MMDS 2014:

Workshop on Algorithms for Modern Massive Data Sets

Stanley Hall
University of California at Berkeley

June 17–20, 2014

The 2014 Workshop on Algorithms for Modern Massive Data Sets (MMDS 2014) will address algorithmic and statistical challenges in modern large-scale data analysis. The goals of MMDS 2014 are to explore novel techniques for modeling and analyzing massive, high-dimensional, and nonlinearly-structured scientific and internet data sets; and to bring together computer scientists, statisticians, mathematicians, and data analysis practitioners to promote the cross-fertilization of ideas.

Organizers: *Michael Mahoney, Alexander Shkolnik*
Petros Drineas, Reza Zadeh, and Fernando Perez

Workshop Schedule

Tuesday, June 17, 2014: Data Analysis and Statistical Data Analysis

Time	Event	Location/Page
Registration & Opening		**Lobby outside Stanley Hall**
8:00–9:45am	*Breakfast and registration*	
9:45–10:00am	Organizers *Welcome and opening remarks*	
First Session		**Stanley Hall**
10:00–11:00am	Ashok Srivastava, Verizon *Large Scale Machine Learning at Verizon (Tutorial)*	pp. 13
11:00–11:30am	Dan Suciu, University of Washington *Communication Cost in Big Data Processing*	pp. 14
11:30–12:00pm	Gerald Friedland, ICSI *Content-based search in 50TB of consumer-produced videos*	pp. 9
12:00–2:00pm	*Lunch* (on your own)	
Second Session		**Stanley Hall**
2:00–2:30pm	Bill Howe, University of Washington *Myria: Scalable Analytics as a Service*	pp. 10
2:30–3:00pm	Devavrat Shah, MIT *Computing stationary distribution, locally*	pp. 13
3:00–3:30pm	Yiannis Koutis, University of Puerto Rico *Spectral algorithms for graph mining and analysis*	pp. 12
3:30–4:00pm	Jiashun Jin, Carnegie Mellon University *Fast Network Community Detection by SCORE*	pp. 11
4:00–4:30pm	*Coffee break*	
Third Session		**Stanley Hall**
4:30–5:00pm	David Woodruff, IBM Research *Optimal CUR Matrix Decompositions*	pp. 14
5:00–5:30pm	Jelani Nelson, Harvard University *Dimensionality reduction via sparse matrices*	pp. 12
5:30–6:00pm	Jinzhu Jia, Peking University *Influence sampling for generalized linear models*	pp. 11
Evening Reception		
6:00–9:00pm	*Dinner Reception*	**Lobby outside Stanley Hall**

Wednesday, June 18, 2014: Industrial and Scientific Applications

Time	Event	Location/Page
First Session		**Stanley Hall**
9:00–10:00am	Leon Bottou, Microsoft Research *Counterfactual Reasoning and Learning Systems (Tutorial)*	pp. 8
10:00–10:30am	Sergei Vassilvitskii, Google *Connected Components in MapReduce and Beyond*	pp. 14
10:30–11:00am	*Coffee break*	
Second Session		**Stanley Hall**
11:00–11:30am	Xavier Amatriain, Netflix *Distributing Large-scale Recommendation Algorithms: from GPUs to the Cloud*	pp. 7
11:30–12:00pm	Jeffrey Bohn, State Street *Disentangling sources of risk in massive financial portfolios*	pp. 7
12:00–2:30pm	*Lunch* (on your own)	
Third Session		**Stanley Hall**
2:30–3:00pm	David Gleich, Purdue University *Localized Methods for Diffusions in Large Graphs*	pp. 9
3:00–3:30pm	Ashish Goel, Stanford University *FAST-PPR: Scaling Personalized PageRank Estimation for Large Graphs*	pp. 9
3:30–4:00pm	Michael Mahoney, ICSI and UC Berkeley *Locally-biased and semi-supervised eigenvectors*	pp. 12
4:00–4:30pm	*Coffee break*	
Fourth Session		**Stanley Hall**
4:30–5:00pm	Matan Gavish, Stanford University *Optimal Shrinkage of Fast Singular Values*	pp. 9
5:00–5:30pm	Reza Zadeh, Stanford University *Dimension Independent Matrix Square using MapReduce*	pp. 14

Thursday, June 19, 2014: Novel Algorithmic Approaches

Time	Event	Location/Page
First Session		**Stanley Hall**
9:00–10:00am	Andrew McGregor, University of Massachusetts *Analyzing Big Graphs via Sketching and Streaming (Tutorial)*	pp. 12
10:00–10:30am	Joshua Bloom, UC Berkeley *Large-Scale Inference in Time Domain Astrophysics*	pp. 7
10:30–11:00am	*Coffee break*	
Second Session		**Stanley Hall**
11:00–11:30am	Alek Kolcz, Twitter *Exploring "forgotten" one-shot learning*	pp. 11
11:30–12:00pm	Sreenivas Gollapudi, Microsoft Research *Modeling Dynamics of Opinion Formation in Social Networks*	pp. 10
12:00–12:30pm	Amit Singer, Princeton University *Multi-reference Alignment: Estimating Group Transformations using Semidefinite Programming*	pp. 13
12:30–2:30pm	*Lunch* (on your own)	
Third Session		**Stanley Hall**
2:30–3:00pm	Fernando Perez, UC Berkeley *IPython: a language-independent framework for computation and data*	pp. 13
3:00–3:30pm	Aydin Buluc, Lawrence Berkeley National Lab *Reducing Communication in Parallel Graph Computations*	pp. 8
3:30–4:00pm	Ankur Dave and Joseph Gonzalez, UC Berkeley *Large Scale Graph-Parallel Computation for Machine Learning: Applications and Systems*	pp. 10
4:00–4:30pm	*Coffee break*	
Fourth Session		**Stanley Hall**
4:30–5:00pm	Mark Embree, Virginia Tech *CUR Factorization via Discrete Empirical Interpolation*	pp. 9
5:00–5:30pm	Ilse Ipsen, North Carolina State University *Leverage scores: Sensitivity and an App*	pp. 11
5:30–6:00pm	Vikas Sindhwani, IBM Research *libSkylark: Sketching-based Accelerated Numerical Linear Algebra and Machine Learning for Distributed-memory Systems*	pp. 13
Evening Reception		
6:00–9:00pm	*Dinner Reception and Poster Session*	**Lobby outside Stanley Hall**

Friday, June 20, 2014: Novel Matrix and Graph Methods

Time	Event	Location/Page
First Session		**Stanley Hall**
9:00–10:00am	Matei Zaharia, Databricks and MIT *Large-Scale Numerical Computation Using a Data Flow Engine (Tutorial)*	pp. 14
10:00–10:30am	Eric Jonas, UC Berkeley *Automatic discovery of cell types and microcircuitry from neural connectomics*	pp. 11
10:30–11:00am	*Coffee break*	
Second Session		**Stanley Hall**
11:00–11:30am	Alexandr Andoni, Microsoft Research *Beyond Locality Sensitive Hashing*	pp. 7
11:30–12:00pm	Dorit Hochbaum, UC Berkeley *Combinatorial optimization and sparse computation for large scale data mining*	pp. 10
12:00–12:30pm	Christopher Stubbs, Harvard University *Public Participation in International Security—Open Source Treaty Verification*	pp. 13
12:30–2:30pm	*Lunch* (on your own)	
Third Session		**Stanley Hall**
2:30–3:00pm	Cecilia Aragon, University of Washington *The Hearts and Minds of Data Science*	pp. 7
3:00–3:30pm	Sebastiano Vigna, Universita degli Studi di Milano *The fall and rise of geometric centralities*	pp. 14
3:30–4:00pm	Constantine Caramanis, UT Austin *Mixed Regression*	pp. 8
4:00–4:30pm	Lisa Goldberg, UC Berkeley *No Free Lunch for Stress Testers: Toward a Normative Theory of Scenario-Based Risk Assessment*	pp. 9

Poster Presentations: Thursday, June 19, 2014

Event	Location/Page

Poster Session — **Lobby outside Stanley Hall**

Poster Presentations, continued.

Event	Location/Page
Poster Session	**Lobby outside Stanley Hall**
Xin Tong, University of South Carolina *Screening genome-wide DNA methylation CpG sites via training and testing data utilizing surrogate variables*	pp. 20
Madeleine Udell, Stanford University *Generalized Low Rank Modeling*	pp. 20
Dani Ushizima, Lawrence Berkeley National Laboratory *F3D: a Step Forward in Image Processing at Scale*	pp. 20
Eugene Vecharynski, Lawrence Berkeley National Lab *Fast updating algorithms for latent semantic indexing*	pp. 20
Stefan Wager, Stanford Univeristy *Semiparametric Exponential Families for Heavy-Tailed Data*	pp. 21
Jia Xu, University of Wisconsin-Madison *GOSUS: Grassmannian Online Subspace Updates with Structured-sparsity*	pp. 21
Joongyeub Yeo, Stanford University *Regime Change in Dynamic Correlation Matrices of High-Dimensional Financial Data*	pp. 21
Mahdi Zamani, University of New Mexico *Privacy-Preserving Multi-Party Sorting of Large Data Sets*	pp. 21
Vadim Zipunnikov, Johns Hopkins Bloomberg School of Public Health *Prediction of heart failure hospitalizations with wearable activity monitors*	pp. 21

Talk Abstracts

Distributing Large-scale Recommendation Algorithms: from GPUs to the Cloud

Xavier Amatriain, Netflix

We at Netflix strive to deliver maximum enjoyment and entertainment to our millions of members across the world. We do so by having great content and by constantly innovating on our product. A key issue to optimize both is to use the massive amounts of data from our users to come up with optimal approaches to applications such as content buying or our renowned personalization algorithms. But, in order to learn from this data, we need to be smart on the algorithms we use, and how we scale them to work on a service that operates under high availability constraints in more than 40 countries around the world. In this talk I will describe some of the machine learning algorithms that power our product. I will also describe how we have scaled them by distributing at 3 different levels: (1) across regions or subsets of the population, (2) at the hyperparameter optimization stage, and (3) at the model training level.

Beyond Locality Sensitive Hashing

Alexandr Andoni, Microsoft Research

Locality Sensitive Hashing (LSH) has emerged as a powerful tool for solving the Nearest Neighbor Search in high-dimensional spaces. The original scheme introduced in 1998 by Indyk and Motwani works for the Hamming space and has been later proven to be the optimal LSH algorithm.

We present a new data structure for approximate nearest neighbor search improving the query time Q of Indyk-Mowani to $Q^{7/8}$. Thus, this is not only the first improvement over the original LSH scheme, but it in fact circumvents the LSH lower bound itself!

Unlike previous algorithms, the new algorithm considers *data-dependent* hashing of the space, showing one can gain from a careful exploitation of the geometry of (any worst-case) dataset at hand. We expect that this technique will lead to more improvements in future.

Joint work with Piotr Indyk, Huy Nguyen, and Ilya Razenshteyn.

The Hearts and Minds of Data Science

Cecilia Aragon, University of Washington

Thanks in part to the recent popularity of the buzzword "big data," it is now generally understood that many important scientific breakthroughs are made by interdisciplinary collaborations of scientists working in geographically distributed locations, producing and analyzing vast and complex data sets. The extraordinary advances in our ability to acquire and generate data in physical, biological, and social sciences are transforming the fundamental nature of science discovery across domains. Much of the research in this area, which has become known as data science, has focused on automated methods of analyzing data such as machine learning and new database techniques. Less attention has been directed to the human aspects of data science, including how to build interactive tools that maximize scientific creativity and human insight, and how to train, support, motivate, and retain the individuals with the necessary skills to produce the next generation of scientific discoveries.

In this talk, I will argue for the importance of a human centered approach to data science as necessary for the success of 21st century scientific discovery. Further, I attest that we need to go beyond well-designed user interfaces for data science software tools to consider the entire ecosystem of software development and use: we need to study scientific collaborations interacting with technology as socio-technical systems, where both computer science and sociological approaches are interwoven. I will discuss promising research in this area, describe the current status of the Moore/Sloan Data Science Environment at UW, and speculate upon future directions for data science.

Large-Scale Inference in Time Domain Astrophysics

Joshua Bloom, UC Berkeley

The scientific promise of modern astrophysical surveys - from exoplanets to gravity waves - is palpable. Yet extracting insight from the data deluge is neither guaranteed nor trivial: existing paradigms for analysis are already beginning to breakdown under the data velocity. I will describe our efforts to apply statistical machine learning to large-scale astronomy datasets both in batch and streaming mode. From the discovery of supernovae to the characterization of tens of thousands of variable stars such approaches are leading the way to novel inference. Specific discoveries concerning precision distance measurements and using LSST as a pseudo-spectrograph will be discussed.

Disentangling sources of risk in massive financial portfolios

Jeffrey Bohn, State Street

Quantitative analysis of financial portfolios is predicated on a factor-based framework for assessing volatility-adjusted return. However, volatility is but one measure of risk. While volatility may be an adequate risk measure for an equity portfolio, it is insufficient for the multi-asset class portfolios held by large financial institutions such as pension funds, sovereign wealth funds, insurance companies and commercial banks. For example, these institutions typically hold significant positions in fixed-income securities, which have highly skewed return distributions due to the possibility of default. These institutions also tend to hold illiquid assets

such as real estate, private equity, derivatives and commodities, which incorporate risk not captured by volatility. To the extent that positions lose value at the same time, a portfolio return distribution may exhibit a "fat tail" i.e., a low but not unimaginable probability of losing a substantial amount of portfolio value. The recent financial crisis illustrated this worrisome possibility. Financial portfolios at large and systemically important financial institutions are characterized by the following:

1. Size: The portfolios are massive both in terms of value and number of positions/exposures.
2. Complexity: The assets included in the portfolios are economically and statistically diverse.
3. Downside risk: When neglected or poorly managed, downside risk may increase beyond an institution's risk appetite, and it may lead to unfavorable liquidation of valuable assets in a difficult macro-economic environment.

The recognition of risk beyond volatility is not new. In 1963, Benoit Mandelbrot argued the importance of accounting for tail risk in commodity prices, and commercial Value-at-Risk (VaR) models were available in the early 1990s. However, an accurate, efficient model that measures risk beyond volatility while clearly identifying the factors that drive risk has yet to be developed.

Data availability coupled with recent advances in data-related technologies have made it possible to consider more in-depth risk modeling approaches. We introduce a comprehensive and detailed factor-based, simulation framework implemented with data-driven tools to disentangle sources of financial portfolio risk and provide a simple-enough (but not necessarily simple) characterization of a massive, financial-portfolio's risk profile. We discuss statistical and visual approaches to improve an analyst's understanding. We also introduce some thoughts on how to move from risk management to portfolio optimization.

Counterfactual reasoning and massive data sets

Leon Bottou, Microsoft Research

The analysis of modern massive datasets sometimes has a clear purpose. Online service providers collect massive datasets because they hope to exploit their statistical correlations to improve the service using a suitable combination of machine learning and human engineering. The first part of the presentation explains the nasty surprises that can hit those who believe that observed correlations can be used to predict the outcome of interventions. One must instead understand how the consequences of our interventions traverse the graph of causes and effects to produce measurable results.

The second part of the presentation presents counterfactual reasoning techniques applicable to such problems. These techniques revolve around statements of the form "if we had applied change X to the system when the data was collected, we would have observed outcome Y." Such statements form a sound basis for reasoning about potential interventions. Such statements can also be verified empirically provided that the data is collected from a system that performs a sufficient level of random exploration. The more knowledge we have about the structure of the causal graph, the more efficiently we can leverage small amounts of exploration. Real-life examples will be provided.

The third part of the presentation returns to the general problem of causal discovery in massive datasets and contrasts two approaches. On the one hand, the tabula rasa approach tries to determine which causal relations explain the observed relation. On the other hand, the incremental approach tries to leverage known causal relations to discover new ones or question existing knowledge.

Reducing Communication in Parallel Graph Computations

Aydin Buluc, Lawrence Berkeley National Lab

Most graph algorithms have low computational intensity; hence their execution times are bounded by communication. In addition to improving the running time drastically, reducing communication can also help improve the energy efficiency of graph algorithms. In this talk, I will present three parallel algorithms that reduce the communication costs of certain graph computations. The new parallel communication-avoiding algorithms are all-pairs shortest-paths, breadth-first search, and sparse matrix-matrix multiplication. The last primitive can be used to implement various key graph algorithms, including betweenness centrality and graph contraction.

Mixed Regression

Constantine Caramanis, UT Austin

Mixture models represent the superposition of statistical processes, and are natural in machine learning and statistics. Despite the prevalence and importance of mixture models, little is known in the realm of efficient algorithms with strong statistical guarantees.

We consider the mixed regression problem with two components, under adversarial and stochastic noise. We give a convex optimization formulation that provably recovers the true solution, and we provide upper bounds on the recovery errors for both arbitrary noise and stochastic noise settings. We also give matching minimax lower bounds (up to log factors), showing that under certain assumptions, our algorithm is information-theoretically optimal. Our results represent what is, as far as we know, the only tractable algorithm guaranteeing successful recovery with tight bounds on recovery errors and sample complexity.

CUR Factorization via Discrete Empirical Interpolation

Mark Embree, Virginia Tech

The Discrete Empirical Interpolation Method (DEIM) of Chaturantabut and Sorensen (2010) has proved to be an essential technique in the reduction of large-scale nonlinear dynamical systems. In this talk we will illustrate with a variety of examples the performance of a CUR factorization based on an oblique projection that is inspired by the DEIM point selection algorithm.

This presentation is based on collaboration with Dan Sorensen.

Content-based search in 50TB of consumer-produced videos

Gerald Friedland, ICSI

Consumer-produced videos are the fastest-growing type of content on the Internet. YouTube alone claims that 100 hours of video are uploaded to their Website every minute. These videos provide a wealth of information about the universe. They consist of entertainment, instructions, personal records, and various aspects of life in general as it was when the video was recorded. Information is not only present in the centerpieces of these videos, but also in the incidental and background, visible and audible context. A major prerequisite to making social media videos usable for global-scale "field studies" is efficient and unbiased (e.g., keyword-independent) retrieval. More importantly, retrieval needs to go beyond simply finding objects to detecting more abstract concepts, such as "baby catching a ball" or "animal dancing to music."

In order to make videos accessible for research and algorithm development, ICSI has teamed up with Lawrence Livermore Lab and Yahoo! to release a corpus of 50TB (compressed) social media items (60M images, 700k videos, and annotations) that is freely available for research purposes. Research on such a large corpus of heterogeneous media-items requires the creation of distributed machine learning techniques and methods that exploit as many cues as possible from different modalities. This talk presents our work on the creation of the corpus, the underlying computing architecture used (Cray CS300) as well as our initial results on content-based search.

Optimal Shrinkage of Fast Singular Values

Matan Gavish, Stanford University

Randomized methods make it possible, in principle, to compute an approximate Singular Value Decomposition of huge low-rank matrices. While these Fast SVD methods have promising applications in analysis of massive data sets, they exhibit nontrivial noise sensitivity which may limit their practical usage. We present an asymptotic analysis of the sensitivity of fast singular values to white noise. We then propose a nonlinear shrinkage function for fast singular values that corrects for this noise sensitivity in an asymptotically optimal manner. Joint work with David Donoho (Stanford).

Localized Methods for Diffusions in Large Graphs

David Gleich, Purdue University

We consider stochastic transition matrices from large social and information networks. For these matrices, we will describe a few methods to evaluate PageRank and heat kernel diffusions. We will focus on two results. The first is an algorithmic anti-derivative of the PageRank method where we establish a precise relationship between PageRank and a minimum-cut problem. The second is a set of new methods to compute heat kernel diffusions that, when the networks have a power-law degree distribution, have provably sublinear runtime. We present further experimental evidence on social networks with billions of edges.

FAST-PPR: Scaling Personalized PageRank Estimation for Large Graphs

Ashish Goel, Stanford University

We propose a new algorithm, FAST-PPR, for the Significant-PageRank problem: given input nodes s, t in a directed graph and threshold δ, decide if the Personalized PageRank from s to t is at least δ. Existing algorithms for this problem have a running-time of $\Omega(1/\delta)$; this makes them unsuitable for use in large social networks for applications requiring values of $\delta = O(1/n)$. FAST-PPR is based on a bi-directional search and requires no preprocessing of the graph. It has a provable average running-time guarantee of $O(\sqrt{d/\delta})$ (where d is the average in-degree of the graph). We complement this result with an $\Omega(\sqrt{1/\delta})$ lower bound for Significant-PageRank, showing that the dependence on δ cannot be improved. We perform a detailed empirical study on numerous massive graphs showing that FAST-PPR dramatically outperforms existing algorithms. For example, with target nodes sampled according to popularity, on the 2010 Twitter graph with 1.5 billion edges, FAST-PPR has a 20 factor speedup over the state of the art. Furthermore, an enhanced version of FAST-PPR has a 160 factor speedup on the Twitter graph, and is at least 20 times faster on all our candidate graphs.

No Free Lunch for Stress Testers: Toward a Normative Theory of Scenario-Based Risk Assessment

Lisa Goldberg, UC Berkeley

Stress testing has become ubiquitous in the regulatory and risk management communities in the wake of the global financial crisis. In some quarters, stress tests have become the de facto approach to determining capital adequacy. However, some authors have argued that reliance on one or a few scenarios, developed ex ante by regulators or senior managers, may lead to inconsistent assessment of capital adequacy. In this paper, we formalize some of these arguments mathematically using the No Free Lunch Theorem (NFL) of computer science, which provide a framework for the development of financial stress tests. I will give a simple example

that illustrates the benefits of the NFL framework as it applies to stress testing.

Modeling Dynamics of Opinion Formation in Social Networks

Sreenivas Gollapudi, Microsoft Research

Our opinions and judgments are increasingly shaped by what we read on social media whether they be tweets and posts in social networks, blog posts, or review boards. These opinions could be about topics such as consumer products, politics, life style, or celebrities. Understanding how users in a network update opinions based on their neighbor's opinions, as well as what global opinion structure is implied when users iteratively update opinions, is important in the context of viral marketing and information dissemination, as well as targeting messages to users in the network. In this study, we consider the problem of modeling how users update opinions based on their neighbors' opinions. Further, we analyze online social network data as well as perform a set of online user studies based on the celebrated conformity experiments of Asch.

We show that existing and widely studied theoretical models do not explain the entire gamut of experimental observations we make. This leads us to posit a new, nuanced model that we term the BiasedVoterModel. We present preliminary theoretical and simulation results on the convergence and structure of opinions in the entire network when users iteratively update their respective opinions according to the BiasedVoterModel. We show that consensus and polarization of opinions arise naturally in this model under easy to interpret initial conditions on the network.

Large Scale Graph-Parallel Computation for Machine Learning: Applications and Systems

Ankur Dave and Joseph Gonzalez, UC Berkeley

From social networks to language modeling, the growing scale and importance of graph data has driven the development of graph computation frameworks such as Google Pregel, Apache Giraph, and GraphLab. These systems exploit specialized APIs and properties of graph computation to achieve orders-of-magnitude performance gains over more general data-parallel systems such as Hadoop MapReduce. In the first half of this talk we review several common data mining and machine learning applications in the context of graph algorithms (e.g. PageRank, community detection, recommender systems, and topic modeling). We then survey the common properties of these algorithms and how specialized graph frameworks exploit these properties in data partitioning and engine execution to achieve substantial performance gains.

In the second half of this talk we revisit the specialized graph-parallel systems through the lens of distributed join optimization in the context of Map-Reduce systems. We will show that the recent innovations in graph-parallel systems can be cast as data-partitioning and indexing enabling us to efficiently execute graph computation within a Map-Reduce framework and opening the opportunity to lift tables and graphs to the level of first-class composable views. Finally, we present GraphX, a distributed, fault-tolerant, and interactive system for large-scale graph analytics that is capable of efficiently expressing and executing graph-parallel algorithms while at the same time enabling users to switch between table and graph views of the same data without data-movement or duplication.

Myria: Scalable Analytics as a Service

Bill Howe, University of Washington

We are working to empower non-experts, especially in the sciences, to write and use data-parallel algorithms. To this end, we are building Myria, a web-based platform for scalable analytics and data-parallel programming. Myria's internal model of computation is the relational algebra extended with iteration. As a result, every program is inherently data-parallel, just as every query in a database is inherently data-parallel. But unlike databases, iteration is a first class concept, allowing us to express machine learning tasks, graph traversal tasks, and more. Programs can be expressed in a number of languages and can be executed on a number of execution environments, but we emphasize a particular language called MyriaL that supports both imperative and declarative styles and a particular execution engine called MyriaX that uses an in-memory column-oriented representation and asynchronous iteration. We deliver Myria over the web as a service, providing an editor, performance analysis tools, and catalog browsing features in a single environment. We find that this web-based "delivery vector" is critical in reaching non-experts: they are insulated from irrelevant effort technical work associated with installation, configuration, and resource management. I will describe the Myria system, give a demo, and present some preliminary evidence of uptake in the sciences.

Combinatorial optimization and sparse computation for large scale data mining

Dorit Hochbaum, UC Berkeley

We present here a novel model of data mining and machine learning that is based on combinatorial optimization, solving the optimization problem of "normalized cut prime" (NC'). NC' is closely related to the NP-hard problem of normalized cut, yet is polynomial time solvable. NC' is shown to be effective in image segmentation and in approximating the objective function of Normalized Cut as compared to the spectral technique. Its adaptation as a supervised classification data mining technique is called Supervised Normalized Cut (SNC). In a comparative study with the most commonly used data mining and machine learning methods, including Support Vector Machines (SVM), neural networks, PCA, logistic regression, SNC was shown to deliver highly accurate results within competitive run times.

In scaling SNC to large scale data sets, its use of pairwise similarities poses a challenge since the rate of growth of the

matrix of pairwise comparisons is quadratic in the size of the dataset. We describe a new approach called sparse computation that generates only the significant weights without ever generating the entire matrix of pairwise comparisons. The sparse computation approach runs in linear time in the number of non-zeros in the output and in that it contrasts with known "sparsification" approaches that require to generate, in advance, the full set of pairwise comparisons and thus take at least quadratic time. Sparse computation is applicable in any set-up where pairwise similarities are employed, and can be used to scale the spectral method and the k-nearest neighbors as well. The efficacy of sparse computation for SNC is manifested by its retention of accuracy, compared to the use of the fully dense matrix, while achieving a dramatic reduction in matrix density and thus run times.

Parts of the research presented are joint with: P. Baumann, E. Bertelli, C. Lyu and Y. Yang.

Leverage scores: Sensitivity and an App

Ilse Ipsen, North Carolina State University

Leverages scores were introduced in the 1970s for outlier detection in regression problems. About two decades later, Drineas, Mahoney *et al.* pioneered the use of leverage scores for importance sampling in randomized matrix algorithms.

We present bounds for the sensitivity of leverage scores to perturbations, and introduce a Matlab App, kappa_SQ, designed to facilitate analytical and empirical evaluation of leverage scores. With the help of a user-friendly interface, kappa_SQ makes it easy to compare a variety of probabilistic bounds (for condition numbers of sampled matrices) and to experiment with different sampling strategies.

This is joint work with Thomas Wentworth.

Influence sampling for generalized linear models

Jinzhu Jia, Peking University

We consider sampling problems for the estimates in generalized linear models. We sample the data points with probability proportional to an influence score. Based on this influence score sampling, the solution of the sub-problem is close to the solution to the original problem. This sampling scheme is similar to leverage score sampling for least square estimators which was used for fast computation of least squares. In this paper, we propose fast algorithms for generalized linear models.

Fast Network Community Detection by SCORE

Jiashun Jin, Carnegie Mellon University

Consider a network where the nodes split into K different communities. The community labels for the nodes are unknown and it is of major interest to estimate them (i.e., community detection). *Degree Corrected Block Model* (DCBM) is a popular network model. How to detect communities with the DCBM is an interesting problem, where the main challenge lies in the degree heterogeneity.

We propose **S**pectral **C**lustering **O**n **R**atios-of-**E**igenvectors (SCORE) as a new approach to community detection. Compared to existing spectral methods, the main innovation is to use the entry-wise ratios between the first a few leading eigenvectors for community detection. The central surprise is that the effect of degree heterogeneity is largely ancillary and can be effectively removed by taking such entry-wise ratios.

We have applied SCORE to the well-known web blogs data and the statistics co-author network data which we have collected very recently. We find that SCORE is competitive both in computation and in performance. On top of that, SCORE is conceptually simple and has the potential for extensions in various directions. Additionally, we have identified several interesting communities in statisticians, including what we call the "Object Bayesian community", "Theoretic Machine Learning Community", and the "Dimension Reduction Community".

We develop a theoretic framework where we show that under mild regularity conditions, SCORE stably yields consistent community detection. In the core of the analysis is the recent development on Random Matrix Theory (RMT), where the matrix-form Bernstein inequality is especially helpful.

Automatic discovery of cell types and microcircuitry from neural connectomics

Eric Jonas, UC Berkeley

New techniques produce massive data about neural connectivity, necessitating new analysis methods to discover the biological and computational basis of this connectivity. It has long been assumed that discovering the local patterns of microcircuitry is crucial to understanding neural function. Here we developed a nonparametric Bayesian technique that identifies neuron types and microcircuitry patterns in connectomics data. We show that the approach recovers known neuron types in the retina, reveals interesting structure in the nervous system of *c. elegans*, and automatically discovers the structure of microprocessors. Our approach extracts structural meaning from connectomics, enabling new approaches of deriving anatomical insights from these emerging datasets.

Exploring "forgotten" one-shot learning

Alek Kolcz, Twitter

The recent explosion in the volume of data available for mining and analysis spurred much research into scalable information processing and machine learning. Map-reduce (MR) frameworks such as Hadoop have drawn a lot of attention due to their solid presence in the modern data analytics pipelines. As many noticed, however, MR is not a great match for many machine learning algorithms, which require several passes through the data and requires synchronization or sequential processing. One way to mitigate these

problems is to move away from classic MR, e.g., by keeping all data in memory in between iterations. On the other hand, one could try to focus on algorithms that exploit what MR does best, which is essentially counting. One such algorithm is Naive Bayes, which has its limitations, but there exist algorithms from the early days of pattern recognition, which also share this property, support one-shot learning and can potentially handle large quantities of data in a MR framework. In this talk we will explore their behavior on a number of tasks and contrast them both with simpler and more complex alternatives to provide a perspective on their performance.

Spectral algorithms for graph mining and analysis

Yiannis Koutis, University of Puerto Rico

Spectral algorithms have long been recognized as a significant tool in the analysis and mining of large graphs. However, their adoption remains relatively limited because they are perceived as computationally demanding or non-robust. The talk addresses these two issues. We review recent algorithmic progress that enables the very fast computation of graph eigenvectors in time nearly linear to the size of the graph, making them very appealing from a computational point of view. We also review theoretical results that provide strong arguments in favor of spectral algorithms from a robustness point of view, showing that Cheeger inequalities are rather pessimistic for significant classes of graphs that include real-world networks. We further argue that we have only scratched the surface in understanding the power of spectral methods for graph analysis. We support this claim by discussing non-standard "generalized" graph eigenvectors, and showing that minor modifications of the default spectral partitioning methods have the potential to enhance their efficacy.

Locally-biased and semi-supervised eigenvectors

Michael Mahoney, ICSI and UC Berkeley

The second eigenvalue of a Laplacian matrix and its associated eigenvector are fundamental features of an undirected graph, and as such they have found widespread use in scientific computing, machine learning, and data analysis. In many applications, however, realistic graphs drawn from realistic data applications have several local regions of interest, and the second eigenvector will typically fail to provide information fine-tuned to each local region. For example, one might be interested in the clustering structure of a data graph near a pre-specified "seed set" of nodes; one might be interested in finding partitions in an image that are near a pre-specified "ground truth" set of pixels; or one might be interested in performing semi-supervised inference with popular diffusion-based machine learning algorithms.

Here, we provide a method to construct a locally-biased analogue of the second eigenvalue and its associated eigenvector, and we demonstrate both theoretically and empirically that this localized vector inherits many of the good properties that make the global second eigenvector so appealing in applications. The basic idea is to view the second eigenvector as the optimum of a constrained global quadratic optimization problem, and then add a locality constraint that has a natural geometric interpretation. The solution to this problem can be found with Laplacian linear equation solvers, and the method can be extended to find multiple locally-biased or semi-supervised eigenvectors and thus to construct locally-biased kernels.

In addition, there are nontrivial connections with strongly local spectral partitioning methods as well as implicit regularization methods that point toward a path to scale these methods up to much larger-scale machine learning problem instances. Indeed, this has been accomplished for the second locally-biased eigenvector—a push-based approximation algorithm for personalized PageRank that also implicitly solves an ℓ_1 regularized version of the PageRank objective exactly has been applied successfully to graphs with tens of millions of edges and billions of nodes—and the challenge is to extend these methods to compute multiple semi-supervised at that size scale.

Analyzing Big Graphs via Sketching and Streaming

Andrew McGregor, University of Massachusetts

Early work on data stream algorithms focused on estimating numerical statistics such as quantiles, frequency moments, and heavy hitters given limited memory and a single pass over the input. However, there is now a growing body of work on analyzing more structured data such as graphs. In this talk, we survey recent research in this area.

Dimensionality reduction via sparse matrices

Jelani Nelson, Harvard University

Dimensionality reduction techniques are used to obtain algorithmic speedup and storage savings in high-dimensional computational geometry, numerical linear algebra, compressed sensing, manifold learning, and clustering and several other machine learning problems. A common method for doing dimensionality reduction is to apply a random linear map to the input data (so-called "Johnson Lindenstrauss transforms"). In this talk we discuss ways of designing such linear maps which are extremely sparse but still provide provable guarantees, thus speeding up the time to do the dimensionality reduction.

Based on joint works with Jean Bourgain, Daniel Kane, and Huy Le Nguyen.

IPython: a language-independent framework for computation and data

Fernando Perez, UC Berkeley

As datasets become larger and our computational approaches to explore them increase in complexity, we need better tools to drive this exploratory process, extract insight and communicate it to others in a clear and reproducible manner.

I will describe the architecture of IPython, a system for interactive computing that supports most modern programming languages. The architecture of IPython defines an open protocol for rich clients to communicate with live computational kernels and to capture the results of this process into notebooks, accessed through a web browser. This design lets scientists drive the computation next to where the data is and in any language of their choice, while extracting insight into rich narratives that can then be used for publication, collaboration and education.

Computing stationary distribution, locally

Devavrat Shah, MIT

Computing the stationary distribution of a large finite or countably infinite state space Markov Chain has become central to many problems such as statistical inference and network analysis. Standard methods involve large matrix multiplications as in power iteration, or simulations of long random walks, as in Markov Chain Monte Carlo (MCMC). For both methods, the convergence rate is is difficult to determine for general Markov chains. Power iteration is costly, as it is global and involves computation at every state. In this work, we provide a novel local algorithm that answers whether a chosen state in a Markov chain has stationary probability larger than some $\Delta \in (0, 1)$, and outputs an estimate of the stationary probability for itself and other nearby states. Our algorithm runs in constant time with respect to the Markov chain, using information from a local neighborhood of the state on the graph induced by the Markov chain, which has constant size relative to the state space. The multiplicative error of the estimate is upper bounded by a function of the mixing properties of the Markov chain. Simulation results show Markov chains for which this method gives tight estimates.

libSkylark: Sketching-based Accelerated Numerical Linear Algebra and Machine Learning for Distributed-memory Systems

Vikas Sindhwani, IBM Research

In the first part of the talk, I will introduce libSkylark - a new open source software library for parallel matrix computations, designed to run on distributed memory machines. libSkylark provides a comprehensive set of high-performance sketching primitives for compressing big dense and sparse distributed matrices. These primitives are used to accelerate solvers for least squares regression, low-rank matrix approximation, and convex optimization arising in the context of large-scale machine learning tasks. In the second part of the talk, I will emphasize the need to be able to learn non-parametric models on big datasets. In particular, I will describe a combination of randomization techniques, Quasi-Monte Carlo approximations and specialized distributed ADMM solvers (also implemented in libSkylark) to scale up kernel methods. Strikingly, this results in kernel-based models that match state-of-the-art deep neural networks in terms of classification accuracy on a well known speech recognition benchmark.

Multi-reference Alignment: Estimating Group Transformations using Semidefinite Programming

Amit Singer, Princeton University

Let G be a group of transformations acting on an "object" space X. Suppose we have n measurements of the form $y_i = P * g_i.x + \epsilon_i (i = 1, \ldots, n)$, where x is a fixed but unknown element of X, $g_1, \ldots, g_n$ are unknown elements of G, P is a linear operator from the object space X to the measurement space Y, and ϵ_i are independent noise terms. We refer to the statistical problem of estimating the n group elements as the multi-reference alignment problem. For example, in the case of alignment of signals or images over the unit sphere in d dimensions, $G = O(d)$ (the group of $d \times d$ orthogonal matrices), X are band-limited functions over the sphere, and P is a sampling operator at m fixed points on the sphere. The challenge in obtaining the maximum likelihood estimator (MLE) for multi-reference alignment is that the parameter space is non-convex and is exponentially large in n. We consider a convex relaxation using semidefinite programming (SDP) that is numerically shown to be tight with high probability for a wide range of parameters, that is, the SDP recovers the MLE with high probability. Of particular interest is the case where P is a tomographic projection operator, which is the situation in cryo-electron microscopy. If time permits, we will also consider the application to the shape matching problem in computer graphics. Based on joined works with Afonso Bandeira, Moses Charikar, Yutong Cheng, and Andy Zhu.

Large Scale Machine Learning at Verizon

Ashok Srivastava, Verizon

This talk will cover aspects of the infrastructure and algorithmic developments that are being made at Verizon to support new products, services, and technologies based on large-scale machine learning. We will cover the development of this infrastructure and the system requirements to build and manage the data fabric necessary to support the massive data sets generated on the network.

Public Participation in International Security—Open Source Treaty Verification

Christopher Stubbs, Harvard University

Data collection and analysis in the context of international security has long been the purview of nation states. But the growing ubiquity of sensors (cameras in space as well as in personal devices, accelerometers, chemical sensors, etc.),

in conjunction with widespread data access, provides an opportunity for the public at large to take a direct role in both data generation and analysis. One aspect of this is treaty verification. A concrete example that illustrates both the opportunities and challenges of Public Technical Means (as distinct from National Technical Means, i.e., spy satellites) is a possible CO_2 treaty. I will use this example to illustrate the technical and engineering challenges and opportunities in this arena. I will also touch upon the moral, ethical and legal aspects of this domain. I will also outline the specific engineering problems that must be overcome to empower the public at large to play a larger role in ensuring their own safety and security.

Communication Cost in Big Data Processing

Dan Suciu, University of Washington

This talk discusses the theoretical complexity of database operations in massively distributed clusters. A query is computed in a sequence of super-steps that interleave computations with communications. The major performance parameter for complex queries is the number of communication steps, and the amount of data sent during each step. The talk will present some recent theoretical results on the tradeoff between the amount of communication and the number of communication steps required to compute a full conjunctive query. If the number of rounds is restricted to one, then we prove tight bounds on the amount of communication expressed in terms of the "fractional edge cover" of the hypergraph associated to the full conjunctive query. If the number of rounds is allowed to be some constant, then we give upper and lower bounds, which are almost tight for a restricted class of queries.

Joint work with Paul Beame and Paris Koutris.

Connected Components in MapReduce and Beyond

Sergei Vassilvitskii, Google

Computing connected components of a graph lies at the core of many data mining algorithms, and is a fundamental subroutine in graph clustering. This problem is well studied, yet many of the algorithms with good theoretical guarantees perform poorly in practice, especially when faced with graphs with billions of edges. We design improved algorithms based on traditional MapReduce architecture for large scale data analysis. We also explore the effect of augmenting MapReduce with a distributed hash table (DHT) service. These are the fastest algorithms that easily scale to graphs with hundreds of billions of edges.

The fall and rise of geometric centralities

Sebastiano Vigna, Universita degli Studi di Milano

"Centrality" is an umbrella name for a number of techniques that try to identify which nodes of a graph are more important. Research in the last years has focused on spectral centralities, which are based on the dominant eigenvector of some matrix derived from the adjacency matrix of a graph. In this talk we discuss recent results showing that geometric centralities, which are based on the distances between nodes, provide in many cases a more interesting signal. Moreover, newly discovered algorithms make it possible to estimate such centralities on very large graphs, paving the way for their usage in the analysis of massive data sets.

Optimal CUR Matrix Decompositions

David Woodruff, IBM Research

The CUR decomposition of an $m \times n$ matrix A finds an $m \times c$ matrix C with a small subset of $c < n$ columns of A, together with an $r \times n$ matrix R with a small subset of $r < m$ rows of A, as well as a $c \times r$ low rank matrix U such that the matrix CUR approximates the input matrix A, that is, $\|A - CUR\|_F \leq (1+\epsilon)\|A - A_k\|_F$, where $\|\cdot\|_F$ denotes the Frobenius norm and A_k is the best $m \times n$ matrix of rank k constructed via the SVD of A. We present input-sparsity-time and deterministic algorithms for constructing such a CUR matrix decomposition of A where $c = O(k/\epsilon)$ and $r = O(k/\epsilon)$ and $\text{rank}(U) = k$. Up to constant factors, our construction is simultaneously optimal in c, r, and $\text{rank}(U)$.

Joint work with Christos Boutsidis.

Dimension Independent Matrix Square using MapReduce

Reza Zadeh, Stanford University

We present a method to compute the singular values and vectors of an $m \times n$ tall and skinny ($m \gg n$) sparse matrix A without dependence on m, for very large m. In particular, we give a simple nonadaptive sampling scheme where the singular values of A are estimated within relative error with high probability. Our proven bounds focus on the MapReduce and Spark frameworks, which have become the standard tools for handling such large matrices that cannot be stored or even streamed through a single machine.

Large-Scale Numerical Computation Using a Data Flow Engine

Matei Zaharia, Databricks and MIT

As computer clusters scale up, data flow models such as MapReduce have emerged as a way to run fault-tolerant computations on commodity hardware. Unfortunately, MapReduce is limited in efficiency for many numerical algorithms. We show how new data flow engines, such as Apache Spark, enable much faster iterative and numerical computations, while keeping the scalability and fault-tolerance properties of MapReduce.

In this tutorial, we will begin with an overview of data flow computing models and the commodity cluster environment

in comparison with traditional HPC and message-passing environments. We will then introduce Spark and show how common numerical and machine learning algorithms have been implemented on it. We will cover both algorithmic ideas and a practical introduction to programming with Spark.

Spark started as a research project at UC Berkeley and is now open source at the Apache Software Foundation. It has a very fast-growing community, with over 180 developers and 50 companies contributing, and a quickly expanding numerical and machine learning library. It offers APIs in Java, Scala and Python.

Poster Abstracts

Affect Analysis in Large-Scale Online Text Communication Datasets

Nan-Chen Chen, University of Washington

In recent years, online text communication has become one of the most popular means of communication for people and has received attention from many areas of research, including psychology, social science, and computer science. Across these fields, there has been research on the role of affect in communication and its relation to people's behaviors and interactions. While previous studies have been trying to build computational models for detecting and analyzing affect, the chaotic essence of human communication makes it hard for programs to catch subtle emotional interactions in conversation. While one way to deal with this issue is to improve the models and algorithms, another way is to enable researchers in psychology and social science to leverage their knowledge about humans during the process of analysis. Thus, it becomes necessary to make large-scale datasets manageable for these researchers to process. In this project, we present a visual analytics tool to allow researchers to explore datasets efficiently. Given a dataset, the tool processes the text with existing machine-learning-based affect labeling software named Affect Labeller of Expressions (ALOE), detects emoticons, and presents the results in a time-series interactive visualization. The researchers can then explore the results in the visualization, make comparisons, and see text details. We show the applicability of this visual analytics tool by conducting an analysis of data collected during a collaborative project using Scratch, an online programming community.

Local Case-Control Sampling: Efficient Subsampling in Imbalanced Data Sets

William Fithian, Stanford University

For classification problems with significant class imbalance, subsampling can reduce computational costs at the price of inflated variance in estimating model parameters. We propose a method for subsampling efficiently for logistic regression by adjusting the class balance locally in feature space via an accept-reject scheme. Our method generalizes standard case-control sampling, using a pilot estimate to preferentially select examples whose responses are conditionally rare given their features. The biased subsampling is corrected by a post-hoc analytic adjustment to the parameters. The method is simple and requires one parallelizable scan over the full data set.

Unlike standard case-control sampling, which is inconsistent under model misspecification for the population coefficients, our method is consistent provided that the pilot estimate is. Although in severely imbalanced data sets the chosen subsample may contain a tiny fraction of the full data set, we show that under correct specification and with a consistent, independent pilot estimate, the subsampled estimate has exactly twice the asymptotic variance of the full-sample MLE. Moreover, this factor improves to $1+1/c$ if we multiply the baseline acceptance probabilities by $c > 1$ (and weight points with acceptance probability greater than 1), taking roughly $(1+c)/2$ times as many data points into the subsample. Experiments on simulated and real data show that our method can substantially outperform standard case-control subsampling.

Speeding Large Nonlinear Least Squares Problems by Near-Optimal Data Compression

Jeffrey Hokanson, MD Anderson

Many applications require fitting a nonlinear model to vast quantities of experimental data. Traditionally, reducing computation complexity was done by selecting a subset of data through either down sampling or truncation and then fitting the model to this subsample. If the distribution of noise in the data is known, we can quantify the loss of precision in the resulting model parameters using a multidimensional generalization of Fisher Efficiency. Under this metric, subsampling often performs poorly unless large portions of the data contain negligible information. Our work develops a more sophisticated approach that compresses the data onto a low dimension subspace by means of a dense rectangular matrix, allowing more efficient parameter estimates using fewer dimensions than subsampling. This approach yields substantial improvements for the exponential fitting problem with additive Gaussian white noise. By restricting the compression matrix to contain columns selected from a block diagonal Fourier matrix we bypass the computational expense of multiplying the compression matrix and model function by using closed form expressions. Further, we demonstrate that we can build near optimal compression spaces on the fly and that this approach becomes computationally advantageous for datasets as small as 1024 samples. Numerical experiments suggest that for a fixed efficiency, the necessary compression subspace dimension per parameter is a constant independent of data dimension; i.e., 95% efficiency requires 20 dimensions per parameter. This approach is readily extended to other large problems in system identification and is ripe for application to other nonlinear fitting problems.

Large Scale Analysis of Location Annotated Data

Juan M. Huerta, PlaceIQ

Location and movement data are strong indicators of human intention. The availability of very large volumes of location-annotated data combined with the growth and maturation of open frameworks for large-scale distributed data processing has provided the ideal conditions for large scale location analytics to emerge. Today, there are multiple important commercial applications based on location-annotated data;

one of the most important application domains and dominant source of location-annotated data are location-enabled mobile-device applications.

In this poster I will describe the nature of the data available, specifically the data observed in the mobile advertisement ecosystem, as well as the specialized frame of reference, as well as modeling and algorithmic approaches we have developed in order to create our location-enabled analytic platform. I will describe the problems that arise from location data idiosyncrasies and how we have addressed these in order to make our analysis tractable. I will also describe the applications where this location-enabled analysis is currently applied, as well as what we believe are the most promising directions of location analytics in the future.

The Generalized Haar-Walsh Transform

Jeff Irion, UC Davis

We present a new multiscale transform for data on graphs and networks which is a generalization of the classical Haar and Walsh-Hadamard Transforms. Using a recursive partitioning of the graph, the transform generates an overcomplete dictionary of piecewise-constant orthonormal bases. We adapt the best-basis selection algorithm to this setting, allowing us to choose a basis most suitable for the task at hand. We conclude with some results from a signal denoising experiment.

Analyzing Portfolios with Highly Skewed, Fat-Tailed Return Distributions

Nate Jensen, State Street

Multi-asset-class portfolios at most large financial institutions are characterized by skewed, fat-tailed return distributions. These large financial portfolios are massive with typically between 10,000 to 100,000 (and often in the millions of non-aggregated) positions or exposures with 20 to 200 attributes per exposure—resulting in difficult and resource-intensive stochastic calculations. Until recently, computational constraints have made it difficult to do more than aggregated, or index-based portfolio risk analyses. Most risk and optimization analyses regarding these massive portfolios can be significantly improved with greater calculation granularity i.e., focusing on "bottom-up" as opposed to "top-down" portfolio analyses. In particular, without a large number (usually greater than 1 million) of simulation iterations, contribution to tail-risk measures at the exposure level can become overly sensitive to slight changes in underlying latent-risk factors, which can lead to substantial estimation error. In order to drill in to these tail-risk measures and their drivers at a suitable level of granularity with an acceptable level of estimation error, more simulation iterations are needed requiring better computational tools to facilitate the requisite number of simulation iterations using available computing resources.

Recent advances in computational tools has created the opportunity to satisfactorily address challenges with these tail-risk measures greatly improving the usefulness of the calculations. Specifically, developments in the context of massively parallel processing have ushered in a new set of tools and possibilities for improving portfolio risk analyses and producing better optimized portfolios. This project looks at what is required for adopting such tools and considerations for effective implementation.

Frequency domain methods for functional time series

Lukasz Kidzinski, Universite libre de Bruxelles

In many fields of science data are sampled from processes that can be described most naturally as functional. Examples include growth curves, temperature curves, curves of financial transaction data and patterns of pollution data. Functional data analysis is concerned with the statistical analysis of such data. In this setting we often deal with temporal dependence, which occurs, for example, if the data consist of a continuous time process which has been cut into segments, e.g. days. We are then in the context of functional time series.

We investigate applications of frequency domain methods for functional time series. In particular we study dimension reduction and estimation of dynamic functional linear models. Our approach borrows ideas from the pioneering work of Brillinger in the vector setting. This presentation is based on joint projects with Siegfried Hormann, Marc Hallin and Piotr Kokoszka.

Relaxation Methods for Functions of Matrices Including the Exponential

Kyle Kloster, Purdue University

Functions such as the exponential, the resolvent, and p^{th} roots reveal important properties of networks. We present fast algorithms, related to the Gauss-Seidel linear solver, for approximating the action of these matrix functions on a vector. Assuming a power-law degree distribution on the underlying graph, we show the methods are sub-linear for these functions, and as a corollary we have that the action of such functions must be local.

X-Representativeness in Network and Vector Data

Milos Kudelka, VSB-Technical University of Ostrava

In the poster, a novel x-representativeness measure is presented. This measure is intended for weighted networks and it is also simply applicable to un-weighted networks and vector data. The x-representativeness is focused on measuring the local importance and takes into account the local density of the data and the nearest neighbors of individual nodes or data objects. An important feature of the proposed approach is its natural scalability resulting from the fact that the calculations are executed only in the surroundings of individual data objects. As a part of the poster, experiments

with large-scale real-world datasets are presented. The aim of these experiments is to show that the x-representativeness can be used to deterministically reduce the datasets to differently sized samples of representatives, while maintaining the topological properties of the original datasets.

Analyzing Like-Minded Communities for measuring the satisfaction and Loyalty of classified customers using Big Data Analytics

Ajay Kumar, Indian Institute of Technology Delhi

We present a framework that identifies like minded communities and analyze the social interaction patterns. The main goal of this research is measuring the customer satisfaction and clustering of customers into valuable potential and influential category using social network analysis and machine learning techniques utilizing big data analytics. In first phase using telecom data sets we are identifying the variables which are affecting the consumer's preferences and then we will identify the strongest and weakest service providers and compare their relative efficiency using DEA (Data Envelopment Analysis). In telecom social network graph we will make like minded communities based on questions like "Who contact Whom, How often and How long." and "Who influence Whom, How much or churn for which viral marketing can be very effective."

Global and Local Connectivity Analysis of Galactic Spectra

David Lawlor, Duke University

Over the past decade, the astronomical community has begun to utilize machine learning tools for understanding and interpreting the vast quantities of data collected by large-scale observational surveys such as the Sloan Digital Sky Survey (SDSS). We add to this literature by examining the connectivity of a large set of spectroscopic data through its embedding in certain diffusion spaces. We are able to interpret our embeddings in physical terms as well as to identify certain rare galaxy types and outliers due to errors in the preprocessing pipeline. We also discuss a local analogue of these diffusion embeddings that allows one to focus on a particular region of interest in the data, and demonstrate their utility in a downstream classification task.

Exact post-selection inference with the lasso

Jason Lee, Stanford University

We develop a framework for post-selection inference with the lasso. At the core of our framework is a result that characterizes the exact (non-asymptotic) distribution of linear combinations/contrasts of truncated normal random variables. This result allows us to (i) obtain honest confidence intervals for the selected coefficients that account for the selection procedure, and (ii) devise a test statistic that has an exact (non-asymptotic) uniform in $(0, 1)$ distribution when all relevant variables have been included in the model. The method is not specific to the lasso and can be also applied to marginal screening (t-test screening) and forward stepwise regression (orthogonal matching pursuit).

The Algorithmic Convergence Rate of Random Forests

Miles Lopes, UC Berkeley

When using majority vote as a prediction rule, it is natural to ask "How many votes are needed to obtain a reliable prediction?" In the context of ensemble classifiers, this question specifies a trade-off between computational cost and statistical performance. Namely, by a paying a larger computational price for more classifiers, the prediction error the ensemble tends to decrease and become more stable. Conversely, by sacrificing some statistical efficiency, it is possible to train the ensemble and make new predictions more quickly. In this paper, we quantify this trade-off for binary classification with Bagging or Random Forests, provided that the base classifiers obey some simplifying assumptions. To be specific, let the random variable Err_t denote the false positive rate of a randomly generated ensemble of t classifiers, trained on a fixed dataset. Then as t tends to infinity, we prove a central limit theorem for Err_t and obtain explicit formulas for its limiting mean and variance—as functions of t. As a consequence, it is possible to make a precise trade-off between the number of classifiers and the width of a confidence interval for Err_t.

Delving into R Analytics for Image Analysis

Talita Perciano, Lawrence Berkeley National Laboratory

Scientific imaging facilities have produced massive amounts of images and are eager to make sense of them. We have investigated technologies in R for image analysis, and we deployed a package for R Image Processing (RIPA) [1] recently. We are now testing and validating mechanisms to allow user-friendly interaction between R image analysis packages and R high performance packages, such as bigmemory [2]. To the best of our knowledge there are no current works probing solutions to deal with image processing at scale in R. This poster lays out the potential opportunities to leverage valuable packages, it illustrates some of the gaps we found in face of high-resolution datasets, and it points out some ideas in how to advance in the Big Image Data direction. This work is supported by CAMERA, the Center for Applied Mathematics for Energy Research Applications at LBNL.

[1] Perciano et al.
`http://cran.r-project.org/web/packages/ripa/index.html`

[2] Kane et al.
`http://cran.r-project.org/web/packages/bigmemory/index.html`

DeepWalk: Online Learning of Social Representations

Bryan Perozzi, Stony Brook University

We present DeepWalk, a novel approach for learning latent representations of vertices in a network. These latent representations encode social relations in a continuous vector space, which is easily exploited by statistical models. DeepWalk generalizes recent advancements in language modeling and unsupervised feature learning (or deep learning) from sequences of words to graphs. DeepWalk uses local information obtained from truncated random walks to learn latent representations by treating walks as the equivalent of sentences. We demonstrate DeepWalk's latent representations on several multi-label network classification tasks for social networks such as BlogCatalog, Flickr, and YouTube. Our results show that DeepWalk outperforms challenging baselines which are allowed a global view of the network, especially in the presence of missing information. DeepWalk's representations can provide F1 scores up to 10% higher than competing methods when labeled data is sparse. DeepWalk is also scalable. It is an online learning algorithm which builds useful incremental results, and is trivially parallelizable. These qualities make it suitable for a broad class of real world applications such as network classification, and anomaly detection.

(Joint work with Rami Al-Rfou and Steven Skiena)

Random Projections of Convex Programs

Mert Pilanci, UC Berkeley

Random projection is a classical technique for reducing storage, computational and communication costs. We analyze random projections of convex programs, in which the original optimization problem is approximated by the solution of a lower-dimensional problem. Such dimensionality reduction is essential in computation and communication limited settings, since the complexity of general convex programming can be quite high (e.g., cubic for quadratic programs, and substantially higher for semidefinite programs). In addition to computational savings, random projection is also useful for distributed computation, reducing memory usage, and has useful properties for privacy-sensitive optimization. We prove that the approximation ratio of this procedure can be bounded in terms of the geometry of constraint set. For a broad class of random projections, including those based on various sub-Gaussian distributions as well as randomized Hadamard and Fourier transforms, the data matrix defining the cost function can be projected down to the statistical dimension of the tangent cone of the constraints at the original solution, which is often substantially smaller than the original dimension. We illustrate consequences of our theory for various cases, including unconstrained and ℓ_1-constrained least squares, support vector machines, low-rank matrix estimation, classical and sparse principal component analysis and discuss implications on distributed and privacy-sensitive optimization and some connections with denoising and compressed sensing.

Interactive Streaming Algorithms for the Exact Frequency Moments

Ved Prakash, National University of Singapore

We consider a model for streaming algorithms where the data stream is processed by a third party, who provides the answer and a proof of correctness after the stream has ended. We can view the third party as the helper to whom we delegate the computations on the data stream. The proof provided by the third party is usually short and inexpensive to check. The space needed by the algorithm should be small.

The exact computation of the number of distinct elements is a fundamental problem in the study of data streaming algorithms. We denote the length of the data stream by n where each symbol is drawn from a universe of size m. Recently, we gave a streaming interactive algorithm with $\log m$ rounds for the exact computation of the number of distinct elements in a data stream. Our algorithm has complexity polylogarithmic in m and n, which is an exponential improvement from previous works.

Nearly Linear-Time Model-Based Compressive Sensing

Ludwig Schmidt, MIT

Compressive sensing is a method for recording a length-n signal x with (possibly noisy) linear measurements of the form $y = Ax$, where the $m \times n$ matrix A describes the measurement process. Seminal results in compressive sensing show that it is possible to recover a k-sparse signal x from $m = O(k \log n/k)$ measurements and that this is tight.

The model-based compressive sensing framework overcomes the lower bound and reduces the number of measurements to $m = O(k)$. This improvement is achieved by limiting the supports of x to a structured sparsity model, which is a subset of all $\binom{n}{k}$ possible k-sparse supports. This approach has led to measurement-efficient recovery schemes for a variety of signal models, including tree-sparsity and block-sparsity.

While model-based compressive sensing succeeds in reducing the number of measurements, the framework entails a computationally expensive recovery process. In particular, existing recovery algorithms perform multiple projections into the structured sparsity model. For several sparsity models, the best known model-projection algorithms run in time $\Omega(n^k)$, which can be too slow for large n and k.

Our work offers a way to overcome this obstacle by allowing the model-projection algorithms to be approximate. We illustrate our extension of the model-based compressive sensing framework with fast approximation algorithms for the tree-sparsity model. Our algorithms give the asymptotically fastest recovery scheme for the tree-sparsity model and run in nearly-linear time. Moreover, our algorithms are practical and show competitive performance on real data.

Joint work with Chinmay Hegde and Piotr Indyk. Based on papers accepted to SODA'14, ISIT'14, and ICALP'14.

Analyzing forest cover from airborne LiDAR elevation data using regional shape descriptors

Zehra Shah, UC Davis

Global warming is on the rise, and the importance of forest ecosystems for sustaining the planet has never before been felt more acutely. Management and monitoring of this resource is critical. This research presents a novel method of efficiently analyzing remotely sensed Light Detection and Ranging (LiDAR) data from a forest, with the goal of identifying patterns of interest in the forest structure as well as for estimating forest parameters like biomass. LiDAR technology provides a cost-effective means of collecting three-dimensional point data representing large study areas. The method described in this work focuses on analyzing regional point neighborhoods within the LiDAR data by summarizing the neighborhood into a two dimensional shape descriptor histogram. This representation, along with a suitable distance metric, allows for direct comparison between different regions within the data. Its applicability is demonstrated in two ways: clustering, to identify underlying forest structure, and regression, to estimate forest parameters like biomass. Results indicate this is a credible approach that can be explored further within the context of forest monitoring.

Screening genome-wide DNA methylation CpG sites via training and testing data utilizing surrogate variables

Xin Tong, University of South Carolina

Screening Cytosine-phosphate-Guanine dinucleotide (CpG) DNA methylation sites in association with single-nucleotide polymorphisms (SNPs), or covariate of interest, and/or their interactions is desired before performing more complicated analyses due to high dimensionality. It is possible the variation in methylation cannot be fully explained by SNPs and covariates of interest and thus important to account for variations introduced by other unknown factors. Furthermore, CpG sites screened from one data set may be inconsistent with those from another data set and equally important to improve the reproducibility of the selected CpG sites. A user-friendly R package, training-testing screening method (ttScreening), was developed to achieve these goals which provides users the flexibility of choosing different screening methods: proposed training and testing method, a method controlling false discovery rate (FDR), and a method controlling the significance level corrected by use of the Bonferroni method.

Generalized Low Rank Modeling

Madeleine Udell, Stanford University

Principal components analysis (PCA) is a well-known technique for approximating a data set represented by a matrix by a low rank matrix. Here, we extend the idea of PCA to handle arbitrary data sets consisting of numerical, Boolean, categorical, ordinal, and other data types. This framework encompasses many well known techniques in data analysis, such as nonnegative matrix factorization, matrix completion, sparse and robust PCA, k-means, k-SVD, and maximum margin matrix factorization. The method handles heterogeneous data sets, and leads to coherent schemes for compressing, denoising, and imputing missing entries across all data types simultaneously. It also admits a number of interesting interpretations of the low rank factors, which allow clustering of examples or of features. We propose a number of large scale and parallel algorithms for fitting generalized low rank models, which allows us to find low rank approximations to large heterogeneous datasets.

F3D: a Step Forward in Image Processing at Scale

Dani Ushizima, Lawrence Berkeley National Laboratory

F3D: a Step Forward in Image Processing at Scale using Free Software F3D plugin is designed to accelerate key image processing algorithms for fast filtering, enabling segmentation and analysis of high resolution image datasets, as those that comes from imaging instruments, such as X-ray micro-tomography and FIB-SEM. These imaging modalities require parallel-capable algorithms to accommodate tens of gigabytes in data sizes and to allow real-time feedback. F3D contains refactored 3D image processing algorithms, such as 3D grayscale-based mathematical morphology operators with variable structuring elements, non-linear filters and tools for high-level pipeline definition of image processing tasks, minimizing I/O. Our tool is implemented in Java and works as part of Fiji framework. Our filter kernels are implemented in OpenCL and they can be called from the Java code through JOCL. This work is supported by CAMERA, the Center for Applied Mathematics for Energy Research Applications at LBNL.

Fast updating algorithms for latent semantic indexing

Eugene Vecharynski, Lawrence Berkeley National Lab

We present a few algorithms for updating the approximate Singular Value Decomposition (SVD) in the context of information retrieval by Latent Semantic Indexing (LSI) methods. A unifying framework is considered which is based on Rayleigh-Ritz projection methods. First, a Rayleigh-Ritz approach for the SVD is discussed and it is then used to interpret the Zha–Simon algorithms [SIAM J. Scient. Comput. vol. 21 (1999), pp. 782-791]. This viewpoint leads to a few alternatives whose goal is to reduce computational cost and storage requirement by projection techniques that utilize subspaces of much smaller dimension. Numerical experiments show that the proposed algorithms yield accuracies comparable to those obtained from standard ones at a much lower computational cost.

Semiparametric Exponential Families for Heavy-Tailed Data

Stefan Wager, Stanford Univeristy

We propose a semiparametric method for estimating the mean of a heavy tailed population given a relatively small sample from that population and a larger sample from a related background population. We model the tail of the small sample as an exponential tilt of the better-observed large-sample tail using a robust sufficient statistic motivated by extreme value theory, and give both theoretical and empirical evidence that our method outperforms the sample mean and the Winsorized mean. If the small and large samples are drawn from regularly varying distributions with the same tail index, we show under mild additional conditions that our method achieves a better rate of convergence than the optimally Winsorized mean. Applying our method to both simulated data and a large controlled experiment conducted by an internet company, we exhibit substantial efficiency gains over competing methods.

Joint work with Will Fithian.

GOSUS: Grassmannian Online Subspace Updates with Structured-sparsity

Jia Xu, University of Wisconsin-Madison

We study the problem of online subspace learning in the context of sequential observations involving structured perturbations. In online subspace learning, the observations are an unknown mixture of two components presented to the model sequentially – the main effect which pertains to the subspace and a residual/error term. If no additional requirement is imposed on the residual, it often corresponds to noise terms in the signal which were unaccounted for by the main effect. To remedy this, one may impose "structural" contiguity, which has the intended effect of leveraging the secondary terms as a covariate that helps the estimation of the subspace itself, instead of merely serving as a noise residual. We show that the corresponding online estimation procedure can be written as an approximate optimization process on a Grassmannian. We propose an efficient numerical solution, GOSUS, Grassmannian Online Subspace Updates with Structured-sparsity, for this problem. GOSUS is expressive enough in modeling both homogeneous perturbations of the subspace and structural contiguities of outliers, and after certain manipulations, solvable via an alternating direction method of multipliers (ADMM). We evaluate the empirical performance of this algorithm on two problems of interest: online background subtraction and online multiple face tracking, and demonstrate that it achieves competitive performance with the state-of-the-art in near real time. Codes and data are available at: `http://pages.cs.wisc.edu/ jiaxu/projects/gosus/`.

Regime Change in Dynamic Correlation Matrices of High-Dimensional Financial Data

Joongyeub Yeo, Stanford University

We propose a new computational method to estimate the correlation structure of high-dimensional financial data. We use free random variable techniques and minimize the Kullback-Leibler distance between the theoretical spectral density of a model and the spectral density obtained from data. By doing this, we estimate factor model parameters with moving windows, which shows regime changes in residual spaces. The comparison between the parameters derived from our method of estimation and the mean-reversion time estimated from an Ornstein-Uhlenbeck model gives consistent results for regime-changes. We discuss applications of these techniques in algorithmic trading.

Privacy-Preserving Multi-Party Sorting of Large Data Sets

Mahdi Zamani, University of New Mexico

Just as Big Data lays out many promises, it lays out many questions and challenges when it comes to privacy. For example, how will that data get used and for what purpose, or who owns the data. To answer these challenges, we need new tools and technologies for private analysis, for anonymizing data, for running queries over private data, and for managing and sharing our own personal data. In this poster, we present a scalable privacy-preserving algorithm for multi-party sorting of large data sets. In multi-party sorting, a set of parties, each having a private input, want to sort their inputs without revealing the inputs to each other or any other authority. Such an algorithm is a critical component of many applications like privacy-preserving statistical analysis (e.g., private set intersection and private top-k queries), large-scale anonymous communication, and large-scale distributed intrusion detection. These applications often require algorithms that can efficiently perform statistical analysis over large sets of inputs provided by a large number of parties while the inputs are guaranteed to remain private to the providers. Using our algorithm, the number of bits sent and the number of operations performed by each party scales poly-logarithmically. To the best of our knowledge, the best algorithm for achieving the same goals scales linearly for each party. To compare the algorithms in practice, we conducted microbenchmarks to simulate a network with about 30 million parties. The results show that for sorting about 5 Gigabytes of data, our algorithm requires each party to send about 5 Kilobytes of data per sorted element while the best previous algorithm requires each party to send about 2 Megabytes of data per sorted element.

Prediction of heart failure hospitalizations with wearable activity monitors

Vadim Zipunnikov, Johns Hopkins Bloomberg School of Public Health

Sixty one patients of the Advanced Cardiac Care Center of Columbia University Medical Center diagnosed with congestive heart failure (CHF) wore Actical, an accelerometer device that continuously recorded physical activity over a seven to nine months period. Over the course of the study, twenty two subjects were either hospitalized or had an emergency room visit. We explore whether ambulatory monitoring of physical activity with accelerometers predicts clinically relevant adverse events in CHF patients. We introduce novel actigraphy summaries and identify prevalent pre-event patterns that explain roughly 60% of all CHF related episodes. These patterns can be detected as early as two months to three weeks prior to an episode.

Acknowledgements

Sponsors

The Organizers of MMDS 2014 and the MMDS Foundation would like to thank the following institutional sponsors for their generous support:

- **Ayasdi**
- **Databricks**
- **iCME,** the Institute for Computational and Mathematical Engineering, Stanford University
- **Department of Statistics,** at UC Berkeley
- **AMPLab,** the Algorithms, Machines and People Lab, at UC Berkeley
- **Simons Institute for the Theory of Computing,** at UC Berkeley