Data Science Down Under Worksop

8-12 December 2019, Newcastle, Australia

Invited Talks

1. **Michael Mahoney, University of California – Berkeley, US**

   Time: Monday, 9 Dec 2019, 11:50am-12:40pm

   Title: Minimax Experimental Design, Exact Expressions for Double Descent, and Implicit Regularization in RandNLA Algorithms

   Abstract: Computer scientists adopt what may be termed an algorithmic perspective on their data. This is very different than the perspective adopted by statisticians, scientific computers, machine learners, and others who work on what may be broadly termed statistical data analysis. Informally, the former involves considering (e.g., providing worst-case or numerical running time bounds for algorithms on) "today's data," and the latter involves considering (e.g., inferential guarantees for) "tomorrow's data." Within the context of RandNLA, in particular, nearly all prior work adopts the algorithmic perspective and provides running time guarantees in one form or another, but the usefulness of RandNLA methods often comes from the randomness in the algorithm interacting in non-trivial ways with the randomness in the data, thereby providing strong statistical connections. I'll describe several recent RandNLA results aimed at bridging the gap between the statistical and worst-case approaches to algorithms. First, I'll describe a framework for the experimental design problem that involves a general response model that generalizes both the traditional statistical as well as the worst-case algorithmic perspectives. Second, I'll describe how we can use connections to determinantal point processes to get exact expressions for the mean squared error in both the overdetermined and underdetermined regime of least squares regression, illustrating that a so-called "double descent" phenomenon exists even for least squares regression problems. Third, I'll describe a precise sense in which RandNLA sampling and solving of (unregularized) least squares problems implicitly leads to the solutions of ridge regularized least squares problems.

2. **David Woodruff, Carnegie Mellon University, US**

   Time: Monday, 9 Dec 2019, 2:00pm-2:50pm

   Title: Towards a Zero-One Law for Column Subset Selection

   Abstract: There are a number of approximation algorithms for NP-hard versions of low rank approximation, such as finding a rank-k matrix minimizing the sum of absolute values of differences to a given n x n matrix A, or more generally finding a rank-k matrix which minimizes the sum of p-th powers of absolute values of differences to A. Many of these algorithms are linear time columns subset selection algorithms, returning a subset of columns whose cost is no more than a poly(k) factor larger than the cost of the best rank-k matrix. The above error measures are special cases of the following general entrywise low rank approximation problem: given an arbitrary function f, find a rank-k matrix B which minimizes |A-B|_f. A natural question is which functions admit efficient approximation algorithms? Indeed, this is a central question of recent work studying generalized low rank models. In this work we give approximation algorithms for every function which is approximately monotone and satisfies an approximate triangle inequality, and we show both of these conditions are necessary. Further, our algorithm is efficient if the function admits an efficient approximate regression algorithm. Our approximation algorithms handle functions which are not even scale-invariant, such as the Huber loss function, which we show have very different structural properties than norms, e.g., one can show the lack of scale-invariance causes any column subset selection algorithm

to provably require a factor sqrt{log n} larger number of columns than p-norms; nevertheless we design the first efficient column subset selection algorithms for such error measures.

3. **Kenneth Clarkson, IBM, US**

   Time: Monday, 9 Dec 2019, 2:50pm-3:40pm

   Title: Quantum-inspired-inspired Algorithms for Data Analysis

   Abstract: We describe algorithms for matrix rank computation, reduction to rank (that is, selection of linearly independent columns), leverage-score sampling, least-squares regression. The algorithms have modest space requirements, and their running times are modestly improved over prior work. We also describe a data structure supporting certain random queries on a low-rank approximation to an input matrix, targeted for application to recommender systems. For an m by n input matrix A with M nonzero entries, our algorithms take linear time, that is O(M), or within logarithmic factors of O(M), plus some *sublinear* work whose time bound is independent of M, m, and n. Our motivation is recent work on *quantum-inspired* algorithms: these are classical (that is, non-quantum) algorithms whose input-output model is based on that used by several quantum algorithms; the latter start with input data arranged in simple data structures and use those data structures to quickly prepare quantum states, to begin quantum processing. Quantum-inspired algorithms use those same data structures, to build data structures as output; the overall running times include a linear O(M) term, for building the data structures, plus a sublinear term, corresponding to the quantum processing. Here, we keep in the same general framework, but allow a broader range of linear-time operations, keeping to O(M) time or close to it, and obtain speedups in the sublinear part.

4. **Michael E. Houle, National Institute of Informatics (NII), Japan**

   Time: Monday, 9 Dec 2019, 4:10pm-5:00pm

   Title: Local Intrinsic Dimensionality: A Practical Foundation for Dimensionally-aware Data Analysis

   Abstract: Researchers have long considered the analysis of similarity applications in terms of the intrinsic dimensionality (ID) of the data. This presentation is concerned with a generalization of a discrete measure of ID, the expansion dimension, to the case of smooth functions in general, and distance distributions in particular. A local model of the ID of smooth functions is first proposed and then explained within the well-established statistical framework of extreme value theory (EVT). Moreover, it is shown that under appropriate smoothness conditions, the cumulative distribution function of a distance distribution can be completely characterized by an equivalent notion of data discriminability. As the local ID model makes no assumptions on the nature of the function (or distribution) other than continuous differentiability, its generality makes it ideally suited for the learning tasks that often arise in data mining, machine learning, and other AI applications that depend on the interplay of similarity measures and feature representations. An extension of the local ID model to a multivariate form will also be presented, that can account for the contributions of different distributional components towards the intrinsic dimensionality of the entire feature set, or equivalently towards the discriminability of distance measures defined in terms of these feature combinations. The talk will conclude with a discussion of recent applications of local ID to deep learning.

5. **Deanna Needell, University of California – Los Angeles, US**

   Time: Tuesday, 10 Dec 2019, 9:00am-9:50am

   Title: Simple Approaches to Complicated Data Analysis

Abstract: Recent advances in technology have led to a monumental increase in large-scale data across many platforms. One mathematical model that has gained a lot of recent attention is the use of sparsity. Sparsity captures the idea that high dimensional signals often contain a very small amount of intrinsic information. Using this notion, one may design efficient low-dimensional representations of large-scale data as well as robust reconstruction methods for those representations. Binary, or one-bit, representations of data for example, arise naturally in many applications, and are appealing in both hardware implementations and algorithm design. In this talk, we provide a brief background to sparsity and 1-bit measurements, and present new results on the problem of data classification with low computation and resource costs. We illustrate the utility of the proposed approach on recently acquired data about Lyme disease.

6. **Matt Wand, University of Technology Sydney, Australia**

   Time: Tuesday, 10 Dec 2019, 10:20am-11:10am

   Title: Streamlined Variational Inference for Random Effects Models

   Abstract: Variational inference offers fast approximate inference for graphical models arising in computer science and statistics. However, for models containing random effects, direct application of variational inference principles is not sufficient for fast inference due to the sizes of the relevant design matrices. We explain how the notion of matrix algebraic streamlining is crucial for making variational inference practical for models containing very high numbers of random effects. Both nested higher level and crossed random effect structures are discussed.

7. **Peter Taylor, University of Melbourne, Australia**

   Time: Tuesday, 10 Dec 2019, 11:10am-12:00pm

   Title: Some Thoughts About a Distributed Solution of the PageRank Equation

   Abstract: The success of Google is widely attributed to its initial use of the PageRank Algorithm developed in the late 1990s by Brin and Page. Although its primary purpose was to rank web pages, the PageRank Algorithm can be used to rank nodes on any directed graph and, in particular, it can be applied to social networks of all kinds. I shall start by discussing work that I did with some vacation students quite a few years ago in which we experimented with using PageRank to rank academic papers according to their citation graph. While I am not a believer in using citation data to assess the quality of papers, the PageRank algorithm does have some advantages that are not enjoyed by procedures that just count citations. The PageRank Algorithm really just solves an eigenvalue equation for a stochastic matrix. However, there are practical advantages in being able to do this in a distributed manner using only local information. In this talk, I shall propose a random procedure for doing this.

8. **Kate Smith-Miles, University of Melbourne, Australia**

   Time: Wednesday, 11 Dec 2019, 9:00am-9:50am

   Title: Party Tricks with Numerical Linear Algebra and the Quest for Trust

   Abstract: Establishing trustworthiness of algorithms in the minds of skeptical humans is a challenge that relies on improvements in transparency and elimination of any perception of bias. But trust is also earned, or easily destroyed, when we test an algorithm's performance on instances with known correct responses. If an algorithm obtains reliable results for enough test instances, then we are likely to feel comfortable with its decision-making ability; if it produces incorrect responses for even just a few critically chosen instances, then it will be dismissed quickly as unreliable and therefore untrustworthy. The choice of test instances is clearly critical to gaining trust, but their quality and sufficiency is rarely examined comprehensively, with each study usually inheriting a collection of benchmarks from previous studies. Furthermore, summarizing

algorithm performance "on average" across a database is not nuanced enough to really understand its strengths and weaknesses, and suitability. Our recent advances in instance space analysis has now enabled the sufficiency of test instances to be established, with consideration of key instance properties – diversity, unbiasedness, discriminatory power and real-world-like structures –offering visual evidence that we can be confident that an algorithm has been stress-tested under the widest range of scenarios. In this talk we revisit our study from a decade ago that proposed a numerical linear algebra algorithm for facial age estimation which, when tested on a popular database of faces, was shown to be superior to existing methods. From the fresh perspective offered by instance space analysis, we now challenge those conclusions, and expose the weaknesses of our method, and the strengths and weaknesses of competitor algorithms. These insights also lead to a new algorithm idea. The looming ethical implications of facial image analysis will also be briefly discussed, further justifying the urgency for methodologies to establish trustworthy algorithms.

9. Kerri Mengersen, Queensland University of Technology, Australia

Time: Wednesday, 11 Dec 2019, 9:50am-10:40am

Title: Bayesian Statistical Analysis of Large Images

Abstract: Images pervade almost all areas of science and society. As the generators of such images, such as cameras, satellites, medical scanning devices and digital text, become more sophisticated and accessible, the resultant images grow in size and complexity. The analysis of large images is now a research field in its own right and is characterised by a wide range of computational mathematical, statistical and machine learning solutions. Randomised numerical linear algebra, the focus of the DSDU Workshop in 2019, plays an important role in this field, as do Bayesian approaches. In this presentation, I will discuss some of our forays into Bayesian computational methods for the analysis of large images. These methods include scalable approximate algorithms and pre-processing (led by Dr Matthew Moores, now at University of Wollongong) and sparse matrix factorisation and nonlocal singular value thresholding (led by Dr Hongbo Xie at QUT). Discussion will also touch on the real-world problems that have motivated this research, including using satellite data to achieve UN Sustainable Development Goals (SDGs), improving medical imaging for radiation therapy and enhancing computer vision and pattern recognition.