MMDS 2012:

Workshop on Algorithms for Modern Massive Data Sets

Cubberley Auditorium
Stanford University

July 10–13, 2012

The 2012 Workshop on Algorithms for Modern Massive Data Sets (MMDS 2012) will address algorithmic and statistical challenges in modern large-scale data analysis. The goals of MMDS 2012 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, Alex Shkolnik, Gunnar Carlsson, Petros Drineas*

Workshop Schedule

Tuesday, July 10, 2012: Data Analysis and Statistical Data Analysis

Time	Event	Location/Page
Registration & Opening		**Cubberley Auditorium**
8:00–9:45am	*Breakfast and registration*	
9:45–10:00am	Organizers *Welcome and opening remarks*	
First Session		**Cubberley Auditorium**
10:00–11:00am	Jiawei Han, University of Illinois, Urbana-Champaign *A Meta Path-Based Approach for Similarity Search and Mining of Heterogeneous Information Networks* *(Tutorial)*	pp. 9
11:00–11:30am	Alexander Gray, Georgia Institute of Technology *Faster Learning for Massive Datasets*	pp. 8
11:30–12:00pm	Christopher Re, University of Wisconsin, Madison *Hazy: Making Data-driven Statistical Applications Easier to Build and Maintain*	pp. 12
12:00–2:00pm	*Lunch* (on your own)	
Second Session		**Cubberley Auditorium**
2:00–3:00pm	Peter Bartlett, University of California, Berkeley, and QUT *Model Selection and Recent Results for Large Scale Problems* *(Tutorial)*	pp. 6
3:00–3:30pm	Noureddine El Karoui, University of California, Berkeley *On Robust Regression Estimators in High-dimension*	pp. 7
3:30–4:00pm	Jure Leskovec, Stanford University *Affiliation Network Models for Densely Overlapping Communities in Networks*	pp. 9
4:00–4:30pm	*Coffee break*	
Third Session		**Cubberley Auditorium**
4:30–5:00pm	Haesun Park, Georgia Institute of Technology *Nonnegative Matrix Factorizations for Clustering*	pp. 12
5:00–5:30pm	Fan Chung Graham, University of California, San Diego *Vectorized Laplacians for Dealing with High-dimensional Data Sets*	pp. 7
5:30–6:00pm	Joydeep Ghosh, University of Texas, Austin *Actionable Mining of Large, Multi-relational Data using Localized Predictive Models*	pp. 8
Evening Reception		
6:00–9:00pm	*Welcome Dinner Reception*	**New Guinea Garden**

Wednesday, July 11, 2012: Industrial and Scientific Applications

Time	Event	Location/Page
First Session		**Cubberley Auditorium**
9:00–10:00am	DJ Patil, Greylock Partners *When Algorithms Go Wrong: How Product Design Can Save Algorithmic Limitations (Tutorial)*	pp. 12
10:00–10:30am	Sean Fahey, Johns Hopkins Applied Physics Laboratory *Big Data and Analytics for National Security*	pp. 7
10:30–11:00am	*Coffee break*	
Second Session		**Cubberley Auditorium**
11:00–11:30am	Petros Drineas, Rensselaer Polytechnic Institute *Leverage Scores, the Column Subset Selection Problem, and Least-squares Problems*	pp. 7
11:30–12:00n	David Woodruff, IBM Research, Almaden *Low Rank Approximation and Regression in Input Sparsity Time*	pp. 14
12:00–12:30pm	Michael W. Mahoney, Stanford University *Implementing Randomized Matrix Algorithms in Parallel and Distributed Environments*	pp. 10
12:30–2:30pm	*Lunch* (on your own)	
Third Session		**Cubberley Auditorium**
2:30–3:30pm	Rick Stevens, Argonne National Laboratory *The Biological, Algorithmic and Computational Challenges of Systems Biology (Tutorial)*	pp. 13
3:30–4:00pm	Tiankai Tu, DE Shaw Research *Fault-Tolerant Parallel Analysis of Millisecond-Scale Molecular Dynamics Trajectories*	pp. 14
4:00–4:30pm	*Coffee break*	
Fourth Session		**Cubberley Auditorium**
4:30–5:00pm	Alexander Szalay, Johns Hopkins University *Current Statistical Challenges in Large Astronomical Surveys*	pp. 14
5:00–5:30pm	Joseph Richards, University of California, Berkeley *Astronomical Time Series Analysis for the Synoptic Survey Era*	pp. 12
5:30–6:00pm	Tony Cass, CERN *Data Handling for LHC: Plans and Reality*	pp. 7
Evening Reception		
6:00–9:00pm	*Dinner Reception and Poster Session*	**Cubberley Courtyard**

Thursday, July 12, 2012: Novel Algorithmic Approaches

Time	Event	Location/Page
First Session		**Cubberley Auditorium**
9:00–10:00am	Michael Mitzenmacher, Harvard University *Peeling Arguments: Invertible Bloom Lookup Tables and Biff Codes (Tutorial)*	pp. 11
10:00–10:30am	Frederic Chazal, INRIA *Detection and Approximation of Linear Structures in Metric Spaces*	pp. 7
10:30–11:00am	*Coffee break*	
Second Session		**Cubberley Auditorium**
11:00–11:30am	Ping Li, Cornell University *Probabilistic Hashing for Efficient Search and Learning on Massive Data*	pp. 10
11:30–12:00n	Ashish Goel, Stanford University *Real Time Social Search and Related Problems*	pp. 8
12:00–12:30pm	Andrew Goldberg, Microsoft Research, Silicon Valley *Hub Labels in Databases: Shortest Paths for the Masses*	pp. 8
12:30–2:30pm	*Lunch* (on your own)	
Third Session		**Cubberley Auditorium**
2:30–3:00pm	Theodore Johnson, AT&T Research Labs *Data Stream Warehousing*	pp. 9
3:00–3:30pm	Josh Wills, Cloudera, Inc *Experimenting at Scale*	pp. 14
3:30–4:00pm	Hang Li, Huawei *Large Scale Machine Learning for Query Document Matching in Web Search*	pp. 10
4:00–4:30pm	*Coffee break*	
Fourth Session		**Cubberley Auditorium**
4:30–4:50pm	Blair Sullivan, Oak Ridge National Labs *Branching Out: Quantifying Tree-like Structure in Complex Networks*	pp. 13
4:50–5:10pm	Mahdi Soltanolkotabi, Stanford University *A Geometric Analysis of Subspace Clustering with Outliers*	pp. 12
5:10–5:30pm	Bahman Bahmani, Stanford University *Scalable K-Means++*	pp. 6
5:30–6:00pm	Steve Bartel, Dropbox *Analytics at Dropbox*	pp. 6

Friday, July 13, 2012: Novel Matrix and Graph Methods

Time	Event	Location/Page
First Session		**Cubberley Auditorium**
9:00–10:00am	Yi Ma , Microsoft Research, Asia *The Pursuit of Low-dimensional Structures in High-dimensional Data (Tutorial)*	pp. 10
10:00–10:30am	Edoardo Airoldi, Harvard University *Graphlets Decomposition of a Weighted Network*	pp. 6
10:30–11:00am	*Coffee break*	
Second Session		**Cubberley Auditorium**
11:00–11:30pm	Yiannis Koutis, University of Puerto Rico, Rio Piedras *SDD Solvers: Bridging the Gap Between Theory and Practice*	pp. 9
11:30–12:00n	Art Owen, Stanford University *Bootstrapping r-fold Tensor Data*	pp. 11
12:00–12:30pm	Kamesh Madduri, Pennsylvania State University *Algorithms and Tools for Scalable Graph Analytics*	pp. 11
12:30–2:30pm	*Lunch* (on your own)	
Third Session		**Cubberley Auditorium**
2:30–3:00pm	Shaowei Lin, University of California, Berkeley *Studying Model Asymptotics with Singular Learning Theory*	pp. 10
3:00–3:30pm	David Bindel, Cornell University *Communities, Spectral Clustering, and Random Walks*	pp. 6
3:30–4:00pm	Ali Pinar, Sandia National Laboratories *The Block Two-Level Erdos-Renyi (BTER) Graph Model*	pp. 12
4:00–4:30pm	*Coffee break*	
Fourth Session		**Cubberley Auditorium**
4:30–5:00pm	Xiao-Li Meng, Harvard University *Preprocessing, Multiphase Inference, and Massive Data in Theory and Practice*	pp. 11
5:00–5:30pm	Alfred Hero, University of Michigan *Hub Discovery in Large Correlation Networks*	pp. 9
5:30–6:00pm	Dan Feldman, Massachusetts Institute of Technology *Google Your Life: Learning Sensors Data*	pp. 8

Poster Presentations: Wednesday, July 11, 2012

Event	Location/Page
Poster Session	**Cubberley Courtyard**
Santanu Das, UARC / Nasa Ames *Discovery of Safety Incidents in Airborne Systems*	pp. 15
S. Mukhopadhyay (Deep), Texas A & M University *Nonparametric Quantile based Large Scale Correlation Learning*	pp. 15
Joseph Gonzalez, Carnegie Mellon University *Large-Scale Graph-Parallel Computation on Natural Graphs*	pp. 15
Rishi Gupta, Stanford University *Sparse Recovery for Earth Mover Distance*	pp. 15
Toke Jansen Hansen, Technical University of Denmark *Semi-supervised Eigenvectors as a Fast Exploratory Tool for fMRI Analysis*	pp. 16
John Holodnak, North Carolina State University *Singular Values and Randomized Matrix Multiplication*	pp. 16
Wolfgang Kraske, Docomo Innovations *Scale-Free Graph Representation and Analysis with P-adic Lifting*	pp. 16
Aapo Kyrola, Carnegie Mellon University *Disk-based Large-scale Graph Computation*	pp. 16
Liangda Li, Georgia Institute of Technology *Scalar Block Coordinate Descent Algorithm for Non-Negative Matrix Factorization with Bregman Divergences*	pp. 16
Rahul Mazumder, Stanford University *Improved Matrix Completion via Warm-Started SVDs*	pp. 16
Xiangrui Meng, Stanford University *LSRN: A Parallel Iterative Solver for Strongly Over- or Under-Determined Systems*	pp. 17
Mario Morales, Simulmedia Inc. / Stanford University *Adaptive Quality Control for Streaming Datasets: Techniques Applied to Set-Top Box Data*	pp. 17
Bryan Perozzi, Stony Brook University *Cloud-based Large Scale Graph Clustering*	pp. 17
Vamsi K. Potluru, University of New Mexico *Parallel Coordinate Descent for Linear SVM*	pp. 17
Emile Richard, Ecole Normale Superieure de Cachan *Graph Denoising*	pp. 17
Guillermo Rodriguez-Cano, KTH - Royal Institute of Technology *Modeling Online Social Networks*	pp. 17
Mahdi Soltanolkotabi, Stanford University *A Geometric Analysis of Subspace Clustering with Outliers*	pp. 18
Ju Sun, Columbia University *Efficient Point-to-Subspace Query in ℓ^1 with Application to Robust Face Recognition*	pp. 18
Vishal Verma, University of North Carolina, Chapel Hill *Computing Geodesic Shortest Paths on Large Terrains Using Less Memory*	pp. 18
Thomas Wentworth, North Carolina State University *The Affects of Coherence on Randomized Least Squares Algorithms*	pp. 18

Talk Abstracts

Graphlets Decomposition of a Weighted Network

Edo Airoldi, Harvard University

We introduce the graphlet decomposition of a weighted network, which encodes a notion of social information based on social structure. We develop a scalable inference algorithm, which combines EM with Bron-Kerbosch in a novel fashion, for estimating the parameters of the model underlying graphlets using one network sample. We explore some theoretical properties of the graphlet decomposition, including computational complexity, redundancy and expected accuracy. We demonstrate graphlets on synthetic and real data. We analyze messaging patterns on Facebook and criminal associations in the 19th century.

Scalable K-Means++

Bahman Bahmani, Stanford University

Over half a century old and showing no signs of aging, k-means remains one of the most popular data processing algorithms. As is well-known, a proper initialization of k-means is crucial for obtaining a good final solution. The recently proposed k-means++ initialization algorithm achieves this, obtaining an initial set of centers that is provably close to the optimum solution. A major downside of k-means++ is its inherent sequential nature, which limits its applicability to massive data: one must make k passes over the data to find a good initial set of centers. In this talk, I will introduce an algorithm called k-means|| and show how it drastically reduces the number of passes needed to obtain, in parallel, a good initialization. I will prove that k-means|| obtains a nearly optimal solution after a logarithmic number of passes, and then show that in practice a constant number of passes suffices. Finally, I will present experimental results on real-world large-scale data demonstrating that k-means|| outperforms k-means++ in both sequential and parallel settings.

Model Selection and Recent Results for Large Scale Problems

Peter Bartlett, University of California, Berkeley, and QUT

The talk will first review model selection methods, which aim to automatically determine a model complexity for a prediction problem that best trades off approximation and estimation properties. The talk will focus on complexity penalization methods, which minimize the sum of an empirical risk term and a complexity penalty. It will review oracle inequalities, that is, results that show that the accuracy of a prediction rule produced by a model selection method is almost as good as the best bound that would have been achieved by any model complexity. The simplest results of this form are for complexity penalties based on bounds on the maximal deviation between empirical and true risks. There are many situations, however, where such complexity penalties give loose upper bounds on the risk, including regression with strongly convex loss functions, and pattern classification where the probability distribution satisfies a low noise condition. The talk will review model selection methods with smaller complexity penalties that satisfy oracle inequalities with faster rates in these situations.

The second part of the talk will cover recent results on large-scale model selection problems. In these problems, performance is limited by computational resources rather than sample size, and the classical oracle inequalities, which show a near-optimal trade-off between approximation error and estimation error for a given sample size, are no longer relevant. We formulate the problem of model selection under computational constraints: is it better to spend our computation gathering more data and estimating a simpler model, or gathering less data and estimating a more complex model? We present methods that satisfy computational oracle inequalities, that is, devoting all of our computational budget to the best model would not have led to a significant performance improvement. We also show that our methods and results extend to those cases where faster rates are possible.

Analytics at Dropbox

Steve Bartel, Dropbox, Inc.

In the internet era, information is distributed across many devices and stored in various locations. Dropbox tackles the task of organizing the information most important to you — your own files. We now have petabytes of data that can be aggregated to gather insights about data storage, access patterns, and more. The first half of this presentation will describe this data set as well as problems we would like to solve in the upcoming years. In the second half, we will open the floor to questions about our experiences, analytics, and Dropbox.

Communities, Spectral Clustering, and Random Walks

David Bindel, Cornell University

A community in a network is a subgraph that is unusually tightly connected. But what does it mean to be "unusually tightly connected"? In this talk, we describe three approaches to thinking about communities via block models, optimization problems, and random walks; and we show how each of these approaches leads naturally to eigenvectors and invariant subspaces. After discussing some standard spectral ideas for detecting disjoint communities, we discuss a new approach to finding overlapping communities by a constrained l1 minimization problem over an invariant subspace. Motivated by the dynamics of random walks, we then describe how these methods can be made more effective by stopping the usual iterations used to compute invariant

subspaces before they have fully converged. We illustrate our approach on several real and artificial test networks.

Data Handling for LHC: Plans and Reality

Tony Cass,CERN

The four major experiments at the Large Hadron Collider-ALICE, ATLAS, CMS and LHCb-together generate roughly 20PB of data each year. Although the experiments are hosted at CERN, data handling and analysis, just like the user community, is distributed worldwide. This presentation will describe the Worldwide LHC Computing Grid that was established to organise the computing effort, covering initial plans, current experience and likely future directions.

Detection and Approximation of Linear Structures in Metric Spaces

Frederic Chazal, INRIA

In many real-world applications data come as discrete metric spaces sampled around 1-dimensional linear structures (metric trees or graphs). Building on elementary tools of the theory of δ-hyperbolic spaces introduced by M. Gromov in 1987, we provide a framework in which one can measure how much data can be approximated by 1-dimensional metric structures and we provide algorithms to reconstruct such approximating structures. We will also illustrate the performances of our algorithm on various data sets.

This is a joint work with Jian Sun (Tsinghua University).

Vectorized Laplacians for Dealing with High-dimensional Data Sets

Fan Chung Graham, University of California, San Diego

We consider a generalization of graph Laplacian which acts on the space of functions which assign to each vertex a point in d-dimensional space. The eigenvalues of such connection Laplacian are useful for examining vibrational spectra of molecules as well as vector diffusion maps for analyzing high dimensional data. We will discuss algebraic, probabilistic and algorithmic methods in the study of the connection spectra. Edge ranking algorithms and Graph sparsification algorithms can be generalized to approximate and extract the global structure of information networks arising in signal and data processing.

Leverage Scores, the Column Subset Selection Problem, and Least-squares Problems

Petros Drineas, Rensselaer Polytechnic Institute

In this talk we will discuss the notion of leverage scores: a simple statistic that reveals columns (or rows) of a matrix that lie well within the subspace spanned by the top principal components. Sampling with respect to leverage scores has been used to speed up solving least-squares problems, as well as to approximately solve variants of the column subset selection problem. Finally, leverage scores are deeply connected to the so-called effective resistances of the edges of undirected, positively weighted graphs; effective resistances have been critical in approximately solving systems of linear equations with Laplacian matrices as inputs, in time nearly linear to the number of non-zero entries in the input matrix.

On Robust Regression Estimators in High-dimension

Noureddine El Karoui, University of California, Berkeley

In this talk I will describe the asymptotic behavior of robust regression estimators in the high-dimensional setting where n, the number of observations, is of the same order of magnitude as p, the number of predictors.

Based on this characterization, one can find the optimal objective function to use in high-dimensional regression, as a function of certain aspects of the statistical model.

Interestingly, the optimal objective depends on the asymptotic value of p/n. Generally, approaches based on maximum likelihood ideas are suboptimal.

Big Data and Analytics for National Security

Sean Fahey, Johns Hopkins Applied Physics Laboratory

Over the past few decades the United States has moved from a national security environment where data were relatively scarce and the focus of development efforts was on improving sensors and collection methods to the current environment where the government has access to large volumes of data from various sensors and sources and the development emphasis is on algorithms and methods to analyze the data. The ongoing revolution in big data storage and analysis led by the commercial and open source internet community offers new tools that the federal government can bring to bear against national security challenges. While many of the tools and approaches can be applied directly to the national security environment, some aspects of the national security environment pose challenges that will require modification of current tools and development efforts to meet government needs. The goal of this talk is to explore:

- What big data challenges are being faced in the national security environment?
- How federal government agencies using big data approaches for national security challenges?
- What is unique/different about the big data in the national security setting?
- What solutions are currently being sought by the government to address national security challenges with big data?

Google Your Life: Learning Sensors Data

Dan Feldman, Massachusetts Institute of Technology

Your smart-phone can collect real-time data about your life using its sensors. This includes where you are, what you see, and what you hear in any given moment. Tera-bytes of such data are written to smart-phones all over the world in every given second.

We are developing algorithms and systems to turn the sensors data into information in the form of a readable diary. The diary enables the automatic creation of an autobiography for the user. Additionally, the system provides real-time summaries of daily activities (e.g., which restaurants you go to and when, how you spend your leisure and work time, etc.) that can be shared with friends and family.

How can we turn noisy sensors signals into a searchable text? How to learn our users statistics while preserving their privacy? We answer these questions by maintaining a semantic compression of the streaming data (called sketch or core-set). This core-set represents the original data in the sense that running queries or fitting models on the semantic compression will yield a similar result when applied to the original data set, under natural assumptions (intuitively, that you are not doing a random walk on the planet).

Combining map-and-reduce techniques with our core-sets yields a system capable of compressing in parallel a stream of $O(n)$ samples using space and update time that is only $O(\log n)$. We present the current state of our application, experiments and theoretic results.

Actionable Mining of Large, Multi-relational Data using Localized Predictive Models

Joydeep Ghosh, University of Texas, Austin

Many large datasets associated with modern predictive data mining applications are quite complex and heterogeneous, possibly involving multiple relations, or exhibiting a dyadic nature with associated "side-information". For example, one may be interested in predicting the preferences of a large set of customers for a variety of products, given various properties of both customers and products, as well as past purchase history, a social network on the customers, and a conceptual hierarchy on the products. This talk will introduce a broad framework for effectively tackling such scenarios using a simultaneous problem decomposition and modeling strategy that can exploit the wide variety of information available.

Real Time Social Search and Related Problems

Ashish Goel, Stanford University

To answer search queries on a social network rich with user-generated content, it is desirable to give higher ranking to content that is closer to the individual issuing the query. To solve this problem exactly requires either prohibitively large pre-processing (e.g. maintaining a separate index of the entire corpus for every user) or prohibitively large effort for every query (e.g. a full breadth first search). In this talk, we will present an efficient approximate solution for this problem. Our solution requires maintaining a small number of indexes of the entire corpus. The pre-processing phase of our algorithm performs a small number (poly-logarithmic in N, where N is network size) of breadth-first search operations over the underlying social network, and is efficiently implementable offline. The indexing and querying operations can be performed efficiently (just two network calls in the typical scenario) in real-time on modern distributed stream processing platforms. The distance of the result returned is within an $O(\log N)$-factor of the closest result in the worst case, but our experimental evaluation shows that we typically find a closest result for realistic graphs. Time permitting, we will also describe an efficient distributed implementation of Locality Sensitive Hashing, which is another potentially useful tool in social search.

This is joint work with Bahman Bahmani, and appeared in WWW 2012.

Hub Labels in Databases: Shortest Paths for the Masses

Andrew Goldberg, Microsoft Research, Silicon Valley

We introduce HLDB, the first practical system that can answer exact spatial queries on continental road networks entirely within a database. HLDB is based on hub labels (HL), the fastest point-to-point algorithm for road networks, and its queries are implemented (quite naturally) in standard SQL. Within the database, HLDB can answer exact distance queries and retrieve full shortest-path descriptions in real time, even on networks with tens of millions of vertices. Moreover, the basic algorithm can be extended in a natural way (still in pure SQL) to answer much more sophisticated queries, such as finding the ten closest fast-food restaurants or minimizing the detour for stopping at a gas station on the way home. By taking advantage of special properties of HLDB (instead of using a distance oracle as a black box), even these sophisticated queries can be answered in real time. As databases are external memory by design, so is HLDB.

Joint work with Ittai Abraham, Daniel Delling, and Renato Werneck.

Faster Learning for Massive Datasets

Alexander Gray, Georgia Institute of Technology

I will describe new approaches for online learning and stochastic programming, which achieve both tighter theoretical bounds across the board and significant empirical gains over state-of-the-art approaches including stochastic gradient descent and mirror descent. I will then present a scheme for distributed online learning exhibiting first-of-a-kind theoretical and empirical gains. For nonlinear kernelized methods, kernel matrix multiplications and summations become a bottleneck. I will show fast algorithms which provably reduce computation times from quadratic to linear time, with

corresponding empirical runtime results, demonstrated on over 10,000 cores.

A Meta Path-Based Approach for Similarity Search and Mining of Heterogeneous Information Networks

Jiawei Han, University of Illinois, Urbana-Champaign

Objects in the real world are interconnected, often forming complex heterogeneous but structured or semi-structured information networks. Different from many studies on social networks where friendship networks or web page networks form homogeneous information networks, heterogeneous information networks reflect complex and structured relationships among multiple typed objects. For example, in a university network, objects of multiple types, such as students, professors, courses, departments, and multiple typed relationships, such as teach and advise are intertwined together, providing rich information.

We explore methodologies on mining such structured information networks and introduce meta path-based approach for similarity search and mining of heterogeneous information networks. We show that structured information networks are informative, and link analysis on such networks becomes powerful at uncovering critical knowledge hidden in large networks. The tutorial will present a few examples with this new methodology and suggest some promising research directions.

Hub Discovery in Large Correlation Networks

Alfred Hero, University of Michigan

One of the most important problems in large scale inference problems is the identification of variables that are highly dependent on several other variables. When dependency is measured by partial correlations these variables identify those rows of the partial correlation matrix that have several entries with large magnitudes; i.e., hubs in the associated partial correlation graph. This talk will present theory and algorithms for discovering such hubs from a few observations of these variables. The theory is applied to discovering hubs in gene expression networks.

This is joint work with Bala Rajaratnam.

Data Stream Warehousing

Theodore Johnson, AT&T Research Labs

Data stream processing is a well-studied area that has led to many commercial offerings. Stream processing systems typically store only a small history of a stream, and provide services such as real-time alerting and visualization, and data reduction for downstream processing. However many applications also require that long term (e.g. 2 year) histories as well as real-time data be provided to analysts. A stream warehouse combines aspects of data stream processing (continual data ingest) with data warehousing (long-term storage and materialized views). I will discuss special issues and opportunities that arise in stream warehousing, and stream warehouse applications within AT&T Labs - Research.

SDD Solvers: Bridging the Gap Between Theory and Practice

Yiannis Koutis, University of Puerto Rico, Rio Piedras

Symmetric diagonally dominant (SDD) linear system solvers are central in what has been described as an "incipient revolution in the theory of graph algorithms." Indeed, fast SDD solvers are the key subroutine of the asymptotically fastest known approximation algorithms for several problems, including image denoising and the max-flow problem.

But will this recent theoretical progress extend to practice? The answer obviously depends on the success of SDD solver implementations. In the first part of this talk I will briefly review "combinatorial multigrid" (CMG), a graph decomposition-based SDD solver, which is indeed able to solve very quickly massive linear systems. Its key limitation is sparsity: the underlying approach is not known to extend to non-sparse systems. In the second part of the talk I will discuss our recent efforts—partly motivated by the desire to extend the usability of CMG—to design practical spectral sparsification algorithms. A by-product of this work is an essentially $O(m)$ time algorithm for solving slightly dense SDD systems, improving upon the previous $O(m \log m)$ time algorithm.

Affiliation Network Models for Densely Overlapping Communities in Networks

Jure Leskovec, Stanford University

Networks are a general language for describing social, technological and biological systems. Nodes in such networks organize into densely linked and overlapping clusters that correspond to communities in social networks, functionally related proteins in biological networks, or topically related webpages in information networks. Identifying such clusters is crucial to the understanding of the structural and functional roles of networks.

Our work stems from an intuitive observation that the probability of an edge between a pair of nodes increases with the number of shared cluster affiliations, which means that cluster overlaps are more densely connected that their non-overlapping parts. We discuss a model-based network community detection method that builds on bipartite node-community affiliation networks and can detect dense cluster overlaps. The approach allows for modeling overlapping, non-overlapping as well as hierarchically nested clusters. We develop a set of model inference techniques and accurately identify clusters in networks ranging from biological protein-protein interaction networks to social, collaboration and information networks. The results show imply that while networks organize into overlapping communities, globally networks also exhibit a nested core-periphery structure, which

arises as a consequence of overlapping parts of communities being more densely connected.

Large Scale Machine Learning for Query Document Matching in Web Search

Hang Li, Huawei Labs

In this talk, I will introduce our recent work on large scale machine learning for query document matching in web search. I will start my talk by pointing out that query-document mismatch is one of the biggest challenges in web search. I will then describe the necessity of employing machine learning techniques to overcome the challenge. Next, I will specifically explain two approaches, namely topic modeling and projection into latent space. In the former approach queries and documents are matched in the topic space learned from documents, and in the latter approach queries and documents are matched in the latent space learned from click-through data. Both approaches need to be applied to very large scale data sets. I will describe in details about the learning techniques which we have developed for enhancing the scalability.

Joint work with Wei Wu, Quan Wang, Jun Xu, and Zhengdong Lv.

Probabilistic Hashing for Efficient Search and Learning on Massive Data

Ping Li, Cornell University

Modern applications in the context of search often encounter massive high-dimensional binary data. For example, using n-gram representations, documents are often parsed to be binary (0/1) vectors in billions, trillions, quadrillions, or even higher dimensions. How to efficiently store, transmit, and search these data is a very interesting research topic with numerous applications in the industry. This talk focuses on a probabilistic hashing method named b-bit minwise hashing, which stores only the lowest b bits (for small b) of each hashed value after applying the standard minwise hashing procedure. Theoretically, it can be shown that b-bit minwise hashing improves minwise hashing at least by 21.3-fold when the threshold similarity (i.e., resemblance) is 0.5. More interestingly, we realize that b-bit minwise hashing can be seamlessly integrated with (i) logistic regression and SVM to solve extremely large predictive learning problems, and (ii) locality sensitive hashing (LSH) for sub-linear time near neighbor search. Extensive experiments will be presented.

Studying Model Asymptotics with Singular Learning Theory

Shaowei Lin, University of California, Berkeley

Singular statistical models occur frequently in machine learning and computational biology. An important problem in the learning theory of singular models is determining their asymptotic behavior for massive data sets. In this talk, we give a brief introduction to Sumio Watanabe's Singular Learning Theory, as outlined in his book "Algebraic Geometry and Statistical Learning Theory." We will also explore the rich algebraic geometry and combinatorics that arise from studying the asymptotics of Bayesian integrals.

The Pursuit of Low-dimensional Structures in High-dimensional Data

Yi Ma, Microsoft Research, Asia

In this talk, we will discuss a new class of models and techniques that can effectively model and extract rich low-dimensional structures in high-dimensional data such as images and videos, despite nonlinear transformation, gross corruption, or severely compressed measurements. This work leverages recent advancements in convex optimization for recovering low-rank or sparse signals that provide both strong theoretical guarantees and efficient and scalable algorithms for solving such high-dimensional combinatorial problems. These results and tools actually generalize to a large family of low-complexity structures whose associated regularizers are decomposable. We illustrate how these new mathematical models and tools could bring disruptive changes to solutions to many challenging tasks in computer vision, image processing, and pattern recognition. We will also illustrate some emerging applications of these tools to other data types such as web documents, image tags, microarray data, audio/music analysis, and graphical models.

This is joint work with John Wright of Columbia, Emmanuel Candes of Stanford, Zhouchen Lin of Peking University, and my students Zhengdong Zhang, Xiao Liang of Tsinghua University, Arvind Ganesh, Zihan Zhou, Kerui Min and Hossein Mobahi of UIUC.

Implementing Randomized Matrix Algorithms in Parallel and Distributed Environments

Michael W. Mahoney, Stanford University

Motivated by problems in large-scale data analysis, randomized algorithms for matrix problems such as regression and low-rank matrix approximation have been the focus of a great deal of attention in recent years. These algorithms exploit novel random sampling and random projection methods; and implementations of these algorithms have already proven superior to traditional state-of-the-art algorithms, as implemented in Lapack and high-quality scientific computing software, for moderately-large problems stored in RAM on a single machine. Here, we describe the extension of these methods to computing high-precision solutions in parallel and distributed environments that are more common in very large-scale data analysis applications.

In particular, we consider both the Least Squares Approximation problem and the Least Absolute Deviation problem, and we develop and implement randomized algorithms that take advantage of modern computer architectures in order to achieve improved communication profiles. Our

iterative least-squares solver, LSRN, is competitive with state-of-the-art implementations on moderately-large problems; and, when coupled with the Chebyshev semi-iterative method, scales well for solving large problems on clusters that have high communication costs such as on an Amazon Elastic Compute Cloud cluster. Our iterative least-absolute-deviations solver is based on fast ellipsoidal rounding, random sampling, and interior-point cutting-plane methods; and we demonstrate significant improvements over traditional algorithms on MapReduce. In addition, this algorithm can also be extended to solve more general convex problems on MapReduce.

Algorithms and Tools for Scalable Graph Analytics

Kamesh Madduri, Pennsylvania State University

Graph-theoretic abstractions are at the core of data-intensive problems arising in social and technological network analysis, scientific computing, and security applications. Due to their large memory footprint, fine-grained computational granularity, and low degrees of spatial locality, massive graph problems pose serious challenges on current parallel machines. In this talk, we present novel data-centric algorithms and methods for enabling large-scale network analysis. Our parallel implementations on leading supercomputers and massively multithreaded platforms achieve significant parallel speedup for traversal, connectivity, and clustering problems on graph instances with billions of vertices and edges. We will present recent results related to three data-intensive problems: community identification in social network analysis, de Bruijn graph-based genome assembly, and pattern matching queries on terascale semantic web data.

Preprocessing, Multiphase Inference, and Massive Data in Theory and Practice

Xiao-Li Meng, Harvard University

Analyses of massive datasets are often built upon preprocessed data (e.g., microarray experiments) or constitute a form of preprocessing themselves (e.g., dimensionality reduction or feature extraction). Such preprocessing is often vital, but it is rife with subtleties and pitfalls. When such steps are taken, the data analysis effectively becomes a collaborative endeavor by all parties involved in data collection, preprocessing and curation, and downstream inference. Each party does not and often cannot have a perfect understanding of the entire phenomenon at hand; the final results will inevitably contain some combination of their judgments, and some preprocessing can irreversibly destroy information from the raw data. Furthermore, even if each party has done their absolute best given the information and resources available to them, the final result may still fall short of the best possible when it is evaluated in the traditional single-phase framework due to the problem of uncongeniality (Meng, 1994, Statistical Science). We therefore need a core of statistical theory for such multiphase inference problems. This talk presents some building blocks for a multiphase theoretical framework, illustrated by some applied examples. Our work highlights the importance of providing information beyond optimal estimators for downstream analyses; however, such information need not correspond to sufficient statistics, even in theory.

This is joint work with Alex Blocker.

Peeling Arguments: Invertible Bloom Lookup Tables and Biff Codes

Michael Mitzenmacher, Harvard University

The analysis of several recent algorithms and data structures make use of "peeling arguments": repeatedly find an element you can process (greedily) and keep going. We start by examining some known examples of peeling arguments in coding and hashing.

We then present Invertible Bloom Lookup Tables (IBLTs), a version of the Bloom filter data structure that supports not only the insertion, deletion, and lookup of key-value pairs, but also allows a complete listing of its contents with high probability, as long the number of key-value pairs is below a designed threshold. Our analysis of IBLTs depends on peeling arguments. We discuss some applications in network and databases, and provide a detailed example of how IBLTs can be used to develop what we call Biff Codes, a very fast and simple error-correcting coding scheme for large data sets.

The talk includes joint work with Michael Goodrich and George Varghese.

Bootstrapping r-fold Tensor Data

Art Owen, Stanford University

The famous Netflix data is a sparsely sampled table with rows for customers and columns for movies (or vice versa). Both movies and rows are naturally modeled as random effects. Bootstrapping such data is problematic: no proper bootstrap can exist, according to a theorem of Peter McCullagh. Resampling rows and columns independently is effective though slightly conservative.

Computerized data gathering frequently produces data sets with three-way or even higher order data tables. We present a bootstrap for such tensor valued data. Our version uses independent weights instead of multinomial ones. It remains mildly conservative. Poisson weights are close to the original bootstrap, but binary weights have computational and statistical advantages. Under certain conditions a single bootstrap replicate suffices to give a variance estimate. We apply our method to compare the length of comments made by Facebook users in the US and the UK.

This work is joint with Dean Eckles, Facebook.

Nonnegative Matrix Factorizations for Clustering

Haesun Park, Georgia Institute of Technology

Nonnegative matrix factorization (NMF) provides two nonnegative lower rank factors whose product approximates a nonnegative matrix. We show that many of the promising algorithms for NMF can be explained using a framework of block coordinate descent method. We then discuss some capabilities and shortcomings of NMF as a clustering method and propose Symmetric NMF (SymNMF), which computes a lower rank nonnegative factor $H = \text{argmin}||S - HH^T||_F$ that approximates a similarity matrix S, as a general method for graph clustering. We propose efficient algorithms for SymNMF and give an intuitive explanation of why SymNMF naturally captures the cluster structure embedded in a graph representation. Our experiments on text corpus, object recognition, and image segmentation demonstrate substantially enhanced clustering results from SymNMF over spectral clustering and NMF.

When Algorithms Go Wrong: How Product Design Can Save Algorithmic Limitations

DJ Patil, Greylock Partners

All technologies and algorithms have limitations. Combined with the broad range of use cases that can take place on consumer internet sites, this can lead to disastrous product experiences. In this talk I'll walk through some lessons on how to minimize the impact from technical limitations.

The Block Two-Level Erdos-Renyi (BTER) Graph Model

Ali Pinar, Sandia National Laboratories

Despite their growing importance, our understanding of graphs is still limited. Most notably we do not have models that can characterize these graphs. Such models are crucial, since they can provide insights into generative processes, properties, and evolution of these graphs; enable anomaly detection; and guide statistical sampling. Moreover, due to limitations in sharing real graphs, generative models are critical for benchmarking high-performance computing systems and developing better algorithms at various scales and properties. We propose the Block Two-level Erdos-Renyi (BTER) graph model, which is motivated by two basic observations about real-world graphs: they have skewed degree distributions and high clustering coefficients. Our experiments showed that graphs generated by BTER are strikingly similar to the originals, making the BTER model a strong candidate for benchmarking high-performance computing platforms.

Joint work with C. Seshadhri and Tamara G. Kolda.

Hazy: Making Data-driven Statistical Applications Easier to Build and Maintain

Christopher Re, University of Wisconsin, Madison

The main question driving my group's research is: how does one deploy statistical data-analysis tools to enhance data-driven systems? Our goal is to find abstractions that one needs to deploy and maintain such systems. In this talk, I describe my group's attack on this question by building a diverse set of statistical-based data-driven applications: a system whose goal is to read the Web and answer complex questions, a muon detector in collaboration with a neutrino telescope called IceCube, and a social-science applications involving rich content (OCR and speech data). Even in this diverse set, my group has found common abstractions that we are exploiting to build and to maintain statistical data analysis systems.

Astronomical Time Series Analysis for the Synoptic Survey Era

Joseph Richards, University of California, Berkeley

We have entered the Synoptic Survey Era of observational astronomy, where data rates are quickly reaching several terabytes per night. A growing army of telescopes monitor, nightly, the luminosities of millions of objects, and soon that number will reach upwards of a billion. Real-time analysis of these data is critical to determine which objects and events require timely observations with expensive follow-up resources. To maximize the scientific returns from these massive projects, sophisticated machine-learning tools must be used. Our group has been on the cutting edge of the methodological and algorithmic development for time-domain astronomical data analysis. I will describe several problems in which we have made great strides, including real-time ML detection and classification of transient events, photometric supernova typing, and probabilistic classification of variable stars from long-baseline time series. In this talk I will outline some of the methodology we have developed for these problems, including the use of manifold learning for time-series feature extraction, active learning to overcome sample-selection biases, and semi-supervised learning to detect anomalies in data streams. I will also describe our newly released Machine-learned ASAS Classification Catalog (MACC, www.bigmacc.info) and discuss the future of astronomical source catalogs.

A Geometric Analysis of Subspace Clustering with Outliers

Mahdi Soltanolkotabi, Stanford University

One of the most fundamental steps in data analysis and dimensionality reduction consists of approximating a given dataset by a single low-dimensional subspace, which is classically achieved via Principal Component Analysis (PCA). However, in many applications, the data often lie near a union of low-dimensional subspaces, reflecting the multiple categories or classes a set of observations may belong to. In this talk we discuss the problem of clustering a collection of

unlabeled data points assumed to lie near a union of lower dimensional planes. As is common in computer vision or unsupervised learning applications, we do not know in advance how many subspaces there are nor do we have any information about their dimensions. We present a novel geometric analysis of an algorithm named sparse subspace clustering (SSC) [Elhamifar and Vidal 2009], which significantly broadens the range of problems where it is provably effective. For instance, we show that SSC can recover multiple subspaces, each of dimension comparable to the ambient dimension. We also show that SSC can correctly cluster data points even when the subspaces of interest intersect. Further, we develop an extension of SSC that succeeds when the data set is corrupted with possibly overwhelmingly many outliers. Underlying our analysis are clear geometric insights, which may bear on other sparse recovery problems. We will demonstrate the effectiveness of these methods by various numerical studies.

The Biological, Algorithmic and Computational Challenges of Systems Biology

Rick Stevens, Argonne National Laboratory

Breakthroughs in biology are being powered by advanced computing capabilities that enable researchers to manipulate, explore and compare massive datasets. The speed at which any given scientific domain advances will soon depend on how well its researchers collaborate with one another, and with computer scientists and mathematicians. The ability to rapidly analyze new data, relate it to existing knowledge and derive new predictions to inform the next step is the key to accelerating scientific discovery.

In this talk I will discuss three things. First I'll provide an overview of the new approach to biological research known as "systems biology" that addresses the problem of reverse engineering complex biological systems and using that information to build integrative models of cells, pathways and networks and how we use these models to say things about the systems behavior of organisms and communities. I'll then talk about a specific project to build a "systems knowledge base" known as the "KBase" project for the Department of Energy that is aimed at advancing predictive systems biology in microbes, microbial communities and plants. The KBase project is integrating data from many existing sources, building tools and services that will support complex workflows enabling modeling of microbes, reconciling experimental data with computational predictions, and providing a large-number of computational services that go beyond existing integrated biological databases. KBase will be deployed on a purpose-built infrastructure spanning four laboratories that collectively house multiple petabytes of data, and that will support scalable computing resources on both cloud and cluster environments. End users will be able to access many thousands of public genomes and related datasets for microbes. They will also gain access to tens of thousands of metagenomic samples and dozens of plant genomes and phenotype datasets. In addition to providing web and programmatic interfaces to these data, the KBase will enable users to upload their own private data and virtually integrate it with the public datasets for comparative analysis and development of models. The KBase is aiming to enable collaborative workflows and multiple ways of sharing. Finally, I'll dive down into a couple of interesting examples addressing novel data organization and new algorithms that will be required to more effectively analyze the large volumes of data that are being assembled to support genome scale biological research. I'll touch on alignment-free methods for sequence comparison, computing on compressed representations of closely related genomes and the problem of representing genome assemblies as graphs rather than linear sequences.

Branching Out: Quantifying Tree-like Structure in Complex Networks

Blair Sullivan, Oak Ridge National Labs

A significant challenge in analyzing large complex networks has been understanding the "intermediate-scale structure"—those properties not captured by metrics which are very local (e.g., clustering coefficient) or very global (e.g., degree distribution). It is often this structure which governs the dynamic evolution of the network and the behavior of diffusion-like processes on it. Although there is a large body of empirical evidence suggesting that complex networks are often "tree-like" at intermediate to large size-scales (e.g., work of Boguna et al. in physics, Kleinberg on internet routing, and Chung and Lu on power-law graphs), it remains a challenge to quantify and take algorithmic advantage of this structure in data analysis.

We present preliminary computational results showing the successes and failings of existing metric-based and cut-based measures of tree-likeness on real world data and discuss impact on several downstream applications. This will include new comprehensive calculations of the Gromov four-point hyperbolicity profile of networks from SNAP/Facebook (as well as synthetic networks for comparison). These profiles allow easy distinction between planar-like networks such as the power grid, random homogeneous expanders, and more typical social/power-law networks. Additionally, we provide data on the performance of several popular heuristics for finding low-width tree-decompositions and contrast with the hyperbolicity results. Finally, we will discuss new analysis using the k-core decomposition, a measure of connectivity which is calculated by recursively "peeling" off low degree nodes from the graph, to infer various structural differences. Finally, we observe that the depth of decomposition and the connections between the layers give insight into the behavior of other algorithms (e.g., spectral diffusions). A short discussion of downstream applications such as improved algorithms for noisy covariance estimation and sparse random projections will be included as time allows.

Current Statistical Challenges in Large Astronomical Surveys

Alex Szalay, Johns Hopkins University

The talk will present some of the challenges arising in today's large astronomical surveys. In particular, we will discuss various techniques of estimating galaxy distances from their broad-band colors, ranging from template fitting to base-learners, KNN and random forests. Galaxy spectra also represent interesting challenges, and we will show how heuristic traditional astronomical estimators can be objectively derived from statistical considerations.

Fault-Tolerant Parallel Analysis of Millisecond-Scale Molecular Dynamics Trajectories

Tiankai Tu, DE Shaw Research

Anton is a special-purpose supercomputer that has enabled the first molecular dynamics (MD) simulations of proteins on millisecond timescales. The unique characteristics of the resulting trajectories-a large number (millions) of relatively small (megabyte-sized) frames-have posed significant data analysis challenges when it comes to algorithm design and implementation, data storage and retrieval, and overall performance and scalability.

Our first attempt to support the analysis of very long MD trajectories on commodity clusters led to the development and deployment of an MPI-based MapReduce library called HiMach. While HiMach was able to deliver the necessary functionality and performance in most cases, it was susceptible to certain hardware and software failures such as network connectivity glitches and network file system hiccups. To resolve the robustness and usability issues, we adopted a second approach using a client-server model, and built a fault-tolerant parallel execution engine called Pitstop. This talk provides an overview of our progression from HiMach to Pitstop, and explains how we use Pitstop to implement both relatively simple interactive applications and more sophisticated statistical algorithms for analyzing millisecond-scale MD trajectories.

Experimenting at Scale

Josh Wills, Cloudera, Inc

A/B tests and multivariate experiments are widely used by web companies to make decisions about everything from user interfaces to backend system architectures to machine learning algorithms. The enormous volume of web data, the pace of product changes, and the financial stakes involved have required us to change the way we think about some seemingly basic concepts in experiment design and analysis. In this talk, we will discuss how to design a system for rapidly and reliably evaluating the impact of new product ideas at massive scale.

Low Rank Approximation and Regression in Input Sparsity Time

David Woodruff, IBM Research, Almaden

We improve the running times of algorithms for least squares regression and low-rank approximation to account for the sparsity of the input matrix. Namely, if $nnz(A)$ denotes the number of non-zero entries of an input matrix A:

- we show how to solve approximate least squares regression given an $n \times d$ matrix A in $nnz(A) + \text{poly}(d \log n)$ time;
- we show how to find an approximate best rank-k approximation of an $n \times n$ matrix in $nnz(A) + n * \text{poly}(k \log n)$ time.

All approximations are relative error. Previous algorithms based on fast Johnson-Lindenstrauss transforms took at least $nd \log d$ or $nnz(A) * k$ time. We have implemented our algorithms, and preliminary results suggest the algorithms are competitive in practice.

Joint work with Ken Clarkson.

Poster Abstracts

Discovery of Safety Incidents in Airborne Systems

Santanu Das, UARC / Nasa Ames

We present a recently developed knowledge discovery process that enables discovery of precursors to operationally significant aviation safety events through the analysis of large fleetwide heterogeneous data sources. In this approach each aircraft is monitored on its own so that one can observe significant changes in its unique operations and also across all aircrafts in the fleet to help determining how unique each system really is relative to the remaining systems in the fleet. Our method properly takes into account the sequential nature of the data and also works with discrete and continuous sequences simultaneously and in a realistic way. This allows us to not only find anomalous events as deviations from the normal, but also abnormal sequences where each snapshot is normal, but their consecutive occurrence is not. The identified patterns are further analyzed for risk-assessment based on additional information and domain expert's feedback. Domain expert feedback plays an important role to corroborate the operational significance of statistical outliers. We tested our method on 10TB of data and obtained operationally significant results that have been validated by a team of experts within NASA, airline carriers, and experts in aviation safety.

Joint work with Bryan Matthews, Kanishka Bhaduri, Kamalika Das, Ashok Srivastava, Rodney Martin, Nikunj Oza, John Stutz.

Nonparametric Quantile based Large Scale Correlation Learning

S. Mukhopadhyay (Deep), Texas A & M University

Suppose we observe $X_1, X_2, \ldots, X_p$. Goal is to detect the pair of random variables having interesting patterns. Finding highly correlated pairs is an interesting and challenging problem which has an enormous application starting from classification, regression, clustering, interaction network, etc. Motivated by the recent Science article on "Detecting Novel Associations in Large Data Sets" by Reshef brothers, we propose a unified framework for large scale correlation learning. Our strategy of building the detector, LPINFOR, enable us to go beyond exploratory phase and address the question of modeling. Unique features (i) completely data analytic and adaptive; (ii) Extremely robust, can handle multiple source of experimental noise and outliers; (iii) detect non-linear, non-monotonic relations; (iv) rich unified framework encompassing all know measures. We can systematically compute all traditional measures under one algorithmic setup and finally (v) run time, almost 50 fold faster than other competing measures dcor and MIC for massive data.

Extensive numerical comparison demonstrate the superiority of our method in terms of robustness and efficiency. Finally we utilize this association mining tool to a large scale Pharmacogenomics study to understand how genes and their variations impact drug response—to aid in the process of drug discovery and to provide a rationale for selection of therapy on the basis of molecular characteristics of a patients tumor.

The goal to describe both the aspect of Beauty (unified methodology) and Applicability (algorithm and computational ease) of LPINFOR—A nonparametric information theoretic measure to detect highly depended pair in massive data.

Large-Scale Graph-Parallel Computation on Natural Graphs

Joseph Gonzalez & Yucheng Low, Carnegie Mellon University

Two years ago we introduced GraphLab to address the critical need for a high-level abstraction for large-scale graph structured computation in machine learning. Since then, we have implemented GraphLab on multicore and cloud systems, evaluated its performance on a wide range of applications, developed new ML algorithms, and fostered a growing community of users. Along the way, we have identified new challenges to the abstraction, our implementation, and the important task of fostering a community around a research project. However, one of the most interesting and important challenges we have encountered is large-scale distributed computation on natural power-law graphs.

To address the unique challenges posed by natural graphs, we introduce GraphLab2, a fundamental redesign of the GraphLab abstraction which provides a much richer computational framework. In this work we characterize the challenges of large-scale computation on natural graphs. We then present a simple set of design techniques which can be widely applied to existing systems. Using these techniques we derive GraphLab2 and unify the GraphLab and Pregel frameworks. Finally, we present an empirical evaluation of GraphLab2 on large-scale real-world problems.

Joint work with Jay Haijie, Danny Bickson, Carlos Guestrin.

Sparse Recovery for Earth Mover Distance

Rishi Gupta, Stanford University

We initiate the study of sparse recovery problems under the Earth Mover Distance (EMD). Specifically, we design a distribution over $m \times n$ matrices A, for $m \ll n$, such that for any x, given Ax, we can recover a k-sparse approximation to x under the EMD distance. We also provide an empirical evaluation of the method that, in some scenarios, shows its advantages over the usual compressive sensing methods.

Joint with Piotr Indyk and Eric Price.

Semi-supervised Eigenvectors as a Fast Exploratory Tool for fMRI Analysis

Toke Jansen Hansen,Technical University of Denmark

Spectral clustering has recently been proposed as a methodology for analysis of high dimensional Functional Magnetic Resonance Imaging (fMRI) data. Although eigenvector-based methods are popular in machine learning, eigenvectors are inherently global quantities, thus limiting their applicability in situations where one is interested in very local properties of a data graph. In terms of fMRI, one might be interested in the clustering structure of data "nearby" a prespecified spatial "seed set". We provide a methodology to construct *semi-supervised eigenvectors* of a graph Laplacian, and we illustrate how these locally-biased eigenvectors can be used to perform fast locally-biased clustering on resting state fMRI data.

Singular Values and Randomized Matrix Multiplication

John Holodnak, North Carolina State University

When applied to AA^T, randomized matrix multiplication algorithms produce an approximation WW^T. We present a lower bound on the error of such algorithms in terms of the singular values of A and also upper bounds derived from matrix concentration inequalities. Using the bounds, we draw conclusions about the accuracy of the algorithms. In addition, we present numerical experiments that support our conclusions.

Scale-Free Graph Representation and Analysis with P-adic Lifting

Wolfgang Kraske, Docomo Innovations

An algorithm to recursively lift hierarchical representations of social network graphs is introduced. Elaboration of a natural analysis approach is summarized as a representation to facilitate advanced machine learning algorithms. Essentially attributed graphs are represented with a recursive p-adic lifting directed by scale-free network characteristics. The confluence of p-adic and scale free representation of attributed graphs provides a convenient approach to improve the accuracy and detail of analysis results. Traditionally p-adic and scale-free analyses have been pursued independently to improve accuracy and social measures respectively. Anonymized social graphs are mapped with the representation and reduced with variations of the support vector machine algorithm.

Disk-based Large-scale Graph Computation

Aapo Kyrola, Carnegie Mellon University

GraphLab and Pregel are recent systems implementing the vertex-centric model of computation. To work on very large problems, these systems require a distributed cluster. Cluster computing on graphs is challenging for many reasons: one needs to partition the graph, manage a distributed computing cluster, and handle failures. In this work, we present an alternative: a system that can run vertex-centric computation on just a single personal computer, by using disk (SSD or hard drive) as a memory extension. We compare its performance to existing distributed systems, and show that it can solve as big problems as they, in reasonable time, but with a fraction of the cost. Our work brings large-scale graph computation available to anyone with a modern PC.

Scalar Block Coordinate Descent Algorithm for Non-Negative Matrix Factorization with Bregman Divergences

Liangda Li, Georgia Institute of Technology

We propose a fast NMF algorithm that is applicable to general Bregman divergences. The algorithm uses the scalar block coordinate descent in conjunction with Taylor series expansion of the Bregman divergences, which reveals Bregman divergences relationship with Euclidean distance. The proposed algorithm generalizes several recently proposed methods and suggests new Bregman divergence optimization that is computationally much faster than existing alternatives.

Joint work with Guy Lebanon and Haesun Park.

Improved Matrix Completion via Warm-Started SVDs

Rahul Mazumder, Stanford University

Low-rank matrix modeling is an active area of research that has received considerable importance in statistics, machine learning with important applications in collaborative filtering, movie recommender systems (for example the Netflix Prize competition), bio-informatics, image processing, among others. In particular for collaborative filtering problems, for example, nuclear norm relaxations of the combinatorially hard rank constraints are often used leading to convex optimization problems. Of late there has been a considerable interest in developing large scale convex optimization algorithms for these problems. First order convex optimization techniques are often used for them. These rely on evaluating a proximal map for every iteration k (say) of the iterative algorithm. This requires computing the top few singular vectors and values of a large matrix X_k — this operation is computationally very expensive. State of the art practice uses iterative SVD solvers (for eg PROPACK) for computing a low-rank SVD for matrices $\{X_k\}_{k\geq 1}$ at the kth iteration of the algorithm. Unfortunately, these solvers are black box in the sense that they do not exploit the fact that the difference of the successive iterates converge to zero ie $X_k - X_{k-1} \to 0$. We propose a methodology which takes into account this property. Our proposed method of warm-started SVDs when used as a module by first order convex

optimization algorithms, significantly enhances the state-of-the art and leads to many fold speedups over algorithms that do not exploit warm-start information. The main idea is in viewing the low-rank SVD problem as an optimization problem and using alternating least squares (Orthogonal QR iterations) with up-sampling and Reitz acceleration in presence of warm-start information. We analyze convergence properties of the resultant procedure and support our findings via experiments

Joint work with Trevor Hastie.

LSRN: A Parallel Iterative Solver for Strongly Over- or Under-Determined Systems

Xiangrui Meng, Stanford University

We describe a parallel iterative least squares solver named `LSRN` that is based on random normal projection. `LSRN` computes the min-length solution to $\min_{x \in \mathbb{R}^n} \|Ax - b\|_2$, where $A \in \mathbb{R}^{m \times n}$ with $m \gg n$ or $m \ll n$, and where A may be rank-deficient. Tikhonov regularization may also be included. Since A is only involved in matrix-matrix and matrix-vector multiplications, it can be a dense or sparse matrix or a linear operator, and `LSRN` automatically speeds up when A is sparse or a fast linear operator. The preconditioning phase consists of a random normal projection, which is embarrassingly parallel, and a singular value decomposition of size $\lceil \gamma \min(m,n) \rceil \times \min(m,n)$, where γ is moderately larger than 1, e.g., $\gamma = 2$. We prove that the preconditioned system is well-conditioned, with a strong concentration result on the extreme singular values, and hence that the number of iterations is fully predictable when we apply LSQR or the Chebyshev semi-iterative method. As we demonstrate, the Chebyshev method is particularly efficient for solving large problems on clusters with high communication cost. Numerical results demonstrate that on a shared-memory machine, `LSRN` outperforms LAPACK's DGELSD on large dense problems, and MATLAB's backslash (SuiteSparseQR) on sparse problems. Further experiments demonstrate that `LSRN` scales well on an Amazon Elastic Compute Cloud cluster.

Adaptive Quality Control for Streaming Datasets: Techniques Applied to Set-Top Box Data

Mario Morales, Simulmedia Inc/Stanford University

We explore techniques to filter low quality observations from a streaming dataset. Our focus is based in the application of quality control process monitoring under different windows of time with the goal of detecting outliers and missing values under an adaptive scheme. This model is a building block in a current production environment of a TV Set-Top Box analytical system.

Cloud-based Large Scale Graph Clustering

Bryan Perozzi, Stony Brook University

Finding clusters in graphs is a problem which has applications in fields ranging from biology to sociology. There are a variety of existing methods for clustering small to medium sized networks, but these techniques are typically ill-suited to deal with graphs that have billions of nodes and edges. We present results from our ongoing experiments with PAPR, a BSP and MapReduce based clustering technique designed from the start to scale to very large graphs.

Parallel Coordinate Descent for Linear SVM

Vamsi K. Potluru, University of New Mexico

The linear Support Vector Machine formulation has seen successful application in a wide range of classification problems. We extend the previous state-of-the art coordinate descent method called DCD by parallelizing the updates. Also, we give convergence rate guarantees for our updates both in the sequential and parallel settings scaling as $O(1/T)$ where T is the number of passes over the data. We validate our approach by presenting experimental work on various real-world datasets.

Joint work with S. Sathiya Keerthi.

Graph Denoising

Emile Richard, Ecole Normale Superieure de Cachan

Our work is motivated by real-world applications for denoising graph data. We show that this problem amounts to solving matrix recovery for an adjacency matrix which is both sparse and low-rank under a random perturbation matrix which is sparse. We formulate the problem as the minimization of a regularized convex objective with an L1 loss. We present a Douglas-Rachford splitting algorithm to compute the estimator and another method using matrix factorization and rank-one updates which offers more scalability. Numerical experiments conrm the relevance of the approach compared to state-of-the-art methods such as robust PCA.

Joint work with Pierre-Andre Savalle (presenter).

Modeling Online Social Networks

Guillermo Rodriguez-Cano, KTH Royal Institute of Technology

Social networks are a popular topic in many areas of research. From the human and behavioral perspective that social science investigates in general, and sociology focuses in particular, to the mathematical and algorithmic side that computer science considers. These networks often are, if not primarily, mathematically modeled as graphs, one of the most known and used combinatorial structures in computer

science. However, the inner essence of the individuals interacting in these systems and the information exchanged over time might not be truly depicted with one-to-one relationships. The semantics of the formation of relationships between individuals, but also the information exchanged, go beyond the limitations of these associations, which current implementations overcome, for instance, by replicating relationships when an individual is associated with various others at once. When moving to a digital world, Online Social Networks (OSNs), such as Facebook, with almost a billion individuals, performance and scalability but also efficiency in storage becomes important when modeling such large graph. In this work we propose an alternative model for OSNs based on hyper-graphs, a generalization of graphs, aiming at representing more coherently the one-to-many and many-to-many associations occurring in these networks. We also take advantage of some Social Network Analysis (SNA) centrality metrics for a finer understanding of the communities formed in the network. Lastly, we look into the depiction of nodes and edges in OSNs as the rapidly increasing types of data suggests that the typical association of nodes with individuals and edges with their relationships might not be sufficiently expressive for a large scale OSN.

A Geometric Analysis of Subspace Clustering with Outliers

Mahdi Soltanolkotabi, Stanford University

One of the most fundamental steps in data analysis and dimensionality reduction consists of approximating a given dataset by a single low-dimensional subspace, which is classically achieved via Principal Component Analysis (PCA). However, in many applications, the data often lie near a union of low-dimensional subspaces, reflecting the multiple categories or classes a set of observations may belong to. In this talk we discuss the problem of clustering a collection of unlabeled data points assumed to lie near a union of lower dimensional planes. As is common in computer vision or unsupervised learning applications, we do not know in advance how many subspaces there are nor do we have any information about their dimensions. We present a novel geometric analysis of an algorithm named sparse subspace clustering (SSC) [Elhamifar and Vidal 2009], which significantly broadens the range of problems where it is provably effective. For instance, we show that SSC can recover multiple subspaces, each of dimension comparable to the ambient dimension. We also show that SSC can correctly cluster data points even when the subspaces of interest intersect. Further, we develop an extension of SSC that succeeds when the data set is corrupted with possibly overwhelmingly many outliers. Underlying our analysis are clear geometric insights, which may bear on other sparse recovery problems. We will demonstrate the effectiveness of these methods by various numerical studies.

Joint work with Emmanuel J. Candes.

Efficient Point-to-Subspace Query in ℓ^1 with Application to Robust Face Recognition

Ju Sun, Columbia University

Motivated by applications to face and object recognition, we consider the following problem: given a collection of low-dimensional linear subspaces of a high-dimensional ambient (image) space, and a new query image, efficiently determine the closest subspace to the query in ℓ^1 norm. We show in theory this problem can be solved efficiently with a simple two-stage algorithm: (1) random Cauchy projection of query and subspaces into low-dimensional spaces followed by efficient distance evaluation (ℓ^1 regression); (2) getting back to the high-dimensional space with very few candidates and performing exhaustive search. We present preliminary experiments on face recognition to corroborate our theory.

Computing Geodesic Shortest Paths on Large Terrains Using Less Memory

Vishal Verma, University of North Carolina, Chapel Hill

For computing shortest paths on large meshes, the bottleneck is often the memory access rather than the theoretical time complexity. Two frameworks that explicitly address how memory is used are I/O efficient and streaming algorithms. We use the streaming mesh framework for computing shortest paths in-core.

Since shortest path algorithms process the mesh in the order of distance from source, to efficiently use the memory we reorder the input streaming mesh using an approximate shortest path algorithm. *E.g.* for the MMP (aka continuous Dijkstra) algorithm, which computes geodesic distances, we order the input stream using Dijkstra's algorithm. We also propose a continuous version of the A* algorithm. The input stream to this algorithm is reordered using the standard A* algorithm. For meshes containing millions of triangles we see a two orders of magnitude improvement in the memory usage. This allows us to compare geodesic distances to distances along the edges of the mesh.

The Affects of Coherence on Randomized Least Squares Algorithms

Thomas Wentworth, North Carolina State University

Coherence is a matrix property that is of particular interest to randomized methods for approximating least squares and other problems. In this poster we present how coherence is related to other matrix properties. In addition we present an efficient orthogonal transformation which minimizes the coherence of a given matrix. Finally we use this transformation to examine the potential performance of the Blendenpik algorithm.

Acknowledgements

Sponsors

The Organizers of MMDS 2012 and the MMDS Foundation would like to thank the following institutional sponsors for their generous support:

- **AFOSR:** the Air Force Office of Scientific Research
- **iCME:** the Institute for Computational and Mathematical Engineering, Stanford University
- **LBL:** the Lawrence Berkeley National Laboratory
- **Dropbox, Inc.**
- **eBay, Inc.**
- **Cloudera, Inc.**