On Retrieval Order of Statistics Information from OpenFlow Switches to Locate Lossy Links by Network Tomographic Refinement

On Retrieval Order of Statistics Information from OpenFlow Switches to Locate Lossy Links by Network Tomographic Refinement

Takemi Nakamura, Masahiro Shibata and Masato Tsuru

Abstract To maintain service quality and availability in managed networks, detecting and locating high loss-rate links (i.e., lossy links that are likely congested or physically unstable) in a fast and light-weight manner is required. In our previous study, we proposed a framework of network-assisted location of lossy links on OpenFlow networks. In the framework, a measurement host launches a series of multicast probe packets traversing all full-duplex links; and then the controller retrieves statistics on the arrival of those probe packets at different input ports on different switches and compares them to locate high loss-rate links. The number of accesses to switches required to locate all lossy links strongly depends on the retrieval order in collecting the statistics and should be small as much as possible. Therefore, in this paper, to minimize the necessary number of accesses, we develop a new location scheme with an appropriate retrieval order using a Bayesian-based network tomography to refine candidates for lossy links. The results of numerical simulation on a real-world topology demonstrate the effectiveness of the new location scheme.

Takemi Nakamura
Computer Science and System Engineering, Kyushu Institute of Technology, Fukuoka, Japan
e-mail: nakamurat@infonet.cse.kyutech.ac.jp

Masahiro Shibata
Computer Science and System Engineering, Kyushu Institute of Technology, Fukuoka, Japan
e-mail: shibata@cse.kyutech.ac.jp

Masato Tsuru
Computer Science and System Engineering, Kyushu Institute of Technology, Fukuoka, Japan
e-mail: tsuru@cse.kyutech.ac.jp

1 Introduction

The recent proliferation of cloud and edge-could computing technologies requires flexible, dynamic, and reliable networking among geographically distributed but centrally managed servers and sites. Therefore, SDN (Software defined network) in general and OpenFlow technology in particular have been applied not only to data centers but also to enterprise networks and wide area [1].

To maintain service quality and availability in such networks, detecting and locating high loss-rate links that are likely congested or physically unstable in a fast and light-weight manner is essential. In general, network operators constantly monitor the communication performance and the internal status of links by either passive or active measurement. Passive measurement in OpenFlow networks is quite useful. In addition to traditional SNMP monitoring, per-flow statistics can be monitored and collected by FlowStats function. However, although passive measurement itself does not incur additional load on the data plane, frequent accesses to switches for accurate and timely monitoring may incur additional load on the control plane.

Active measurement in OpenFlow networks is also attractive due to the capability of per-flow flexible routing. Probe packets can flow on any designed routes to measure the packet loss, delay, the round-trip-time (RTT), and so on. Furthermore, in networks connecting geographically-wider locations, a "link" between two nodes is not always physical but virtual (e.g., tunneling). So passive measurement only on managed nodes is not enough and active measurement is essential to monitor the entire network. However, probing at a high sending rate for precise and reliable monitoring incurs unnecessary load on switches and the data plane. An infrastructure was proposed to monitor RTT with suppressing the number of flow entries and probe packets [2]. A delay monitoring that covers all links in both directions with minimizing flow entries on switches was also studied [3].

Network-tomographic approaches have been studied to infer network link states without directly monitoring those links [4]. Original network tomography monitors packet-level correlations among measurement paths, which was thought as too costly in practice. Then the Boolean network tomography was proposed that only monitors performance-level correlations among measurement paths to infer the location of bad links [5] and followed by a number of studies because of its practicality (e.g.,[6]). The impact of the capability of routing of probe packets has also been studied in localizing failed nodes based on Boolean network tomography [7].

In our previous study [8], we proposed a framework of detecting and locating lossy links on OpenFlow networks with a light load on both the data and control planes by a collaboration of switches and controller with measurement host. In the framework, the retrieval order in collecting the statistics from switches strongly impacts the number of accesses to switches required to locate all lossy links, i.e., the load incurred on the control plane. Therefore we adopted a simple Boolean network-tomographic inference of highly lossy links to design an appropriate retrieval order. However, this approach is efficient only when the number of lossy links is small or lossy links are commonly shared by many measurement paths.

In this paper, therefore, to minimize the necessary number of accesses, we develop a new location scheme on retrieval order using a Bayesian-based network tomographic refinement of the candidates for lossy links. It uses correlation of loss events of each probe packet that can be monitored by an extension of per-flow statistics of OpenFlow. This extension is feasible and light-weight by using ID field of IP packet. It also requires the prior loss probability of link calculated by using past measurement results. Note that it can be tolerant in inference-error because the aim is not to precisely infer the link loss rates but to promptly locate lossy links. To reduce the computational cost of posterior probabilities, a series-reduced tree is used instead of the original multicast measurement tree.

Section 2 explains the framework and the basic location scheme we previously proposed. Section 3 proposes a new location scheme. Simulation evaluation of both basic and proposed schemes on a real-world network topology is performed in Section 4, followed by the concluding remarks in the last section.

2 Overview and the basic location scheme

The framework assumes a network comprising OpenFlow controller (OFC) and OpenFlow switches (OFS). The process starts when the measurement host (MH) sends a request to the OFC as illustrated in Fig. 1. Next, the OFC gets the network topology, calculates probe packet routes, and installs them to OFSs.

Then, the MH launches a series of multicast probe packets traversing all links once and only once (separately in each direction of the full duplex link) to minimize the load on the data plane incurred by probe packets. The probe packets are discarded at "leaf port" on the last OFS of the measurement path. Then, the OFC retrieves statistics on the arrival of those probe packets (i.e., the number of probe packets arriving) at different input ports on different OFSs and compares them to locate high loss-rate links. The number of lost packets on a link (or series of links) between the two switch ports can be calculated by taking the difference in the number of probe packets arriving at those ports.

An example of the route configuration is shown in Fig. 2. The root port is a switch port connected to the MH, and the leaf port is a switch port for discarding probe packets. A path of the probe packet (i.e., measurement flow) from the root port to a leaf port is called "terminal path".

In our framework, measurement routes through which probe packets flow is designed by three steps. Please see more details in [8].

- Generate the shortest path tree in the downward direction from the root (blue dashed lines in Fig. 3).
- Complement unused links not on the shortest path tree (green dotted lines).
- Add return links in the upward direction bound for the root (red lines).

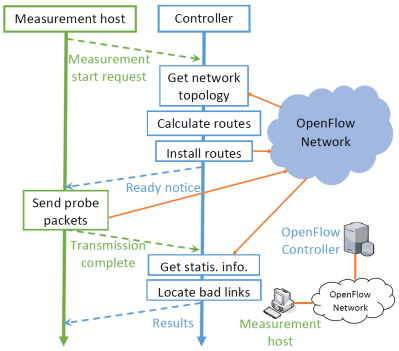

Fig. 1 Measurement process [8]

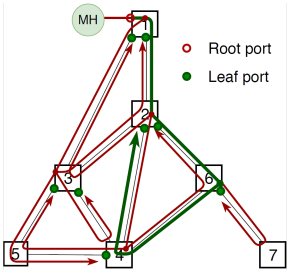

Fig. 2 Route scheme example [8]

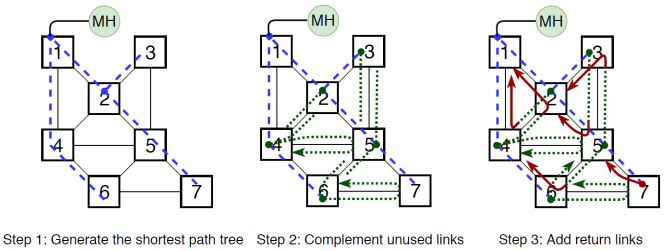

Fig. 3 Route scheme design [8]

After a series of probe packets flows on all links, the MH informs the OFC of the probing completion. Then, in our previous study, the following algorithm (i.e., the basic location scheme) is used to find lossy links, which determines an appropriate access order to selected OFSs to collect flow statistics. A link is considered as lossy if and only if its loss rate exceeds a threshold value h that is a design parameter representing the target link quality to maintain and depends on the target applications. The packet loss rate of a segment (a series of links) from ports i to j, PLR, can be computed by r_i and r_j that are the numbers of probe packets arriving at switch ports i and j, respectively.

$$PLR = 1 - \frac{r_j}{r_i} \tag{1}$$

First, the OFC accesses to the root port (port 0) and each (j) of all leaf ports to retrieve r_0 and r_j, respectively, in order to calculate the PLR of each terminal path, using equation (1). If the PLR of a terminal path is less than h, it does not include any lossy link. If the PLR of a terminal path exceeds h, this terminal path is likely to include one or more lossy links.

Then, we narrow the search range, i.e., the expected locations of lossy links, and determine the retrieval order. If a terminal path is lossy (i.e., its PLR value exceeds threshold h) and there are no other lossy terminal paths, the lossy links are located within a segment between the leaf port and the nearest parent port on the considered lossy terminal path, as illustrated by the dashed line in Case 1 of Fig. 4.

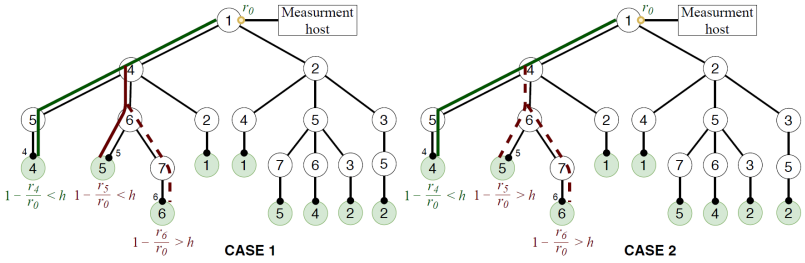

Fig. 4 Basic location scheme

The ports along this segment are queried in a binary-search manner until lossy links in this segment are located. If there are multiple lossy terminal paths, the port most commonly shared by those paths is queried first to collect the number r_j of arrival probe packets, as illustrated by the dashed line in Case 2 of Fig. 4. The access and packet loss rate are checked from the most upstream port in the common ports. This procedure generates separated sub-trees, and the same procedure is performed on each sub-tree recursively until all lossy segments are identified in the trees.

3 Proposed location scheme

We propose a new location scheme with an appropriate retrieval order using a Bayesian-based network tomography to probabilistically refine candidates for lossy links. The packet-level loss event correlation among terminal paths is monitored. A unique ID is assigned to each probe packet at the MH, which is essential for the OFC to monitor the packet-level loss event correlation. Then, to compute the posterior loss probability (i.e., expected packet loss rate) on each link based on the monitored events on terminal paths, we need the prior loss probability of link.

Here, a link that often exhibits a high packet loss rate, i.e., at high failure level, lately is assumed to cause a packet loss with a high probability in near future. Therefore, the prior loss probability of link is calculated based on the measured loss rates in the past measurements.

Let l be the current measurement cycle, m be the number of links of the network, and b_{kl} be the prior loss probability on each link $k\,(1 \leq k \leq m)$ in cycle l. The value b_{kl} reflects the latest measured loss rate p_k of link k by using the Exponential Moving Average (EMA) with smoothing factor α. $b_{k(l+1)}$ is updated as follows.

$$b_{k(l+1)} = (1-\alpha)b_{kl} + \alpha p_k \tag{2}$$

In case that p_k is not measured in cycle l, $b_{k(l+1)} = b_{kl}$. In the initial cycle 1, b_{k1} is a given initial value identical for any link k.

The posterior loss probability q_k is the conditional probability of packet loss on link k given correlated loss events occurred over measurement paths. Here, L_k is the

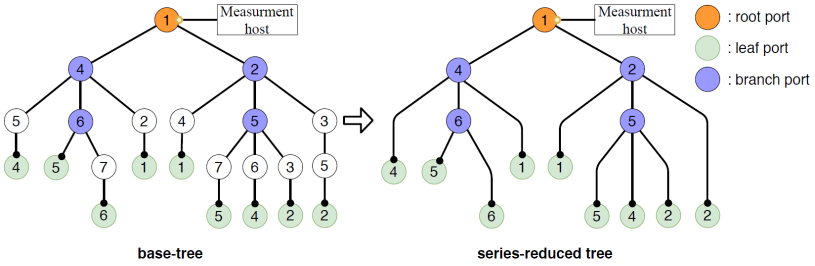

Fig. 5 Making series-reduced tree

number of packet losses on link k for n probe packets, Let Y_i be 1 if a probe packet is lost on terminal path $i\,(1 \leq i \leq z)$ and be 0 otherwise. And $M_j = (Y_1, ..., Y_z)$ is the correlated loss events over z paths for the j-th probe packet: $1 \leq j \leq n$.

Based on measured $M_1, ..., M_n$ and the prior loss probabilities of all links, q_k is calculated by the following equation:

$$q_k = \frac{E[L_k | M_1, ..., M_n]}{n} \tag{3}$$

By letting X_k be 1 if a probe packet is lost on link k; be 0 if the packet is not lost; and be "*" if unknown (the packet does not arrive at k), the following equation holds.

$$\frac{E[L_k | M_1, ..., M_n]}{n} = \frac{\Sigma_{j=1}^{n} P(X_k = 1 | M_j)}{n} = \frac{\Sigma_{j=1}^{n} \frac{P(X_k = 1 \cap M_j)}{P(M_j)}}{n} \tag{4}$$

$P(M_j)$ is calculated based on possible combinations of packet loss occurrences $\mathbf{X} = (X_1, ..., X_m)$ consistent with M_j, and $P(X_k = 1 \cap M_j)$ is calculated based on possible combinations of packet loss occurrences consistent with $X_k = 1$ and M_j.

Here, the probability of $X_k = 1$ is set to b_{kl}, and the probability of $X_k = 0$ is set to $(1 - b_{kl})$ to obtain q_k. Link k with a value of q_k exceeding h is regarded as suspected link that is likely a lossy link; it is used for the algorithm.

However, as the number of links inevitably increases in a large-scale topology, the number of combinations of $\mathbf{X}$ to be estimated becomes enormous, and the calculation time of the posterior probability increases, resulting in a large delay to locate lossy links. Therefore, we introduce series-reduced tree to solve this problem.

A path tree for transmitting probe packets is called "base-tree". By directly connecting the root port, leaf ports, and branch ports of a base-tree, a path tree reduced in scale is created and called a series-reduced tree like Fig. 5. Here, the link in the series-reduced tree is called "reduction link" k'. Note the series-reduced tree is used only to calculate the posterior probability and to narrow down the lossy links at first.

After calculating and updating b_{kl}, the prior loss probability $g_{k'}$ on reduction link k' is calculated. If reduction link k' comprises t links $(1 \leq k \leq t)$ of the base-tree, $g_{k'}$ in a given cycle l is calculated from b_{kl} as follows.

$$g_{k'} = 1 - (1 - b_{1l})(1 - b_{2l})...(1 - b_{tl}) \quad (5)$$

Then the posterior loss probability $q_{k'}$ of reduction link k' is calculated from $g_{k'}$, which is used to narrow a possibly-lossy segment as "range".

A reduction link with a value of the posterior loss probability $q_{k'}$ exceeding h is likely to include lossy links, and can be an initial range. We start from the reduction link with the higher posterior loss probability. The actual packet loss rate of the range is checked using equation (1). If it exceeds h, the process proceeds to identify one or more lossy links within the range.

Here, the following two cases of the prior loss probabilities in a range are considered. If all b_{kl} values are the same, binary search is performed. Otherwise, the prior loss probability-based narrowing is performed.

A. Binary search
The statistics is obtained from the center port of the range to be narrowed down, and the range divided based on the port, then the packet loss rate is checked for each range.

B. The prior loss probability-based refinement
The link with the highest prior loss probability within the range is regarded as a suspected link. The statistics is obtained from the port on the downstream side of the suspected link, and the range is divided at the port, then the packet loss rate of each range is checked. When the packet loss rate in the upstream range exceeds h, the statistics is obtained from the upstream port of the suspected link, and the packet loss rate of the link is checked.

After the narrowing within selected reduction link is completed, the process moves to another reduction link which has the next higher posterior loss probability.

4 Simulation evaluation

To evaluate the proposed location scheme compared with the basic location scheme, numerical simulation are performed on a real-world network topology, from a topology database [9], shown in Fig. 6 (MH is connected).

There are 19 OFSs and 24 links (48 links in both directions); the number n of probe packets is 1000, threshold h of lossy link is 0.01, smoothing factor α is 0.3, initial value b_{k1} of prior loss probabilities is 0.001, and one simulation runs for 10 measurement cycles. The packet loss rate of each link is randomly set to 0.05∼0.1 (high failure level), 0.001∼0.01 (low failure level), or 0.001(normal link). The path tree is shown in Fig. 7. The number in each circle is an ID of input (downstream) port of a link illustrated by an arrow-line, and the number also identifies the link.

Here, low loss-rate links are positioned on (6, 45, 7, 32, 21) in both two patterns. Fig. 8 (resp., Fig. 9) shows the number of accesses to switches required to retrieve statistics of all terminal paths (yellow-coloerd); and at additional three (resp., four) stages before all lossy links are detected in each measurement cycle (differently

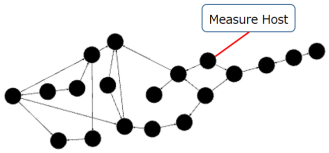

Fig. 6 Simulated network topology

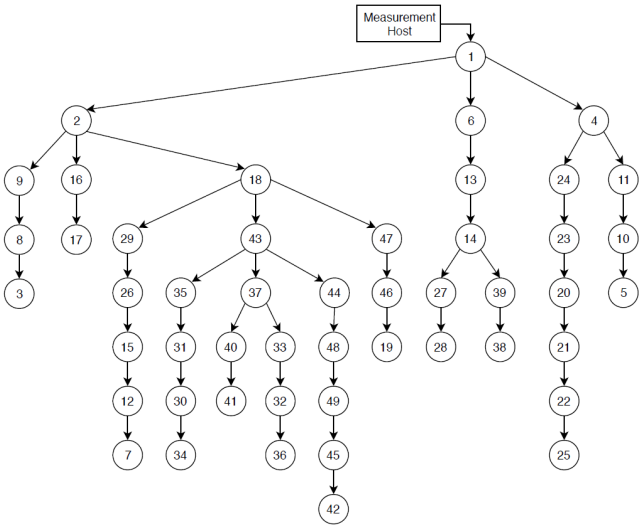

Fig. 7 Generated path tree

colored). Note that each high loss-rate link always becomes lossy (i.e., its actual loss rate exceeds h) in every measurement cycle and each low loss-rate link unlikely but sometimes becomes lossy.

Pattern 1 : high loss-rate links (2, 4, 13) (Fig. 8)

In Pattern 1, a few number of high loss-rate links are set and positioned in the most common parts of all terminal paths. In such case, those terminal paths are observed as lossy and correlated at the packet-level loss. Then the basic scheme always simply checks the common part first, which can result in the most efficient retrieval order of port accesses in this case. Therefore no advantage by the proposed scheme is expected. Furthermore, if a low loss-rate link accidentally becomes lossy in some cycle, it may result in an inefficient retrieval order in the proposed scheme as shown in the 3, 4, and 7th cycles in Fig. 8. This is because the prior loss probability of such a low loss-rate link is generally low and it decreases the search priority of that low loss-rate link (but it is actually lossy in this cycle).

Pattern 2 : high loss-rate links (10, 20, 39, 27, 18, 16, 8) (Fig. 9)

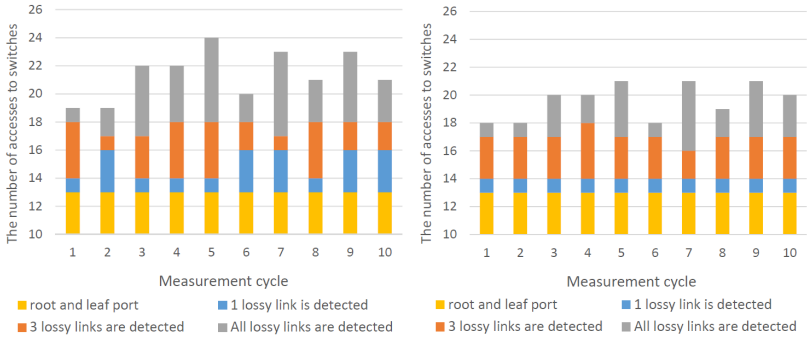

Fig. 8 Pattern 1 (left: proposed, right: basic)

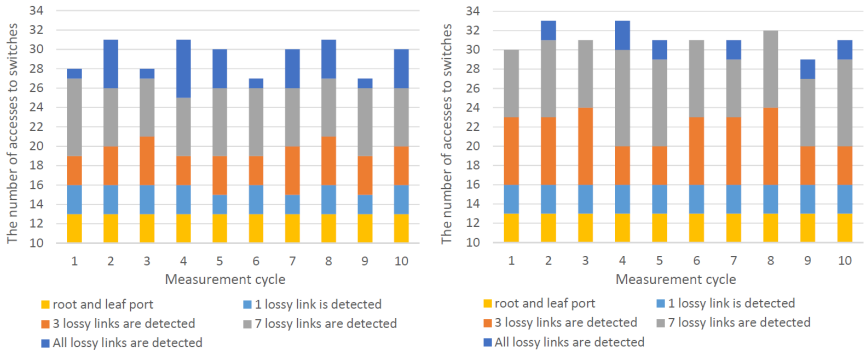

Fig. 9 Pattern 2 (left: proposed, right: basic)

In Pattern 2, a more number of high loss-rate links are set and positioned in less-common and downstream parts of terminal paths. In such case, those terminal paths are observed as lossy but almost independent (uncorrelated) at the packet-level loss. Then, again, the basic scheme always simply checks the common part first but it is not the actual lossy part in this case. On the other hand, even in the first measurement cycle, the proposed scheme utilizes the packet-level loss event correlation among terminal paths to infer which reduction link is more likely to be lossy. It computes a high posterior loss probabilities of the reduction links on which the high loss-rate links are included, and likely results in a better retrieval order. In the second and later cycles, since the positions of high loss-rate links are unchanged, both the prior and posterior loss probabilities of high loss-rate links are kept high, and thus a better retrieval order is kept as well.

However, similarly to pattern 1, if a low loss-rate link accidentally becomes lossy, it may result in a disadvantage of the proposed scheme as shown in the 2, 4, 5, 7, and 10th cycles in Fig. 9.

Table 1 shows the average number of retrievals to find all lossy links in pattern 2 over 10 trials of simulation; the proposed scheme outperforms the basic at any cycle.

Table 1 Average number of retrievals to find all lossy links in pattern 2

measurement cycle	1	2	3	4	5	6	7	8	9	10
propose	30.2	29.2	30.0	28.9	30.1	29.4	29.6	28.6	28.3	29.2
basic	31.3	31.6	32.2	31.1	32.0	32.0	31.2	31.3	30.6	31.2

5 Concluding remakes

Based on our previously proposed framework to monitor and locate lossy links on OpenFlow networks, we have developed a new location scheme to reduce the necessary number of accesses in collecting the statistics from switches. A Bayesian-based network tomography on a series-reduced tree is adopted to refine the search range with an acceptable computational cost. The simulation results show the new scheme can efficiently locate all lossy links with a fewer number (less than the 60% of the total number of links) of accesses to switches even in case that the basic scheme is not efficient.

Acknowledgements The research results have been achieved by the "Resilient Edge Cloud Designed Network (19304)," NICT, and by JSPS KAKENHI JP16K00130 and JP17K00135, Japan. We thank Mr. Suguru Goto and Mr. Yuki Fujimura for assistance.

References

1. S. Jain, A. Kumar, et al., "B4: Experience with a globally-deployed software defined WAN," Proc. ACM SIGCOMM'13, pp. 3–14, 2013.
2. A. Atary and A. Bremler-Barr, "Efficient round-trip time monitoring in OpenFlow networks," Proc. IEEE INFOCOM, pp. 1-9, 2016.
3. M. Shibuya, A. Tachibana, and T. Hasegawa, "Efficient active measurement for monitoring link-by-link performance in OpenFlow networks," IEICE Trans. Commun., vol. E99B, no. 5, pp. 1032–1040, 2016.
4. R. Castro, M. Coates, G. Liang, R. Nowak, B. Yu, "Network Tomography: Recent Developments," Statist. Sci. 19(3):499–517, 2004.
5. N. Duffield, "Network Tomography of Binary Network Performance Characteristics," IEEE Transactions on Information Theory, vol. 52, no. 12, pp. 5373-5388, 2006.
6. A. Tachibana, S. Ano, T. Hasegawa, M. Tsuru, Y. Oie, "Locating Congested Segments over the Internet Based on Multiple End-to-End Path Measurements," IEICE Trans. Commun. Vol. E89-B, no.4, pp. 1099–1109, 2006.
7. L. Ma, T. He, A. Swami, D. Towsley, and K. K. Leung, " Network capability in localizing node failures via end-to-end path measurements, " IEEE/ACM Trans. on Networking, vol. 25, no. 1, pp. 434–450, 2017.
8. Nguyen Minh Tri, Masato Tsuru,"Locating Deteriorated Links by Network-Assisted Multicast Proving on OpenFlow Networks," Proc. the 24th IEEE Symposium on Computers and Communications (ISCC 2019), 6 pages, to appear in July 2019.
9. The Internet Topology Zoo, http://www.topology-zoo.org/ (accessed on Feb 15, 2019.)