MASTERS
FINAL ORAL EXAMINATION

Wednesday, November 16th
12:01p.m. @ 223 Atanasoff

Chenguang He

Major Professor: Wensheng Zhang

Sampling Over Distributed Streaming Data

Streaming data is a data type that has been widely used in a variety of real-time, data-intensive applications. This data type appears as endless and continuous sequence of data items. Due to this nature, it is not realistic to store the sequence in memory or external storage; instead, the data sequence can be examined typically in just one single pass. Furthermore, in a distributed environment where sequences of streaming data are produced in different sites and multiple distributed sequences need to be collected for analysis, it is not realistic to transfer the sequences themselves directly over the network. Rather, random sampling should be conducted on each sequence to get a smaller set of samples, and the sample sets are then collected and analyzed. Hence, how to efficiently conduct random sampling over distributed data streams becomes an important problem.

In this creative component, I have designed a new sampling algorithm that extends Vitter's reservoir sampling algorithm to sample over distributed data streams. The proposed algorithm uses random gap sampling to reduce the processing cost while ensuring the obtained set of samples to be selected from the original data sequence uniformly at random. I have also implemented the algorithm, evaluated and justified the randomness of the obtained samples, and compared this algorithm with the state-of-the-art algorithms to demonstrate its improvement in efficiency.

IOWA STATE UNIVERSITY
Department of Computer Science