

# Asmt 4: Frequent Items

Gopal Menon

Turn in through Canvas by 2:45pm:

Wednesday, March 22

## 1 Streaming Algorithms

**A: (20 points)** 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.

The estimated counts for streams  $S1$  and  $S2$  are given in tables 1 and 2 respectively.

Table 1: Misra-Gries Counter Outputs for stream  $S1$

c	a	b	o	v	f	p
105,715	195,715	155,715	2	1	1	1

Table 2: Misra-Gries Counter Outputs for stream  $S2$

b	c	a	h	l	j	w	r
135,715	175,715	245,715	1	1	1	1	1

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.

For any item  $q$ , the actual frequency  $f_q$  and the frequency  $\hat{f}_q$  reported by the algorithm are related by the equation

$$f_q - \frac{m}{k} \leq \hat{f}_q$$

where  $m = 1,000,000$  is the size of the stream, and  $k = 10$  is the number of counters. Substituting these values into the above equation, we get

$$f_q - \frac{1,000,000}{10} \leq \hat{f}_q$$

$$f_q - 100,000 \leq \hat{f}_q$$

$$f_q - \hat{f}_q \leq 100,000$$

This means that the maximum possible undercounting is by 100,000. Given that 20% of 1,000,000 is 200,000, any label with count more than 200,000, must occur more than 20% of the time since overcounting is not possible. So label  $a$  in stream  $S2$  must occur more than 20% of the time. Any label with count between 100,000 and 200,000 might occur more than 20% of the time. So labels  $a$ ,  $b$  and  $c$  in stream  $S1$  and labels  $b$  and  $c$  in stream  $S2$  might occur more than 20% of the time.

**B: (20 points)** 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 probably  $1 - \delta = \frac{31}{32}$ , might occur more than 20% of the time?

The estimated counts for streams  $S1$  and  $S2$  are given in table 3.

Table 3: Count-Min Sketch Counter Outputs			
Stream	a	b	c
$S1$	250,000	243,121	160,000
$S2$	290,000	206,917	220,000