		Case 1	Case 2	Case 3	Total
IP138	DB	46/55	9/10	34/41	89/106
	Improve	52/55	9/10	34/41	95/106
QQWry	DB	48/55	8/8	33/43	89/106
	Improve	53/55	8/8	33/43	94/106
IPcn	DB	44/55	8/10	28/41	80/106
	Improve	52/55	9/10	28/41	89/106
MaxMind	DB	N/A	N/A	N/A	57/106

We show the numbers of the Unicom landmarks that are accurately located by the geoIP databases and by our geo-clustering methodology in Table 4.

6 Related Work

Spring *et al.* [9] propose Rocketfuel to probe ISPs' networks with public traceroute servers. Rocketfuel improves measurement efficiency by avoiding traceroutes through the same ingress and egress interfaces of the target ISP. The iPlane project [10], [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 [12]. On the other hand, CAIDA/Ark [13] works with dedicated monitors, dividing its monitors into teams for workload reduction. Several algorithms are proposed for improving the measurement efficiency on dedicated platforms: Donnet *et al.* [14] propose the Double Tree algorithm to avoid redundant probing packets by heuristically probing forward and backward to a vantage point and a target from the mid-point. Beverly *et al.* [15] propose a scheme that optimally selects the probing packets that can fully cover the entire interface topology obtained in the previous measurement cycle. Our collaborative tracerouting scheme differs from these works in that it is more suitable for a public platform of PlanetLab nodes and traceroute servers, and is more effective than the approaches used in Rocketfuel and iPlane; in particular, our approach avoids probes if an earlier interface path was found from the vantage point to the target. In addition, we leverage the block nesting in BGP snapshots [16] to

improve the measurement coverage without introducing many unnecessary targets.

Limitations of the standard traceroute tool are recognized and addresses in recent years. Paris traceroute [17] customizes probe packets to avoid the inconsistent routes returned by traceroute due to flow-based load-balancers; while Reverse Traceroute [18] allows to discover an asymmetric route from destination to source, by applying a variety of collaborative measurement techniques on tens to hundreds of vantage points. Nevertheless, we use standard traceroute in our study, as these techniques generally require root privileges on vantage points, which we do not have on the looking glass servers; in addition, we do not need to probe the reverse route as in our geo-clustering heuristic, only the links within the backbone and the "down backbone" links are included in the interface topologies.

For mapping the Internet, interfaces are typically clustered to routers and PoPs in order to reveal the Internet structure [9], [10], [13], [19]. 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 work, we instead group the interfaces into geo-clusters, which reveals the internal structure of the major Chinese ISPs.

Many automatic IP address geolocation techniques based on landmarks and active delay measurement have been proposed in recent years [7], [20], [21], [22], [8]. However, Li *et al.* [23] 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.* [24] propose to use PoP-level topologies, which are derived from delay measurements [19], to compare and evaluate geoIP database services. In this paper, we exploit the location information in geoIP databases to cluster the interfaces from a traceroute measurement; we then use the resulting geo-clusters to improve the completeness and accuracy of these databases.

There have only been a few studies focused on China's Internet. Yin *et al.* [25] analyze the Chinese Internet AS topology by combining various data sources including traceroutes and BGP snapshots. Guo *et al.* propose Structon [26], which mines and extracts location information from web pages in order to provide a geolocation service within China. Our work differs from [25] in that we focus on ISP's internal structure instead of the AS topology, and our geolocation approach differs from Structon in that we first obtain interface topologies from traceroute measurements, then combine the interface topology with the partially correct locations from commercial geoIP databases. Structon uses a prefix partitioning rule and location voting. We instead use traceroutes, which directly reveal the underlying network structure, to infer IP addresses' locations; also, instead of using locations found on webpages, we use commercial geoIP databases that provide richer and more accurate location information to drive our heuristic.

REFERENCES

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

[2] P. Uria-Recio, "China telecommunications panorama," 2006, http://globthink.com/2009/08/12/china-telecommunications-panorama/.

[3] P. Dhungel, K. W. Ross, M. Steiner, Y. Tian, and X. Hei, "Xunlei: Peer-assisted download acceleration on a massive scale," in *Proc. of PAM'12*, Vienna, Austria, Mar. 2012.

[4] D. Feldman and Y. Shavitt, "An optimal median calculation algorithm for estimating Internet link delays from active measurements," in *Proc. of Workshop on End-to-End Monitoring Techniques and Services*, Munich, Germany, May 2007.

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

[6] 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.

[7] V. N. Padmanabhan and L. Subramanian, "An investigation of geographic mapping techniques for Internet host," in *Proc. of SIGCOMM'01*, San Diego, CA, USA, Aug. 2001.

[8] 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, USA, Mar. 2011.

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

[10] "iPlane: An information plane for distributed services," http://iplane.cs.washington.edu/.

[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] 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, Nov. 2002.

[13] "Archipelago measurement infrastructure," http://www.caida.org/projects/ark/.

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

[15] 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.

[16] 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.

[17] B. Augustin, X. Cuvellier, B. Orgogozo, F. Viger, T. Friedman, M. Latapy, C. Magnien, and R. Teixeira, "Avoiding traceroute anomalies with paris traceroute," in *Proc. of IMC'06*, Rio de Janeiro, Brazil, Oct. 2006.

[18] E. Katz-Bassett, H. V. Madhyastha, V. K. Adhikari, C. Scott, J. Sherry, P. van Wesep, T. Anderson, and A. Krishnamurthy, "Reverse traceroute," in *Proc. of USENIX NSDI'10*, San Jose, CA, USA, Apr. 2010.

[19] 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, USA, Mar. 2010.

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

[21] 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.

[22] B. Eriksson, P. Barford, J. Sommersy, and R. Nowak, "A learning-based approach for IP geolocation," in *Proc. of PAM'10*, Zurich, Switzerland, Apr. 2010.

[23] D. Li, J. Chen, C. Guo, Y. Liu, J. Zhang, Z.-L. Zhang, and Y. Zhang, "IP-geolocation mapping for moderately-connected Internet regions," *IEEE Trans. Parallel and Distributed Systems*, 2012, http://doi.ieeecomputersociety.org/10.1109/TPDS.2012.136.

[24] Y. Shavitt and N. Zilberman, "A geolocation databases study," *IEEE J. on Selected Areas in Communications*, vol. 29, no. 10, pp. 2044 – 2056, 2011.

[25] 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.

[26] 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, Apr. 2009.