SEEM5020: Individual Project

Total weight: 20% **Due date:** 2026.04.18 11:59:59

Overview

This semester's project focuses on **Frequency Estimation under the Strict Turnstile Model with α-Bounded Deletion property**.

To understand the scope of the project, we formalize the stream updates and define the underlying data stream models. Let the data stream be a sequence of updates (e_t, v_t) at time t, where $e_t \in \mathcal{U}$ is an element from a universe and v_t is the update amount (weight). The frequency f_e of an element e is the sum of all v_t such that $e_t = e$.

1. Standard Data Stream (Insertion-Only): All update amounts are strictly positive ($v_t > 0$). Elements only arrive.

2. Turnstile Model: Generalizes the standard stream by allowing both positive and negative update amounts ($v_t > 0$ for insertions, $v_t < 0$ for deletions).

3. **Strict Turnstile Model**: For this project, we explicitly assume a strict turnstile setting. This means that while elements can be inserted and deleted (i.e., v_t can be positive or negative), the true frequency f_e of any individual element e can never drop below zero at any point in the stream ($f_e \geq 0$ for all e at all times).

Additionally, we assume the stream satisfies the **α-Bounded Deletion property** (also known as the α-property). The α-bounded deletion model itself does not inherently dictate a strict turnstile; rather, it is a property of the stream's volume. As defined in the original paper, a stream satisfies the L_1 α-property if the total absolute volume of all updates is at most α times the total final weight of the stream, where $\alpha \geq 1$. Formally, $\sum_t |v_t| \leq \alpha \sum_e |f_e|$. For theoretical background, refer to arxiv:1803.08777.

Your task is to extend classical frequency estimation algorithms to accommodate this model, rigorously evaluate their performance under various conditions, and document your findings.

Tasks

(1) Algorithm Implementation and Extension [6 marks]

- Implement the following classic frequency estimation algorithms and propose/implement extensions for them to explicitly handle the **Strict Turnstile Model with α-Bounded Deletion property**:
  - Misra-Gries algorithm

  - Space-Saving algorithm
  - Count-Min Sketch
  - Count-Sketch
- *Note:* Count-Min and Count-Sketch natively support turnstile updates, but consider if the α-bounded property can be used to optimize them or tighten their error bounds. Misra-Gries and Space-Saving are originally insertion-only, so extending them to handle deletions under the strict turnstile and α-bounded properties is a primary challenge. You may want to do literature research and look into recent variants of Space-Saving for inspiration; be careful of alternative definitions of the α-bounded deletion property.

(2) Parametric Evaluation [5 marks]

- Conduct experiments to compare the performance of your four extended algorithms.
- Evaluate the **relative error** and **space consumption** by varying the following parameters:
  - **Stream Length:** How does the performance scale as the number of total operations in the stream increases?
  - α **parameter:** How do varying levels of stream churn (e.g., $\alpha = 1.5, 2, 4, 8$) impact the accuracy and space requirements?

(3) Diverse Dataset Evaluation [5 marks]

- Evaluate your implementations on a variety of datasets to analyze robustness under different data characteristics. The input datasets must include:
  - **Skewed Synthetic Data:** e.g., generated using a Zipfian distribution.
  - **Balanced Synthetic Data:** e.g., generated using a uniform distribution.
  - **Real-world Datasets:** Find and use at least one real-world dataset (such as network packet traces, NLP text corpora, or web clickstream data) and simulate updates on it such that the strict turnstile and α-Bounded Deletion properties are satisfied.

(4) Report on Experimental Results & Advanced Design [4 marks]

- Document your findings in a comprehensive report that includes:
  - **Results Presentation:** Clear presentations of your experimental results (using charts, graphs, and tables).
  - **Analysis:** A summary of the advantages, disadvantages, and trade-offs of each extended method under different α values and data distributions.

- **Advanced Solution Design:** Propose and conceptually design an enhanced solution or novel optimization that could outperform the baseline extended methods (in terms of smaller error or less space) specifically within the Strict Turnstile Model with the α-Bounded Deletion property.

Submission Guidelines

- Please submit one of the following to Blackboard. A zip file of your source code and the report pdf **OR** a pointer to your project github repository
- Ensure your code is well-commented, modularized, and includes a `README` with instructions on how to reproduce the experiments.