ESnet
ENERGY SCIENCES NETWORK

Probabilistic Heavy Hitters Fun with SumStats

Jim Mellander

Cybersecurity Engineer, ESnet

Lawrence Berkeley National Laboratory

BroCon 2018

Arlington, VA

October 11, 2018

U.S. DEPARTMENT OF ENERGY
Office of Science

BERKELEY LAB

Problem: Tracking Key & Amount in a Memory-efficient manner.

- SumStats Framework summarizes a set of observations, in order to communicate the largest amount of information as simply as possible. (Wikipedia)
  - Limit memory usage, streaming observations.
- Existing probabilistic SumStats plugins (HyperLogLog & TopK) track cardinality only = Updates by 1.
- What if you want weighted updates, e.g. track amounts such as bytecounts by IP, and extract the Heavy Hitters in realtime?
  - Can use the Sum plugin, but memory utilization quickly becomes a problem, since we're keeping sums for many IPs that will never make the Heavy Hitters list (and SumStats has no way of purging)
  - There's got to be a better way!?

ESnet

Enter: Modified Misra-Gries Summary Algorithm

- Reference: "A High-Performance Algorithm for Identifying Frequent Items in Data Streams", Anderson, Bevin, Lang, Liberty, Rhodes, Thaler, 2017
  - https://conferences.sigcomm.org/imc/2017/papers/imc17-final255.pdf
  - Based on: "Finding repeated elements", Misra, Gries, 1982
- Idea is to keep a fixed size table, keyed by item with amounts.
- Smart purging to retain heavy hitters
- Probabilistic guarantees.

ESnet

Modified Misra-Gries Summary Algorithm

127.0.0.1 555	10.0.0.1 666	172.16.0.1 833
10.0.0.2 1277	192.168.7.8 72	172.19.1.1 777
192.168.1.3 2222	172.17.1.1 68944	10.1.1.1 314159

Initial Fill of Table

ESnet

Modified Misra-Gries Summary Algorithm

127.0.0.1 7777	10.0.0.1 6661	172.16.0.1 8332
10.0.0.2 12773	192.168.7.8 724	172.19.1.1 7775
192.168.1.3 22222	172.17.1.1 6894466	10.1.1.1 31415926

Updates to Items in Table

Simple addition to table values

ESnet

Modified Misra-Gries Summary Algorithm

127.0.0.1 7777	10.0.0.1 6661	172.16.0.1 8332
10.0.0.2 12773	192.168.7.8 724	172.19.1.1 7775
192.168.1.3 22222	172.17.1.1 6894466	10.1.1.1 31415926

192.168.1.15
271828

------?---->

No Room to Add.

What to Do?

ESnet

Modified Misra-Gries Summary Algorithm

192.168.1.15
271828

------?---->

127.0.0.1 7777	10.0.0.1 6661	172.16.0.1 8332
10.0.0.2 12773	192.168.7.8 724	172.19.1.1 7775
192.168.1.3 22222	172.17.1.1 6894466	10.1.1.1 31415926

Median = 8332

- Calculate Median of Sampled Entries (not all)
- Reduce all values by Median
- Delete all entries now <= 0
- Add new entry (if > Median)

ESnet

Modified Misra-Gries Summary Algorithm

192.168.1.15
271828-8332

---------->

127.0.0.1 7777-8332 **Free!**	10.0.0.1 6661-8332 **Free!**	172.16.0.1 8332-8332 **Free!**
10.0.0.2 12773-8332	192.168.7.8 724-8332 **Free!**	172.19.1.1 7775-8332 **Free!**
192.168.1.3 22222-8332	172.17.1.1 6894466- 8332	10.1.1.1 31415926- 8332

Median = 8332

- Average number of slots freed is 50%
- Keep track of total medians subtracted for final result
- Can use a Lazy delete algorithm to free one slot at a time

ESnet

Modified Misra-Gries Summary Algorithm

- Keep track of total median subtracted.
- Can also easily keep track of Grand Total.

192.168.1.15 263496	Free!	Free!
10.0.0.2 4441	Free!	Free!
192.168.1.3 13890	172.17.1.1 6886134	10.1.1.1 31407594

After Freeing space & subtracting median

ESnet

Results: Modified Misra-Gries Summary Algorithm

- Final results are the contents of the table, with total median added.
- Obviously, cannot rely on the more recent entries to the table as much, since ~50% of the table is replaced each median calculation.
  - Just looking for heavy hitters, though. Top 10% empirically reliable.
    - As in any probabilistic algorithm, there are some data patterns that will produce inaccurate results.
  - Statistical guarantees, see paper for details.

ESnet

Where to get more information: Modified Misra-Gries Summary Algorithm

- mg.bro installed as plugin in bro/share/bro/base/frameworks/sumstats/plugins
  - Includes efficient TopK sort function.
  - Can also be installed in site directory & loaded.
- Demo Programs
  - heavyhitters.bro – uses Sum plugin (memory problems!)
  - mg-heavyhitters.bro – using mg plugin – much better memory footprint
  - Both use a “long-running connection” strategy.
- Where to get: TBD (soon!)
  - Need to put in boilerplate UC copyright notice
  - Put up on GitHub

ESnet