Algorithms for Big Data

This course will discuss about algorithms that process massive amounts of data that does not fit in a computer's storage. This calls for a model-centric view of algorithms where even basic algorithmic problems, like counting the number of distinct elements, selection, and sorting become challenging. We will cover such topics under some broad heads and pick a subset of topics from them --- (i) Streaming and sketching; (ii) Sub-linear algorithms with some ideas about handling big data from different domains in practical situations;

(a) Topics:

(i) Streaming and sketching:

Sketching and Streaming algorithms for basic statistics: Distinct elements, heavy hitters, frequency moments, p-stable sketches

Graph streaming algorithms: connectivity, cut/spectral sparsifiers, spanners, matching, graph sketching

Lower bounds for Sketching and Streaming: communication complexity: Equality, Index and Set-Disjointness

Dimensionality Reduction: Johnson Lindenstrauss lemma, lower bounds and impossibility results

Locality Sensitive Hashing: similarity estimation, approximate nearest neighbor search, data dependent hashing

Geometric streaming algorithms: Coresets; The min-enclosing-ball problem, Metric streams; Clustering: *k*-center, *k*-median, *k*-means.

(ii) Sub-linear algorithms: Approximating the diameter of a point set; Approximating the number of connected components, the average degree, the number of edges in a graph; Notions of property testing for sortedness, connectivity, Minimum Spanning Tree weight and Connected components estimation

(b) Pre-requisites/co-requisites, if any: Design and Analysis of Algorithms, Probability is essential; Topics in Algorithms and Complexity is desirable

(c) Number of lectures (tutorials if any) per week: Four lectures per week

(d) Percentage weights for theory and programming/software assignments/mini projects (if any): Theory: 60% and assignment (including problems / exercises / programming assignments / project/mini project): 40%

(e) List of references:

The course will draw its contents from various surveys, lecture notes, papers, apart from the following references:

1. Foundations of Data Science by A. Blum, J. Hopcroft and R. Kannan.
2. Probabilistic Methods by Noga Alon and Joel H. Spencer, Wiley
3. Introduction to Property Testing by Oded Goldreich, Cambridge University Press
4. Design and Analysis of Algorithms: A Contemporary Perspective by S. Sen and A. Kumar, Cambridge University Press

(f) Supplementary information: This course was designed with inputs from the various courses listed at https://www.sketchingbigdata.org/