Algorithms For Big Data Analysis

Rationale

Traditional analysis of algorithms generally assumes full storage of data and considers running times polynomial in input size to be efficient. Operating on massive-scale data sets such as those of tech companies such as Google, Facebook, etc., or on indefinitely large data streams, such as those generated by sensor networks and security applications, leads to fundamentally different algorithmic models. MapReduce/Hadoop in particular has seen widespread adoption in industry.

Course Description

This course addresses algorithmic problems in a world of big data, i.e., problems in settings where the algorithm's input [the data] is too large to fit within a single computer's memory. Traditional analysis of algorithms generally assumes full storage of data and considers running times polynomial in input size to be efficient. Operating on massive-scale data sets such as those of tech companies such as Google, Facebook, etc., or on indefinitely large data streams, such as those generated by sensor networks and security applications, leads to fundamentally different algorithmic models. In previous decades, DBMS settings where the data sets on a machine's disk but not in memory motivated the external memory or I/O model (e.g. ex- ternal mergesort and B-trees). More recently, models such as MapReduce/Hadoop have appeared for computing on data distributed across many machines (e.g. PageRank computation or matrix multiplication).

Topic List

The topics may include but are not limited to:

- MapReduce/Hadoop

- Streaming and Sketching

- Finding Matchings

- Counting Triangles

- Deciding Bipartiteness and Connectivity

- The Secretary problem

- Matrix Multiplication

- Compressed Sensing

Learning goals

The student will learn what the modern models for massive-scale data algorithms are, how to analyze algorithms in these models, and when to use them. In particular, the student will gain experience with MapReduce/Hadoop and with a number of streaming/sketching algorithms.

Assessment

Several bi-weekly problem sets and a course project. The project can take the form of either an expository paper, a nontrivial implementation project, or performing original research on a problem.