From *SIAM News*, Volume 42, Number 1, January/February 2009

MMDS 2008: Algorithmic and Statistical Challenges in Modern Large-Scale Data Analysis, Part I

By Michael W. Mahoney, Lek-Heng Lim, and Gunnar E. Carlsson

The 2008 Workshop on Algorithms for Modern Massive Data Sets (MMDS 2008), held at Stanford University, June 25–28, had two goals: first, to explore novel techniques for modeling and analyzing massive, high-dimensional, and nonlinearly structured scientific and Internet data sets, and second, to bring together computer scientists, statisticians, mathematicians, and data analysis practitioners to promote the cross-fertilization of ideas. The workshop was sponsored by NSF, DARPA, Yahoo! and LinkedIn.

MMDS 2008 grew out of discussions of our vision for what the algorithmic, mathematical, and statistical analyses of large-scale, complex data sets should look like a generation from now. These discussions occurred in the wake of MMDS 2006, which had been motivated by the complementary perspectives brought by the numerical linear algebra and theoretical computer science communities to combinatorial, numerical, and randomized algorithms in modern informatics applications (see www.siam.org/news/news.php?id=1019). As with the 2006 meeting, the MMDS 2008 program was intensely interdisciplinary, with close to 300 participants representing a wide spectrum of research in modern large-scale data analysis.

Diverse Approaches to Modern Data Problems

A common way to model a large social or information network is with an *interaction graph model* $G = (V, E)$, in which nodes in the vertex set V represent "entities" and the edges (whether directed or undirected, weighted or unweighted) in the edge set E represent "interactions" between pairs of entities. Alternatively, because an $m \times n$ real-valued matrix A provides a natural structure for encoding information about m objects, each described by n features, these and other data sets can be modeled as matrices. Because of their large size, their extreme sparsity, and their complex and often adversarial noise properties, data graphs and data matrices arising in modern informatics applications present considerable challenges and opportunities for interdisciplinary research. These algorithmic, statistical, and mathematical challenges were the focus of MMDS 2008.

Historically, very different perspectives have been brought to such problems. Computer scientists interested in data mining and knowledge discovery commonly view the data in a database as an accounting or a record of everything that happened in a particular setting. The database might consist of all customer transactions over the course of a month, or of all friendship links between members of a social networking site. From this perspective, the goal is to tabulate and process the data at hand to find interesting patterns, rules, and associations. An example of an association rule is the proverbial "People who buy beer between 5 PM and 7 PM buy diapers at the same time." The performance or quality of such a rule is judged by the fraction of the database that satisfies the rule exactly, and the problem then boils down to that of finding frequent item sets. This is a computationally hard problem, and much algorithmic work has been devoted to its exact or approximate solution under different models of data access.

When it comes to Web-scale data analysis, an algorithm that is expensive in floating-point cost but readily parallelizable is often a better choice than a less expensive algorithm that is not parallelizable.

A very different way to view the data, more common among statisticians, is as a particular random instantiation of an underlying process describing unobserved patterns in the world. In this case, the goal is to extract information about the world from the noisy or uncertain data that is observed. To achieve this, one might posit the model $data \sim F_\theta$ and $\text{mean}(data) = g(\theta)$, where F_θ is a distribution that describes the random variability of the data around the deterministic model $g(\theta)$ of the data. Using this model, researchers would analyze the data to make inferences about the underlying processes and predictions about future observations. From this perspective, modeling the noise component or variability is as important as modeling the mean structure, in large part because understanding the former is necessary for understanding the quality of the predictions made. This approach even permits predictions about events not observed before—it is possible to give, for example, the probability that a given user on a given Web site will click on a given advertisement presented at a given time of day, even if this particular event does not exist in the database.

Of course, we have many indications that the two perspectives are not incompatible: Statistical and probabilistic ideas are central to many recent efforts to develop improved approximation algorithms for matrix problems; otherwise-intractable optimization problems on graphs and networks yield to approximation algorithms when assumptions are made about the network participants; much recent work in machine learning draws on ideas from both areas; and in boosting, a statistical technique that fits an additive model by minimizing an objective function with a method like gradient descent, the computation parameter, i.e., the number of iterations, also serves as a regularization parameter.

Given the diversity of possible perspectives, MMDS 2008 was loosely organized around six hour-long tutorials that introduced participants to the major themes of the workshop; each is briefly summarized either here or in Part II of this article.

Large-Scale Informatics: Problems, Methods, and Models

Christos Faloutsos of Carnegie Mellon University opened the tutorial "Graph Mining: Laws, Generators and Tools" by describing a wide range of applications in which graphs arise naturally. Large graphs that arise in modern informatics applications, he pointed out, have structural properties very different from those of traditional Erdős–Rényi random graphs; an example is the heavy-tailed behavior of degree distributions, eigenvalue distributions, and other statistics that result from subtle correlations.

These structural properties have been studied extensively in recent years and have been used to develop numerous well-publicized models; Faloutsos also described empirically observed properties that are not well reproduced by existing models. Most models, for example, predict that over time a graph should become sparser and the diameter should grow as $O(\log N)$, or perhaps $O(\log \log N)$, where N is the number of nodes at the current timestep; empirically, though, many of these networks have been observed to densify over time and to shrink in diameter. To explain these phenomena, Faloutsos described a model based on Kronecker products and another in which edges are added via an iterative "forest fire" mechanism. With appropriate choice of parameters, both models can be made to reproduce a range of static and dynamic properties much wider than those of previous generative models.

Building on this modeling foundation, Faloutsos described several graph-mining applications of current interest: methods for finding nodes that are central to a group of individuals; use of the singular value decomposition and recently developed tensor methods to identify anomalous patterns in time-evolving graphs; modeling information cascades in the blogosphere as virus propagation; and novel methods for fraud detection.

Other developments in Web-scale data analysis were the subject of Edward Chang's tutorial, "Mining Large-scale Social Networks: Challenges and Scalable Solutions." After reviewing emerging applications—such as social network analysis and personalized information retrieval—that have arisen as we transition from Web 1.0 (links between pages and documents) to Web 2.0 (links between documents, people, and social platforms), Chang, of Google Research, covered four applications in detail: spectral clustering for network analysis, frequent-itemset mining, combinatorial collaborative filtering, and parallel support vector machines (SVMs) for personalized search. In all these cases, he emphasized, the main performance requirements are "scalability, scalability, scalability."

Modern informatics applications like Web search afford easy parallelization—for example, the overall index can be partitioned in such a way that even a single query can use multiple processors. Moreover, the peak performance of a machine is less important than the price–performance ratio. In this environment, scaling up to petabyte-sized data often means working in a software framework that, like MapReduce or Hadoop, supports data-intensive distributed computations running on large clusters of hundreds, thousands, or even hundreds of thousands of commodity computers. This differs substantially from the scalability issues that arise in traditional applications of interest in scientific computing. A recurrent theme in Chang's presentation was that an algorithm that is expensive in floating-point cost but readily parallelizable is often a better choice than a less expensive algorithm that is not parallelizable.

As an example, Chang considered SVMs: Although widely used, mainly because of their empirical success and attractive theoretical foundations, they suffer from well-known scalability problems in both memory use and computational time. Chang described a parallel SVM algorithm that addresses these problems—it reduces memory requirements by performing a *row-based* incomplete Cholesky factorization (ICF) and by loading only essential data to each of the parallel machines, and it reduces computation time by intelligently reordering computational steps and performing them on parallel machines. The traditional *column-based* ICF, he pointed out, is better in a single-machine setting but is not suitable for parallelization across many machines.

Part II of this article will appear in an upcoming issue of SIAM News.

Michael Mahoney (mmahoney@cs.stanford.edu) is a research scientist in the Department of Mathematics at Stanford University. Lek-Heng Lim (lekheng@math.berkeley.edu) is a Charles Morrey Assistant Professor in the Department of Mathematics at the University of California, Berkeley. Gunnar Carlsson (gunnar@math.stanford.edu) is a professor in the Department of Mathematics at Stanford University.