On Information Value of Top N Statistics

INTERNATIONAL CONFERENCE ON IT CONVERGENCE AND SECURITY 2016

Wednesday 28th September, 2016

Tomáš Jirsík
Milan Čermák, Pavel Čeleda

T A
Č R

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.

Motivation

- Brace yourself, IoT is coming.
- Large volume of network data data to analyse.
- Nearly limitless number of primary or derived statistics to compute and analyze.
- Resource intensive task.

To measure, or not to measure

–

that is the question.

How about Top N?

Why Top N?

- Widely used in network security, network accounting
- Overview over most important events.
- Top talker identification.
- Widely supported by tools for network traffic analysis (e.g., nfdump, fbitdump, ntop, ...)

We focus on ...

- ... nature of Top N statistics,
- ... characteristics of information provided by Top N statistics with respect to ...
- ... suitability of host identification from network traffic.

All about Top N

General Definition

Top N of X sorted by Y, over period of time P

e.g., Find 3 IP addresses that transferred the most bytes during last five minutes

Top N computation

1. Select data from period P.
2. Selected data are aggregated according return characteristics X and compute aggregated characteristics of Y.
3. Sort data by aggregated values of Y characteristics.
4. Cut off first N records from sorted list.

Top N for host identification

Host identification from network data

- Seems easy, is it really?
- MAC Address - unusable network monitoring
- IP Address - could be used, but
  - Network address translation
  - Dynamic addressing

Data sources

- Deep packet inspection
- Network flows
  - Abstraction of network connection
  - Aggregation of information from packets with same flow keys

Top N for host identification

Return characteristics X

- L2 - useless, lost after next hop
- L3/4
  - src/IP address, src/dst port - enough combination, but....
  - protokol nubmer - useless
- L7 - application information
  - e.g. HTTP protocol - Host, URI,

Sorting characteristics Y

- Number of flows
- Number of unique pairs

Experimental Evaluation

Evaluation metrics for Top N statistics

- **General**
  - Availability - is the statistics available
  - Time stability - how does the statistics behave in time
- **Host identification**
  - Uniqueness - how unique Top N is for a given host
  - TP/FP rates

Dataset

	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

Availability Evaluation

P = 5 minutes		P = 1 hour		P = 1 day	
# of obs.	% of IP	# of obs.	% of IP	# of obs.	% of IP
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

Time Stability Evaluation

	P = 1 hour			P = 1 day		
	% of IP addresses					
Equal rec.	**DstIP**	**DstPort**	**HTTP**	**DstIP**	**DstPort**	**HTTP**
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

Uniqueness Evaluation

Two Top N statistics are similar, when Jaccard is greater than 0.25 (i.e. approx. 4 equal records in two Top 10 statistics).

	P = 1 hour			P = 1 day		
	% of statistics					
U(s)	DstIP	Dst-Port	HTTP	DstIP	Dst-Port	HTTP
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

Host Identification Evaluation

TP - a host is within a set of identified hosts.

Period	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

Host Identification Evaluation

Cardinality of identified set

		% 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

Conclusions

- We need to choose, which characteristics are measured.
- We showed behavior of Top N statistics for individual hosts.
- The experimental evaluation on real-world data showed that a period P correlates with availability and time stability of the statistics.
- The uniqueness has been highest for Top N of DstIP statistics and increased with longer period.
- Statistic has a limited application on host identification problem. It could be enhanced by combining more types of Top N statistics together.

ON INFORMATION VALUE OF TOP N STATISTICS

Tomáš Jirsík

jirsik@ics.muni.cz