October 2010

MADALGO seminar by Qin Zhang, Aarhus University

Optimal Sampling from Distributed Streams

Abstract:

A fundamental problem in data management is to draw a sample of a large data set, for approximate query answering, selectivity estimation, and query planning. With large, streaming data sets, this problem becomes particularly difficult when the data is shared across multiple distributed sites. The challenge is to ensure that a sample is drawn uniformly across the union of the data while minimizing the communication needed to run the protocol on the evolving data. At the same time, it is also necessary to make the protocol lightweight, by keeping the space and time costs low for each participant.

In this paper, we present communication-efficient protocols for sampling (both with and without replacement) from *k* distributed streams. These apply to the case when we want a sample from the full streams, and to the sliding window cases of only the *W* most recent elements, or arrivals within the last *w* time units.

We show that our protocols are optimal (up to logarithmic factors), not just in terms of the communication used, but also the time and space costs for each participant.

Joint work with Graham Cormode S. Muthukrishnan and Ke Yi