Prof. Micheal Mahoney

Title: Algorithmic and statistical perspectives on large-scale data analysis

Abstract:

Computer scientists and natural scientists have historically adopted quite different views on data and thus on data analysis. For example, the former tend to view the data as noiseless and focus on algorithms with bounds on worst-case running time, independent of the input; while the latter often have, either explicitly or implicitly, an underlying statistical model in mind. In recent years, however, ideas from statistics and scientific computing have begun to interact in increasingly sophisticated and fruitful ways with ideas from computer science and the theory of algorithms to aid in the development of improved worst-case algorithms that are also useful in practice for solving large-scale scientific and Internet data analysis problems.

After reviewing these two complementary perspectives on data, I will describe two recent examples of improved algorithms that used ideas from both areas in novel ways. The first example has to do with improved methods for structure identification from large-scale DNA SNP data, a problem which can be viewed as trying to find good columns or features from a large data matrix. The second example has to do with selecting good clusters or communities from a data graph, or demonstrating that there are none, a problem that has wide application in the analysis of social and information networks. Understanding how statistical ideas are useful for obtaining improved algorithms in these two applications may serve as a model for exploiting complementary algorithmic and statistical perspectives in order to solve applied large-scale scientific and Internet data analysis problems more generally.