On Information Value of Top N Statistics

Tomáš Jirsík, Milan Čermák, Pavel Čeleda
Institute of Computer Science
Masaryk University
Brno, Czech Republic
Email: {jirsik, cermak, celeda}@ics.muni.cz

***Abstract*—In the era of Internet of Things (IoT), the volume of the monitored data from IoT network is enormous. However, not all data provide sufficient or relevant information. Since the analysis of big data is both resource and time exhausting, only relevant information should be analysed. In this paper, we scrutinize the widely used *Top N* statistics and evaluate its information value with respect to gathering information about individual hosts in the network. All theoretical discussions are evaluated on the real-world data. Moreover, we provide an assessment of statistic's suitability for identifying a host in network traffic. The results of the paper should assist data analyst of IoT network data.**

I. Introduction

An emerging era of Internet of Things (IoT) affects all aspects of our future lives. Network-connected devices are going to be omnipresent and responsible for a number of tasks including the critical ones. This fact has already brought more attention to network security monitoring of IoT. The focus leads to significant improvement of processes for computer network information gathering (diverse data sources, a higher level of data visibility and granularity) and data storage enhancements (scalable, distributed data warehouses, etc.). The new technologies result into a flood of monitored data. The volume of the data, however, makes extraction of security-relevant information more challenging.

In our work, we focus on the gathering information of individual hosts in a network. The information of a host in the network is widely used in many areas such as network security, network accounting, law enforcement etc. There exists a variety of both raw and derived statistics that can be gathered about individual hosts. However, the research lacks focus on the characteristics of the information provided by statistics and their properties. Therefore, we focus on the *Top N* statistics and try to describe properties of information that can be obtained by *Top N* statistics.

Our research can be summarized in the following research questions:

- *What are the characteristics of information provided by Top N statistics?*
- *Is Top N statistic suitable for identifying an entity in network traffic?*

In this paper, we will describe in detail the *Top N* statistics itself and identify statistic's parameters and their influence on the statistics' outcome. We choose to evaluate the *Top N* statistics with regard to following perspectives: *availability*, *uniqueness of the information* and *time stability*. These perspectives provide an insight into the characteristics of information provided by *Top N* characteristics. The *availability* is necessary to be able to obtain the results. The *time stability* representing the variability of provided information in time will be assessed. The *uniqueness* represents the similarity between the *Top N* statistics of different hosts. After the detailed analysis of the statistics, its suitability for entity identification will be scrutinized. We will discuss *Top N* each property variable and describe the implications on the entity identification. The theoretic discussion is then validated on experiments based on the real world data from university campus network.

The paper focuses on the readers which analyze network data to mine relevant information. The paper provides a deep insight into the *Top N* statistics and on the real-world example shows the possible information that can be harvested by this statistics. As a result, the information in this paper should ease the decision whether *Top N* statistics is suitable for a proposed task and the reader should be informed about possible pitfalls.

The rest of the paper is organized as follows. *Top N* statistics obtainable from network traffic and their properties are discussed in Section II. In Section III, we shall assess the *Top N* statistics based on their suitability from an entity identification point of view. Section IV provides an experimental evaluation of the *Top N* information value on real-world data and assess the suitability of the statistics for a host identification. Section V concludes the paper.

II. Top N statistics

"Find 3 IP addresses that transferred the most bytes during last 5 minutes" is a typical query for *Top N* statistics. The statistics is an internal part of tools for analysing network data, such as *nfdump* [1], *fbitdump* [2] or *ntop* [3]. Therefore, it is widely used for various applications in network traffic monitoring, such as identifying top talkers, providing an overview of the most important events in network traffic, port utilization statistics or discovering popular network applications. Its results are used for optimising network performance, identifying abnormal events [4], security monitoring or for management reports [5]. In following paragraphs, we will specify *Top N* statistics, scrutinize its parameters and examine its computational requirements.

978-1-5090-3765-0/16/$31.00

A full specification of *Top N* query is the following:

$$\textit{Top N of X sorted by Y, over period of time P,} \quad (1)$$

where N is the number of output records of return characteristic X. The records of X are sorted descending by the variable Y and counted over a period of time P. From the sorted list, the first N records are returned. A *Top N* query processing consists of four basic operations. First, data from defined period P is selected. Second, the selected data is aggregated by characteristic X and aggregated values of characteristic Y are computed. Third, the aggregated records are sorted descending by variable Y. Lastly, the first N records from the sorted list are returned. All parameters have a significant effect on results of the *Top N* statistic. We discuss each of them further in the paper.

The first parameters to determine are the return characteristic X, and sorting variable Y. The choice of both X and Y depends on the purpose for which the *Top N* statistic is used. The characteristic X defines the return of the statistics (e.g., an IP address is used as X for top talkers identification). Sorting variable Y needs to be defined on a totally ordered set. A totally ordered set ensures that records of X can be sorted by Y. For example, the sorting variable Y can represent the number of transferred bytes, packets or flows, duration, and the number of occurrences. Derived characteristics can also be used, such as the average number of connections, maximum packet size or the number of distinct web pages visited.

The next parameter to set is the period P for *Top N* statistic computation. The period influences the amount of information which is processed and consequently determines the aggregation level of the *Top N* statistic. Short periods are chosen when detailed data is needed, whereas long periods are used for getting an overview. Since the period affects the amount of information processed and aggregated, it also affects the computational resources needed to compute the statistic. The longer the period is, the more information is processed and the more computational resources are needed.

The last parameter to set is the number of returned records, N. This parameter plays the role of a cut-off. Only information which passes the cut-off, is presented. Therefore, the proper setting of this parameter is crucial. N depends only on the reason *Top N* statistics are used for. When we want to identify the most active host in a network, N equal to one is sufficient. This is not the case when we want to create a report on port usage in a network. When N is set to low, only a little information is returned.

Let's consider the computational requirements of the *Top N* statistic. The statistic computation includes aggregation and sorting operations, which are computationally demanding. The aggregation process aggregates variable Y by return characteristic X. The aggregation process that covers a longer period or large scale network may result in the need to keep billions of records in memory. There are approaches to decrease the amount of memory needed. One approach leverages *map and reduce* technique [6], where partial *Top N* are computed in the map phase and only results which pass a predefined threshold are passed to the reduce phase. Another approach leveraging the statistical properties of network traffic is presented in [4]. The aggregation process is succeeded by a sorting operation. The sorting operation adds significantly to the time complexity of *Top N* statistic computation. Depending on the choice of sorting algorithm, the comparison-based sorting algorithms can not perform better than $O(n \log n)$.

III. Host Identification using Top N

In this section, we research a suitability of *Top N* statistics for host identification. We briefly describe state-of-the-art of host identification in a network and then we discuss the *Top N* suitability for host identification.

The host can be identified in the network via its MAC/IP address. This identifiers are not however always available (MAC address) or reliable (IP addresses in dynamic addressed networks, NATs). Therefore, other approaches to host identification are developed. In general, these approaches are looking for a unique key, based on which a host could be identified. A key could be imprints of an host in observed data or can leverage an entity's characteristics, e.g., ciphersuites [7]. We will discuss a suitability of *Top N* statistics to generate such an identification key.

The parameters of *Top N* statistic are affected by the available data, which we use to compute the statistics. Given our data source, network traffic, we can use, in theory, any information that is transmitted via a network. There are two approaches to information retrieval from network traffic: *deep packet inspection* (DPI) and *network flow* [8]. DPI enables us to obtain any information available from the packet. Costs for this universality are high computational requirements and a limited throughput as the whole packet needs to be processed. Therefore, DPI can be used to retrieve information only on limited scale. A concept of network flows has been introduced for retrieving information from large-scale networks with high throughput. A network flow [9] represents an abstraction of a network connection. A flow carries only information from aggregated packets belonging to the flow. All other information is lost during the aggregation. Only information from packet headers is aggregated usually. Researchers are aware of these limitations and methods for enriching flow information while preserving high-throughput have been introduced, e.g., in [8]. We prefer to use network flow monitoring for retrieving information. This approach enables us to retrieve information from large-scale and high-speed networks. An overview of basic information which can be retrieved from network flows is provided in [10].

Next, we need to identify the return characteristic X and the sorting variable Y such that the computed *Top N* statistic is able to identify a host. We believe that a host in a network is identified by the trace it leaves in network traffic. *Top N* statistics based on X and Y need to be chosen such that it transforms the trace into as much of a unique statistics results as possible. Information from the link layer (L2) can be used only in LAN. Hence, we focused on L3 - L7 layer information. L3/4 layers of the OSI/ISO model provide information about

communication partners, ports, and the protocol used. Since the number of distinct protocols that can be observed in network traffic is low compared to the number of entities, it is impossible to create enough variations of *Top N* statistic achieve uniqueness. This leaves us with information about communication partners and ports used. There are 65536 distinct ports in total, which allow us to create $\frac{65536!}{(65536-N)!}$ different variations of the *Top N* statistic. Assuming $N = 10$, this results in 1.46×10^{48} unique variations. However, the distribution of port utilization in a network is not uniform. A group of ports exists which are used more often than others (e.g., ports up to 1000). The amount of unique variations of actually used ports is then much lower. Nevertheless, a port seems to be a suitable return characteristic.

The communication partner is represented by the destination IP address. The number of distinct IP addresses is 2^{32} for IPv4 and 2^{128} for the IPv6 protocol, which ensures a high number of different *Top N* statistics. The set of a host's communication partners to identify is, however, much smaller than the theoretical maximum, therefore less distinct statistics exist. Still, the number of communication partners is much higher than the number of entities, therefore we consider communication partner as a suitable return variable. The L7 layer is much more information rich than the previously discussed layers. We can retrieve information from HTTP, DNS, SMTP, FTP protocols and many others. The information from the layers usually provides enough variability and a deep insight into host's behaviour. Therefore, the results of *Top N* are likely to be unique. Considering L7 layer as an information source is, however, limited by the increasing portion of encrypted network traffic. When network traffic is encrypted, only information from L3 and L4 can be used.

IV. Experimental Evaluation of Top N Statistic

In this section, we shall provide an experimental evaluation of *Top N* information value. We will describe the data set used for evaluation, evaluate the statistics with respect to *availability*, *uniqueness* and *time stability* and provide results of a host identification suitability experiment.

A. Data set

The data set contains network traffic captured from a university campus network. The data set is divided into two subsets: training and testing. The training data set is used for *Top N* characteristics evaluation and creating host's signatures. The signatures sets are then used on data from the testing data set to assess the suitability for host identification. Table I provides a general description of the data sets. We choose to capture information about communication partners (destination IP), destination ports and the HTTP host information field. A host to detect is represented by the source IP address. To overcome problems of IP address assignment, we chose only such networks where only static addressing is permitted and no proxies or NAT devices are present. The granularity of captured information is 5 minutes. Every 5 minutes, for each source IP address, we retrieve a set of communication partners with a given IP address (DstIP), a set of destination ports (DstPort) the address communicated to, and a set of web pages the address visited in the interval (HTTP_host).

TABLE I
Data sets description.

	Training DS	**Testing DS**
Observation Period	05 - 11/10/2015	19 - 25/10/2015
Unique IP Address	497	507
Total Flows	3 711 378	3 357 389
Total Bytes	36.6 GB	29.4 GB
Total Packets	236.4 M	228.6 M

B. Top N properties evaluation

In this subsection, we shall evaluate the *Top N* statistics properties. First, we inspect a choice of data used for computing the statistic. We examine the information availability in the data. Second, we compute *Top N* statistics and observe their behaviour during the time to check the *time stability* requirement. Next, we compare the *Top N* statistic of each distinct IP addresses with each other to assess the *uniqueness* requirement. Lastly, we compute *Top N* statistics and evaluate their characteristics on the real world data set.

TABLE II
Statistics availability in time.

***P* = 5 minutes**		***P* = 1 hour**		***P* = 1 day**	
# of observations	**% of IP addresses**	**# of observations**	**% of IP addresses**	**# of observations**	**% of IP addresses**
0-288	25.506	0-24	14.575	1	1.417
288-576	36.235	24-48	34.413	2	1.417
576-864	21.053	48-72	19.838	3	7.085
864-1152	11.741	72-96	20.648	4	15.992
1152-1440	2.429	96-120	6.478	5	19.231
1440-1728	1.417	120-144	1.417	6	15.789
1728-2016	1.417	144-168	2.632	7	36.032

First, we investigated the availability of the statistics in time. For each source IP address, we computed *Top N* statistics with a different setting of period P and counted the non-empty results. Values of P were set to 5 minutes (the minimum value), 1 hour and 1 day. The results of the analysis are presented in Table II. The table shows, that a 5 minute period is not suitable for harvesting host information as there are only a few observations at the majority of IP addresses (25% of addresses are present in less than 288 observations from 2016 in total). The longer a period P for *Top N* computation is, the more IP addresses are observed in a higher portion of observations.

Secondly, we investigated the stability of information provided by *Top N* statistic over time. For P equals one hour and one day, we computed a relevant number of *Top 10* statistics on the whole data set (e.g., for $P = 1$ *hour* we computed $7 * 24$ *Top N* statistics). Next, we counted a number of *Top N* statistics per IP address, in which the 10 most frequent results of the *Top N* statistic were presented. Regarding the hour

period, all of the 10 most frequent results were observed in less than 42% of observations in 69% of IP addresses, which covers less than 15% of total observations. This indicates a higher variability in the data. The day period setting provides better results as the 10 most frequent results of *Top N* statistics are observed more than in 57% of observations in 57.8% of IP addresses.

TABLE III
Top N TIME STABILITY.

	P = 1 hour			P = 1 day		
	% of IP addresses					
Equal records	**DstIP**	**Dst-Port**	**HTTP _host**	**DstIP**	**Dst-Port**	**HTTP _host**
0 - 2	11.0	11.7	4.6	7.1	13.1	2.3
3 - 4	66.1	51.7	62.4	38.5	30.2	18.6
5 - 6	21.3	31.9	31.3	44.8	38.5	56.8
7 - 8	1.6	4.3	1.5	9.4	15.8	21.8
9 - 10	0.0	0.4	0.2	0.2	2.3	0.4
Jaccard	% of IP addresses					
0 - 0.2	45.2	2.0	28.4	22.3	4.0	6.6
0.2 - 0.4	51.3	5.5	66.4	61.3	25.8	56.8
0.4 - 0.6	3.3	27.0	5.0	15.6	36.7	33.9
0.6 - 0.8	0.2	33.7	0.2	0.8	23.5	2.8
0.8 - 1	0.0	31.7	0.0	0.0	10.0	0.0

To capture the variability of *Top N* statistics for a particular host over time, we compared consecutive *Top N* statistics for each source IP address and counted the similarity of these *Top N* statistics. The higher the similarity is, the more stable the statistics are over time. We chose $N = 10$ for the comparison. We measured the similarity of *Top 10* statistics by the number of equal records and by their Jaccard index [11]. To provide an overview of the whole data set, we computed average values for each similarity measure per IP and showed a frequency histogram of the averages. The results are shown in Table III. We observed, that the majority of IP addresses have 3-6 equal records in two consecutive *Top N* statistics. Regarding Jaccard similarity measure, DstPort characteristics showed more similarity than other characteristic for the one-hour interval. However, the similarity of the DstPort was been low when an equal record count similarity is used. The divergence in the similarity measures is explained by the high number of IP addresses which use less than 10 ports to communicate. The low number of ports decreases the similarity when using count of equal record similarity measure. However, the Jaccard similarity can handle this situation and provides unbiased results. HTTP_host performed well in both measures which proves the stability in users behaviour. Generally, the similarity is higher in one day period than in one hour period.

The test for *uniqueness* requirement was also based on similarity. We used *Top 10* statistics to generate statistics for all IP addresses. The statistics results were then compared with each other. For comparison, we used Jaccard similarity measure. We set a threshold to 0.25 and mark two results similar when the Jaccard was greater than or equal 0.25 (i.e., approx. 4 equal records in two *Top 10* statistics). The results of the experiment are presented in Table IV. DstPort characteristic did not meet *uniqueness* requirement as *Top N* statistics of the majority of the statistics were similar to more than 100 other ones for both periods. In general, period $P =$ 1 day performed better as there were more unique statistics. The greater aggregation implies, that more information is captured in the statistics than in the case the aggregation is low. Therefore the more aggregated the statistics is, the more likely it is unique. The DstIP provided most unique *Top N* statistics (51.9 % of statistics are unique). Hence, it should be the most suitable for host identification.

TABLE IV
Top N UNIQUENESS.

	P = 1 hour			P = 1 day		
	% of statistics					
U(s)	**DstIP**	**Dst-Port**	**HTTP _host**	**DstIP**	**Dst-Port**	**HTTP _host**
0	34.5	2.6	16.3	51.9	0.6	28.9
1 - 9	31.3	3.4	25.3	33.9	2.8	44.2
10 - 99	34.0	21.4	51.0	14.2	15.0	26.4
>= 100	0.2	72.6	5.4	0.0	81.7	0.0

C. Top N suitability for host identification

We computed *Top N* statistics for each host in the training data set. The statistics were then applied to the testing data set. The testing data set consists of the same entities as the training data set, which enables us to evaluate the results of the host identification process. Since a statistics consists of a number of records, a host was identified by a given statistics based on Jaccard similarity of *Top N* statistics. We set $M = 30$, period $T = 7$ *days* and period $P \in \{$*one hour, one day*$\}$ and Jaccard to 0.2 (approx. 10 equal records out of 30). The results are shown in Table V.

TABLE V
EXPERIMENTAL EVALUATION OF *Top N* STATISTICS.
($M = 30$, JACCARD = 0.2).

P	**Variable**	**TP (%)**	**FP (%)**	**Not Found (%)**
one hour	DstIP	3.04	0.61	96.36
	DstPort	34.01	21.86	44.13
	HTTP_host	8.35	2.09	89.56
one day	DstIP	20.45	7.89	71.66
	DstPort	44.13	25.91	29.96
	HTTP_host	59.50	15.66	24.84

True positive rate (TP) shows, how many of the hosts is correctly identified, i.e. the searched host is within the set of hosts identified by the statistics. False positive rate (FP) says how many hosts has been misclassified, i.e. the searched host is not within the set of identified hosts. We observed, that the day period had a higher TP rate than the hour period, which proved our latter conclusions. The highest TP rate was achieved by HTTP_host characteristic. In total, 59.5 % of the hosts from the testing data set were successfully identified by the statistics based on this variable. The DstIP characteristic should be used when we prefer the precision of identification to identification rate as it had the lowest FP rate for both P. We also inspected

different values for Jaccard for statistics' match. We observed, that with decreasing Jaccard, the TP rate increased and more hosts were identified as the similarity needed for the match was lower and more hosts were matched. The decrease of Jaccard also leads to higher FP rate as more hosts were mismatched due to decreased level similarity of statistics.

We further evaluated cardinality of a set of identified hosts to determine the uniqueness of a statistics $U(s)$. For each key that correctly identified a host we measured cardinality of the set of identified hosts. Table VI shows the distribution of statistics with regard to statistics' uniqueness $U(s)$[1].

TABLE VI
EXPERIMETAL EVALUATION OF *Top N* STATISTICS UNIQUENESS.

		% of hosts			
P	**Variable**	***U(s)=1***	***U(s)≤5***	***U(s)≤10***	***U(s)≤50***
one hour	DstIP	86.67	100.00	-	-
	DstPort	1.19	9.52	13.69	24.40
	HTTP_host	85.00	100.00	-	-
one day	DstIP	77.23	93.07	96.04	100.00
	DstPort	4.59	10.55	18.35	39.91
	HTTP_host	36.49	72.98	85.61	100.00

We observed, that in the one hour period, the uniqueness of the statistics was larger, as the majority of the statistics was *unique* ($U(s) = 1$). The maximum cardinality of the set of identified hosts based on the DstIP and HTTP_host characteristics was 5. The statistics based on DstPort characteristic did not prove to be unique as the majority of the statistics identified more than 50 of hosts.

V. CONCLUSION

This paper describes the information value of *Top N* statistics. We investigated the *availability* and *time stability* of the statistics and evaluated *uniqueness* of its outcomes. The *Top N* statistics was then applied to the testing data and the suitability for host identification was evaluated. We identified parameters of the *Top N* statistics and described their impact statistics outcome.

The experimental evaluation on real-world data showed that a period P correlates with *availability* and *time stability* of the statistics. The longer the period is, the more available and stable the statistics. The *uniqueness* has been highest for *Top N* of DstIP statistics and increased with longer period.

Moreover, we discovered that a single *Top N* statistic has a limited application on host identification problem. We were able to identify at maximum 60 % of hosts in the network traffic. However, the setting of Jaccard index threshold, which determined the equality of the statistics, was rather strict (two keys belonged to the same host when at least 10 records out of 30 were equal). If we relaxed the setting, we would identify a higher portion of hosts (but it would also increase the FP rate). Nevertheless, once we were able to identify an host the host was identified with high precision when we used the DstIP or HTTP_host characteristics (77.23% and 36.49% of the hosts were identified unambiguously).

The statistics identification capabilities could be enhanced by combining more types of *Top N* statistics together. The host could be represented by records generated by both DstIP and HTTP_host statistics. This would increase the *uniqueness* of the compound statistics while preserving the *time stability* of the statistics. Moreover, we could use information from another L7 protocols for statistics (e.g., DNS protocol). Both improvements are left for the future work.

The host identification based on *Top N* statistic can be used for identifying a set of hosts which are similar to the searched host. Such identification can be used for law enforcement to identify a set of suspects for further investigation or in network security monitoring for identification of IoT devices in a network that need detailed surveillance.

[1] $U(s) = 1$ reads as statistics is similar to only one other statistics.

ACKNOWLEDGEMENT

This research was supported by the Technology Agency of the Czech Republic under No. TA04010062 *Technology for processing and analysis of network data in big data concept*.

REFERENCES

[1] P. Haag, "NFDUMP," Web page, December 2014, accessed August 6, 2015. [Online]. Available: http://nfdump.sourceforge.net/

[2] P. Velan, "fbitdump," Web page, December 2015, accessed August 6, 2015. [Online]. Available: https://github.com/CESNET/ipfixcol/tree/master/tools/fbitdump

[3] L. Deri and S. Suin, "Practical network security: experiences with ntop ," *Computer Networks*, vol. 34, no. 6, pp. 873 – 880, 2000, pioneering Tomorrow's Internet: Selected papers from the {TARENA} Networking Conference 2000. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1389128600001584

[4] X. Cao, W. Feng, Y. Dou, Z. Lei, and H. Yu, "A space-saving method for aggregate Top-N flow statistics with high accuracy," in *Broadband Network and Multimedia Technology (IC-BNMT), 2011 4th IEEE International Conference on*, Oct 2011, pp. 407–411.

[5] I. Cisco Systems, "Cisco ios master command list, all releases," Online, January 2014, accessed December 6, 2015. [Online]. Available: https://www.cisco.com/c/en/us/td/docs/ios-xml/ios/mcl/allreleasemcl/all-book.pdf

[6] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," *Commun. ACM*, vol. 51, no. 1, pp. 107–113, Jan. 2008. [Online]. Available: http://doi.acm.org/10.1145/1327452.1327492

[7] M. Husak, M. Cermak, T. Jirsik, and P. Celeda, "Network-Based HTTPS Client Identification Using SSL/TLS Fingerprinting," in *Availability, Reliability and Security (ARES), 2015 10th International Conference on*, Aug 2015, pp. 389–396.

[8] P. Velan and P. Celeda, "Next Generation Application-Aware Flow Monitoring," in *Monitoring and Securing Virtualized Networks and Services*, ser. Lecture Notes in Computer Science, A. Sperotto, G. Doyen, S. Latr, M. Charalambides, and B. Stiller, Eds., vol. 8508. Springer Berlin Heidelberg, 2014, pp. 173–178. [Online]. Available: http://dx.doi.org/10.1007/978-3-662-43862-6_20

[9] B. Claise, B. Trammell, and P. Aitken, "Specification of the IP Flow Information Export (IPFIX) Protocol for the Exchange of Flow Information," RFC 7011 (INTERNET STANDARD), Internet Engineering Task Force, Sep. 2013. [Online]. Available: http://www.ietf.org/rfc/rfc7011.txt

[10] A. Moore, M. Crogan, and D. Zuev, "Discriminators for use in flow-based classification," Queen Mary, University of London, Tech. Rep., 2005. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.101.7450&rep=rep1&type=pdf

[11] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, Cloth, Ed. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2005, vol. 1.