# On the Retention Rate of Top Talkers for Sampled NetFlow

Shou-Chuan Lai[1], Jo-Chuan Cheng[2], Yu-Hsiu Chuang[3]

[1] Department of Information and Telecommunications Engineering, Ming Chuan University, Taiwan
[2] Research Department, Gentrice Tech Co., Ltd., Taiwan
[3] Information Center, National Health Research Institution, Taiwan
sclai@mail.mcu.edu.tw, shelly@gentrice.net, chuang@nhri.edu.tw

## Abstract

Traditionally, we can analyze network traffic by capturing packets on the network and making statistics. However, as network bandwidth and network traffic increase, we will need more computing resources to complete the traffic analysis job on time. Using NetFlow technology may reduce the resources required for traffic analysis. However, as network bandwidth keeps growing, the time it takes to accomplish traffic analysis may increase dramatically. When resources and time are limited, applying the sampling technique to NetFlow generation may reduce the amount of time and resources required. Nowadays, NetFlow data are often used to generate various statistical reports. Thus, we must fully understand whether the sampling technique will affect the statistical results before applying it to the NetFlow generation. In this paper, 28 days of NetFlow data obtained from the Taiwan Academic Network were studied. The differences in the IP address list and top talkers for different sampling rates are examined. The results show that sampling NetFlow does affect the retention rates of IP addresses and top talkers in ranking lists, and the higher the sampling rate is, the greater the impact is.

**Keywords:** NetFlow, Sampled NetFlow, Retention rate, Top talkers

## 1 Introduction

With the development of network technology and cloud services, more and more traffic is transmitted through the Internet. To keep a network run smoothly and efficiently, the network manager must know how the network is used. To figure out the transmission status of a network, the network manager may make use of tools like Tcpdump [1] or WireShark [2] to capture and analyze network packets. If we capture all packets of a network for a day, we can calculate the total amount of data sent by each sender and find the top N talkers for that day. However, with rapid growth in network bandwidth and widespread use of network applications, the amount of data generated per day increase gradually. Thus, the amount of storage space, as well as computation power, required to analyze captured packets increase even more dramatically. And eventually, the cost for acquiring computation and storage equipment may become unrealistic.

In 2001, the backbone bandwidth of the Taiwan Academic Network (TANet) was 1 Gbps. Four years later, the bandwidth was upgraded to 10 Gbps. Then, it was further upgraded to 100 Gbps in 2015. According to the development of bandwidth increment, the amount of storage space required to store captured packets for daily analysis was increased from 10.8 TB to 108 TB and further increased to 1.08 PB. Also, to find the top N talkers of a given date, we need to find the senders and sort each sender according to the total bytes or packets of the sender. Sorting such a huge amount of data requires lots of computation and memory resources. Therefore, packet capture gradually becomes an unpractical solution for daily traffic analysis. To reduce the amount of data required for network traffic analysis, many network managers turn to adopt the NetFlow solution [3-4]. By taking only the header information of network packets, NetFlow may effectively reduce the amount of data required for traffic analysis and obtain the same statistical results as packet capture does. However, due to the continuous growth of network traffic and the capacity limitation of NetFlow export equipment, some network managers begin to deploy sampling NetFlow [5-10] to further reduce the amount of data required for traffic analysis and to ensure that the network traffic analysis can be completed in time.

Currently, NetFlow data is often used to generate various types of network traffic statistical reports. Among these reports, the reports of top N talkers sorted by byte counts and sorted by packet counts are commonly used by network managers to profile network traffic behavior [11]. However, these top N talkers reports are originally generated from unsampled NetFlow data. If the deployment of sampling NetFlow is a must, it is important for network managers to know whether or not these statistical reports, especially these

*Corresponding Author: Shou-Chuan Lai; E-mail: sclai@mail.mcu.edu.tw
DOI: 10.3966/160792642021032202019

top N talkers reports generated from sampled NetFlow data are consistent with the original ones generated from unsampled NetFlow data. Having the knowledge of the differences between these reports generated from unsampled and sampled NetFlow data, network managers may make better decisions and avoid unintended consequences.

In the next section, we will explain how NetFlow works and review its sampling-related research. The sampling method and experimental design will be presented in the third section. In the fourth section, we will discuss the results obtained from these experiments. The fifth section is the conclusion of this study.

## 2 Background

This study is based on NetFlow technology, so in this section, we will first introduce what NetFlow is and its related techniques used in this study. Then we will review the related research on applying NetFlow sampling techniques to traffic analysis.

NetFlow technology was first developed by Cisco in 1996. It is originally an experimental network specification for collecting IP transmission information and has been set as an Internet standard [4]. Nowadays, NetFlow has been widely deployed on the Internet for network traffic analysis. NetFlow is based on the characteristics that when data are sent from a source to a destination, the packets of these data will be continuously transmitted. NetFlow will collect these packets periodically and generate a flow to represent that traffic. A flow is a record that contains the source IP address, the destination IP address, the source port number, the destination port number, the number of bytes transmitted, the number of packets transmitted, and other information of that traffic. A data sent from a source to a destination may generate more than one flow records.

With the rapid increase of network traffic, the number of flows generated by NetFlow also increases dramatically. Foreseeing the need to reduce the number of flows generated, both commonly used versions of NetFlow format, namely version 5 and 9, have sampling-related fields in place. The version 5 format of NetFlow has defined the sampling mode and sampling ratio fields. Both are used in conjunction to implement NetFlow sampling. The version 9 format has defined six sampling-related fields and supports sampled NetFlow by packet sampling or flow sampling.

Today, the backbone network bandwidth of the TANet is over 100 Gbps, and the number of flows generated per unit of time is quite huge. Considering the amount of data and the limitation of NetFlow export capability, Taiwan Academic Network currently uses NetFlow with a sample rate of 1:64 for data analysis.

Many researchers have studied the sampling techniques related to NetFlow. The research results of [7] show that the estimation accuracy is not easy to guarantee when NetFlow is sampled statically. In addition, when NetFlow is sampled and then classified by supervised learning method, the accuracy of classification will be seriously affected [8-9]. And in [10], it shows that when NetFlow is sampled, the estimated file size will be different from the original one. These studies did not focus on the accuracy of finding the top N talkers after sampling NetFlow. However, considering sampling NetFlow has gradually become a necessity in today's network operation and the top N talkers are essential information for network management, it needs to be known whether the content of the top N talkers list will be changed due to NetFlow sampling.

## 3 The Retention Rate

In order to know whether the contents of the top N talkers list will change due to sampling, we will compare the IP address list of the top N talkers generated from the unsampled NetFlow data with the same list generated from the sampled NetFlow data. There are several types of top N talkers lists. In this study, we will use the IP address of a sender (hereinafter referred to as *srcip*), the total number of data bytes (hereinafter referred to as *byte*) and the total number of data packets (hereinafter referred to as *packet*) sent by a sender as the objects of statistical analysis.

To obtain the top N talkers lists, firstly, we shall find out the IP addresses of all senders and put each of them in the IP address list from unsampled NetFlow data. After all IP addresses are found, sort these IP addresses by the total number of data bytes and data packets sent by each IP address, and then take out the top 10, 50, and 100 places as the top 10, 50, and 100 talkers lists, respectively. These top N talkers lists for byte and packet will serve as the basis for comparison after sampling.

Next, the original NetFlow data are sampled to generate the sampled NetFlow data. In this study, we will discuss the results of the sampling rate of $1{:}2^1$, $1{:}2^2$, and so on to $1{:}2^{10}$. Find the IP address list for the sampled data and sort these IP addresses by the total number of data bytes and data packets transmitted from these IP addresses. Then, take out the top 10, 50, and 100 places as the sampled the top N talkers lists for follow-up discussion.

### 3.1 Sampling Method

The NetFlow data used in this study are generated from network traffic passed through a network node of the TANet. These data contain traffic records for 28 consecutive days and have been fully de-identification. The NetFlow data for a whole day is stored in a single file and the file which stores the data for the *i*-th day is referred to as $F_i$. For each file $F_i$, if the sampling rate is $1{:}2^j$, every $2^j$-th flow is sampled and then put into the new sampled NetFlow file $SF_{i,j}$. This sampling method which periodically selects every *k*-th flow is referred to as systematic sampling. The pseudo-code for systematic sampling method is shown in Figure 1.

```
for i in {1, 2, ..., 28}
   for j in {0, 1, 2, ..., 10}
     k = 0
     rate = 2^j
     while not end of F_i
        read flow
        if k == 0
           output flow to SF_i,j
        k = (k + 1 ) % rate
     end
   end
end
```

**Figure 1.** Pseudo-code for systematic sampling

After the sampled NetFlow file $SF_{i,j}$ is produced, we proceed to examine every flow in this file and find the *srcip*, *byte*, and *packet* for each flow. And then, firstly, check whether *srcip* already exists in the list of IP addresses $SET_{i,j}$. If it is not found in $SET_{i,j}$, add *srcip* to $SET_{i,j}$. Secondly, add the *byte* and *packet* for *srcip* to $SUM_{i,j,byte}\{srcip\}$ and $SUM_{i,j,packet}\{srcip\}$ respectively to find the total bytes and packets sent by *srcip*. After file $SF_{i,j}$ has been examined, we shall sort the lists $SUM_{i,j,byte}$ and $SUM_{i,j,packet}$ according to its value in descending order to find the sorted list $SORT_{i,j,byte}$ and $SORT_{i,j,packet}$ respectively. We then derive the $TOP_{i,j,k,t}$ lists from the top *k* entries of the sorted list $SORT_{i,j,t}$. The pseudo-code for obtaining $SET_{i,j}$ and $TOP_{i,j,k,t}$ for $SF_{i,j}$ is shown in Figure 2.

```
for i in {1, 2, ..., 28 }
   for j in {0, 1, 2, ..., 10}
     while not end of SF_i,j
        read flow
        get srcip, byte, packet from flow
        if srcip not in SET_i,j
           add srcip to SET_i,j
        SUM_i,j,byte{srcip} += byte
        SUM_i,j,packet{srcip} += packet
     end
     for t in {byte, packet}
        SORT_i,j,t = sort SUM_i,j,t
        for k in {10, 50, 100}
           TOP_i,j k,t = top k of SORT_i,j,t
        end
     end
   end
end
```

**Figure 2.** Pseudo-code for obtaining the top talkers lists

According to the pseudo-codes in Figure 1 and Figure 2, when *j*=0, the obtained IP address table $SET_{i,0}$ and top talkers lists $TOP_{i,0,k,t}$ will be used as a reference group. And, when *j*={1,2,…,10}, the obtained IP address tables and top talkers lists will be used as observation groups for subsequent difference examination.

## 3.2 IP Address Retention Rate

The retention rate of IP addresses is mainly to find out whether the list of IP addresses before and after sampling is the same. When the sample rate is high, it is likely to happen that flows with short-term activities may not get sampled. In that case, the IP address which appeared in these flows may get missing after sampling. To calculate the IP address retention rate for $SF_{i,j}$, we shall check every IP address in $SET_{i,0}$ to see if it still exists in $SET_{i,j}$. The IP address retention rate is the percentage of the number of IP addresses that exist after sampling to the number of IP addresses before sampling. The pseudo-code for calculating the IP address retention rate is shown in Figure 3. In Figure 3, the IP address retention rate for $SET_{i,j}$ is denoted by $RET_{i,j}$.

```
for i in {1, 2, ..., r }
   for j in {1, 2, ..., 10}
      counter = 0
   while not end of SET_i,0
      read ip-addr from SET_i,0
      if ip-addr in SET_i,j
         counter ++
      end
      RET_i,j = counter / size of SET_i,0
   end
end
```

**Figure 3.** Pseudo-code for calculating the IP address retention rate

## 3.3 Retention Rate for Top Talkers

The second observation that we want to make in this study is whether the list of top talkers is still the same before and after sampling. After examining a small portion of flow data, we found that the activities of a top talker are likely to be high density in a short term. It is possible that, if the sample rate is high, some activities of the top talker may not get sampled and the top talker may not be able to enter the top talkers list after sampling. The retention rate of top talkers is the percentage of the original top talkers who entered the list of top talkers after sampling. To calculate the retention rate of a top talkers list, we initially set the retention counter to zero, and then check all the IP addresses in the original top talkers list one by one. If the IP address does enter the top talkers list after sampling, we increase the retention counter by one. After all the IP addresses are checked, the retention rate is the percentage of the number of the retention counter to the number of IP addresses in the top talkers list. The pseudo-code for calculating the retention rate of top talkers is shown in Figure 4. In Figure 4, the retention rate for $TOP_{i,j,k,t}$ is denoted by $RET_{i,j,k,t}$.

```
for t in { byte, packet }
   for i in {1, 2, ..., r }
      for j in {1, 2, ..., 10}
         for k in {10, 50, 100}
            counter = 0
            while not end of TOP_{i,0,k,t}
               read ip-addr from TOP_{i,0,k,t}
               if ip-addr in TOP_{i,j,k,t}
                  counter++
            end
            RET_{i,j,k,t} = counter / k
         end
      end
   end
end
```

**Figure 4.** Pseudo-code for calculating the retention rate for the top talkers

## 4 The Experimental Results

The NetFlow data used in our experiment are generated by the traffic passed through a network node of the TANet for 28 consecutive days. According to the evaluation methods introduced in the previous section, for each day, we can compute IP address retention rates for different sampling rates ranging from $1{:}2^1$ to $1{:}2^{10}$. Also, for every day's top talkers by bytes and by packets, we can calculate retention rates for different lengths of lists and various sampling rates. Based on these results, we shall study the following issues: the IP address retention rates under different sampling rates, the retention rates for top talkers by bytes, and by packets after sampling.

### 4.1 Different Lengths of Reference Periods

Before we begin to explore the impact of sampling on statistical results, we need to understand the impact of the length of the reference period on the issues under discussion. Therefore, we will take the duration of 1, 6, 12, and 24 hours in a day as the reference periods to observe the impact on the retention rates of IP addresses and top talkers by bytes and packets.

In the following experiments, the period around noon of a day is taken as the reference period for statistical calculation, and the average of 28 consecutive days of the same periods was taken as the result. That is, NetFlow data from 12:00 to 13:00 per day are used as the reference period for one hour; NetFlow data from 10:00 to 16:00 per day are used as the reference period for six hours; NetFlow data from 8:00 to 20:00 per day are used as the reference period for 12 hours; NetFlow data from 0:00 to 24:00 per day are used as the reference period for 24 hours. Table 1 shows the periods used for examining different lengths of reference periods.

**Table 1.** The lengths of reference periods

| Duration (hour) | Period |
|---|---|
| 1 | 12:00-13:00 |
| 6 | 10:00-16:00 |
| 12 | 08:00-20:00 |
| 24 | 00:00-24:00 |

The IP address retention rates for different reference periods are shown in Figure 5. The results show that IP address retention rates which use 6, 12, and 24 hours as the reference periods are similar and the result which uses 1 hour as the reference period has only a slight difference. For example, when the sampling ratio is $1{:}2^6$, the IP address retention rates by reference periods of 1, 6, 12, and 24 hours are 0.22, 0.26, 0.27, and 0.28 respectively. It means that the longer the reference period is, the smaller the difference is. But it also shows that when the reference period is longer than 6 hours, the results are similar.

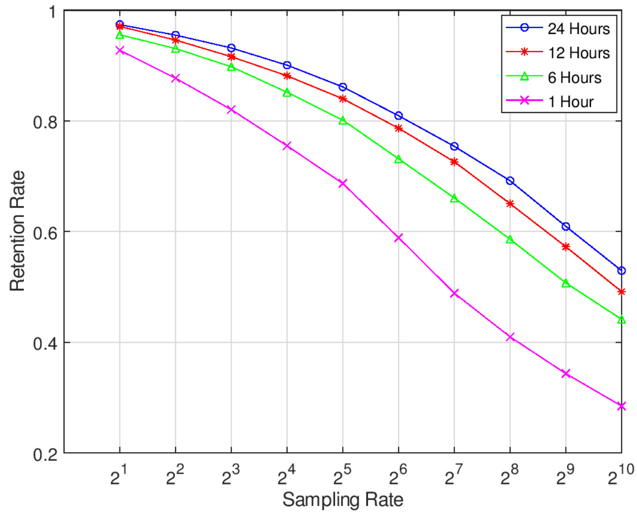


**Figure 5.** The retention rates of IP addresses for different lengths of reference periods

The retention rates of top talkers by byte for different lengths of reference periods are shown in Figure 6. The results show that retention rates by reference periods of 6, 12, and 24 hours are alike and the result which uses 1 hour as the reference period has a larger difference. For example, when the sampling rate is $1{:}2^6$, the retention rates by reference periods of 1, 6, 12, and 24 hours are about 0.53, 0.73, 0.79, and 0.81 respectively. It shows that the length of reference periods has little impact on the retention rate and the longer the reference period is, the larger the retention rate is.

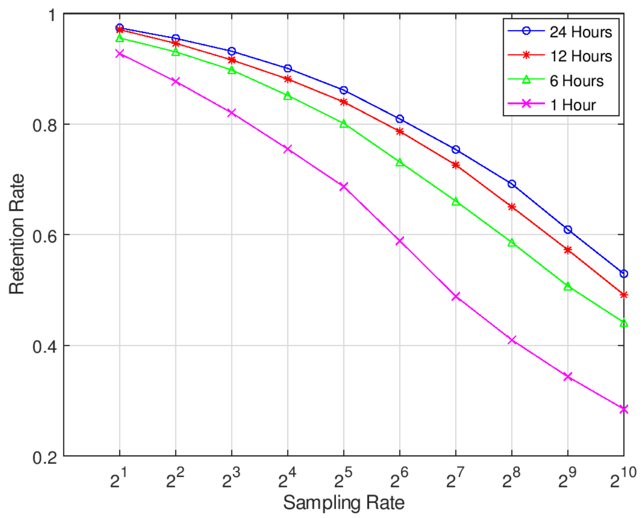


**Figure 6.** The retention rates of top talkers by bytes for different lengths of reference periods

Figure 7 shows the retention rates of top talkers by packets for different lengths of reference periods. The results show that retention rates by reference periods of 6, 12, and 24 hours are also alike and the result which uses 1 hour as a reference period has a larger difference. For example, when the sampling rate is $1{:}2^6$, the retention rates by reference periods of 1, 6, 12, and 24 hours are 0.53, 0.73, 0.79, and 0.81 respectively. It shows that the length of the reference period has little impact on the retention rate and the longer the reference period is, the larger the retention rate is.

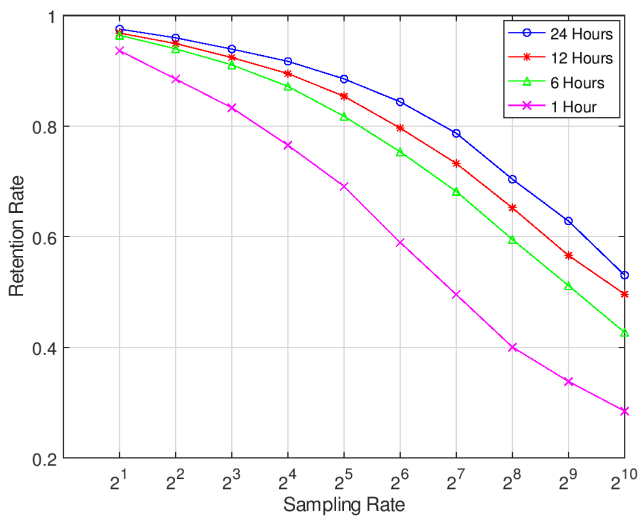


**Figure 7.** The retention rates of top talkers by packets for different lengths of reference periods

Based on the above results, we will use a whole day as the reference period for the following statistical calculation and discuss the impact of different sampling rates on the retention rates of IP addresses and top talkers by bytes and by packets.

## 4.2 Impact on IP Address Retention Rate

To study how the sampling rate affects the IP address retention rate, we calculate the IP address retention rates for different sampling rates on a daily basis. The average value for 28 consecutive days in a month for a specific sampling rate is taken as the result of that sampling rate. The sampling rates that we studied are $1{:}2^1$, $1{:}2^2$, …, $1{:}2^{10}$, and the result is shown in Figure 9.

The experimental result shows that sampled NetFlow will significantly affect the retention rate of IP addresses. From Figure 8, we can see that with a larger sampling rate, the IP address retention rate will decrease significantly. For example, when the sampling rate is $1{:}2^6$, the IP address retention rate will be about 0.28, which means lots of IP addresses will be missing after sampling.

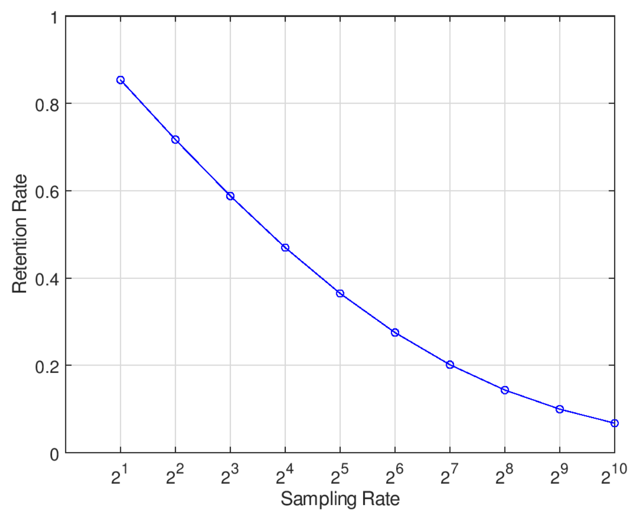


**Figure 8.** The impact of sampling rate on IP address retention rate

## 4.3 Impact on Retention Rate of Top Talkers by Bytes

After knowing the impact of sampled NetFlow on the retention rates of IP addresses, we continue to study the retention rates of top talkers by bytes for sampled NetFlow. We calculate the retention rates of top talkers by bytes for different sampling rates on a daily basis. The average value for 28 consecutive days in a month for a specific sampling rate is taken as the result of that sampling rate. The sampling rates that we study are $1{:}2^1$, $1{:}2^2$, …, $1{:}2^{10}$ and the lengths of the top talkers lists examined are 10, 50, and 100. Figure 9 shows the results.

It shows that, after sampling NetFlow, the top 10, 50, and 100 talkers by bytes may become quite different from the ones before sampling. The results are similar for different lengths of the top talkers' list. From Figure 10, we can see that the bigger the sampling rate is, the smaller the retention rate is. For example, when the sampling rate is $1{:}2^6$, the retention rate of top 100 talkers is about 0.80, that is to say, after sampling, 20 IP addresses originally on the top 100 talkers list will

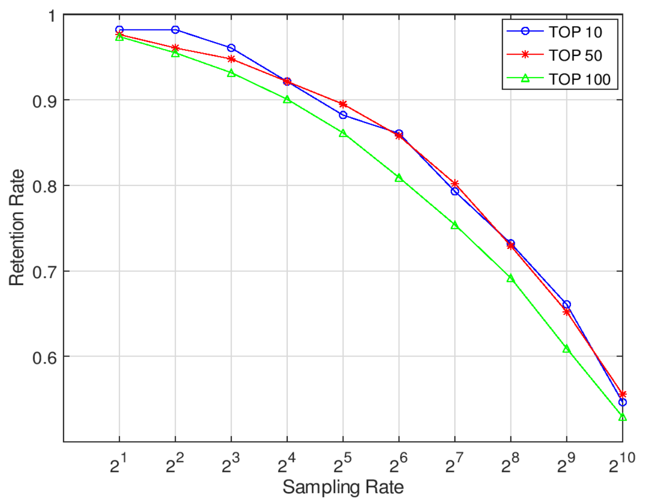


**Figure 9.** The impact of sampling rate on retention rates of top talkers by bytes

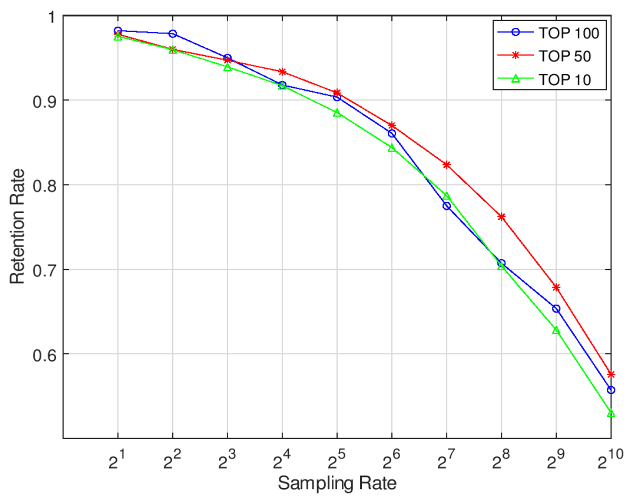


**Figure 10.** The impact of sampling rate on retention rates of top talkers by packets

be missing, and when the sampling rate is $1:2^{10}$, 47 IP address originally on the top 100 talkers list will be missing.

## 4.4 Impact on Retention Rate of Top Talkers by Packets

To study how NetFlow sampling affects the top talkers by packets, we do a similar experiment as we do for top talkers by bytes. The only difference is that the talkers are sorted by packets instead of bytes. The result is shown in Figure 10.

The result is similar to one of the top talkers by bytes. It shows that, after sampling NetFlow, the top 10, 50, and 100 talkers by packets may become quite different from the ones before sampling and the results are similar under various lengths of the top talkers' list. From Figure 10, we can also see that the bigger the sampling rate is, the smaller the retention rate is. For example, when the sampling rate is $1:2^6$, the retention rate of the top 100 talkers is about 0.84, which means, after sampling, 16 IP addresses originally on the top 100 talkers list will be missing. And when the sampling rate is $1:2^{10}$, 47 IP addresses originally on the top 100 talkers list will be missing.

## 5 Conclusion

This study explores the differences in lists of IP addresses and top talkers caused by sampling technology. It is discussed whether IP addresses will disappear from the list because of sampling and whether the top talkers by bytes and by packets will change because of sampling. Based on the NetFlow data of 28 consecutive days in a network node of TANet, we examined the impact of different lengths of reference period and different sampling rates on the retention rates of IP addresses, top talkers by bytes, and top talkers by packets. From the experimental results, we have the following findings:

(1) Increasing the length of the reference period can increase the retention rates of IP addresses, top talkers by bytes, and top talker by packets.

(2) With the increase of sampling rate, the IP address retention rate will be significantly decreased after sampling.

(3) With the increase of sampling rate, the retention rates for top talkers by bytes and by packets are decreased greatly.

(4) With the same sampling rate and length of the top list, the result for top talkers by bytes and the result for top talkers by packets are similar.

(5) Under the same sampling rate, the results for different lengths of top talkers list by bytes and top talkers list by packets are similar.

The results show that when the sampling rate is 1:64, about 28% of the original IP addresses will remain in the list of IP addresses after sampling and about 81% of the top 100 talkers by bytes will still be the top 100 talkers by bytes after sampling. The results also show that if it is necessary to control the IP address retention rate to be more than 70%, the sampling rate should be limited to less than 1:4. And if the retention rate of the top 100 talkers by bytes is to be controlled over 90%, the sampling rate should be limited below 1:16.

## References

[1] The Tcpdump Group, TCPDUMP/LIBPCAP public repository, http://www.tcpdump.org/.

[2] Wireshark Foundation, Wireshark · Go Deep, http://www.wireshark.org/.

[3] Cisco Systems, Inc., *Introduction to Cisco IOS NetFlow*, NetFlow White Papers Document ID: 1518925696681207, May, 2012, https://www.cisco.com/c/en/us/products/collateral/ios-nx-os-software/ios-etflow/prod_white_paper0900aecd80406232.pdf.

[4] B. Claise, G. Sadasivan, V. Valluri, M. Djernaes, *Cisco Systems NetFlow Services Export Version 9*, RFC 3954, October, 2004.

[5] C. Estan, G. Varghese, New Directions in Traffic Measurement and Accounting, *ACM SIGCOMM Computer Communication Review*, Vol. 32, No. 4, pp. 323-336, October, 2002.

[6] C. Estan, K. Keys, D. Moore, G. Varghese, Building a Better NetFlow, *ACM SIGCOMM Computer Communication Review*, Vol. 34, No. 4, pp. 245-256, October, 2004.

[7] B.-Y. Choi, S. Bhattacharyya, On the Accuracy and Overhead of Cisco Sampled NetFlow, *ACM SIGMETRICS Workshop on Large Scale Network Inference (LSNI)*, Banff, Canada, June, 2005, pp. 1-6.

[8] V. Carela-Español, P. Barlet-Ros, J. Solé-Pareta, *Traffic Classification with Sampled NetFlow*, Technical Report, UPC-DAC-RR-CBA-2009-6, Universitat Politecnica de Catalunya, February, 2009.

[9] V. Carela-Español, P. Barlet-Ros, A. Cabellos-Aparicio, and J. Solé-Pareta, Analysis of the Impact of Sampling on NetFlow Traffic Classification, *Computer Networks*, Vol. 55, No. 5, pp. 1083-1099, April, 2011.

[10] J.-Z. Huang, *On the Accuracy of Traffic Analysis based on NetFlow Sampling*, M.S. Thesis, Department of Information and Telecommunications Engineering, Ming Chuan University, Taipei, Taiwan, 2012.

[11] N-Partner Technology Co., Ltd., *N-Reporter 6 Data Sheet*, October, 2020, http://www.npartnertech.com/upload/Download/N-Reporter6-DS-EN.pdf.


## Biographies

**Shou-Chuan Lai** received a Ph.D. degree in computer science from the National Tsing Hua University, Hsinchu, Taiwan in 2000 and is an assistant professor in the Department of Information and Telecommunications Engineering at Ming Chuan University, Taoyuan, Taiwan. His research interests include computer networks, network security, and network management.

**Jo-Chuan Cheng** received the B.S. and M.S. degrees from Ming Chuan University of Information and Telecommunication Engineering, Taoyuan, Taiwan, in 2017 and 2018. She is currently a NetFlow Engineer in Gentrice Tech Co., Ltd., New Taipei, Taiwan. Her research interests include NetFlow analysis, network management, network traffic monitoring, and information security.

**Yu-Hsiu Chuang** received a Ph.D. degree in the Institute of Management from the National Taiwan University of Science and Technology, Taipei, Taiwan in 2014 and is a director of Information Center in the Department of Administration at National Health Research Institutes, Miaoli, Taiwan. His research interests include administrative computerization, Internet, network security, and network management.