Technical Perspective on 'Better Differentially Private Approximate Histograms and Heavy Hitters using the Misra-Gries Sketch'

Graham Cormode
University of Warwick, UK
G.Cormode@warwick.ac.uk

The topics of private data analysis and streaming data management have both been separately the focus of much study within the data management community for many years. However, more recently there have been studies which bring these two previously isolated topics together.

Although data streams and privacy might not seem to have much in common, it turns out that they share a symbiotic relationship. Within data streams, a common pattern is to design a compact data structure that summarizes the input that has been seen so far, and which can be updated quickly in order to reflect a small change. Within privacy, it is often helpful to build an intermediate representation of a data set, so that a small change to the input does not change the representation much. Moreover, both areas make use of the tools of randomized algorithms and probability to prove that their mechanisms give strong guarantees. Thus, data stream summaries can sometimes inspire private algorithms, and vice-versa.

A few prior examples of this phenomenon include analyzing the (differential) privacy of randomized projections via the Johnson-Lindenstrauss lemma [3] and the popular Flajolet-Martin count-distinct summaries [4]. This allows a variety of downstream data analysis to be performed using these data summaries, which are both compact and private. A major application of private data analysis is to find popular items among a large collection of observations, and tools for this deployed in major operating systems have made use of stream summaries (Bloom filters and Count-min sketches) [1, 2].

This work, by Lebeda and Tětek, first published in PODS 2023, also considers frequencies, and asks whether a popular streaming data structure can be made private. Specifically, it considers the Misra-Gries (MG) sketch, a simple and effective data stream summary that allows the frequency of each item to be estimated up to a $1/k$ additive error while using space proportional to k.

One perspective on the MG sketch is that it allows us to extract a sparse vector representation of the input. If we think of the input as defining a stochastic frequency vector x (i.e., $\|x\|_1 = 1$), then the MG summary provides a vector y such that the maximum error in any coordinate is bounded ($\|x - y\|_\infty \leq 1/k$) and it is sparse ($\|y\|_0 \leq k$). This is tantalizing, since given x in full there is an easy differentially private way to obtain a y that is sparse: add noise to each entry of x, and retain only those noisy entries that exceed a threshold. However, under the constraint of small space and streaming updates, it is not at all straightforward to obtain such a representation.

The approach here is very effective: essentially, with a simple post-processing step, the MG summary can be made (ϵ, δ)-differentially private. The algorithm is almost identical to the naive idealized algorithm: add noise to each entry of the sparse MG vector, and retain only those noisy entries that exceed a (noisy) threshold. The heavy lifting comes in the proof, which involves an intricate and clever case analysis to argue that the MG summary of two close inputs can only vary in a bounded way. Then, by relating the noise to this bound on deviation, differential privacy can be shown.

The implications of this result are powerful: we have a deeper understanding of this classical streaming algorithm, and a space-efficient technique to capture private frequency statistics of streams of data. Applications which process large volumes of data can form private representations based on the efficiently computable MG sketch.

The paper opens the door to further studies of questions at the intersection of streaming and privacy. What other practical and widely-used summaries can be made private by similar carefully calibrated noise addition? More broadly, what other models of computation can be made privacy-aware—in particular, computations that handle large volumes of data in support of "big data" and AI/ML workflows, and associated monitoring tasks? The nascent areas of federated analytics and federated learning in particular, which seek to support private evaluation of complex queries and modeling over distributed data, are ripe for making use of compact data summaries to minimize communication costs.

1. REFERENCES

[1] Apple Differential Privacy Team. Learning with privacy at scale. *Apple ML Research*, 2017.

[2] Ú. Erlingsson, V. Pihur, and A. Korolova. RAPPOR: randomized aggregatable privacy-preserving ordinal response. In *CCS*, 2014.

[3] K. Kenthapadi, A. Korolova, I. Mironov, and N. Mishra. Privacy via the Johnson-Lindenstrauss transform. *J. Priv. Confidentiality*, 5(1), 2013.

[4] A. D. Smith, S. Song, and A. Thakurta. The Flajolet-Martin sketch itself preserves differential privacy: Private counting with minimal space. In *Neural Information Processing Systems*, 2020.