TEL AVIV UNIVERSITY
Department of Computer Science
0368.3239 – Foundations of Data Mining
Fall Semester, 2017/2018

Homework 1, November 2, 2017

- **Due on Thursday November 16 23:59 IST.**
- **Submission instructions: We are using `https://gradescope.com`. Gradescope entry code for the course is: MYVDZ5**
  Please prepare a PDF file with each problem starting on a new page. When uploading, you will need to indicate locations of each problem/section.
- **You may consult any sources or people but you must write and submit the solution yourself and state your collaborators.**

1. **The Misra Gries Sketch and Stream Order:** Consider the application of a Misra Gries sketch of size $k = 2$ to a stream of 90 elements with 5 distinct keys $(x_1, x_2, x_3, x_4, x_5)$ with respective frequencies $(50, 10, 10, 10, 10)$.

   a. (5 points) Which arrangement of the stream will maximize the count of x_1 and what would be this count? Prove your answer.

   b. (5 points) From the final sketch you got in (a), and the length of the stream (m=90), what guarantee we get on the actual frequency of x_1?

   c. (5 points) Which arrangement of the stream will minimize the count of x_1 and what would be this count? Prove your answer.

   d. (5 points) From the final sketch you got in (c), and the length of the stream (m=90), what guarantee we get on the actual frequency of x_1?

2. **Flipper's Randomized Counter:** A stream of elements is counted with the following approximate counter that has a parameter $p \in (0, 1]$:

   Initialize: $s \leftarrow 0$.

   Process element: Let $u \sim U[0, 1]$. If $u \leq p$ then $s \leftarrow s + 1$

   a. (5 points) Compute an unbiased estimator for the number n of elements

   b. (5 points) Express the variance and the coefficient of variation (CV) (in terms of n)

c. (5 points) Write the expression for the likelihood function (a function of n and s that is the probability having the counter value s after n increments)

d. (5 points) Compute an expression (as a function of p) for the Maximum Likelihood estimator (MLE) for $s = 1$

e. (5 points) For $n = 1$, express (in terms of p) the bias of the MLE estimate (defined as $E[\hat{n}_{\text{MLE}}] - n$), and the Mean Squared Error (MSE).

f. (5 points) Are Flipper counters composable? If so, show how to merge two counters.

g. (10 points) What are the advantages of Morris counters over Flipper's counters? Compare the CV and the expected counter size.

3. **Reservoir Sampling (Vitter's)**

a. (10 points) Describe an algorithm for merging of two reservoir samples S_1 and S_2 of size k. The sample S_i is of a set of elements E_i that has size $n_1 = |E_1|$. The result is a uniform sample S of size k of $E_1 \cup E_2$.

Assume that the number of elements n_1 and n_2 are provided with S_1 and S_2.

b. (5 points) Prove that the final inclusion probability of each element $e \in E_1 \cup E_2$ in S is $\min\left\{1, \frac{k}{n_1+n_2}\right\}$.

4. **Bloom filter with deletion:**

a. (10 points) How would you modify the Bloom filter structure so that it supports insertions and deletions. Assume that each element can be inserted only if it is not currently stored in the data structure and that it can be deleted only if it currently stored. That is, the sequence of operations on element a must be a non-empty prefix of a sequence of pairs

$$\texttt{Bloom.insert}(a), \texttt{Bloom.delete}(a)$$

Note that you cannot insert an element that is already stored in the data structure but you can re-insert an element that was deleted. .

Hint: Replace each bit of the Bloom filter with a counter.

b. (5 points) Suppose we insert m elements into a Bloom filter of size n (with n counters) with parameter k.

Explain how to compute a bound on the length of each counter (in bits) so that the probability of an overflow is at most δ.

By an overflow we mean that there are not enough bits to count. Recall that we need $\lceil \log_2(M) \rceil$ bits to count up to M.

Hint: Use a multiplicative Chernoff bound.
Hint: First bound the probability that one counter overflows. Then use an approximation for the probability that none overflows.

5. **Maximum likelihood estimator** Consider a k-mins minhash sketch of a set with n distinct keys.

a. (5 points) Derive the Maximum Likelihood Estimators (MLE) for estimating $1/n$ and n^2

b. (5 points) What are the corresponding sufficient statistics?