Project Progress Report #3
Local Model Heavy Hitters

Yueran Cao, Kang Zhao

April 28, 2025

What have we done so far?

Since the last update, we made significant strides in completing our project. We finished the data preprocessing pipeline for the March 2025 English Wikipedia clickstream dataset, including filtering and sampling the raw data and generating the synthetic client inputs for each algorithm. We have fully implemented both the Trie Heavy Hitters (TrieHH) algorithm and Apple's Sequence Frequency Puzzle (SFP) algorithm in our codebase, simulating the local differential privacy setting for each. After integrating the algorithms with the preprocessed data, we ran extensive experiments (with multiple Monte Carlo runs) to evaluate their performance. We successfully generated the F1-score vs. K plots for each algorithm, complete with 95

Deviations from original plan

Our current progress largely aligns with the project plan, but there were a few deviations. First, instead of using purely synthetic data with various distributions as initially envisioned, we pivoted to using a real-world dataset (the *Wikipedia clickstream*) to drive our evaluation. This decision was made to demonstrate the algorithms on realistic data; however, it meant that we did not explore as many distribution scenarios (e.g., uniform vs. skewed) as originally planned. We also focused on a single privacy setting ($\varepsilon = 4$ with a very small δ on the order of 10^{-12}) for the majority of our tests, rather than experimenting with multiple ε values, due to time constraints. Implementing the SFP algorithm turned out to be more complex than expected, which required some additional debugging and slightly adjusted our timeline. Despite these minor deviations, the core goal of comparing TrieHH and SFP under local differential privacy was achieved.

Next steps / Progress plan

In the coming days, our main focus is preparing for the final project presentation. We will:

- Finalize the analysis and visualization: We plan to polish the F1-vs-K plot and ensure our interpretation of the results is clear and backed by the data. If time permits, we might perform a brief additional experiment (for example, relaxing certain parameters) to further validate our findings.
- Prepare presentation slides: We will create slides summarizing our project motivation, approach, results, and key insights. Emphasis will be on the heavy hitters problem, how TrieHH and SFP work, and the comparison of their performance on the clickstream data.
- Practice and refine the talk: We intend to rehearse our 15-minute presentation to ensure it fits the time limit and that our explanations are concise and accessible. We will refine our talking points about any challenges we faced (such as implementing SFP or handling the large dataset) and highlight what we learned through this project.
- Incorporate feedback: As we finalize our written report (approximately 1.5 pages as required), we will incorporate any feedback received on the draft. We will double-check that the report clearly communicates our methods and findings in a formal academic style.

By following this plan, we aim to deliver a clear and informative final presentation and submit a polished final report that accurately reflects our work on discovering heavy hitters under local differential privacy.