

# Asmt 4: Frequent Items \*

Jesus Zarate

March 22, 2017

## 1 Streaming Algorithms

- (A) Run the Misra-Gries Algorithm (see L11.3.1) with  $(k - 1) = 9$  counters on streams S1 and S2. Report the output of the counters at the end of the stream. In each stream, from just the counters, report how many objects might occur more than 20% of the time, and which must occur more than 20% of the time

Note: Because of the way I implemented the algorithm I simply remove the label once they get to 0

Output of counters:

S1.txt = 'c': 105715, 'a': 195715, 'b': 155715, 'v': 1, 'f': 1, 'p': 1, 'o': 2, 's': 0, 'l': 0

S2.txt = 'b': 135715, 'l': 1, 'c': 175715, 'a': 245715, 'j': 1, 'w': 1, 'r': 1, 'f': 0, 'h': 1

S1 must: None

S1 might: 'a', 'b', 'c'

S2 must: 'a'

S2 might: 'c', 'b'

---

\*CS6140-Data Mining; Spring 2017

- (B) Build a Count-Min Sketch (see L12.1.1) with  $k = 10$  counters using  $t = 5$  hash functions. Run it on streams S1 and S2.

For both streams, report the estimated counts for objects a, b, and c. Just from the output of the sketch, which of these objects, with probability  $1 - \delta = 31/32$ , might occur more than 20% of the time?

S1:

a - 282879

b - 226455

c - 176507

a and b might occur more than 20%

S2:

a - 316651

b - 193402

c - 233373

a and c might occur more than 20%

- (C) How would your implementation of these algorithms need to change (to answer the same questions) if each object of the stream was a word seen on Twitter, and the stream contained all tweets concatenated together?

The way the streaming is done would change. Instead of parsing the character we would have to parse the words.

For Misra-Gries the labels would have to change in order to fit the word and not just a character.

For the Count-Min Sketching nothing comes to mind that would have to change because the hashing of the word would still be able to get hashed, as a word just as a single character.

- (D) Describe one advantage of the Count-Min Sketch over the Misra-Gries Algorithm.

Subtraction is allowed. Count-Min Sketch is a Turnstyle Model, which means we can add and subtract things rather than just always putting things in.