Sketch-based Querying of Distributed Sliding-Window Data Streams
by O. Papapetrou, M. Garofalakis, and A. Deligiannakis

1 Summary

In this paper authors considered the problem of answering complex queries over distributed and high dimensional data streams, in the sliding window model. They propose an ECM-sketch, which is a new data structure that allows to store statistics about single and distributed streams. ECM-sketches are hybrid of Min-Count sketches and Exponential Histograms. Authors considered also Deterministic Waves and Randomized Waves as possible substitutions for Exponential Histograms. In addition, they defined algorithms to aggregate ECM-sketches directly and in a hierarchical setting with respect to evaluated queries.

The ECM-sketches have been extensively analyzed, and proved to provide probabilistic accuracy guarantees for the quality of the results. Types of problems used in such analyzes were: point queries, self-join queries, finding heavy hitters, computing quantiles, and answering range queries over sliding windows. ECM-sketches were also heavily evaluated in a set of experiments, using two large real-world datasets, and considering both centralized and hierarchically distributed settings.

2 Strong Points

- The paper contains detailed analyzes of performance complexities for considered settings.
- Authors presented not only exponential histograms, but also other implementations of sliding window counters in their ECM-sketches using deterministic waves and randomized waves.
- Authors examined two types of data aggregation topologies while testing ECM-sketches, i.e., a single aggregator (star topology) and hierarchical aggregation (tree topology).
- The paper presents a relation between approximation errors and complexities, e.g., by calculating values of approximation error thresholds, for which complexity of query replies is minimal. Analytical calculations have been verified in experiments.

3 Weak Points

- The paper does not describe neither deterministic nor randomized waves in details. Authors referred to original papers presenting such waves.
- Authors presented only examples of queries and problems that can be solved using ECM-sketches. They did not characterize groups or classes of problems that can and cannot be solved using their approach.
- While describing experimental results authors mentioned a few times about insufficient memory to finish computations for some settings, but in the setup they did not specify how much memory was available for computations. Therefore, it is difficult to judge if lack of memory for such settings is caused by complexity of a tested setting or limited hardware resources.

4 Questions and Discussion Points

1. It would be interesting to reverse the way authors mixed components of ECM-sketches, i.e., adapt count-min sketches to be used in exponential histograms. Would it be a better/worse solution than the presented one?

2. Authors did not specify what problems cannot be addressed by ECM-sketches. It would be interesting to find limits of such data structure.

Paper Presentations – Preferences.

1. Summarization and Matching of Density-Based Clusters in Streaming Environments by D. Yang, E. A. Rundensteiner, and Matthew O. Ward

2. Sketch-based Querying of Distributed Sliding-Window Data Streams by O. Papapetrou, M. Garofalakis, and A. Deligiannakis