Inferring POP-Level ISP Topology through End-to-End Delay Measurement*

Kaoru Yoshida[1], Yutaka Kikuchi[2], Masateru Yamamoto[3], Yoriko Fujii[4], Ken'ichi Nagami[5], Ikuo Nakagawa[5], and Hiroshi Esaki[1]

[1] Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
[2] Kochi University of Technology
[3] Cyberlinks co.,LTD
[4] Keio University
[5] Intec Netcore, Inc.

Abstract. In this paper, we propose a new topology inference technique that aims to reveals how ISPs deploy their layer two and three networks at the POP level, without relying on ISP core network information such as router hops and domain names. This is because, even though most of previous works in this field leverage core network information to infer ISP topologies, some of our measured ISPs filter ICMP packets and do not allow us to access core network information through traceroute. And, several researchers point out that such information is not always reliable. So, to infer ISP core network topology without relying on ISP releasing information, we deploy systems to measure end-to-end communication delay between residential users, and map the collected delay and corresponding POP-by-POP paths. In our inference process, we introduce assumptions about how ISPs tend to deploy their layer one and two networks. To validate our methodology, we measure end-to-end communication delay of four nationwide ISPs between thirteen different cities in Japan and infer their POP-level topologies.

Keywords: End-to-end measurement, network tomography, communication delay, Japanese Internet.

1 Introduction

When inferring ISP topologies and identifying locations of their network elements (such as routers), researchers often rely on ISP core network information, such as the domain names of the router interfaces. Indeed many previous works (e.g., [1,2]) rely on ISP core network information to infer ISP topologies or identify network element locations, but we observe that some of measured Japanese ISPs filter ICMP packets, therefore we cannot even have access to their core network information. And also, Zhang *et al.* points out that the domain names sometimes do not represent accurate geographical locations of network elements[3].

* This work has been supported in part by Ministry of Internal Affairs and Communications in Japan.

S.B. Moon et al. (Eds.): PAM 2009, LNCS 5448, pp. 35–44, 2009.

Moreover, ISPs sometimes outsource designs and deployments of their layer two networks, and therefore, they do not even know where their layer two links are laid down.

In this paper, we propose a new topology inference technique that aims to reveal how ISPs deploy their layer two and three networks at the POP (Point Of Presence) level, without relying on ISP core network information. To infer such topological properties of ISP networks, we deploy systems to measure communication delay between residential users and map the collected delay and corresponding geographical POP-level paths. Our approach is based on an assumption that communication delay between users closely depends on the length of their communication path over optical fibers or copper cables. Even though it is true that a transmission delay in an access network is relatively large due to its lower speed, we could eliminate the factor under some circumstances (described in Sec. 2). By eliminating access delays, we try to map core network delays, which are derived from end-to-end delays and access delays, and their corresponding POP-level paths.

Since how carrier services lay their optical fibers is one of the important issues to map them in practice, we introduce Japan specific circumstances that major carrier services (e.g., KDDI[1] and SBTM[2]) were established as part of other infrastructure services such as railroads or expressways and optical fibers are presumably laid along those infrastructures. Through leveraging distances derived from those infrastructures, we try to map the core network delays and corresponding geographical paths. The result reveals that the Japanese Internet has the following two characteristics: 1)Some of the measured ISPs have hub-and-spoke topologies where hubs are the most populated cities in Japan such as Tokyo and Osaka; 2)All of the ISPs exchange their customer traffic at the cities.

The rest of this paper is organized as follows. In Section 2, we briefly describe our inference methodology. Section 3 shows our measurement environment and approach, in practice. In Section 4, we classify Japanese ISPs and infer POP-level ISP topologies based on a classification. We then present related works in Section 5 and finally summarize the discussion of this paper in Section 6.

2 Inference Methodology

Since our motivation of this work is to explore where ISPs deploy their POPs and how POPs are connected with each other, through end-to-end delay measurements, we briefly describe a communication path between residential users. When residential users communicate with each other, communication paths between them consist of both access (layer two) and ISP core (layer three) networks. Therefore, an end-to-end communication delay between residential users can be described as below.

$$delay(src, dst) = ad_{src} + ad_{dst} + CD(src, dst) + E_{src,dst} \quad (1)$$

[1] `http://www.kddi.com`

[2] `http://www.softbanktelecom.co.jp/`

Here, src and dst are nodes connected to the Internet, and communicate with each other; $delay(src, dst)$ denotes an end-to-end communication delay between src and dst; ad_{src} is the access delay at src and $CD(src, dst)$ is the delay of ISP core networks between src and dst; $E_{src,dst}$ is the measurement error of the delay.

If Internet access services are served by LECs (Local Exchange Carriers), especially in case that LECs provide DSL (Digital Subscriber Line) services, a detailed communication path can be described as follows: (1) measurement node $\leftrightarrow$ BRAS (Broadband Remote Access Server) that aggregate user sessions from the Internet access services; (2) BRAS $\leftrightarrow$ ISP CE (Customer's Edge) router that is located in the closest POP to users; (3) ISP core network (shown in Fig. 1). Although all the customer sessions are aggregated at BRAS not depending on which ISP users connect to, each ISP's CE routers can be deployed anywhere based on ISP policies. ISPs that serve the access services by themselves (e.g., CATVs) also deploy CE routers, which are the same as customers' default routers. So, through measuring the end-to-end communication delay and the access delays individually, we are able to derive the core delay from (1).

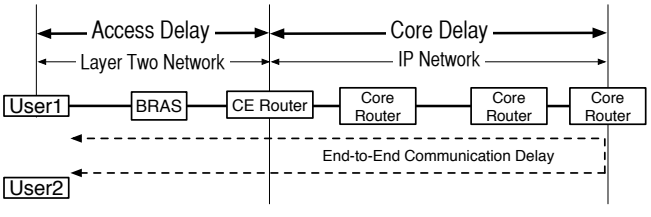


Fig. 1. Delay Model between Residential Users

To explore where ISPs deploy their POPs and how they are connected with each other, we need to map the core delay and corresponding POP-level paths. If we are able to select candidate POP locations and links among them, (1) can be transcribed into the following set of simultaneous equations.

$$delay(src, dst) = ad_{src} + ad_{dst} + \sum_{p,q \in N} x_{p,q} \times cd_{p,q} + E_{src,dst} \qquad (2)$$

Here, N denotes a set of candidate POP locations of a measured ISP; p and q satisfy $\{p, q \mid p, q \in N\}$; $cd_{p,q}$ denotes a core delay between p and q; $x_{p,q} = 1$ if a direct path between p and q exists and the path is used to connect between src and dst, otherwise $x_{p,q} = 0$. $delay(src, dst), ad_{src}$ and ad_{dst} are measurable through end-to-end measurements and $cd_{p,q}$ can be derived leveraging the distance between p and q.

In the equation, $x_{p,q} \times cd_{p,q}$ denotes the path between p and q. Since the maximum number of path patterns is almost $2^{|N|^2}$ where $|N|$ is the number of POPs, we must reduce complexities under the practical conditions of the real world. We are able to shrink the possible path patterns with the major

restriction that is operations and management (OAM) cost as follows: 1) Link cost: Traffic aggregation cuts both layer one and two cost because of getting shorter optical fibers; 2) Node cost: Aggregating layer three elements (routers) reduces the maintenance costs. We will show the possible patterns in Japan according to the restriction above in Sec.4.2, which make (2) solvable.

Moreover, since it is hard to have access to layer one and two network information in general, we introduce an assumption that those networks are usually deployed along other infrastructure services, e.g. railroads and expressways, as we described in Sec. 1.

3 Measurement Environment

For a nationwide delay measurement in Japan, we deploy thirteen measurement nodes in different cities. Some of the selected cities are the most populated cities in Japan (e.g., Tokyo and Osaka), and we also choose cities that are junctions of railroads and expressways and possibly the POP locations in Japan. Each node connects to four nationwide ISPs (ISP X, Y, Z and W) through "NTT Flet's Service" that is a nationwide layer two network service for connecting end users to their ISPs via PPPoE over optical fibers.

We implement a measurement system to measure the communication delay between IP addresses attached to measurement nodes. Each node runs measurement UNIX daemons bound to four PPPoE interfaces, respectively, and each daemon measures the communication delay between the IP address and IP addresses attached to other measurement nodes. The daemon uses Linux raw socket and libpcap[3] for both sending and receiving measurement packets. The communication delay is measured by an echo request/reply method using 64-byte UDP packets. In order to minimize the queuing delay caused by network congestion, the daemon sends three train packets to each destination every ten seconds and only retains the minimum delay of them[4].

To measure the access delay, the daemon generates and send a special packet whose source and destination IP addresses are the IP address bound to the PPPoE interface. The delay measured in this manner directly corresponds to the access delay in Sec.2.

4 ISP Topology Inference

In this section, we apply our methodology to the Japanese Internet and infer POP-level ISP topologies with measured delay data. We introduce some preconditions to make our inference more accurate: 1)The velocity of light in an optical fiber becomes 60-70% compared to it in vacuum[5]. So, the velocity of light in an optical fiber cable becomes $C' = 2/3 \times C[km/sec]$; 2) There is some overhead to process the measurement packets, since we use PPPoE sessions to connect ISPs and libpcap for capturing the packets. And, there also exist queuing delays when

[3] `http://www.tcpdump.org`

the packets go through network elements such as layer two switches, layer three routers and BRAS. we define queueing delay(qd) caused by network elements as $> 1[msec]$ (this value is derived from our preliminary experiments).

4.1 Analysis of Access Delay

Figure 2 shows the minimum access delays of measured cities and ISPs in January 2008. As it shows, the access delay trends of the ISP X, Y and W are relatively similar, while the trend of the ISP Z is quite different from them. This indicates that ISP CE routers of the ISP X, Y and W are located in the almost same places. And, most of the access delays connecting to the ISPs are around $1[msec]$, we can estimate that CE routers of these ISPs are located in the same prefecture where measurement nodes are located. On the other hand, most cities except Tokyo and Osaka in the ISP Z network has long access delays that are more than 5[msec], and this implies that the ISP Z does not deploy its CE router in each prefecture. Since the access delays of the ISP Z at Tokyo and Osaka are almost the same values of the other ISPs', we can estimate that the ISP Z deploys its POP at Tokyo and Osaka, at least.

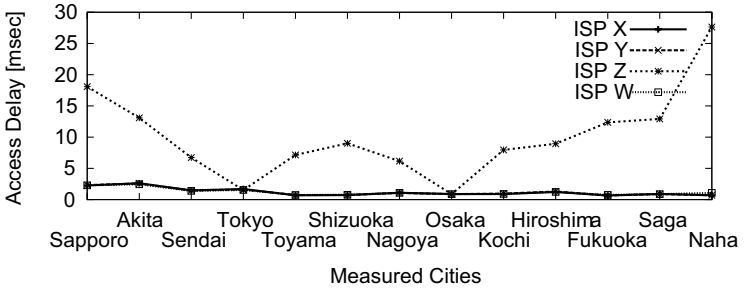


Fig. 2. Access Delay of Each City/ISP [msec]

4.2 Analysis of Core Delay

Based on the location information of the ISP CE routers in Sec.4.1, we infer ISP core network topologies. In the inference process, we use railroad distances between POPs and analyze where ISP POPs are located and how they are connected to each other. In this section, we, first, propose topology models of the Japanese Internet. Then, we try to solove (2) with the measured delay data set for the purpose of POP-level ISP topology inferences.

Topology Models of Japanese ISPs. We propose some ISP topology models of the Japanese Internet to solve (2). This intends to disclose which pairs of (p, q) make $x_{p,q} = 1$ under $E_{src,dst} < qd$ for each pair of src and dst in (2). Here, we adapt the restriction described in Sec.2 to introduce the models. And, there is one more factor that is specific to Japan. The population concentrates in Tokyo,

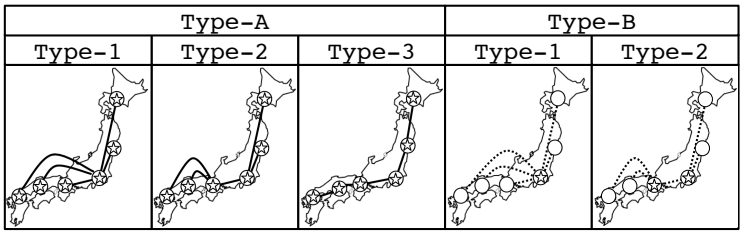


Fig. 3. ISP Topology Classifications

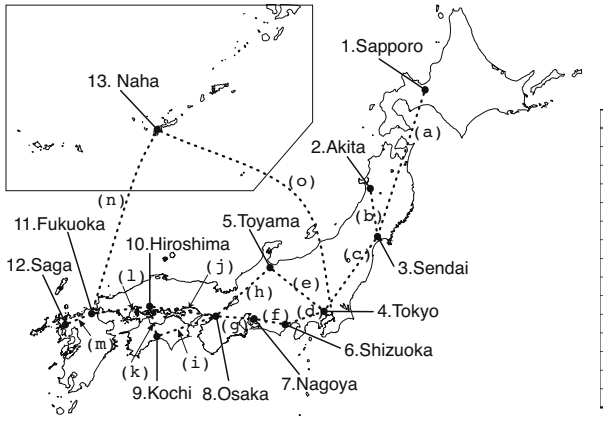
Osaka, and cities along the Pacific coast between Tokyo and Osaka, therefore ISPs have planned the link and node aggregation based on these conditions.

We classify ISP topologies into the followings in Fig.3. Here, a star denotes a layer three router and solid lines between routers are an ISP backbone network. And a circle denotes a layer two switch and dashed lines between switches are an ISP access network that is same as the NTT Flet's service.

The difference between Type-1, 2 and 3 is how ISP backbone networks are structured. A Type-1 network simply aggregates all the customer traffic to Tokyo, and a Type-2 network aggregates them to Tokyo and Osaka. A Type-3 network, on the other hand, deploys routers in more or every prefecture and connects next to each other. Layer three nodes are completely reduced in a Type-1 network, and a Type-3 network maximizes link aggregation. A Type-2 network is an intermediate one of the Type-1 and Type-3. It would depend on the operation policy of a ISP's management.

The difference between Type-A and Type-B is how ISPs rely on NTT Flet's service for their layer two networks. In case of Type-A network, an ISP deploys its layer three routers in each prefecture and connects to NTT Flet's service at there. In case of Type-B networks, on the other hand, an ISP aggregates its customer traffic through NTT Flet's service and only deploys its layer three routers in the most populated prefectures such as Tokyo and Osaka. A Type-B3 network does not exist, because it indicates that all the customer traffic is exchanged through a layer two service.

Note that there are following two communication restrictions of NTT Flet's service; 1)All the PPPoE sessions connected from a measurement node are terminated at a single PPPoE accommodation called BRAS which is managed by NTT Flet's service. 2)Even though NTT Flet's service provides layer two networks for ISPs, users who connect to the same ISP at the same prefecture cannot communicate with each other through the layer two network by the nature of the service. So, if two users connect to a Type-B1 ISP, traffic between them always goes through Tokyo even if they are in the same prefecture. Since we only focus on ISPs that apply NTT Flet's service as their layer two networks in this paper, the possible network structures are covered by the above classifications. This is because NTT Flet's service should deploy layer two switches in every prefecture due to legal regulations and ISPs can only construct their network based on the layer two structure.



Leg	Distance [km]
(a)	855.2
(b)	310.8
(c)	351.8
(d)	180.2
(e)	401.2
(f)	185.5
(g)	186.6
(h)	327.0
(i)	363.4
(j)	341.6
(k)	340.6
(l)	280.7
(m)	53.6
(n)	864.0
(o)	1584.0

Fig. 4. Location of Measurement Nodes and Distance between Cities

Core Delay between POPs. The rest of the restrictions to solve the equations(2) is to determine the link delay denoted as $cd(p,q)$. We assume $cd(p,q) = RD(p,q)/C'$ where C' is light speed in optical fiber described in Sec.4, and $RD(p,q)$ is derived from geographical information shown in Fig.4[4]. Here, numerical symbols denote cities we set up measurement nodes and alphabetical symbols denote railroad distance RD between POPs. Since the RD is a distance between stations, there is some distance between stations and residences. We assume that the distance between users is approximately 10% longer than RD.

Inferring POP-Level ISP Topologies. Solving the simultaneous equations with the measured delays, we infer the four ISP topologies as follows.

ISP X: We use the same data set in Sec. 4.1 and Table 1 shows the end-to-end delays of the ISP X. The ISP X is a Tokyo centric network, that is a TYPE-A1 network, because the ISP deploys its CE routers in each prefecture, and the core delays between Tokyo and most of the other cities closely correlate with the distances.

One exception we find interesting is that the end-to-end delay between Tokyo and Shizuoka is larger than the corresponding delay between Tokyo and Nagoya, while the distance of the former is shorter than it of the latter. To figure out the reason, we introduce an assumption that the path between them goes through POPs away from both Tokyo and Shizuoka. The core delay between Tokyo and Shizuoka is about 8.0$[msec]$, therefore the fiber length would be around 800$[km]$ where the length between Tokyo and Shizuoka is about 180$[km]$. Since NTT Flet's Service in Shizuoka is operated by NTT West[5] and NTT West has a huge data center in Osaka, one possible assumption is that the link detours via Osaka.

[4] JTB Timetable (Sep., 2008, ISBN:4910051250988)

[5] `http://ntt.flets-web.com/en/west/`

Table 1. End-to-End Delays of the ISP X [msec]

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	2.32	36.25	27.52	23.21	35.36	31.53	27.62	29.64	36.78	34.51	37.69	39.04	52.85
2	36.37	2.63	21.08	16.92	28.39	24.50	21.40	23.19	30.73	28.34	31.36	34.38	45.70
3	27.44	20.94	1.45	9.76	20.95	16.16	14.25	16.05	20.35	21.23	24.23	32.25	37.79
4	23.21	17.23	9.85	1.58	16.99	13.14	9.88	11.67	18.78	16.63	19.74	20.81	34.32
5	34.90	27.83	20.83	16.72	0.73	22.89	21.01	24.10	27.22	27.93	31.02	39.14	44.58
6	31.43	24.62	16.03	12.58	22.87	0.77	17.93	20.97	24.05	24.70	26.04	36.01	45.41
7	27.70	21.42	14.13	9.66	21.18	17.93	1.09	16.20	23.23	20.92	24.23	25.37	39.42
8	29.48	23.30	16.00	11.62	24.08	20.57	16.13	0.94	25.30	23.11	26.21	27.50	41.89
9	36.63	30.53	20.18	18.67	27.27	24.09	23.20	25.37	0.95	30.20	30.43	34.61	45.77
10	34.57	28.27	21.15	16.52	27.92	24.75	20.93	23.17	30.19	1.25	31.07	32.32	46.36
11	37.63	31.47	24.31	19.61	31.04	26.09	24.05	26.27	30.44	31.05	0.75	37.36	47.68
12	38.99	34.37	32.37	20.95	39.12	36.08	25.36	27.62	34.65	32.37	37.52	0.98	53.10
13	52.77	45.62	37.63	33.89	44.71	45.41	39.38	42.07	45.75	46.39	47.78	53.10	0.68

So, we can conclude that the layer two network between Tokyo and Shizuoka goes through Osaka, even though Shizuoka and Osaka are not adjacent with each other in the layer three network.

ISP Y: The ISP Y has characteristics of Type-A2 network except the network aggregates traffic not only to Tokyo and Osaka but also to Fukuoka that is the largest city in Kyushu Island. We infer this as follows. Sapporo and Sendai are the neighbors of Tokyo and Akita connects to Sendai. Toyama, Shizuoka, Nagoya, Kochi and Hiroshima are the neighbors of Osaka. Tokyo, Nagoya and Osaka connect to each other. And, Saga and Naha are the neighbors of Fukuoka. Since the core delay between Tokyo and Shizuoka is larger than the expected value derived from the geographical distance, the ISP Y also has a detour path between them.

ISP Z: Since we classify the ISP Z network as a Type-B network in Sec.4.1, we only infer where the ISP Z deploys its CE routers and how each node connects to CE routers. Based on the collected delay data, we infer that the ISP Z deploys its CE routers in Tokyo and Osaka. And, Tokyo is the neighbor of Sapporo, Akita and Sendai and Osaka is the neighbor of Tokyo and Toyama, Shizuoka, Nagoya, Kochi, Hiroshima, Fukuoka, Saga and Naha. Therefore the layer three network of the ISP Z only exists between Osaka and Tokyo.

ISP W: Solving the simultaneous equation with the delay data, we infer the ISP W topology is a TYPE-A3 as follows. Most cities connect to the neighbor cities. Here, the neighbor of Toyama is Tokyo; the neighbors of Kochi are both Osaka and Hiroshima; and the neighbor of Naha is Fukuoka.

Figure 5 shows topological properties of ISP networks that we infer through the above processes. These ISP topology inferences are convinced by traceroute results and anonymous network operator sights. The results indicate that communication delay between users even in the same ISP differs depending on which ISP users connect to.

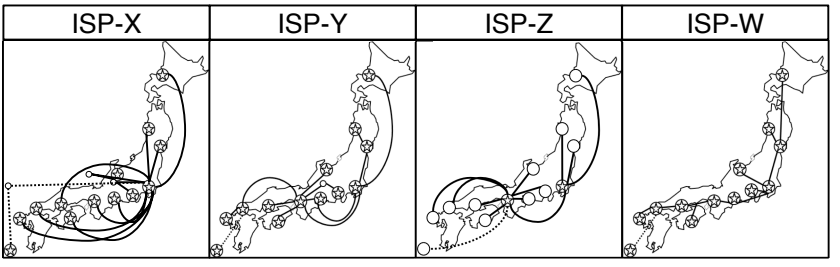


Fig. 5. ISP Topologies inferred by Our Approach

In addition, according to the result in case of ISP-X, the dissociations between layer two and three networks cannot be disclosed if we only use traceroute or other layer three information based measurements.

5 Related Work

Spring *et al.* propose Rocketfuel[1] to infer router level ISP topologies. Rocketfuel uses traceroute and aims to explore adjacency relationships between routers. Teixeira *et al.* point out that POP level topologies inferred by Rocketful have significant path diversity, and therefore they introduce the heuristic approach to improve the accuracy of the inferred Rocketfuel topologies[2]. In [6], Augustin *et al.* propose Paris traceroute to explore more accurate routing path compared to existing traceroutes. Different from these works, our approach aims to infer POP-level ISP topologies without relying on the ISP core information.

Network tomography is a research field that aims to figure out network characteristics through end-to-end measurements. Coates *et al.* introduce an overview of tomographic approaches for inferring link-level network performance[7]. Since their approach basically analyzes network characteristics from a single source point of view, Rabbat *et al.* propose new approaches that explore network characteristics through multiple source measurements[8]. We also investigate link-level network characteristics through multiple source measurements in [9] and furthermore infer ISP topologies with the same collected delay data set.

6 Conclusion

In this paper, we present a new approach for inferring POP-level ISP topologies. Since our approach leverages the end-to-end communication delay between residential users, layer one and two information and candidate POP locations for topology inferences, it is different from any previous works that require ISP core network information such as domain names of routers. Considering that ISPs tend to hide both their layer two and three structures including domain names of their equipments, our approach should become one of realistic approaches to exploring ISP topologies. Since it has been common that ISPs independently construct their layer two and three networks in these days, taking account of end-to-end communication characteristics is necessary to infer ISP topologies even for ISPs.

We apply round-trip delay measurements to infer ISP topologies based on an assumption that a round-trip path is identical. Even though this assumption is true in this paper, we need take the fact that there are asymmetric paths between users into account as our future work.

References

1. Spring, N., Mahajan, R., Wetherall, D., Anderson, T.: Measuring ISP topologies with rocketfuel. IEEE/ACM Trans. Netw. 12(1), 2–16 (2004)
2. Teixeira, R., Marzullo, K., Savage, S., Voelker, G.M.: In search of path diversity in ISP networks. In: IMC 2003: Proceedings of the 3rd ACM SIGCOMM conference on Internet measurement, pp. 313–318. ACM, New York (2003)
3. Zhang, M., Ruan, Y., Pai, V., Rexford, J.: How dns misnaming distorts internet topology mapping. In: ATEC 2006: Proceedings of the annual conference on USENIX 2006 Annual Technical Conference, Berkeley, CA, USA, USENIX Association, pp. 34–34 (2006)
4. Jacobson, V.: pathchar — a tool to infer characteristics of Internet paths, MSRI Presentation (April 1997)
5. Okamoto, K.: Fundamentals of Optical Waveguides. Academic Press, San Diego (2000)
6. Augustin, B., Cuvellier, X., Orgogozo, B., Viger, F., Friedman, T., Latapy, M., Magnien, C., Teixeira, R.: Avoiding traceroute anomalies with Paris traceroute. In: IMC 2006: Proceedings of the 6th ACM SIGCOMM conference on Internet measurement, pp. 153–158. ACM, New York (2006)
7. Coates, A., Hero, Nowak, R., Yu, B.: Internet tomography. Signal Processing Magazine 19(3), 47–65 (2002)
8. Rabbat, M., Coates, M., Nowak, R.D.: Multiple-Source Internet Tomography. IEEE Journal on Selected Areas in Communications 24(12), 2221–2234 (2006)
9. Yoshida, K., Fujii, Y., Kikuchi, Y., Yamamoto, M., Nagami, K., Nakagawa, I., Esaki, H.: A Trend Analysis of Delay and Packet Loss in Broadband Internet Environment through End Customers View. IEICE Transaction on Communications J91-B(10), 1182–1192 (2008) (in Japanese)