MMDS 2008: Workshop on Algorithms for Modern Massive Data Sets

Stanford University
June 25–28, 2008

The 2008 Workshop on Algorithms for Modern Massive Data Sets (MMDS 2008) will address algorithmic, mathematical, and statistical challenges in modern large-scale data analysis. The goals of MMDS 2008 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 cross-fertilization of ideas.

Organizers: Gunnar Carlsson, Lek-Heng Lim, Michael Mahoney

Conference Schedule

Wednesday, June 25, 2008: Data analysis and data applications

Time	Event	Location/Page
Registration & Opening		**Annenberg Auditorium**
9:00–9:45am	*Breakfast and registration*	
9:45–10:00am	Organizers *Welcome and opening remarks*	
First Session		**Annenberg Auditorium**
10:00–11:00am	Christos Faloutsos, Carnegie Mellon University *Graph mining: laws, generators and tools* (tutorial)	pp. 9
11:00–11:30am	Deepak Agarwal, Yahoo! Research, Silicon Valley *Predictive discrete latent models for large incomplete dyadic data*	pp. 7
11:30–12:00n	Chandrika Kamath, Lawrence Livermore National Laboratory *Scientific data mining: why is it difficult?*	pp. 11
12:00–2:00pm	*Lunch* (on your own)	
Second Session		**Annenberg Auditorium**
2:00–3:00pm	Edward Chang, Google Research, Mountain View *Challenges in mining large-scale social networks* (tutorial)	pp. 8
3:00–3:30pm	Sharad Goel, Yahoo! Research, New York *Predictive indexing for fast search*	pp. 10
3:30–4:00pm	James Demmel, University of California, Berkeley *Avoiding communication in linear algebra algorithms*	pp. 9
4:00–4:30pm	*Coffee break*	
Third Session		**Annenberg Auditorium**
4:30–5:00pm	Jun Liu, Harvard University *Bayesian inference of interactions and associations*	pp. 12
5:00–5:30pm	Fan Chung, University of California, San Diego *Four graph partitioning algorithms*	pp. 8
5:30–6:00pm	Ronald Coifman, Yale University *Diffusion geometries and harmonic analysis on data sets*	pp. 8
6:00–9:30pm	*Welcome reception*	**New Guinea Garden**

Thursday, June 26, 2008: Networked data and algorithmic tools

Time	Event	Location/Page
First Session		**Annenberg Auditorium**
9:00–10:00am	Milena Mihail, Georgia Institute of Technology *Models and algorithms for complex networks, with network elements maintaining characteristic profiles* (tutorial)	pp. 13
10:00–10:30am	Reid Andersen, Microsoft Research, Redmond *An algorithm for improving graph partitions*	pp. 7
10:30–11:00am	*Coffee break*	
Second Session		**Annenberg Auditorium**
11:00–11:30am	Michael W. Mahoney, Yahoo! Research, Silicon Valley *Community structure in large social and information networks*	pp. 12
11:30–12:00n	Nikhil Srivastava, Yale University *Graph sparsification by effective resistances*	pp. 14
12:00–12:30pm	Amin Saberi, Stanford University *Sequential algorithms for generating random graphs*	pp. 13
12:30–2:30pm	*Lunch* (on your own)	
Third Session		**Annenberg Auditorium**
2:30–3:00pm	Pankaj K. Agarwal, Duke University *Modeling and analyzing massive terrain data sets*	pp. 7
3:00–3:30pm	Leonidas Guibas, Stanford University *Detection of symmetries and repeated patterns in 3D point cloud data*	pp. 10
3:30–4:00pm	Yuan Yao, Stanford University *Topological methods for exploring pathway analysis in complex biomolecular folding*	pp. 14
4:00–4:30pm	*Coffee break*	
Fourth Session		**Annenberg Auditorium**
4:30–5:00pm	Piotr Indyk, Massachusetts Institute of Technology *Sparse recovery using sparse random matrices*	pp. 10
5:00–5:30pm	Ping Li, Cornell University *Compressed counting and stable random projections*	pp. 12
5:30–6:00pm	Joel Tropp, California Institute of Technology *Efficient algorithms for matrix column selection*	pp. 14

Friday, June 27, 2008: Statistical, geometric, and topological methods

Time	Event	Location/Page
First Session		**Annenberg Auditorium**
9:00–10:00am	Jerome H. Friedman, Stanford University *Fast sparse regression and classification* (tutorial)	pp. 9
10:00–10:30am	Tong Zhang, Rutgers University *An adaptive forward/backward greedy algorithm for learning sparse representations*	pp. 15
10:30–11:00am	*Coffee break*	
Second Session		**Annenberg Auditorium**
11:00–11:30am	Jitendra Malik, University of California, Berkeley *Classification using intersection kernel SVMs is efficient*	pp. 13
11:30–12:00n	Elad Hazan, IBM Almaden Research Center *Efficient online routing with limited feedback and optimization in the dark*	pp. 10
12:00–12:30pm	T.S. Jayram, IBM Almaden Research Center *Cascaded aggregates on data streams*	pp. 11
12:30–2:30pm	*Lunch* (on your own)	
Third Session		**Annenberg Auditorium**
2:30–3:30pm	Gunnar Carlsson, Stanford University *Topology and data* (tutorial)	pp. 8
3:30–4:00pm	Partha Niyogi, University of Chicago *Manifold regularization and semi-supervised learning*	pp. 13
4:00–4:30pm	*Coffee break*	
Fourth Session		**Annenberg Auditorium**
4:30–5:00pm	Sanjoy Dasgupta, University of California, San Diego *Random projection trees and low dimensional manifolds*	pp. 8
5:00–5:30pm	Kenneth Clarkson, IBM Almaden Research Center *Tighter bounds for random projections of manifolds*	pp. 8
5:30–6:00pm	Yoram Singer, Google Research, Mountain View *Efficient projection algorithms for learning sparse representations from high dimensional data*	pp. 14
6:00–6:30pm	Arindam Banerjee, University of Minnesota, Twin Cities *Bayesian co-clustering for dyadic data analysis*	pp. 7
6:30–9:30pm	*Reception and poster session*	**Old Union Club House**

Saturday, June 28, 2008: Machine learning and dimensionality reduction

Time	Event	Location/Page
First Session		**Annenberg Auditorium**
9:00–10:00am	Michael I. Jordan, University of California, Berkeley *Sufficient dimension reduction* (tutorial)	pp. 11
10:00–10:30am	Nathan Srebro, University of Chicago *More data less work: SVM training in time decreasing with larger data sets*	pp. 14
10:30–11:00am	*Coffee break*	
Second Session		**Annenberg Auditorium**
11:00–11:30am	Inderjit S. Dhillon, University of Texas, Austin *Rank minimization via online learning*	pp. 9
11:30–12:00n	Nir Ailon, Google Research, New York *Efficient dimension reduction*	pp. 7
12:00–12:30pm	Satyen Kale, Microsoft Research, Redmond *A combinatorial, primal-dual approach to semidefinite programs*	pp. 11
12:30–2:30pm	*Lunch* (box lunch provided)	**Annenberg Courtyard**
Third Session		**Annenberg Auditorium**
2:30–3:00pm	Ravi Kannan, Microsoft Research, India *Spectral algorithms*	pp. 11
3:00–3:30pm	Chris Wiggins, Columbia University *Inferring and encoding graph partitions*	pp. 14
3:30–4:00pm	Anna Gilbert, University of Michigan, Ann Arbor *Combinatorial group testing in signal recovery*	pp. 9
4:00–4:30pm	*Coffee break*	
Fourth Session		**Annenberg Auditorium**
4:30–5:00pm	Lars Kai Hansen, Technical University of Denmark *Generalization in high-dimensional matrix factorization*	pp. 10
5:00–5:30pm	Holly Jin, LinkedIn *Exploring sparse nonnegative matrix factorization*	pp. 11
5:30–6:00pm	Elizabeth Purdom, University of California, Berkeley *Data analysis with graphs*	pp. 13
6:00–6:30pm	Lek-Heng Lim, University of California, Berkeley *Ranking via Hodge decompositions of digraphs and skew-symmetric matrices*	pp. 12
6:30–8:00pm	*Reception and poster session*	**Annenberg Courtyard**

Friday, June 27, 2008: Posters

Event	Location/Page
Poster Session	**Old Union Club House**
Christos Boutsidis, Rensselaer Polytechnic Institute *An improved approximation algorithm for the column subset selection problem*	pp. 16
Pavel Dmitriev, Yahoo! Inc. *Machine learning approach to generalized graph pattern mining*	pp. 16
Charles Elkan, University of California, San Diego *An experimental comparison of randomized methods for low-rank matrix approximation*	pp. 16
Sreenivas Gollapudi, Microsoft Research, Silicon Valley *Estimating PageRank on graph streams*	pp. 16
Mohammad Al Hasan, Rensselaer Polytechnic Institute *Uniform generation and pattern summarization*	pp. 17
Ling Huang, Intel Research *Approximate support vector machines for distributed system*	pp. 17
Xuhui Huang, Stanford University *Convergence of protein folding free energy landscape via application of generalized ensemble sampling methods on a distributed computing environment*	pp. 17
Prateek Jain, University of Texas, Austin *Metric/kernel learning with applications to fast similarity search*	pp. 17
Pentti Kanerva, Stanford University *Random indexing: simple projections for accumulating sparsely distributed frequencies*	pp. 17
Simon Lacoste-Julien, University of California, Berkeley *DiscLDA: discriminative learning for dimensionality reduction and classification*	pp. 18
Taesup Moon, Stanford University *Discrete denoising with shifts*	pp. 18
Ramesh Nallapati, Stanford University *Joint latent topic models for text and citations*	pp. 18
Hariharan Narayanan, University of Chicago *Annealed entropy, heavy tailed data and manifold learning*	pp. 18
Stefan Pickl, Naval Postgraduate School, Monterey and Bundeswehr University, Munich *Algorithmic data analysis with polytopes: a combinatorial approach*	pp. 19
Jian Sun, Stanford University *Global and local intrinsic symmetry detection*	pp. 19
John Wu, Lawrence Berkeley National Lab *FastBit: an efficient indexing tool for massive data*	pp. 19
Zuobing Xu, University of California, Santa Clara *Dirichlet compound multinomial models for retrieval and active relevance feedback*	pp. 19

Abstracts of talks

Predictive discrete latent models for large incomplete dyadic data

Deepak Agarwal, Yahoo! Research, Silicon Valley

We propose a novel statistical model to predict large scale dyadic response variable in the presence of covariate information. Our approach simultaneously incorporates the effect of covariates and estimates local structure that is induced by interactions among the dyads through a discrete latent factor model. The discovered latent factors provide a predictive model that is both accurate and interpretable. To further combat sparsity that is often present in large scale dyadic data, we enhance the discrete latent factors by coupling it with a factorization model on the cluster effects. In fact, by regularizing the factors through L_2 norm constraints, we are able to allow larger number of factors that capture dyadic interactions with greater precision compared to a pure factorization model, especially for sparse data. The methods are illustrated in a generalized linear model framework and includes linear regression, logistic regression and Poisson regression as special cases. We also provide scalable generalized EM-based algorithms for model fitting using both "hard" and "soft" cluster assignments. We demonstrate the generality and efficacy of our approach through large scale simulation studies and analysis of datasets obtained from certain real-world movie recommendation and internet advertising applications.

Modeling and analyzing massive terrain data sets

Pankaj K. Agarwal, Duke University

With recent advances in terrain-mapping technologies such as Laser altimetry (LIDAR) and ground based laser scanning, millions of georeferenced points can be acquired within short periods of time. However, while acquiring and georeferencing the data has become extremely efficient, transforming the resulting massive amounts of heterogeneous data to useful information for different types of users and applications is lagging behind, in large part because of the scarcity of robust, efficient algorithms for terrain modeling and analysis that can handle massive data sets acquired by different technologies and that can rapidly detect and predict changes in the model as the new data is acquired.

This talk will review our on-going work on developing efficient algorithms for terrain modeling and analysis that work with massive data sets. It will focus on an I/O-efficient algorithm for computing contour maps of a terrain. A few open questions will also be discussed.

Efficient dimension reduction

Nir Ailon, Google Research, New York

The Johnson-Lindenstrauss dimension reduction idea using random projections was discovered in the early 80's. Since then many "computer science friendly" versions were published, offering constant factor but no big-O improvements in the runtime. Two years ago Ailon and Chazelle showed a nontrivial algorithm with the first asymptotic improvement, and suggested the question: What is the exact complexity of J-L computation from d dimensions to k dimensions? An $O(d \log d)$ upper bound is implied by A-C for k up to $d^{1/3}$ (in the $L^2 \to L^2$) case. In this talk I will show how to achieve this bound for k up to $d^{1/2}$ combining techniques from the theories of error correcting codes and probability in Banach spaces.

This is based on joint work with Edo Liberty.

An algorithm for improving graph partitions

Reid Andersen, Microsoft Research, Redmond

The minimum quotient cut problem is a fundamental and well-studied problem in graph partitioning. We present an algorithm called Improve that improves a proposed partition of a graph, taking as input a subset of vertices and returning a new subset of vertices with a smaller quotient cut score. The most powerful previously known method for improving quotient cuts, which is based on parametric flow, returns a partition whose quotient cut score is at least as small as any set contained within the proposed set. For our algorithm, we can prove a stronger guarantee: the quotient score of the set returned is nearly as small as any set in the graph with which the proposed set has a larger-than-expected intersection. The algorithm finds such a set by solving a sequence of polynomially many S-T minimum cut problems, a sequence that cannot be cast as a single parametric flow problem. We demonstrate empirically that applying Improve to the output of various graph partitioning algorithms greatly improves the quality of cuts produced without significantly impacting the running time.

Joint work with Kevin Lang.

Bayesian co-clustering for dyadic data analysis

Arindam Banerjee, University of Minnesota, Twin Cities

In recent years, co-clustering has emerged as a powerful data mining tool that can analyze dyadic data connecting two entities. However, almost all existing co-clustering techniques are partitional, and allow individual rows and columns of a data matrix to belong to only one cluster. Several current applications, such as recommendation systems, market basket analysis, and gene expression analysis, can substantially benefit from a mixed cluster assignment of rows and columns. In this talk, we present Bayesian co-clustering (BCC) models, that allow mixed probabilistic assignments of rows and columns to all clusters. BCC maintains separate Dirichlet priors over the mixed assignments and assumes each observation to be generated by an exponential family distribution corresponding to its row and column clusters. The model is designed to naturally handle sparse matrices as only the non-zero/non-missing entries are assumed to be

generated by the model. We propose a fast variational algorithm for inference and parameter estimation in the model. In addition to finding co-cluster structure in observations, the model outputs a low dimensional co-embedding of rows and columns, and predicts missing values in the original matrix. We demonstrate the efficacy of the model through experiments on both simulated and real data.

Topology and data

Gunnar Carlsson, Stanford University

Topological methods have in recent years shown themselves to be capable of identifying a number of qualitative geometric properties of high dimensional data sets. In this talk, we will describe philosophical reasons why topological methods should play a role in the study of data, as well as give several examples of how the ideas play out in practice.

Challenges in mining large-scale social networks

Edward Chang, Google Research, Mountain View

Social networking sites such as Orkut, MySpace, Hi5, and Facebook attract billions of visits a day, surpassing the page views of Web Search. These social networking sites provide applications for individuals to establish communities, to upload and share documents/photos/videos, and to interact with other users. Take Orkut as an example. Orkut hosts millions of communities, with hundreds of communities created and tens of thousands of blogs/photos uploaded each hour. To assist users to find relevant information, it is essential to provide effective collaborative filtering tools to perform recommendations such as friend, community, and ads matching.

In this talk, I will first describe both computational and storage challenges to traditional collaborative filtering algorithms brought by aforementioned information explosion. To deal with huge social graphs that expand continuously, an effective algorithm should be designed to 1) run on thousands of parallel machines for sharing storage and speeding up computation, 2) perform incremental retraining and updates for attaining online performance, and 3) fuse information of multiple sources for alleviating information sparseness. In the second part of the talk, I will present algorithms we recently developed including parallel PF-Growth [1], parallel combinational collaborative filtering [2], parallel LDA, parallel spectral clustering, and parallel Support Vector Machines [3].

[1,2,3] Papers and some codes are available at `http://infolab.stanford.edu/~echang/`

Four graph partitioning algorithms

Fan Chung, University of California, San Diego

We will discuss four partitioning algorithms using eigenvectors, random walks, PageRank and their variations. In particular, we will examine local partitioning algorithms, which find a cut near a specified starting vertex, with a running time that depends on the size of the small side of the cut, rather than on the size of the input graph (which can be prohibitively large). Three of the four partitioning algorithms are local algorithms and are particularly appropriate for applications for modern massive data sets.

Tighter bounds for random projections of manifolds

Kenneth Clarkson, IBM Almaden Research Center

The Johnson-Lindenstrauss random projection lemma gives a simple way to reduce the dimensionality of a set of points while approximately preserving their pairwise Euclidean distances. The most direct application of the lemma applies to a finite set of points, but recent work has extended the technique to affine subspaces, smooth manifolds, and sets of bounded doubling dimension; in each case, a projection to a sufficiently large dimension k implies that all pairwise distances are approximately preserved with high probability. Here the case of random projection of a smooth manifold (submanifold of $\mathbb{R}^m$) is considered, and a previous analysis is sharpened, giving an upper bound for k that depends on the surface area, total absolute curvature, and a few other quantities associated with the manifold, and in particular does not depend on the ambient dimension m or the reach of the manifold.

Diffusion geometries and harmonic analysis on data sets

Ronald Coifman, Yale University

We discuss the emergence of self organization of data, either through eigenvectors of affinity operators, or equivalently through a multiscale ontology. We illustrate these ideas on images and audio files as well as molecular dynamics.

Random projection trees and low dimensional manifolds

Sanjoy Dasgupta, University of California, San Diego

The curse of dimensionality has traditionally been the bane of nonparametric statistics (histograms, kernel density estimation, nearest neighbor search, and so on), as reflected in running times and convergence rates that are exponentially bad in the dimension. This problem is all the more pressing as data sets get increasingly high dimensional. Recently the field has been rejuvenated substantially, in part by the realization that a lot of real-world data which appears high-dimensional in fact has low "intrinsic" dimension in the sense of lying close to a low-dimensional manifold. In the past few years, there has been a huge interest in learning such manifolds from data, and then using the learned structure to transform the data into a lower-dimensional space where standard statistical methods generically work better.

I'll exhibit a way to benefit from intrinsic low dimensionality without having to go to the trouble of explicitly

learning its fine structure. Specifically, I'll present a simple variant of the ubiquitous k-d tree (a spatial data structure widely used in machine learning and statistics) that is provably adaptive to low dimensional structure. We call this a "random projection tree" (RP tree).

Along the way, I'll discuss different notions of intrinsic dimension — motivated by manifolds, by local statistics, and by analysis on metric spaces — and relate them. I'll then prove that RP trees require resources that depend only on these dimensions rather than the dimension of the space in which the data happens to be situated.

This is work with Yoav Freund (UC San Diego).

Avoiding communication in linear algebra algorithms

James Demmel, University of California, Berkeley

We survey recent results on designing numerical algorithms to minimize the largest cost component: communication. This could be bandwidth and latency costs between processors over a network, or between levels of a memory hierarchy; both costs are increasing exponentially compared to floating point. We describe novel algorithms in sparse and dense linear algebra, for both direct methods (like QR and LU) and iterative methods that can minimize communication.

Rank minimization via online learning

Inderjit S. Dhillon, University of Texas, Austin

Minimum rank problems arise frequently in machine learning applications and are notoriously difficult to solve due to the non-convex nature of the rank objective. In this talk, we will present a novel online learning approach for the problem of rank minimization of matrices over polyhedral sets. In particular, we present two online learning algorithms for rank minimization—our first algorithm is a multiplicative update method based on a generalized experts framework, while our second algorithm is a novel application of the online convex programming framework (Zinkevich, 2003). A salient feature of our online learning approach is that it allows us to give provable approximation guarantees for the rank minimization problem over polyhedral sets. We demonstrate that our methods are effective in practice through experiments on synthetic examples, and on the real-life application of low-rank kernel learning.

This is joint work with Constantine Caramanis, Prateek Jain and Raghu Meka.

Graph mining: laws, generators and tools

Christos Faloutsos, Carnegie Mellon University

How do graphs look like? How do they evolve over time? How can we generate realistic-looking graphs? We review some static and temporal 'laws', and we describe the "Kronecker" graph generator, which naturally matches all of the known properties of real graphs. Moreover, we present tools for discovering anomalies and patterns in two types of graphs, static and time-evolving. For the former, we present the 'CenterPiece' subgraphs (CePS), which expects q query nodes (eg., suspicious people) and finds the node that is best connected to all q of them (eg., the master mind of a criminal group). We also show how to compute CenterPiece subgraphs efficiently. For the time evolving graphs, we present tensor-based methods, and apply them on real data, like the DBLP author-paper dataset, where they are able to find natural research communities, and track their evolution.

Finally, we also briefly mention some results on influence and virus propagation on real graphs, as well as on the emerging map/reduce approach and its impact on large graph mining.

Fast sparse regression and classification

Jerome H. Friedman, Stanford University

Regularized regression and classification methods fit a linear model to data, based on some loss criterion, subject to a constraint on the coefficient values. As special cases, ridge-regression, the lasso, and subset selection all use squared-error loss with different particular constraint choices. For large problems the general choice of loss/constraint combinations is usually limited by the computation required to obtain the corresponding solution estimates, especially when non convex constraints are used to induce very sparse solutions. A fast algorithm is presented that produces solutions that closely approximate those for any convex loss and a wide variety of convex and non convex constraints, permitting application to very large problems. The benefits of this generality are illustrated by examples.

Combinatorial group testing in signal recovery

Anna Gilbert, University of Michigan, Ann Arbor

Traditionally, group testing is a design problem. The goal is to construct an optimally efficient set of tests of items such that the test results contains enough information to determine a small subset of items of interest. It has its roots in the statistics community and was originally designed for the Selective Service to find and to remove men with syphilis from the draft. It appears in many forms, including coin-weighing problems, experimental designs, and public health. We are interested in both the design of tests and the design of an efficient algorithm that works with the tests to determine the group of interest. Many of the same techniques that are useful for designing tests are also used to solve algorithmic problems in analyzing and in recovering statistical quantities from streaming data. I will discuss some of these methods and briefly discuss several recent applications in high throughput drug screening.

Predictive indexing for fast search

Sharad Goel, Yahoo! Research, New York

We tackle the computational problem of query-conditioned search. Given a machine-learned scoring rule and a query distribution, we build a predictive index by pre-computing lists of potential results sorted based on an expected score of the result over future queries. The predictive index datastructure supports an anytime algorithm for approximate retrieval of the top elements. The general approach is applicable to webpage ranking, internet advertisement, and approximate nearest neighbor search. It is particularly effective in settings where standard techniques (e.g., inverted indices) are intractable. We experimentally find substantial improvement over existing methods for internet advertisement and approximate nearest neighbors.

This work is joint with John Langford and Alex Strehl.

Detection of symmetries and repeated patterns in 3d point cloud data

Leonidas Guibas, Stanford University

Digital models of physical shapes are becoming ubiquitous in our economy and life. Such models are sometimes designed ab initio using CAD tools, but more and more often they are based on existing real objects whose shape is acquired using various 3D scanning technologies. In most instances, the original scanner data is just a set, but a very large set, of points sampled from the surface of the object. We are interested in tools for understanding the local and global structure of such large-scale scanned geometry for a variety of tasks, including model completion, reverse engineering, shape comparison and retrieval, shape editing, inclusion in virtual worlds and simulations, etc. This talk will present a number of point-based techniques for discovering global structure in 3D data sets, including partial and approximate symmetries, shared parts, repeated patterns, etc. It is also of interest to perform such structure discovery across multiple data sets distributed in a network, without actually ever bring them all to the same host.

Generalization in high-dimensional matrix factorization

Lars Kai Hansen, Technical University of Denmark

While the generalization performance of high-dimensional principal component analysis is quite well understood, matrix factorizations like independent component analysis, non-negative matrix factorization, and clustering are less investigated for generalizability. I will review theoretical results for PCA and heuristics used to improve PCA test performance, and discuss extensions to high-dimensional ICA, NMF, and clustering.

Efficient online routing with limited feedback and optimization in the dark

Elad Hazan, IBM Almaden Research Center

In many decision making scenarios the decision maker has access only to limited information on the success/effect of his previous decisions. For example, an advertiser can observe the success in sales of a certain product, but can hardly infer on results if other strategies were to be used. Another example is network routing, in which the trip time along the chosen path can be measured, but little or no information is available on other links in the network.

We study a more general framework of online linear optimization with "bandit" feedback. The existence of an efficient low-regret algorithm has been an open question addressed in several recent papers. We present the first known efficient algorithm with optimal regret bounds for bandit Online Linear Optimization over arbitrary convex decision sets. We show how the difficulties encountered by previous approaches are overcome by exploiting the surprising exploration-exploitation properties of a self concordant barrier function for the underlying decision set.

Joint work with Jacob Abernethy and Alexander Rakhlin of UC Berkeley.

Sparse recovery using sparse random matrices

Piotr Indyk, Massachusetts Institute of Technology

Over the recent years, a new *linear* method for compressing high-dimensional data (e.g., images) has been discovered. For any high-dimensional vector x, its *sketch* is equal to Ax, where A is an $m \times n$ matrix (possibly chosen at random). Although typically the sketch length m is much smaller than the number of dimensions n, the sketch contains enough information to recover an *approximation* to x. At the same time, the linearity of the sketching method is very convenient for many applications, such as data stream computing and compressed sensing.

The major sketching approaches can be classified as either combinatorial (using sparse sketching matrices) or geometric (using dense sketching matrices). They achieve different trade-offs, notably between the compression rate and the running time. Thus, it is desirable to understand the connections between them, with the ultimate goal of obtaining the "best of both worlds" solution.

In this talk we show that, in a sense, the combinatorial and geometric approaches are based on different manifestations of the same phenomenon. This connection will enable us to obtain several novel algorithms and constructions, which inherit advantages of sparse matrices, such as lower sketching and recovery times.

Joint work with: Radu Berinde, Anna Gilbert, Howard Karloff, Milan Ruzic and Martin Strauss.

Cascaded aggregates on data streams

T.S. Jayram, IBM Almaden Research Center

Let A be a matrix whose rows are given by $A_1, A_2, \ldots, A_m$. For two aggregate operators P and Q, the cascaded aggregate $(P \circ Q)(A)$ is defined to be $P(Q(A_1), Q(A_2), \ldots, Q(A_m))$. The problem of computing such aggregates over data streams has received considerable interest recently. In this talk, I will present some recent results on computing cascaded frequency moments and norms over data streams.

Joint work with David Woodruff (IBM Almaden).

Exploring sparse nonnegative matrix factorization

Holly Jin, LinkedIn

We explore the use of basis pursuit denoising (BPDN) for sparse nonnegative matrix factorization (sparse NMF). A matrix A is approximated by low-rank factors UDV', where U and V are sparse with unit-norm columns, and D is diagonal. We use an active-set BPDN solver with warm starts to compute the rows of U and V in turn. (Friedlander and Hatz have developed a similar formulation for both matrices and tensors.) We present computational results and discuss the benefits of sparse NMF for some real matrix applications.

This is joint work with Michael Saunders of Stanford University.

Sufficient dimension reduction

Michael I. Jordan, University of California, Berkeley

Consider a regression or classification problem in which the data consist of pairs (X, Y), where X is a high-dimensional vector. If we wish to find a low-dimensional subspace to represent X, one option is to ignore Y and avail ourselves of unsupervised methods such PCA and factor analysis. In this talk, I will discuss a supervised alternative which aims to take Y into account in finding a low-dimensional representation for X, while avoiding making strong assumptions about the relationship of X to Y. Specifically, the problem of "sufficient dimension reduction" (SDR) is that of finding a subspace S such that the projection of the covariate vector X onto S captures the statistical dependency of the response Y on X (Li, 1991; Cook, 1998). I will present a general overview of the SDR problem, focusing on the formulation of SDR in terms of conditional independence. I will also discuss some of the popular algorithmic approaches to SDR, particularly those based on inverse regression. Finally, I will describe a new methodology for SDR which is based on the characterization of conditional independence in terms of conditional covariance operators on reproducing kernel Hilbert spaces (a characterization of conditional independence that is of independent interest). I show how this characterization leads to an M-estimator for S and I show that the estimator is consistent under weak conditions; in particular, we do not have to impose linearity or ellipticity conditions of the kinds that are generally invoked for SDR methods based on inverse regression.

Joint work with Kenji Fukumizu and Francis Bach.

A combinatorial, primal-dual approach to semidefinite programs

Satyen Kale, Microsoft Research, Redmond

Algorithms based on convex optimization, especially linear and semidefinite programming, are ubiquitous in Computer Science. While there are polynomial time algorithms known to solve such problems, quite often the running time of these algorithms is very high. Designing simpler and more efficient algorithms is important for practical impact.

In my talk, I will describe applications of a Lagrangian relaxation technique, the Multiplicative Weights Update method in the design of efficient algorithms for various optimization problems. We generalize the method to the setting of symmetric matrices rather than real numbers. The new algorithm yields the first truly general, combinatorial, primal-dual method for designing efficient algorithms for semidefinite programming. Using these techniques, we obtain significantly faster algorithms for approximating the Sparsest Cut and Balanced Separator in both directed and undirected weighted graphs to factors of $O(\log n)$ and $O(\sqrt{\log n})$, and also for the Min UnCut and Min 2CNF Deletion problems. The algorithm also has applications in quantum computing and derandomization.

This is joint work with Sanjeev Arora.

Scientific data mining: why is it difficult?

Chandrika Kamath, Lawrence Livermore National Laboratory

Scientific data sets, whether from computer simulations, observations, or experiments, are not only massive, but also very complex. This makes it challenging to find useful information in the data, a fact scientists find disconcerting as they wonder about the science still undiscovered in the data. In this talk, I will describe my experiences in analyzing data in a variety of scientific domains. Using example problems from astronomy, fluid mixing, remote sensing, and experimental physics, I will describe our solution approaches and discuss some of the challenges we have encountered in mining these datasets.

Spectral algorithms

Ravi Kannan, Microsoft Research, India

While spectral algorithms enjoy quite some success in practice, there are not many provable results about them. In general, results are only known for the "average case" assuming some probabilitic model of the data. We discuss some models and results. Two sets of models have been analyzed. Gaussian mixture models yield provable results on spectral algorithms. The geometry of Gaussians and "isopermetry" play a crucial role here. A second set of models inspired by Random Graphs and Random Matrix theory assume total mutual independence of all entries of the input

matrix. While this allows the use of the classical bounds on eigenvalues of random matrices, it is too restrictive. A Partial Independence model where the rows are assumed to be independent vector-valued random variables is more realistic and recently results akin to totally independent matrices have been proved for these.

Ranking via Hodge decompositions of digraphs and skew-symmetric matrices

Lek-Heng Lim, University of California, Berkeley

Modern ranking data is often incomplete, unbalanced, and arises from a complex network. We will propose a method to analyze such data using combinatorial Hodge theory, which may be regarded either as an additive decomposition of a skew-symmetric matrix into three matrices with special properties or a decomposition of a weighted digraph into three orthogonal components. In this framework, ranking data is represented as a skew-symmetric matrix and Hodge-decomposed into three mutually orthogonal components corresponding to globally consistent, locally consistent, and locally inconsistent parts of the ranking data. Rank aggregation then naturally emerges as projections onto a suitable subspace and an inconsistency measure of the ranking data arises as the triangular trace distribution.

This is joint work with Yuan Yao of Stanford University.

Compressed counting and stable random projections

Ping Li, Cornell University

The method of stable random projections has become a popular tool for dimension reduction, in particular, for efficiently computing pairwise distances in massive high-dimensional data (including dynamic streaming data) matrices, with many applications in data mining and machine learning such as clustering, nearest neighbors, kernel methods etc. Closely related to stable random projections, Compressed Counting (CC) is recently developed to efficiently compute L_p frequency moments of a single dynamic data stream. CC exhibits a dramatic improvement over stable random projections when p is about 1. Applications of CC include estimating entropy moments of data streams and statistical parameter estimations in dynamic data using low memory.

Bayesian inference of interactions and associations

Jun Liu, Harvard University

In a regression or classification problem, one often has many potential predictors (independent variables), and these predictors may interact with each other to exert non-additive effects. I will present a Bayesian approach to search for these interactions. We were motivated by the epistasis detection problem in population-based genetic association studies, i.e., to detect interactions among multiple genetic defects (mutations) that may be causal to a specific complex disease. Aided with MCMC sampling techniques, our Bayesian method can efficiently detect interactions among many thousands of genetic markers.

A related problem is to find patterns or associations in text documents, such as the "market basket" problem, in which one attempts to mine association rules among the items in a supermarket based on customers' transaction records. For this problem, we formulate our goal as to find subsets of words that tend to co-occur in a sentence. We call the set of co-occurring words (not necessarily orderly) a "theme" or a "module". I will illustrate a simple generative model and the EM algorithm for inferring the themes. I will demonstrate its applications in a few examples including an analysis of chinese medicine prescriptions and an analysis of a chinese novel.

Community structure in large social and information networks

Michael W. Mahoney, Yahoo! Research, Silicon Valley

The concept of a community is central to social network analysis, and thus a large body of work has been devoted to identifying community structure. For example, a community may be thought of as a set of web pages on related topics, a set of people who share common interests, or more generally as a set of nodes in a network more similar amongst themselves than with the remainder of the network. Motivated by difficulties we experienced at actually finding meaningful communities in large real-world networks, we have performed a large scale analysis of a wide range of social and information networks. Our main methodology uses local spectral methods and involves computing isoperimetric properties of the networks at various size scales — a novel application of ideas from scientific computation to internet data analysis. Our empirical results suggest a significantly more refined picture of community structure than has been appreciated previously. Our most striking finding is that in nearly every network dataset we examined, we observe tight but almost trivial communities at very small size scales, and at larger size scales, the best possible communities gradually "blend in" with the rest of the network and thus become less "community-like." This behavior is not explained, even at a qualitative level, by any of the commonly-used network generation models. Moreover, this behavior is exactly the opposite of what one would expect based on experience with and intuition from expander graphs, from graphs that are well-embeddable in a low-dimensional structure, and from small social networks that have served as testbeds of community detection algorithms. Possible mechanisms for reproducing our empirical observations will be discussed, as will implications of these findings for clustering, classification, and more general data analysis in modern large social and information networks.

Classification using intersection kernel SVMs is efficient

Jitendra Malik, University of California, Berkeley

Straightforward classification using (non-linear) kernelized SVMs requires evaluating the kernel for a test vector and each of the support vectors. This can be prohibitively expensive, particularly in image recognition applications, where one might be trying to search for an object at any of a number of possible scales and locations. We show that for a large class of kernels, namely those based on histogram intersection and its variants, one can build classifiers with run time complexity logarithmic in the number of support vectors (as opposed to linear in the standard case). By precomputing auxiliary tables, we can construct approximations with constant runtime and space requirements , independent of the number of SV's. These improvements lead to up to 3 orders of magnitude speedup in classification time and up to 2 orders of magnitude memory savings on various classification tasks. We also obtain the state of the art results on pedestrian classification datasets.

This is joint work with Subhransu Maji and Alexander Berg. A paper as well as source code is available at `http://www.cs.berkeley.edu/~smaji/projects/fiksvm/`

Models and algorithms for complex networks, with network elements maintaining characteristic profiles

Milena Mihail, Georgia Institute of Technology

We will discuss new trends in the study of models and algorithms for complex networks. The common theme is to enhance network elements, say nodes, with a small list of attributes describing the profile of the node relative to other nodes of the network. Indeed, this is common practice in real complex networks.

Standard models for complex networks have captured sparse graphs with heavy tailed statistics and/or the small world phenomenon. We will discuss random dot product graphs, which is a model generating networks with a much wider range of properties. These properties include varying degree distributions, varying graph densities, and varying spectral decompositions. The flexibility of this model follows by representing nodes as d-dim vectors (one dimension for each relevant attribute) sampled from natural general distributions, and links added with probabilities proportional to a similarity function over vectors in high dimensions.

Concerning algorithms for complex networks, we argue that endowing local network elements with (a computationally reasonable amount of) global information can facilitate fundamental algorithmic primitives. For example, we will discuss how spectral representations (in particular, a network node being aware of its value according to a principal eigenvector) can facilitate searching, information propagation and information retrieval. However, computing such eigenvector components locally in a distributed, asynchronous network is a challenging open problem.

Manifold regularization and semi-supervised learning

Partha Niyogi, University of Chicago

Increasingly, we face machine learning problems in very high dimensional spaces. We proceed with the intuition that although natural data lives in very high dimensions, they have relatively few degrees of freedom. One way to formalize this intuition is to model the data as lying on or near a low dimensional manifold embedded in the high dimensional space. This point of view leads to a new class of algorithms that are "manifold motivated" and a new set of theoretical questions that surround their analysis. A central construction in these algorithms is a graph or simplicial complex that is data-derived and we will relate the geometry of these to the geometry of the underlying manifold. We will develop the idea of manifold regularization, its applications to semi-supervised learning, and the theoretical guarantees that may be provided in some settings.

Data analysis with graphs

Elizabeth Purdom, University of California, Berkeley

Graphs or networks are common ways of depicting information. In biology, for example, many different biological processes are represented by graphs, such as signalling networks or metabolic pathways. This kind of a priori information is a useful supplement to the standard numerical data coming from an experiment. Incorporating the information from these graphs into an analysis of the numerical data is a non-trivial task that is generating increasing interest.

There are many results from graph theory regarding the properties of an adjacency matrix and other closely related matrices. These results suggest jointly analyzing numerical data and a graph by using the spectral decomposition of the adjacency matrix (or certain transformations of it) to find interesting features of the numerical data that also reflect the structure of the graph. The adjacency matrix representation of graphs also underlies similar kernel methods that have been used to jointly analyze graphs and data.

We will briefly discuss these problems and how these ideas can be used in for prediction of novel graph structure as well as graph-based classification.

Sequential algorithms for generating random graphs

Amin Saberi, Stanford University

Random graph generation has been studied extensively as an interesting theoretical problem. It has also become an important tool in a variety of real world applications including detecting motifs in biological networks and simulating networking protocols on the Internet topology. One of the central problems in this context is to generate a uniform sample from the set of simple graphs with a given degree sequence.

The classic method for approaching this problem is the Markov chain Monte Carlo (MCMC) method. However,

current MCMC-based algorithms have large running times, which make them unusable for real-world networks that have tens of thousands of nodes. This has constrained practitioners to use simple heuristics that are non-rigorous and have often led to wrong conclusions.

I will present a technique that leads to almost linear time algorithms for generating random graphs for a range of degree sequences. I will also show how this method can be extended for generating random graphs that exclude certain subgraphs e.g. short cycles.

Our approach is based on sequential importance sampling (SIS) technique that has been recently successful for counting and sampling graphs in practice.

Efficient projection algorithms for learning sparse representations from high dimensional data

Yoram Singer, Google Research, Mountain View

Many machine learning tasks can be cast as constrained optimization problems. The talk focuses on efficient algorithms for learning tasks which are cast as optimization problems subject to L1 and hyper-box constraints. The end result are typically sparse and accurate models. We start with an overview of existing projection algorithms onto the simplex. We then describe a linear time projection for dense input spaces. Last, we describe a new efficient projection algorithm for very high dimensional spaces. We demonstrate the merits of the algorithm in experiments with large scale image and text classification.

More data less work: SVM training in time decreasing with larger data sets

Nathan Srebro, University of Chicago

Traditional runtime analysis of training Support Vector Machines, and indeed most learning methods, shows how the training runtime increases as more training examples are available. Considering the true objective of training, which is to obtain a good predictor, I will argue that training time should be studied as a decreasing, or at least non-increasing, function of training set size. I will then present both theoretical and empirical results demonstrating how a simple stochastic subgradient descent approach for training SVMs indeed displays such monotonic decreasing behavior.

Graph sparsification by effective resistances

Nikhil Srivastava, Yale University

A sparsifier of a graph G is a sparse subgraph H that approximates it in some natural or useful way. Benczur and Karger gave the first sparsifiers in 1996, in which the weight of every cut in H was within a factor of $(1 \pm \epsilon)$ of its weight in G. In this work, we consider a stronger notion of approximation (introduced by Spielman and Teng in 2004) that requires the Laplacian quadratic forms of H and G to be close — specifically, that $x^T L' x = (1 \pm \epsilon) x^T L x$ for all vectors $x \in \mathbb{R}^n$, where L and L' are the Laplacians of G and H respectively. This subsumes the Benczur-Karger notion, which corresponds to the special case of x in $\{0,1\}^n$. It also implies that G and H are similar spectrally, and that L' is a good preconditioner for L.

We show that every graph contains a sparsifier with $O(n \log n)$ edges, and that one can be found in nearly-linear time by randomly sampling each edge of the graph with probability proportional to its effective resistance. A key ingredient in our algorithm is a subroutine of independent interest: a nearly-linear time algorithm that builds a data structure from which we can query the approximate effective resistance between any two vertices in a graph in $O(\log n)$ time.

We conjecture the existence of sparsifiers with $O(n)$ edges, noting that these would generalize the notion of expander graphs, which are constant-degree sparsifiers for the complete graph.

Joint work with Dan Spielman.

Efficient algorithms for matrix column selection

Joel Tropp, California Institute of Technology

A deep result of Bourgain and Tzafriri states that every matrix with unit-norm columns contains a large collection of columns that is exceptionally well conditioned. This talk describes a randomized, polynomial-time algorithm for producing the desired submatrix.

Inferring and encoding graph partitions

Chris Wiggins, Columbia University

Connections among disparate approaches to graph partitioning may be made by reinterpreting the problem as a special case of one of either of two more general and well-studied problems in machine learning: inferring latent variables in a generative model or estimating an (information-theoretic) optimal encoding in rate distortion theory. In either approach, setting in a more general context shows how to unite and generalize a number of approaches. As an encoding problem, we see how heuristic measures such as the normalized cut may be derived from information theoretic quantities. As an inference problem, we see how variational Bayesian methods lead naturally to an efficient and interpretable approach to identifying "communities" in graphs as well as revealing the natural scale (or number of communities) via Bayesian approaches to complexity control.

Topological methods for exploring pathway analysis in complex biomolecular folding

Yuan Yao, Stanford University

We develop a computational approach to explore the relatively low populated transition or intermediate states in biomolecular folding pathways, based on a topological data analysis tool, Mapper, with simulation data from large-scale distributed computing. Characterization of these transient

intermediate states is crucial for the description of biomolecular folding pathways, which is however difficult in both experiments and computer simulations. We are motivated by recent debates over the existence of multiple intermediates even in simple systems like RNA hairpins. With a proper design of conditional density filter functions and clustering on their level sets, we are able to provide structural evidence on the multiple intermediate states. The method is effective in dealing with high degree of heterogeneity in distribution, being less sensitive to the distance metric than geometric embedding methods, and capturing structural features in multiple pathways. It is a reminiscence of the classical Morse theory from mathematics, efficiently capturing topological features of high dimensional data sets.

An adaptive forward/backward greedy algorithm for learning sparse representations

Tong Zhang, Rutgers University

Consider linear least squares regression where the target function is a sparse linear combination of a set of basis functions. We are interested in the problem of identifying those basis functions with non-zero coefficients and reconstructing the target function from noisy observations. This problem is NP-hard. Two heuristics that are widely used in practice are forward and backward greedy algorithms. First, we show that neither idea is adequate. Second, we propose a novel combination that is based on the forward greedy algorithm but takes backward steps adaptively whenever necessary. We prove strong theoretical results showing that this procedure is effective in learning sparse representations. Experimental results support our theory.

Abstracts of posters

An improved approximation algorithm for the column subset selection problem

Christos Boutsidis, Rensselaer Polytechnic Institute

Given an $m \times n$ matrix A and an integer k, with $k \ll n$, the Column Subset Selection Problem (CSSP) is to select k columns of A to form an $m \times k$ matrix C that minimizes the residual $\|A - CC^+A\|_\xi$, for $\xi = 2$ or F. Here C^+ denotes the pseudoinverse of the matrix C. This combinatorial optimization problem has been exhaustively studied in the numerical linear algebra and the theoretical computer science communities. Current state-of-the-art approximation algorithms guarantee

$$\|A - CC^+A\|_2 \leq O(k^{1/2}(n-k)^{1/2})\|A - A_k\|_2,$$
$$\|A - CC^+A\|_F \leq \sqrt{(k+1)!}\|A - A_k\|_F.$$

Here, we present a new approximation algorithm which guarantees that with high probability

$$\|A - CC^+A\|_2 \leq O(k^{3/4}\log^{1/2}(k)(n-k)^{1/4})\|A - A_k\|_2,$$
$$\|A - CC^+A\|_F \leq O(k\sqrt{\log k})\|A - A_k\|_F,$$

thus *asymptotically* improves upon the previous results. A_k is the best-rank-k approximation to A, computed with the SVD. Also note that if $A = U\Sigma V^T$ is the SVD of a matrix A, then $A^+ = V\Sigma^{-1}U^T$. We also present an example on how this algorithm can be utilized for a low rank approximation of a data matrix.

This is joint work with Michael Mahoney and Petros Drineas

Machine learning approach to generalized graph pattern mining

Pavel Dmitriev, Yahoo! Inc.

There has been a lot of recent interest in mining patterns from graphs. Often, the exact structure of the patterns of interest is not known. This happens, for example, when molecular structures are mined to discover fragments useful as features in chemical compound classification task, or when web sites are mined to discover sets of web pages representing logical documents. Such patterns are often generated from a few small subgraphs (cores), according to certain generalization rules (GRs). We call such patterns "generalized patterns"(GPs). While being structurally slightly different, GPs often perform the same function in the network.

Previously proposed approaches to mining GPs either assumed that the cores and the GRs are given, or that all interesting GPs are frequent. These are strong assumptions, which often do not hold in practical applications. In this paper, we propose an approach to mining GPs that is free from the above assumptions. Given a small number of GPs selected by the user, our algorithm discovers all GPs similar to the user examples. First, a machine learning-style approach is used to find the cores. Second, generalizations of the cores in the graph are computed to identify GPs. Evaluation on synthetic and real-world datasets demonstrates the promise as well as highlights some problems with our approach.

An experimental comparison of randomized methods for low-rank matrix approximation

Charles Elkan, University of California, San Diego

Many data mining and knowledge discovery applications that use matrices focus on low-rank approximations, including for example latent semantic indexing (LSI). The singular value decomposition (SVD) is one of the most popular low-rank matrix approximation methods. However, it is often considered intractable for applications involving massive data. Recent research has addressed this problem, with several randomized methods proposed to compute the SVD. We review these techniques and present the first empirical study of some of them. The projection-based approach of Sarlos appears best in terms of accuracy and runtime, although a sampling approach of Drineas, Kannan, and Mahoney also performs favorably. No method offers major savings when the input matrix is sparse. Since most massive matrices used in knowledge discovery, for example word-document matrices, are sparse, this finding reveals an important direction for future research.

This is joint work with Aditya Krishna Menon, University of California, San Diego.

Estimating PageRank on graph streams

Sreenivas Gollapudi, Microsoft Research, Silicon Valley

This study focuses on computations on large graphs (e.g., the web-graph) where the edges of the graph are presented as a stream. The objective in the streaming model is to use small amount of memory (preferably sub-linear in the number of nodes n) and a few passes.

In the streaming model, we show how to perform several graph computations including estimating the probability distribution after a random walk of length l, mixing time, and the conductance. We estimate the mixing time M of a random walk in $\widetilde{O}(n\alpha + M\alpha\sqrt{n} + \sqrt{Mn/\alpha})$ space and $\widetilde{O}(\sqrt{M/\alpha})$ passes. Furthermore, the relation between mixing time and conductance gives us an estimate for the conductance of the graph. By applying our algorithm for computing probability distribution on the web-graph, we can estimate the *PageRank* p of any node up to an additive error of $\sqrt{\epsilon p}$ in $\widetilde{O}(\sqrt{M/\alpha})$ passes and $\widetilde{O}\Big(\min\Big\{n\alpha + \frac{1}{\epsilon}\sqrt{M/\alpha} + \frac{1}{\epsilon}M\alpha, \alpha n\sqrt{M\alpha} + \frac{1}{\epsilon}\sqrt{\frac{M}{\alpha}}\Big\}\Big)$ space, for any $\alpha \in (0, 1]$. In particular, for $\epsilon = M/n$, by setting $\alpha = M^{-1/2}$, we can compute the approximate PageRank values in $\tilde{O}(nM^{-1/4})$ space and $\widetilde{O}(M^{3/4})$ passes. In comparison, a standard implementation of the PageRank algorithm will take $O(n)$ space and $O(M)$ passes.

Uniform generation and pattern summarization

Mohammad Al Hasan, Rensselaer Polytechnic Institute

Frequent Pattern Mining (FPM) is a mature topic in data mining with many efficient algorithms to mine patterns with variable complexities, such as set, sequence, tree, graph, and etc. But, the usage of FPM in real-life knowledge discovery systems is considerably low in comparison to their potentital. The prime reason is the lack of interpretability caused from the enormity of output-set size. A modest size graph database with thousands graphs can produce millions of frequent graph patterns with a reasonable support value. So, the recent research direction in FPM has shifted from obtaining the total set to generate a representative (summary) set. Most of the summmarization approaches that is proposed recently are post-processor; they adopt some form of clustering algorithm to generate a smaller representative pattern-set from the collection of total set of frequent patterns (FP). But, post-processing requires to obtain all the members of FP; which, unfortunately, is not practically feasible for many real-world data set.

In this research, we consider an alternative viewpoint of representativeness. It is based on uniform and identical sampling, the most fundamental theory in data analysis and statistics. Our representative patterns are obtained with uniform probability from the pool of all frequent maximal patterns. Along this idea, we first discuss a hypothetic algorithm for uniform generation that uses a #P-oracle. However, such an oracle is difficult to obtain as it is related to the counting of maximal frequent patterns, which was shown to be #P-complete. Hence, we propose a random-walk based approach that traverse the FP partial order randomly with prescribed transition matrix. On convergence, the random walk visits each maximal pattern with a uniform probability. Since, the number of maximal patterns, generally, are very low compared to the search space, the idea of importance sampling is used to guide the random walk to visit the maximal patterns more frequently.

Approximate support vector machines for distributed system

Ling Huang, Intel Research

The computational and/or communication constraints associated with processing large-scale data sets using support vector machines (SVM) in contexts such as distributed networking systems are often prohibitively high, resulting in practitioners of SVM learning algorithms having to apply the algorithm on approximate versions of the kernel matrix induced by a certain degree of data reduction. In this research project, we study the tradeoffs between data reduction and the loss in an algorithm's classification performance. We introduce and analyze a consistent estimator of the SVM's achieved classification error, and then derive approximate upper bounds on the perturbation on our estimator. The bound is shown to be empirically tight in a wide range of domains, making it practical for the practitioner to determine the amount of data reduction given a permissible loss in the classification performance.

Convergence of protein folding free energy landscape via application of generalized ensemble sampling methods on a distributed computing environment

Xuhui Huang, Stanford University

The Replica Exchange Method (REM) and Simulated Tempering (ST) enhanced sampling algorithms have become widely used in sampling conformation space of bimolecular systems. We have implemented a serial version of REM (SREM) and ST in the heterogeneous Folding@home distributed computing environment in order to calculate folding free energy landscapes. We have applied both algorithms to the 21 residue Fs peptide, and SREM to a 23 residue BBA5 protein. For each system, we find converged folding free energy landscapes for simulations started from near.native and fully unfolded states. We give a detailed comparison of SREM and ST and find that they give equivalent results in reasonable agreement with experimental data. Such accuracy is made possible by the massive parallelism provided by Folding@home, which allowed us to run approximately five thousand 100ns simulations for each system with each algorithm, and generates more than a million configurations. Our extensive sampling shows that the AMBER03 force field gives better agreement with experimental results than previous versions of the force field.

Metric/kernel learning with applications to fast similarity search

Prateek Jain, University of Texas, Austin

Distance and kernel learning are increasingly important for several machine learning applications. However, most existing distance metric algorithms are limited to learning metrics over low-dimensional data, while existing kernel learning algorithms are limited to the transductive setting and thus do not generalize to new data points. In this work, we focus on the Log Determinant loss due to its computational and modelling properties, and describe an algorithm for distance and kernel learning that is simple to implement, scales to very large data sets, and outperforms existing techniques. Unlike most previous methods, our techniques can generalize to new points, and unlike most previous metric learning methods, they can work with high-dimensional data. We demonstrate our learning approach applied to large-scale problems in computer vision, specifically image search.

Random indexing: simple projections for accumulating sparsely distributed frequencies

Pentti Kanerva, Stanford University

We have applied a simple form of random projections, called Random Indexing, to English text, to encode the contexts of words into high-dimensional vectors. The context vectors for words can be used in natural-language research

and text search, for example. Random Indexing is particularly suited for massive data that keeps on accumulating—it is incremental. It compresses a large sparse $M \times N$ matrix, with M and N growing into the billions, into a much smaller matrix of fixed size with little information loss. Possible applications include network connectivity analysis: who is talking to whom?

DiscLDA: discriminative learning for dimensionality reduction and classification

Simon Lacoste-Julien, University of California, Berkeley

Probabilistic topic models (and their extensions) have become popular as models of latent structures in collections of text documents or images. These models are usually treated as generative models and trained using maximum likelihood estimation, an approach which may be suboptimal in the context of an overall classification problem. In this work, we describe DiscLDA, a discriminative learning framework for such models as Latent Dirichlet Allocation (LDA) in the setting of dimensionality reduction with supervised side information. In DiscLDA, a class-dependent linear transformation is introduced on the topic mixture proportions. This parameter is estimated by maximizing the conditional likelihood using Monte Carlo EM. By using the transformed topic mixture proportions as a new representation of documents, we obtain a supervised dimensionality reduction algorithm that uncovers the latent structure in a document collection while preserving predictive power for the task of classification. We compare the predictive power of the latent structure of DiscLDA with unsupervised LDA on the 20 Newsgroup document classification task.

This is joint work with Fei Sha and Michael Jordan.

Discrete denoising with shifts

Taesup Moon, Stanford University

We introduce S-DUDE, a new algorithm for denoising DMC-corruputed data. The algorithm extends the recently introduced DUDE (Discrete Universal DEnoiser) of Weissman et. al., and aims to compete with a genie that has access to both noisy and clean data, and can choose to switch, up to m times, between sliding window denoisers in a way that minimizes the overall loss. The key issue is to learn the best segmentation of data and the associated denoisers of the genie, only based on the noisy data. Unlike the DUDE, we treat cases of one- and two-dimensional data separately due to the fact that the segmentation of two-dimensional data is significantly more challenging than that of one-dimensional data. We use dynamic programming and quadtree decomposition in devising our scheme for one- and two-dimensional data, respectively, and show that, regardless of the underlying clean data, our scheme achieves the performance of the genie provided that m is sub-linear in the sequence length n for one-dimensional data and $m \ln m$ is sub-linear in n for two-dimensional data. We also prove the universal optimality of our scheme for the case where the underlying data are (spatially) stationary. The resulting complexity of our scheme is linear in both n and m for one-dimensional data, and linear in n and exponential in m for two-dimensional data. We also provide a sub-optimal scheme for two-dimensional data that has linear complexity in both n and m, and present some promising experimental results involving this scheme.

Joint work with Professor Tsachy Weissman (Stanford).

Joint latent topic models for text and citations

Ramesh Nallapati, Stanford University

In this work, we address the problem of joint modeling of text and citations in the topic modeling framework. We present two different models called the Pairwise-Link-LDA and the Link-PLSA-LDA models.

The Pairwise-Link-LDA model combines the ideas of LDA and Mixed Membership Block Stochastic Models and allows modeling arbitrary link structure. However, the model is computationally expensive, since it involves modeling the presence or absence of a citation (link) between every pair of documents. The second model solves this problem by assuming that the link structure is a bipartite graph. As the name indicates, Link-PLSA-LDA model combines the LDA and PLSA models into a single graphical model.

Our experiments on a subset of Citeseer data show that both these models are able to predict unseen data better than the base-line model of Erosheva and Lafferty, by capturing the notion of topical similarity between the contents of the cited and citing documents.

Our experiments on two different data sets on the link prediction task show that the Link-PLSA-LDA model performs the best on the citation prediction task, while also remaining highly scalable. In addition, we also present some interesting visualizations generated by each of the models.

Annealed entropy, heavy tailed data and manifold learning

Hariharan Narayanan, University of Chicago

In many applications, the data that requires classification is very high dimensional. In the absence of additional structure, this is a hard problem. We show that the task of classification can tractable in two special cases of interest, namely **1.** when data lies on a low dimensional submanifold on which it has a bounded density, and **2.** when the data has a certain heavy tailed character.

More precisely, we consider the following questions: **1.** What is the sample complexity of classifying data on a manifold when the class boundary is smooth? **2.** What is the sample complexity of classifying data with Heavy-Tailed characteristics in a Hilbert Space H using a linear separator?

In both cases, the VC dimension of the target function class is infinite, and so distribution free learning is impossible. In the first case, we provide sample complexity bounds that depend on the maximum density, curvature parameters

and intrinsic dimension but are independent of the ambient dimension. In the second case, we obtain dimension independent bounds if there exists a basis $e_1, e_2, \ldots$ of H such that the probability that a random data point does not lie in the span of $e_1, \ldots, e_k$ is bounded above by Ck^{-a} for some positive constants C and a.

Cases **1** and **2** are based on joint work with Partha Niyogi and Michael Mahoney, respectively.

Algorithmic data analysis with polytopes: a combinatorial approach

Stefan Pickl, Naval Postgraduate School, Monterey and Bundeswehr University, Munich

A research area of central importance in computational biology, biotechnology — and life sciences at all — is devoted to modeling, prediction and dynamics of complex expression patterns. However, as clearly understood in these days, this enterprise cannot be investigated in a satisfying way without taking into account the role of the environment in its widest sense: Complex Data Analysis, Statistical Behaviour and Algorithmic Approaches are in the Main Center of Interest: To a representation of past, present and most likely future states, this contribution also acknowledge the presence of measurement errors and further uncertainties.

This poster contribution surveys and improves recent advances in understanding the mathematical foundations and interdisciplinary implications of the newly introduced gene-environment networks. Moreover, it integrates the important theme carbon dioxide emission reduction into the context of our networks and their dynamics. Given the data from DNA microarray experiments and environmental records, we extract nonlinear ordinary differential equations which contain parameters that have to be determined. This is done by some modern kinds of (so-called generalized Chebychev) approximation and (so-called generalized semi-infinite) optimization. After this is provided, time-discretized dynamical systems are studied. Here, a combinatorial algorithm constructing and following polyhedra sequences, allows to detect the region of parametric (in)stability. Finally, we analyze the topological landscape of gene-environment networks in its structural stability. With the example of CO_2-emission control and some further statistical perspectives we conclude.

Global and local intrinsic symmetry detection

Jian Sun, Stanford University

Within the general framework of analyzing the properties of shapes which are independent of the shape Rs pose, or embedding, we have developed a novel method for efficiently computing global symmetries of a shape which are invariant up to isometry preserving transformations. Our approach is based on the observation that the intrinsic symmetries of a shape are transformed into the Euclidean symmetries in the signature space defined by the eigenfunctions of the Laplace-Beltrami operator. We devise an algorithm which detects and computes the isometric mappings from the shape onto itself.

The aforementioned method works very well when the whole object is intrinsically symmetric. However, in practice objects are often partially symmetric. To tackle this problem, we have developed an algorithm that detects partial intrinsic self-similarities of shapes. Our method models the heat flow on the object and uses the heat distribution to characterize points and regions on the shape. Using this method we are able to mathematically quantify the similarity between points at certain scale, and determine, in particular, at which scale the point or region becomes unique. This method can be used in data analysis and processing, visualization and compression.

FastBit: an efficient indexing tool for massive data

John Wu, Lawrence Berkeley National Lab

FastBit is a software tool for searching large read-only datasets. It organizes user data in a column-oriented structure which is efficient for on-line analytical processing (OLAP), and utilizes compressed bitmap indexes to further speed up query processing. We have proven the compressed bitmap index used in FastBit to be theoretically optimal for one-dimensional queries. Compared with other optimal indexing methods, bitmap indexes are superior because they can be efficiently combined to answer multi-dimensional queries whereas other optimal methods can not. In this poster, we show that FastBit answers queries 10 times faster than a commercial database system, and highlight an application where FastBit sped up the molecular docking software hundreds of times.

Dirichlet compound multinomial models for retrieval and active relevance feedback

Zuobing Xu, University of California, Santa Clara

We describe new models, including those based on the Dirichlet Compound Multinomial (DCM) distribution and for information retrieval, relevance feedback, and active learning. We first indicate how desirable properties such as search engine ranking being concave in word repetition can be obtained through the use of the DCM model. We also describe several effective estimation methods. Reduction of the state space is achieved through the use of relevance and non-relevance sets. Effective capture of negative feedback and modeling of overlap terms between negative and positive feedback (relevant) documents is thus enabled. We conclude with a discussion of the computation efficiency and demonstration of the very good performance of these new methods on TREC data sets. This work also enables the efficient use of the original probabilistic ranking approach, while incorporating estimation methods developed in the language model framework.

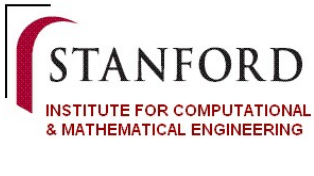
This is joint work with Ramakrishna Akella, University of California, Santa Cruz.

Acknowledgements

Sponsors

The organizers thank the following institutional sponsors for their generous support:

- iCME: the Institute for Computational and Mathematical Engineering, Stanford University
- Department of Mathematics, Stanford University
- Department of Mathematics, University of California, Berkeley
- Stanford Computer Forum
- PIMS: the Pacific Institute for the Mathematical Sciences
- NSF: the National Science Foundation
- DARPA: the Defense Advanced Research Projects Agency
- LinkedIn Inc.
- Yahoo! Inc.



The organizers thank the following persons for their kind assistance:

Events and meeting planning: Victor Olmo, Mayita Romero

Finance: Lisa Ewan, Debbie Lemos

Organization: Petros Drineas

Program design: Sou-Cheng Choi, Michael Saunders

Publicity: Suzanne Bigas

Web: Kuan-Chuen Wu