INDIAN STATISTICAL INSTITUTE

Mid Semestral Examination

M. Tech (CS), 2025-2026 (Semester - I)

Algorithms for Big Data

Date: 10.09.2025 Maximum Marks: 100 Duration: 2.0 Hours

Note: This is a 2-page question paper. Answer all questions.

$E[X]$ and $\mathrm{Var}[X]$ denote the expectation and variance of the random variable X, respectively.

Let $\sigma = \langle a_1, \ldots, a_i, \ldots, a_m \rangle$ be a stream over the universe $\mathcal{U} = [n]$, $[n]$ denotes $\{1, \ldots, n\}$. The frequency of $a_i \in \sigma$ is denoted as f_i.

(QA1) (i) Let X be the distribution of an unbiased estimator for a real quantity Q. Let $\{X_{ij}\}_{i\in[t],j\in[k]}$ be a collection of independent random variables with each X_{ij} distributed identically to X, where

$$t = c \log \frac{1}{\delta} \text{ and } k = \frac{3\,\mathrm{Var}[X]}{\varepsilon^2\, E[X]^2}$$

and c is a universal positive constant. Let

$$Z = \mathrm{median}_{i\in[t]} \left(\frac{1}{k} \sum_{j=1}^{k} X_{ij} \right).$$

Then, show that $\Pr(|Z - Q| \geq \varepsilon Q) \leq \delta$, i.e., Z is an (ε, δ)-estimate for Q.

(ii) Using the result obtained in (i) or otherwise, show that an algorithm $\mathcal{A}$ that acts as a basic estimator can be converted to an (ε, δ) algorithm.

(iii) If $\mathcal{A}$ uses $O(k)$ space, comment on the space usage of the (ε, δ) algorithm. [10+6+4=20]

(QA2) In the FREQUENT problem with a parameter k for a stream σ, one needs to output the set $\{j \mid f_j > \frac{m}{k}\}$. Design a deterministic algorithm for the FREQUENT problem for a stream σ. Comment on the number of passes and space usage of your algorithm. [8+2=10]

(QA3) (i) Describe the Count-Min-Sketch algorithm that counts approximately the frequency of an element by using two operations – increment and count. The algorithm has two parameters – the number of buckets b and the number of hash functions ℓ used. Show how the error parameters, ε in the frequency count and δ in the error probability, play a role in fixing b and ℓ.

(ii) In the *ε-approximate heavy hitters* (ε-HH) problem for a stream σ, based on user-defined parameters k and ε, the problem is to output a list $\mathcal{L}$ of items from σ such that:

- every item that occurs at least $\frac{m}{k}$ times in σ is reported in $\mathcal{L}$.
- every item in $\mathcal{L}$ occurs at least $\frac{m}{k} - \varepsilon m$ times in σ.

Suppose, m, the size of the stream σ is unknown. So, one has no clue on the parameters $\frac{m}{k}$ and εm, during the processing of the stream. Solve (ε-HH) for σ using Count-Min-Sketch or otherwise in this setting. You are allowed to set ε as a function of k. Comment on the space usage in terms of ε in the frequency count and δ in the error probability, and m and n, if needed.

[15+10=25]

(QA4) (i) Define F_k, the k-th frequency moment of a data stream σ.

(ii) Describe the AMS Tug-of-War sketch for computing F_2.

(iii) Analyze the quality of the estimate given by the algorithm. Comment on the space usage of the algorithm by using the median-of-means improvement.

[2+6+(9+3)=20]

(QA5) A graph $G = (V, E)$ is streaming in the *edge arrival* model.

(i) Design and analyze a streaming algorithm for determining if the graph is bipartite. Prove the correctness of the algorithm.

(ii) Design and analyze a streaming algorithm to compute an estimate on the shortest path between any two vertices x, y in G using the idea of spanners. Analyze the quality of the estimate returned by the algorithm. You need not discuss the space complexity.

[(5+5)+(6+9)=25]