Project Proposal

Chenjia Li, Xiang Li, ChinKuan Lin

March 2024

1 Motivation

In the era of big data, efficiently summarizing and extracting insights from vast streams of information is a critical challenge across numerous domains, from social media analytics to real-time network monitoring. Traditional algorithms often struggle to balance accuracy with the constraints of time and space complexity, particularly when dealing with data streams. This project is motivated by both the lecture content and the opportunity to explore and evaluate various frequency count methods that promise to offer a balance between computational efficiency and accuracy. By comparing algorithms like Count Min Sketch and Misra-Gries as the baseline against more recent approaches like BJKST, Tidemark, AMS Sketch, Tug-of-War Sketch, Lossy Counting, and the Flajolet-Martin Algorithm, we aim to deepen our understanding of their operational trade-offs and practical utility.

2 Data(Tentative)

Our evaluation will be conducted using a mix of synthetic and real-world datasets. The synthetic datasets will allow us to control the parameters of the data stream, such as distribution type, skewness, and volume, to systematically assess each algorithm's performance under varied conditions. For real-world data, we will use publicly available datasets like network traffic logs, social media feeds, and transaction records, which are characterized by large volumes and rapid arrival rates. These datasets will help us evaluate the algorithms' effectiveness in practical scenarios, reflecting the challenges faced by industries today. Following is our list of dataset with source.

1. `https://www.kaggle.com/datasets/datagov/usa-names`
2. `https://www.kaggle.com/datasets/mcpenguin/unicode-emoji-frequency-ranks`
3. Self generated
4. TBD

3 Interesting Questions(Tentative)

The foundational analysis for our project will draw upon questions similar to those posed in Homework 2, with a focus on implementing and testing the CountMin Sketch and exploring its estimation

efficiency, accuracy, and other kind of properties. Building on this, we will expand our investigation to include other algorithms and introduce practical considerations alongside theoretical inquiries. Our questions will explore not only the fundamental aspects of these algorithms but also their applicability to real-world data streams and potential for optimization. Here are the revised and expanded questions:

1. Basic Performance Analysis:
   (a) Accuracy vs. Space Trade-off: How does the accuracy of frequency estimations vary with changes in the algorithm's space allocation? This involves varying parameters such as the number of buckets (K) and the number of hash functions (N), examining the trade-offs between memory usage and estimation accuracy.
   (b) Error Distribution: What is the observed distribution of errors in frequency estimations? How does it compare with the theoretical error bounds provided by each algorithm's design? This question seeks to understand the practical error characteristics beyond average case scenarios.
2. Algorithm Comparison:
   (a) Comparative Efficiency: When comparing algorithms like BJKST, Tidemark, AMS Sketch, Tug-of-War Sketch, Lossy Counting, and the Flajolet-Martin Algorithm against Count Min Sketch and Misra-Gries, how do they fare in terms of time complexity and space efficiency for similar accuracy levels?
   (b) Robustness to Data Variability: How do these algorithms perform across different data stream characteristics, such as skewness, volume, and velocity? This question aims to evaluate each algorithm's adaptability to varying data conditions.
3. Advanced Theoretical Questions:
   (a) Beyond the Markov's Inequality Bound: Homework 2 utilized Markov's inequality to bound the probability of overestimation. Can we identify scenarios/dataset where this analysis is not tight and propose more precise bounds or alternative probabilistic analyses for the algorithms under study?
   (b) Theoretical vs. Practical Performance: How well do theoretical predictions of performance and error bounds align with empirical observations across different datasets and algorithm configurations?

4 Rough Contribution Clarification

Everyone in this team will contribute equally to this project. Indeed, everyone will write code, conduct to the final report and presentation.

References

[Araed] N.R. Aravind. Lecture 14: The bjkst algorithm. `https://people.iith.ac.in/aravind/Files-CS5120/pc-lec14-BJKST.pdf`, No Year Provided. Accessed: 03/26/2024.

[Cha19] Amit Chakrabarti. Lecture: Tug of war sketch. https://www.comp.nus.edu.sg/~gilbert/CS5234/2019/lectures/ChakrabartiTugOfWar.pdf, 2019. Accessed: 03/26/2024.

[Flaed] Flajolet-martin algorithm. https://www.geeksforgeeks.org/flajolet-martin-algorithm/, No Year Provided. Accessed: 03/26/2024.

[Kon20] Christian Konrad. Lecture 04: The tidemark algorithm. http://people.cs.bris.ac.uk/~konrad/courses/2020_2021_COMSM0068/slides/04-tidemark.pdf, 2020. Accessed: 03/26/2024.

[Naved] Garg Naveen. Lecture 21: Streaming algorithms. https://www.cse.iitd.ac.in/~naveen/courses/CSL758/scribes/lec21.pdf, No Year Provided. Accessed: 03/26/2024.

[Pri16] Eric Price. Lecture 6: The ams sketch. https://www.cs.utexas.edu/~ecprice/courses/sublinear/fa16/notes/lec6.pdf, 2016. Accessed: 03/26/2024.

[Pro20] No Author Provided. Lecture 06: Data stream algorithms. https://courses.engr.illinois.edu/cs498abd/fa2020/slides/06.pdf, 2020. Accessed: 03/26/2024.

[PVK22] Rameshwar Pratap, Bhisham Verma, and Raghav Kulkarni. Improving *Tug-of-War* sketch using control-variates method, 03 2022.

[Reied] John H. Reif. Lecture: Streaming algorithms. https://users.cs.duke.edu/~reif/courses/alglectures/jain.lectures/StreamingAlgorithms.pdf, No Year Provided. Accessed: 03/26/2024.

[Vog15] Michael Vogiatzis. Frequency counting algorithms over data streams. https://micvog.com/2015/07/18/frequency-counting-algorithms-over-data-streams/, 2015. Accessed: 03/26/2024.

[Zha19] Chihao Zhang. Lecture 4: Data stream algorithms. http://chihaozhang.com/teaching/BDA2019fall/slides/lec4-slides-handout.pdf, 2019. Accessed: 03/26/2024.