Beating CountSketch for heavy hitters in insertion streams

Abstract:

The task of finding heavy hitters is one of the best known and well studied problems in the area of data streams. In a sense, the strongest guarantee available is the L2 guarantee, which requires finding all items that occur at least $\epsilon ||f||$ times in the stream, where the $i$-th coordinate of the vector $f$ is the number of occurrences of $i$ in the stream. The first algorithm to achieve the L2 guarantee was the CountSketch (Charikar, Chen, and Farach-Colton ICALP'02), which, for constant $\epsilon$, requires $O(\log n)$ words of memory and $O(\log n)$ update time. It is known to be space-optimal if the stream includes deletions.

In this talk I will discuss recent improvements that allow us to find L2 heavy hitters in $O(1)$ memory and $O(1)$ update time in insertion only streams. The improvements rely on a deeper understanding of the AMS sketch (Alon, Matias, and Szegedy STOC'96) and similar sketches and draw on the theory of Gaussian processes. This talk is based on joint work with Vladimir Braverman, Nikita Ivkin, Jelani Nelson, Zhengyu Wang, and David P. Woodruff in arxiv:1511.00661 and arxiv:1603.00759.