WARWICK
THE UNIVERSITY OF WARWICK

Review of "Streaming Algorithms for Estimating Quantiles with Novel Error Guarantees"
by Tomáš Domes

The problem of estimating the quantiles of a (massive) stream of values has a long and storied history within computer science, dating back to the work of Munro and Paterson in the late 1970s, who were motivated to study the problem in the context of processing data stored on tape, which could be read only sequentially. In the 21st century, the problem is motivated by the need to analyze truly massive volumes of data that exceed the memory available to store information. Over a long sequence of papers, a number of results have been shown that progressively improve our understanding of the computational complexity of this task (measured by the space needed to obtain a given level of accuracy), and the corresponding usefulness of these algorithms in practical settings. This thesis further extends this sequence of results, and expands our understanding of the complexity of quantile estimation in the streaming model of computation.

The contribution in this thesis is to build on the algorithms presented in the KLL and ReqSketch papers (formal citations in the thesis). A dichotomy in quantile algorithms is whether the target accuracy is measured in absolute or relative terms. Given a stream of items drawn from an ordered universe, and a query location p, the aim is to return an item that is close to p in the cumulative distribution function of item arrivals, i.e., to report q such that $Pr[x < q] \sim= p$, where Pr[.] is the implicit CDF of the input. Under absolute error, it is permitted to report an item that is a small absolute amount e away under the CDF, i.e., an item q such that $|Pr[x < q] - p| < e$. Under relative error, the target is instead to report q such that $(1-e) < Pr[x<q]/p < (1 + e)$. The KLL algorithm is based on building a hierarchical collection of "buffers", along with a randomized procedure to compact these buffers, in order to use the retained information to solve the absolute error quantiles problem. The ReqSketch uses a similar but more involved construction and a more complex compaction schedule (and analysis) in order to provide a relative error guarantee.

The focus of this thesis is to provide a modified algorithm and analysis, leveraging the ReqSketch approach, in order to provide improved guarantees for certain values within the quantile space. The intuition is that in order to give the strong error guarantee needed for relative error across the entirety of the quantile domain, the ReqSketch algorithm may give be loose in its analysis in some regions, and this slack can be used to provide a tighter bound. In particular, it takes some extra parameters: a set of target quantiles of interest (expressed as a set of important ranks, R), and an additional "jaggedness" parameter J, in order to express the bounds. The thesis provides a formal theorem and discussion of the interpretation of the new bounds in terms of these parameters. In short, the error guarantees are never worse than the guarantees from the original ReqSketch algorithm without these parameters; they are equal for some extreme values; but at various points in the domain the guarantees can be improved by factors logarithmic in the size of the input.

The technical progression of the thesis is very clear: the prior work is explained in full detail, so that the new algorithmic approach can be described as a generalization. The key notion of 'jaggedness' is introduced as a way to smoothly modify the pattern of sizes of buffers in order to preserve greater accuracy for certain ranks in the ordered distribution. (Specifically, the pattern is requited to vary according to a polynomial schedule between the "jags"). The bulk of the report is concerned with the mathematical analysis of the behavior of the proposed algorithm. This follows by defining a collection of random variables that describe the error of the algorithm, based on the randomness introduced, and applying the tools of concentration of measure to argue that the with high probability the error is bounded. At a high level, this follows the pattern set in prior work in outline, but requires substantially new and challenging steps in order to adapt to the new algorithm. The discussion proceeds logically, starting by studying the operation of a single compactor in the static setting, when the compactor sizes are fixed for the duration of the algorithm, and then proceeding to generalize to multiple compactors at different levels of the hierarchy, and outlining the changes that are needed when the compactor sizes are allowed to be dynamic (as is required when the length of the input stream N is not fixed at the start of the operation).

Department of Computer Science
The University of Warwick
Coventry CV4 7AL United Kingdom
Tel: +44 24 7652 3987
Email: G.Cormode@warwick.ac.uk
www.warwick.ac.uk

Finally, the thesis reports some empirical results, which demonstrate that the improved analysis also translates to improved performance on realistic data. This is important because the analysis leads to some large constants that are hidden by the big-Oh notation, whereas the experimental study shows absolute values. The conclusion is that the promised improvements in accuracy can indeed be obtained by the new approach.

The thesis is overall of a very high standard, and worthy of the Masters degree for which it is submitted. There are substantial novel and interesting findings from the research work which are suitable for publication in a leading venue on algorithms, such as the APPROX and RANDOM peer-reviewed international conferences. On this basis, I can with high confidence recommend that the thesis be accepted as it is submitted.

I have some minor high-level suggestions to consider, should a revised version be planned for submission for publication, as follows:

- As noted throughout the thesis, the proof of the results follows a similar path to that used in the ReqSketch paper. It would help the reader to highlight the places where there is significant deviation, or where a substantially different approach was needed in order to prove a desired step.
- Similarly, the key difference in the specification of the problem is in the jaggedness factor J and set of important ranks R. It would be helpful to flag the points in the analysis where these are brought into play. This is particularly important for the important ranks R, since it can be hard at times to visually distinguish this from the rank function R() – perhaps different notation would help to separate these two?
- For the experimental study, it is certainly the case that it suffices to consider permutations as the input to test the algorithm's performance. However, it would be of value to also consider permutations that are not selected uniformly at random, since this may be considered as potentially an "easy" case for quantile algorithms to handle. Permutations that are skewed to include higher items at the start or the end, or to have other challenging patterns, may be of interest – note that the prior work cited which includes experiments often includes non-uniform permutations designed to present a challenge for the algorithms studied.

Prof. Graham Cormode
G.Cormode@warwick.ac.uk