Cerca i Anàlisi d'Informació (CAI)
Grau en Ciència i Enginyeria de Dades, UPC

Session 7: Streams and sketches

Exercise List, Fall 2019

Basic comprehension questions.
Check that you can answer them before proceeding

1. Define in your words "sketch" and why it is useful.
2. Of the three V's that people use to characterize Big Data (Volume, Velocity, Variety), which ones do you think are addressed by the sketches seen in class?
3. Explain in one sentence what each of the following sketches computes, and how much memory each one uses: Reservoir sampling, Morris' counter, Hyperloglog, SpaceSaving, Exponential Histograms.

Note: We use Mb, Gb, to denote Megabytes, Gigabytes, etc. We use Mbit and Gbit for Megabits and Gigabits.

log denotes base 2 logarithm.

Exercise 1

You need to keep 10 billion counters each of which can count up to approximately 900. Describe a data structure that does this in the smallest memory you can think of.

You can assume that no counter will be asked to count more than 900 items. Or that after a counter reaches 900, its result doesn't matter any more.

Each count you give is expected to be correct within 20% approximately.

Exercise 2

You are an Internet switch. You see about 1 trillion (10^{12}) packets per day, with each packet containing, besides other info, the IPv6 addresses that are the origin and the destination of the package; an IPv6 address has 128 bits.

You have very limited memory for statistics, say 1Mb. Can you do the following tasks? How?

- **Stat1:** Keep every pair of (origin,destination) addresses that have appeared in at least 0.1% of the packets you've seen.
- **Stat2:** Keep the number of distinct pairs (origin,destination) that you have seen in each of the 24 hours, with precision 10%.

The statistics are reset at the start of each day, and after 24 hours you send them somewhere, so you can forget them and start a new cycle.

Futuristic scenario: Move forward to the 24th century. You are the largest switch for the internet of the United Federation of Planets, located near the huge black hole at the center of the Milky Way - from which, by the way, you draw your energy. You see a sextillion (10^{21}) packets of IPv27 each day. IPv27 was adopted as the pangalactic standard in 2366; an IPv27 address has 1024 bits.

Your job is so trivial for your HyperAI that, out of boredom, you make a bet with your HyperAI buddy in Andromeda: that you can still do both Stat1 and Stat2 in 1Mb. That's the same memory used by the routers of the early 21st century - those primitive things with sub-biological intelligence and infraluminic speed. You once met one in an archaeology museum, and the poor thing couldn't even start the most basic conversation besides work, checksums, and protocol.

Can you win your bet?

Exercise 3

We are the IPv6 switch in the exercise above. Now, at the end of each day, we are asked to report the fraction of packets that we have seen circulating from each country to each other country. Assume that we can know the country of an IP address, and that there are C=200 countries, so C^2 = 40,000 potential pairs (country of origin, country of destination).

For example, if in one day we see 1 trillion packets, and 1 billion of them went from Australia to the US, the report for the pair (Australia,US) should be 0.001. The approximation for each pair should be within 10% of correct.

Use an array of Exponential Histograms (EH) to solve this problem, and answer the following questions:

- What is the length of the sliding window you need to use, n?
- What parameter k do you use for each EH?
- How many buckets will each EH have?

- How much memory does a bucket use?
- So, how much memory does one EH use?
- And so, how much memory does the total data structure use?

It is not guaranteed that you can give totally precise answers to all items. Make reasonable guesses. (Think of your boss asking you "but can I do this with 10Mb? Or 1Gb? Or do I need to buy a 100Gb RAM server?", what do you tell her?)

Exercise 4

Now you are again the 24th century hyperswitch at the center of the Milky Way. You are asked to compute Stat1 and Stat2 daily, but aggregating by traffic among planetary systems. There are a billion $C = 10^9$ inhabited planetary systems in the Federation, so $C^2 = 10^{18}$ possible counts to keep.

Yet, you know that the traffic among the vast majority of the C^2 pairs of planetary system is negligible. So the Federation is only interested in keeping the counts for each pair of systems that are using at least $\theta = 0.0001$ of the bandwidth.

Explain how you approach the problem. You can use one sketch cleverly, or perhaps the combination of several sketches.