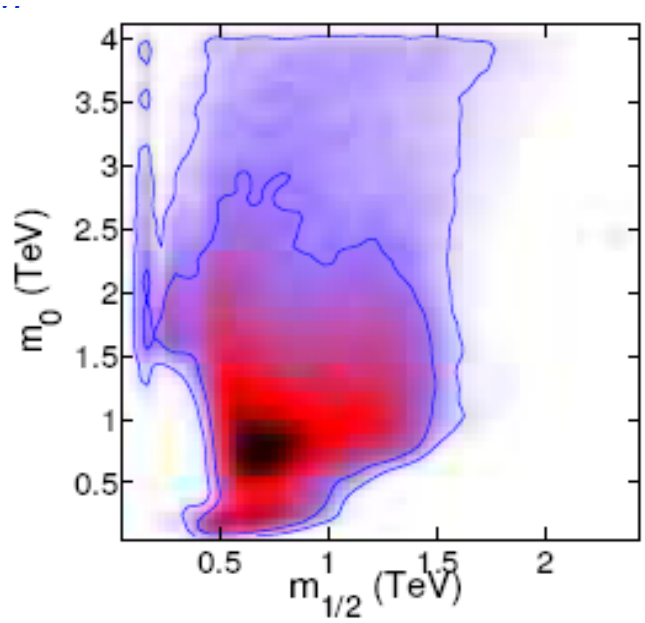
Discovery in Complex or Massive Datasets: Common Statistical Themes

A Workshop funded by the National Science Foundation

October 16-17, 2007



L. Roszkowski, PhyStat 2007

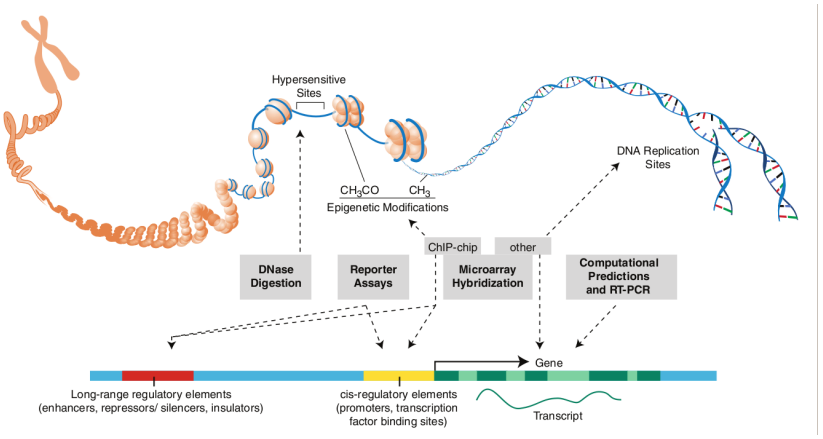


Figure 1: Top: Bayes meets SUSY: A projection to 2 dimensions of a 5-dimensional posterior density, computed for a 5-dimensional model of possible physics beyond the current Standard Model of particle physics. Bottom: Complex data types needed to predict function in the human genome (fragment shows $\geq 10^6$ basepairs.)

Executive Summary

We report on a workshop, "Discovery in Complex or Massive Data Sets: Common Statistical Themes", held in Washington, October 16-17, 2007, funded by NSF's Division of Mathematical Sciences. We connect with a later workshop, "Data Enabled Science in the Mathematical and Physical Sciences" held in Washington, March 29-30, 2010, funded by NSF's Directorate of Mathematical and Physical Sciences.

Research responding to important scientific and societal questions now requires the generation and understanding of vast amounts of often highly complex data. The 2007 workshop dealt with crosscutting issues arising in the analysis of such data sets with a particular focus on the role of statistical analysis. This was done through selected examples matching scientific and societal interests. In particular there were sessions on:

• *Genomics and* other areas of the *biosciences* that play a key role both in fundamental biology and in our current efforts to cure human diseases.

• *Computer models* with an emphasis on modeling in the atmospheric sciences that plays a critical role in climate change forecasting.

• *Finance, economics, and risk management* focusing on problems of financial and other economic forecasting and also on analysis of the flow of potential new regulatory data.

• *Particle and astrophysics* pointing to a plethora of needs and issues, including scientific questions such as solving massive inverse problems as they arise in the study of dark energy, statistical modeling of galactic filamentary structures, and policy issues such as determining resource allocation among expensive experiments.

• *Network modeling* pointing to an old type of data appearing with new complexity and size from many sources: the Internet, ecological networks, biochemical pathways, etc.

In addition, there were two cross cutting sessions,

• *Sparsity*, which reflects how simply we can represent information, has been recognized as the key feature that the new massive data sets must have for us to analyze them at all. Sparsity figures prominently in compressed sensing, now a major topic as the number and types of detectors and the amount of data they can generate has grown exponentially.

• *Machine Learning* developed in computer science and statistics to integrate computational considerations with data modeling. Methods such as clustering look for sparsity or more generally structure in the data. The field's principles are entirely statistical. Its methods play an important role in speech recognition, document retrieval, web-search, computer vision, bioinformatics, neuroscience, and many other areas.

The activity of the 2007 Workshop foreshadowed in its treatment of analysis the 2010 Data Enabled Science Workshop[1], although the latter examined and gave policy recommendations for all divisions in the directorate, rather than focusing on the nature of the science in one sub discipline. But the same themes came up, with all or most divisional sections stressing the growth in size and complexity of data, interdisciplinary collaboration as key to modern progress, and the need for the development of common large databases for analysis. The use of such existing databases in the biomedical sciences and astrophysics was implicit in the presentations of the 2007 workshop. More broadly, advances in statistics and mathematics will be crucial for developments of DES in other disciplines.

In their respective ways both workshops point to the need to support organization and analysis of our massive and high-dimensional data sets as a key to future advances.

[1] and a 2010 E.U. report "Riding the wave: How Europe can gain from the rising tide of scientific data"

Contents

1 Background

This document is the report of a Workshop on Discovery in Complex or Massive Datasets: Common Statistical Themes, held October 16-17, 2007 in Washington, D.C. The idea and funding for the workshop came from Dr. Peter March, Director of the Divison of Mathematical Sciences (DMS) at the National Science Foundation (NSF).

The impetus for the meeting was the observation that interdisciplinary research in statistics engages with so many fields of science that it is neither possible, nor perhaps appropriate, for DMS to fund all of it, either alone, or through partnerships – though successful examples of the latter certainly exist. At the same time, DMS is the primary disciplinary home for statistics within NSF, and so in particular is the primary locus within the Foundation for workforce develpment efforts in statistics. In such an environment, what ideas might guide DMS in its funding of statistics research?

The workshop and report develop the notion of "intersections" – that part of statistical methods and theory that has, or seems likely to have, impact in multiple scientific domains. The intent for the short workshop was to be illustrative rather than encylopedic. It is not, therefore, a report on the 'future of statistics', and deliberately does not contain formal consensus recommendations. However, we hope that the sampling of research areas in this short report illustrates the existence of these intersectional topics and importance of research into their development.

2 Introduction

The amount and complexity of data generated to support contemporary scientific investigation continues to grow rapidly, following its own type of Moore's Law [11]. In domains from genomics to climate science, statisticians are actively engaged in interdisciplinary research teams. In some areas, automated processes collect and process huge amounts of information; in others simulations of complex systems are designed to generate information

about large scale behavior, and in still other areas, the very sources of data are products of the information age.

There is substantial current activity to develop statistical ideas, methods and software in many of these domains, which include astronomy, genomics, climate science, financial market analysis and sensor networks. Statisticians are engaged in (often large) interdisciplinary teams, and frequently receive significant research support from the relevant scientific discipline.

The history of statistics shows that, while frequently initially arising in response to challenges in specific scientific domains, statistical methods and associated theory often achieve broader success and power by being subsequently applied to subjects far remote from those of origin. Well known examples include the analysis of variance, proportional hazard models and the application of sparsity ideas in signal recovery.

We see enormous opportunity, then, in advancing the study of the "intersections" arising from statistical research in today's Age of Information – statistical problems, theories (including probabilistic models), tools and methods that arise in or are relevant to multiple domains of scientific enquiry, and as such, are moving or should move into the "core".

The workshop aimed to enumerate some of today's most intellectually compelling challenges arising out of these intersections, and was guided by the hope of stimulating future research advances that will extend and enhance our data analytic toolkit for scientific discovery.

In order to have a title that both has some focus, and yet is broadly inclusive, we chose "Discovery from Complex or Massive Datasets: Common Statistical Themes". Here "massive" means large relative to existing capability in some way, including, but not restricted to, many cases (sample size), many variables (dimension), or many datasets (sensor networks).

The workshop took a broad view of research in statistics, and included researchers who may not identify themselves as statisticians yet who feel that advances in statistics are central to advances in science and society.

The body of the report contains short summaries of each of the sessions at the workshop. In this introduction, we illustrate three of the themes with brief paragraphs, indicating in parentheses the sessions in which these themes come up explicitly or implicity. We conclude with some reflections on national needs that will be served by a focus on statistical intersections.

Sparsity. [§3.1, 3.2, 3.3, 3.4, 3.6] A preference for parsimony in scientific theories – captured in principles such as "Occam's razor" – has long influenced statistical modeling and estimation. The size of contemporary datasets and the number of variables collected makes the search for, and exploitation of, sparsity even more important. For example, out of a huge list of proteins or genes, only an (unknown) few may be active in a particular metabolic or disease process, or sharp changes in a generally smooth signal or image may occur at a small number of points or boundaries The sparsity of representation may be "hidden": revealed only with the use of new function systems such as wavelets or curvelets.

The theme of sparsity draws upon and stimulates research in many areas of mathematics, statistics and computing: harmonic analysis and approximation theory (for the development and properties of representations), numerical analysis and scientific computation (the associated algorithms), statistical theory and methods (techniques and properties when applied to noisy data).

Sparsity ideas have recently given birth to a new circle of ideas and technologies known collectively as "Compressed Sensing". It is common experience that many images can

be compressed greatly without significant loss of information. So, why not design a data collection, or sensing, mechanism that need collect only roughly the number of bits required for the compressed representation? It has recently be shown that this can be done in a variety of settings, in which sparsity is present, by a judicious introduction of random sampling.

A number of intellectual trends in mathematics and statistics have pointed toward and culminated in the articulation of the Compressed Sensing phenomenon: approximation theory, geometric functional analysis, random matrices and polytopes, robust statistics and statistical decision theory. Once articulated mathematically, CS has stimulated development of new algorithms in fields ranging from magnetic resonance imaging to analog-to-digital conversion to seismic imaging.

Computer and Simulation-Based Models. [§3.1, 3.3, 3.4, 3.5] Mathematical models intended for computational simulation of complex real-world processes are a crucial ingredient in virtually every field of science, engineering, medicine, and business, and in everyday life as well. Cellular telephones attempt to meet a caller's needs by optimizing a network model that adapts to local data, and people threatened by hurricanes decide whether to stay or flee depending on the predictions of a continuously updated computational model.

Growth in computing power and matching gains in algorithmic speed and accuracy have vastly increased the applicability and reliability of simulation—not only by drastically reducing simulation time, thus permitting solution of larger and larger problems, but also by allowing simulation of previously intractable problems.

The intellectual content of computational modeling comes from a variety of disciplines, including statistics and probability, applied mathematics, operations research, and computer science, and the application areas are remarkably diverse. Despite this diversity of methodology and application, there are a variety of common challenges in developing, evaluating and using complex computer models of processes. In trying to predict reality (with uncertainty bounds), some of the key issues that have arisen are: use of model approximations (emulators) as surrogates for expensive simulators, for calibration/prediction tasks and in optimization or decision support; dealing with high dimensional input spaces; validation and utilization of computer models in situations with very little data, and/or functional (possibly multivariate) outputs; non-homogeneity, including jumps and phase changes as we move around the input space; implementation and transference methodology to current practice; efficient MCMC algorithms and prior assessments; optimization and design.

Clustering. [§3.1, 3.6, 3.7] Clustering is another important core problem in data analysis. It is analogous to sparsity in that (1) it involves statistically-sound methods for reducing the dimensionality of data, and (2) it is a nexus for the research efforts of multiple overlapping communities. One general motivation for clustering is that there are often limitations on resources available for data analysis, an issue that is particularly pertinent for massive data sets. Most statistical algorithms run in time that is at least proportional to the number of data points, and many algorithms run in quadratic or cubic time (e.g., linear regression). In terabyte-size data sets, these algorithms may be infeasible, and the only hope is that the data can be broken into smaller clusters that can be processed separately. Thus clustering can be viewed from a computational point of view as an instance of the computational principle of divide-and-conquer. Similarly, there may be bandwidth limitations in the transport of data, and the branch of information theory concerned with compression provides a foundation for the design of clustering algorithms that allow bandwidth limitations to be

surmounted. Another general motivation for clustering arises in exploratory data analysis, where the goal is to find relatively homogeneous subsets of data (or subsets of variables) that correspond to meaningful entities in some problem domain. Many research communities have pursued such an agenda—notable examples include bioinformatics, astronomy, medicine, psychology, marketing, linguistics and artificial intelligence.

The result of this intense effort has been the development of literally hundreds of specific clustering algorithms, including recent contributions from statistical mechanics, graph theory and error-control coding. Statistics has also contributed many specific algorithms (including the prominent K-means algorithm), but even more importantly, statistics provides a general framework for the evaluation of clustering methods, both from a theoretical point of view and am empirical point of view. Such analyses make it possible to expose the tradeoffs involved in clustering, such that appropriate methods can be chosen for specific problems.

Another key contribution of statistics is to provide methods for assessing uncertainty in clustering. As in any area of inference, it is essential to assess uncertainty in order to be able to make statements about whether a phenomenon is likely to be meaningful or could have arisen solely by chance. Assessments of uncertainty are also necessary in order to compare multiple competing models. In fact, one useful view of clustering is as a collection of statistical models, one model for each cluster, where the problem is to decide which models account for which data points. This brings clustering into contact with general problems of model selection and model averaging, areas that have been very active in statistics in recent years. Finally, though, clustering is the area where statistics and subject matter science meet most intimately since the ultimate validation of a cluster has to be science based.

These three common themes are but examples. Other examples of crosscutting themes might include data complexity, modelling at multiple scales, and the tradeoff between computational and optimality considerations.

General Remarks: Complexity and massive data sets go together: enormous numbers of huge vectors of categorical data as in the output of the second-generation sequencing machines in genomics, images in fields ranging from physics, for instance, particle tracks, galactic filaments to neuroscience, for instance CAT scans, to spacetime fields in atmospheric sciences, to general graphical structure in network models. All these types of data are found in combined form creating higher levels ad infinitem.

Statisticians have a long history of modeling complex data types, for instance the proportional hazards model with time varying covariates which has established itself in epidemiology, but the challenges of the current types of data are unprecedented. This theme was stressed in each of the sessions. A consequence is an imperative for statisticians to to work closely with other applied mathematicians and computer scientists with their expertise in data structures and numerical stability of algorithms and also pure mathematicians who have long studied abstract structures for their own sake. A second imperative, again apparent in all sessions, is the need to work closely in the development of methods and models with specialists in substantive fields of interest. Methods can become generic in many fields only if they have proved their success in some field of application.

Another consequence of the data explosion, also apparent in many of the sessions, that has direct impact on theory as well as practice is the need to consider computational efficiency as well as statistical optimality in the construction of new methods.

Comments on national needs: Support for research on "intersectional" topics will advance the capability of statistical theories and methods to contribute to contemporary

challenges of discovery from massive and/or complex datasets, thus enhancing the nation's "methodologic infrastructure" for research.

In conjuction with support for collaboration (and appropriate joint training with specialists), research on 'intersections' offers the prospect of fostering the flow of ideas between disciplines, as analysis methods developed in one domain are transferred in other areas.

The vitality of both these enterprises is and will remain an important component of maintaining the competitiveness, and indeed, the leading character of the U.S. scientific research effort.

Turning to workforce issues, we are in an era of expansion in graduate and undergraduate programs in statistics nationally, while at the same time retirements of faculty hired in the 1960's and 1970's are accelerating. Interdisciplinary research is mostly (though of course not always) done by younger and mid-career faculty. Thus many retirements will deplete the strength of graduate programs in 'core' statistical theory and methods. Support for research on statistical intersections will enhance the pre- and post- doctoral level training of research statisticians who will be critically needed as replacement core faculty members in expanding statistics programs around the country.

There is a second aspect of workforce development worth comment. Many statistics faculty will obtain research support for their cross-disciplinary research from agencies focused on a specific discipline. When such research yields methods or theory of potentially broader scientific utility (an 'intersection'), that funding agency may well not regard support for research into realizing that potential as part of its mission. Yet it is precisely such research that may help early career statistics faculty receive the kind of broader recognition within the statistics community that will help with promotion and career advancement.

3 Session Summaries

3.1 Statistics in Biological and Health Sciences

The biosciences, particularly molecular biology and the fields it has spawned, provide an almost paradigmatic view of the statistical themes that have emerged with the advent of complex, massive data sets.

Over the last several years, biological research has undergone a major transformation as scientists have been assimilating the implications of the genetic revolution, developing new technology for high-throughput genotyping and characterization of the activity of genes, messenger RNAs, and proteins, and studying interplay of genes and the environment, and genes and clinical treatments, in causing human diseases. Technological platforms have advanced to a stage where many biological entities, e.g., genes, transcripts, proteins, lipids and sugars, can be measured on the whole genome scale, yielding massive high-throughput "omics" data, such as genomic, epigenomic, proteomic and metabolomic data. These massive datasets need to be analyzed using biologically meaningful and computationally efficient statistical and computational models with the goals of understanding the mechanisms of experimental and biological systems and to study the associations of genetic and environmental factors and disease phenotypes.

The sequencing of the human genome with its 3 billion basepairs was a landmark preceded by the sequencing of the yeast genome and followed up to the present time by genomes of multiple species. The rapid advance of next generation sequencing technology makes genome-wide sequencing of a large number of subjects feasible in the next few years. The genomes are enormous instruction manuals for producing the complex organisms of life.

To try to determine functions, an enormous variety of data are being generated, such as expression of mRNAs by genes, binding sites of proteins produced by genes, images of hundreds to thousands of individual cells with intensities of various processes measured by fluorescence, sequencing machines and alignment methods etc. The data is being generated by international consortia such as ENCODE (Encylopedia of DNA) for general function, and the 1000 Genomes project for analysis of human variation, accumulating at the rate of terabytes in specialized databases. The high level goal of this enormous activity is to "annotate" genomes, that is, identify and ascribe "function" to the "words," "sentences," "paragraphs," and "topics" of genomes by relating them to each other, to the machinery of the cell, the proteins that they produce and so on up the ladder of complexity. A reference for more details of these activities is [1]. A more immediate goal is to investigate their associations with disease phenotypes [9].

The explosion of information about the human genome presents extraordinary challenges in data processing, integration and analysis. These challenges include (1) manipulating and analyzing high-dimensional "omics" data using advanced and efficient statistical and computational methods that are biologically meaningful; (2) integrating "omics" data from different sources for data analysis and result interpretation; (3) Conducting interdisciplinary research to help synthesize massive existing and rapidly increasing molecular and genetic information to understand biological systems; (4) Developing innovative study designs and analysis and computational tools to study the interaction between genes (nature) and environment (nurture) and hence understand disease etiology and develop effective new disease prevention and intervention strategies. These challenges provide enormous opportunities for statisticians, both in new directions of research and training and call for urgent response as a community.

Data Integration: Biological systems and causes of diseases are complex and are affected by a spectrum of genes and gene products and environmental factors. Different types of data, have to be integrated to understand biological systems and disease processes. For example, genome-wide association studies (GWAS) provide a powerful tool to genotype hundreds of thousands of common genetic variants across the whole genome to study their roles and interactions with the environment in causing human diseases. In addition to the development of statistical and epidemiologic methods for the integration of high dimensional mixed clinical and experimental data, integration with GWAS and other omic information, such as pathways, transcriptions, epigenetics and gene expression and regulatory mechanisms for expression is important. Many public genetic databases have been available rapidly, such as HapMap, UCSC genome browser, dbGaP, Ensemble. Data sharing is now mandated for most of these databases. All these data resources and scientific needs make data integration critical. Statisticians need to play a pivotal role in this endeavor and to incorporate this information in statistical modeling and analysis.

Such integration has long been a part of statistics. Categorical data are combined with numerical data through logistic regression, vectors of numerical data can be combined via canonical correlation analysis and so on. But what makes all this a new enterprise is the complexity, high dimensionality and size of the data and the inability to make real contributions without deep knowledge of or close collaboration with specialists versed in the relevant new biology. There is certainly a very important "computer science" aspect of manipulating data but representation (modeling) and analysis have to be statistical, and insertion of knowledge of biology and genetics is critical. The trade-off between computational effort and power of analysis is also a major challenge which has to be faced.

Stochastic Mechanistic Modeling. Another feature of the new technologies is that they enable us to make measurements of complex quantities dynamically in time at different scales and sometimes at heretofore- unavailable nanoscale resolution. These data call for mixtures of modeling at different levels, purely stochastic at the nanoscale level merging with high dimensional Markov modeling, merging with dynamical systems modeling. An example of mechanistic stochastic modelling at nanoscales explaining phenomena at large scales is [6].

More generally, biological and many other (e.g. financial, geophysical) processes operate at several scales. This is clearly seen in the development of the embryo where stages are physically identified but necessarily correspond to the cellular evolution and interactions of many proteins in many cells. Modeling changes of regime, ("emergent phenomena"), at a coarser scale arising from fine scales, when responses are highly multidimensional is a major and novel challenge faced in modern statistics. This activity links naturally with applied mathematics and our discussion of computer models.

Sparsity: As discussed earlier, biological processes and the data we gather on them involve the temporal evolution of huge numbers of genes and gene products, such as mRNA and proteins in different cell environments. It has become widely accepted in biology that this great complexity is built up out of a relatively small number of building blocks, subunits of proteins, circuits in cellular processes, etc.

For understanding we need to find an appropriate alphabet of parts and procedures out of the mess of different types of high dimensional indirect measurements that we have. This theme of sparsity is pervasive. We hope that for complex diseases, although many different types of mutations may lead to the disease, a relatively small number of genes and pathways are involved.

Redundancy: Another aspect particularly important in biological systems is redundancy – if a pathway fails in most cases there should be an alternative to take up the slack. This can sometimes be interpreted as low dimensionality, "covariates" or "collinearity" of predictors.

Causality and Perturbation Experiments: A weakness of purely statistical analysis of data in the biological context is the lack of ground truth when checking statistical predictions. This can be mitigated through perturbation experiments, e.g. knock out of regions predicted to have functional importance with the expectation of evidence of causation. Unfortunately, masking effects are all too frequent in view of redundancy. One can expect that causal inference, Bayesian networks (i.e. conditional independence models) will be of value here.

Training Issues: Manipulating large data bases analyzing high throughput data successfully requires computational skills which are not typically in the province of statisticians. However these are we view, essential since they are required to: 1) understand the data 2) develop methods which have an impact on the science 3) enable one to make the tradeoff between computational and statistical efficiency needed to make serious progress.

As in most fields, sufficient training in the science in order to communicate with understanding with specialists and generally is of great value.

3.2 Sparsity: Compressed Sensing Using High-Dimensional Thinking

The Shannon sampling theorem is a fundamental tool underlying our modern media-rich era. This theorem prescribes a hard constraint that designers of scientific and engineering systems use daily in designing sensors and data acquisition protocols. For example,

Magnetic Resonance Imaging scanners, now frequently used in medical practice, take in many cases an hour or so to collect enough data to render an image of the inside of the human body. Ultimately the scan takes so long because straightforward calculations using Shannon's theorem suggest that millions of measurements must be made in order to obtain a reconstructed image.

Recently, it has become clear in a number of scientific fields that the Shannon limit, although honored almost universally as a fundamental constraint on data acquisition, can actually be circumvented in some fairly important settings. Thus, one can, in the right setting, 'undersample' – violate the Shannon limit substantially – and still obtain high-quality reconstructions by the right method. As a simple example, at the Society for Magnetic Resonance in Medicine meeting in Berlin, May 2006, results were presented by several teams showing that certain categories of MR imaging tasks could be sped up by factors of 7 over what had previously been considered necessary. This means 7 times as many patients can be served by a facility in the same measurement time.

Such improvements over traditional sampling rates are achieved by a technique often called "Compressed Sensing" [3, 2]. The technique ultimately depends on some fascinating counter-intuitive properties of high-dimensional geometry that are now being systematically explored by statisticians, probabilists, information theorists, and applied mathematicians; ultimately, the source of our understanding can be traced back to investigations by mathematicians in seemingly very remote issues: how many low-dimensional faces does a random polytope have? and how 'thick' is a random low-co-dimensional cross section of a high-dimensional simplex? It is also tied in unexpected ways to work by mathematicial statisticians to understand methods which can resist the influence of outliers.

The potential feasibility of Compressed Sensing can be motivated through a couple of observations:

(1) all images and other media are compressible: they don't need nearly as many bits to represent them as one might expect based on the raw image format. A 1000-by-1000 image indeed has 1 mega-pixels, but as every user of digital cameras, web browsers, cell phones and other modern tools knows, the actual number of bits needed to acheive a reasonably high quality reconstruction of an 8-bit deep image is in the tens or few hundreds of thousands, not 8 million.

(2) since the number of bits is substantially less than the nominal number of bits, we ought to be able to take a number of samples amounting to roughly the number of underlying bits appearing in the compressed representation, not the number of bits that appear in the uncompressed representation.

In short, since media (such as MRI images) are compressible, the sensing process itself ought to be compressible. Possibly, many scientists and engineers have formulated these observations previously, but it is only very recently that a coherent intellectual foundation has emerged.

One foundational explanation goes as follows: Suppose that we have an object x_0 of interest, with N coefficients, which we suppose has a *sparse* representation in a specific basis for $\mathbb{R}^N$—for many media types, this basis could be a suitable wavelet basis. By sparse, we mean that there are relatively few nonzero coefficients. While the typical vector is dense, with all coefficients nonzero – the typical humanly-intelligible media is sparse, with few coefficients nonzero. Sparsity will be the key ingredient allowing us to undersample.

We take measurements $y = Ax_0$, where A is an n by N matrix. Undersampling is

expressed by the fact that $n \ll N$. Our task is to reconstruct x_0 from y; this seems hopeless, as there are fewer equations than unknowns.

However, we have extra information: the object x is sparse, i.e. has at most k significantly nonzero elements, for $k \ll n$. To exploit this knowledge, we reconstruct x_0 by solving a convex optimization problem: minimize $\|x\|_1$ subject to $y = Ax$. In words, we seek the minimal ℓ_1 norm object matching the measurements y. This is a very different goal than that used traditionally, where we ask for the object with minimal ℓ_2 norm.

In this setting, we have the following mathematical result: if the measurements A are random – for example with i.i.d. Gaussian entries, and if x_0 is truly sparse – with k strict nonzeros – then we have exact reconstruction provided the number of measurements exceeds a threshold: $n > 2\log(N/k)k$.

This is a sampling theorem like Shannon's original theorem; however, it requires that the number of samples n be comparable to the number of nonzeros k, rather than the apparent vector dimension N. In short, although undersampling leads to underdetermined systems of equations, when the equations are random, and the solution is sparse, the solution is available by convex optimization.

This is a sampling of the kinds of results which are now available to explain the underlying phenomenon. This particular result follows from properties of high dimensional polytopes subjected to random orthogonal projection onto lower-dimensional space. Other approaches to formulating foundations can be based on properties of minors of random matrices, or on sections of hypercubes.

3.3 Computer and Simulation-Based Models

Computer models are computer codes, often large and deterministic, that simulate complex processes. They can encapsulate a field of knowledge, synthesizing the understanding of many individuals and often seek to answer questions that can not be answered directly with observational data or direct experiments. They are also used to make complex predictions that incorporate a substantial body of scientific understanding. Examples range from a simulation for traffic flow in an urban area (e.g. TRANSIMS) to climate models used to make global projections of future climate change (Fourth Assessment Report, Intergovernmental Panel on Climate Change) to a mechnical deformation model used to test vehicle designs in a crash. Despite the diverse use of computer models in many fields they share common problems in drawing inferences from high dimensional output, combining models with observations and quantifying the model's uncertainty. Statistics provides a framework to solve these problems and hence contribute to the scientific value of these models. Given the now ubiquitous use of computer models throughout science and engineering they merit more attention by the statistical community. We will illustrate with some questions arising in the geosciences.

By representing the interaction among several processes, computer models often yield output that has complex structure and are often difficult to analyze without some form of dimension redution. Data reduction using linear projections derived from sample covariance matrices (known as emipirical orthogonal functions in the geosciences) can miss nonlinear behavior and can be hard to interpret beyond a small number of projections. Statistics can support these efforts by exploiting sparsity of output fields with respect to particular bases. In general, finding efficient, regularized bases that are suited to particular physical processes simulated by computer models is important. Moreover, in the geosciences simulated fields of physical variables have many more spatial locations than temporal or model model

replications and so fit into a "large p small n" context.

Data assimilation refers to combining a numerical model with observations to produce a better estimate of the state of the system. This process is fundamentally a statistical one and, for example, provides the basis for weather forecasting. Some of the most challenging problems in prediction involve the assimilation of massive data streams into computer models with large state vectors. The high dimensionality of both model and data demand that any approach that is computationally feasible must also be approximate. It is uncertain how many of the approximations currently used in data assimilation affect the accuracy of the resulting analysis. In particular several data assimilation approaches can be cast as approximate Bayesian solutions where the posterior is represented through a sample (termed an ensemble in weather prediction). This identification makes connections to general problems of Bayesian inference using Monte Carlo algorithms for computation. An important emerging area in assimliation is estimating parameters in a computer model based on data, also known as model tuning. This activity has the potential to move model development from often subjective and heuristic tuning of parameters to an activity where parameters are estimated from observations using explicit criteria. The parametric components of computer models especially for geophysical problems often address behavior for processes at multiple scales. Thus applications of multiresolution statistical ideas can have an impact in suggesting alternative ways to simulate these processes within a computer model.

Physical models which, at the scales of interest, are not so well understood figure in many geophysical applications. The difficulty that this poor understanding poses is exacerbated by the amount and complexity of the data. This brings to the fore a number of points not limited to these sciences:

1. The models have been developed by geophysicists and applied mathematicians who understand them to be crude approximations The data exhibit substantial systematic biases from the postulated models. To contribute to these fields theoretically or practically statisticians must work closely with physicists and applied mathematicians.

2. Sparsity plays a key role not only in terms of data representation as discussed in the section on sparsity but also in dimension reduction and model selection.

3. Bayesian approaches assist in incorporating scientific knowledge into data analysis. Nonparametric approaches such as ℓ_1 optimization discussed in the sparsity section are similarly important for checking and correcting the sometimes shaky foundations described above.

As computer modeling assumes a mature and central role in many scientific disciplines there will be a tendency to consider families or classes of models, rather than just a single instantiation. From this perspective, one has a space of complex objects, i.e. computer models, that are functions of particular inputs, parameters or different model components. A practical question is: Given several versions of a computer model how well does this limited set of choices represent the possible behavior across the space of possible models? Answering this kind of question reinforces the theme of statistical science's ability to address complex data structures and interpolate/extrapolate a discrete set of information to a continuous set.

3.4 Statistics in Particle & Astrophysics

Experiments at the frontiers of modern physics and astronomy involve measurements of extremely weak signals and the search for extremely rare events. The fact that these experiments produce vast amounts of data, while enhancing the chances of successful detection,

presents a new range of problems of data management and computation.

Examples. *Particle accelerators.* Physical processes in particle accelerators are fundamentally stochastic. Observations of particle tracks collected by a variety of sensors present a filtering problem necessary to reconstruct the tracks. The reconstructed tracks then present a classification problem necessary to identify the particles that created the tracks. High data rates require simple, fast algorithms to do the filtering and the classification and storage limitations require minimal descriptions of each observation; that is, minimal relative to the raw data. The goal of distinguishing among precise, but only subtly different competing models for the data generated suggests the desirability of new, rapidly computable sparse representations of the data that may be much more informative than current representations.

Dark energy. The apparent acceleration in the expansion of the universe has led astrophysicists to hypothesize the existence of "dark energy" that may account for more than two-thirds of the mass-energy of the universe. Evaluation of theories of dark energy involves the solution of nonlinear inverse problems using data in which the signal is very weak. Instability of the inverse and noise in the data make assessing the reliability of the estimates particularly challenging.

Galactic filaments. Galaxies form filamentary structures, the "cosmic web." Existence of these structures is thought to reflect the evolution of the universe at the earliest moments of the big bang. Consequently, assessing their structure is relevant not only to understanding the current form of the universe but also to understanding the formation of the universe. More formally, filaments are sets of one-dimensional curves embedded in a point process. Similar structures appear in seismology, medical imaging, and remote sensing. Methods for locating numerous filaments in the presence of noise and clusters have important applications in all of these areas.

Statistical issues. While the physical sciences present many of same statistical issues as other branches of science, the presence of precise models, known dependencies between quantities of interest and errors in measurements, and different modes of data collection require new methods and new understanding of old methods. The increased collaboration between statisticians and physical scientists that has occurred in recent years should lead to important developments in both fields. Some of the issues that these collaborations must address include

- *Rapid processing of streaming data.* In addition to the obvious computational and data storage problems involved in handling massive amounts of data in real time, there is the problem of deciding what to compute and what to store.

- *Filtering massive data sets.* The effects sought in large physical experiments may be small compared to the scale of the data collected. Detecting the "signal" in the "noise" requires effective, but unbiased models and equally effective computational techniques.

- *Bayesian methods for high-dimensional models and sensitivity to priors.* The risk that the model may bias the outcome in filtering large, noisy data sets is just one example of the need to understand the sensitivity of Bayesian methods to the choice of priors. In addition, there is a critical need to develop well-founded and computationally tractable methods for constructing priors on high-dimensional spaces, so that in some well-defined sense the likelihood function is as dominant as possible relative to the

prior. Important work along those lines has been done, but the extension to models with many parameters remains challenging.

- *Confidence bands for nonparametric estimates.* Meaningful methods for exhibiting the variability of estimates of curves, surfaces, or even higher dimensional objects are needed for describing structure in multidimensional data and for the solution of inverse problems.

- *Cluster identification.* Improved methods for the detection of sparse signals (clusters) in the presence of inhomogeneous Poisson contamination are needed. Problems include evaluation of the sensitivity of the methods and estimation of the background and signal intensities.

- *Modeling resource allocation among expensive experiments.* Beyond the political issue of which scientific questions to address, there is the need to allocate resources to particular experiments. Determining the set of experiments that is most likely to give satisfactory answers requires both accurate physical modeling of the experiments and statistical design under resource constraints.

None of these issues is unique to physical data. As technology emerges in any field that allows the collection of large amounts of data, the opportunity to address increasingly complex and subtle questions arises.

3.5 Economics, Finance and Risk Management

Statistical inferences about extreme events, credit risks and macroeconomic policy modeling and simulations can have profound impact on the well being of society and at the same time pose significant intellectual challenges for academic research. The recent subprime crisis once more demonstrates that innovative statistical tools are urgently needed to contribute to controlling and managing market risks and that quantitative measures are needed for legal regulatory purposes.

Complex systems in finance and economics are inherently high-dimensional. In managing and controlling economic and financial risks, modeling and estimation of extreme events in several hundred dimensions are required. In monetary policy making, large statistical models of the US economy are needed. In managing financial risks, correlation matrices of the order of hundreds and thousands are prominently featured in assessing portfolio risks and portfolio allocations. Understanding systemic risk in the financial industry will have to be based on stochastic network models. These high-dimensional and complex problems share common statistical themes with other biological, physical and social sciences. For example, the collection of housing price indices for all counties or zip codes in the US form high-dimensional time series data, which have some spatial-temporal features and statistical challenges in common with climatological studies.

Extreme events not only occur in size, but also at the level of dependence. Understanding portfolio properties under extreme correlations is crucial in financial risk management, particularly when the analysis of rare extreme events is confronted. Insight into the cause of jumps and spillover effects of markets is important in building a sound financial system. The recent arrival of high-frequency data for a host of financial instruments makes understanding of extreme events more feasible and at the same time poses new statistical challenges of computation, modeling, and inference. In the legal regulatory framework for

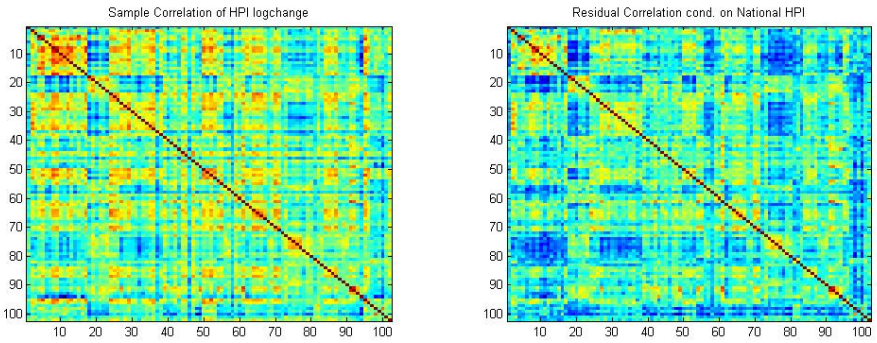


Figure 2: (a) The correlation of HPI among the 100 largest markets in the US. (b) The correlation of HPI among the 100 largest markets in the US after adjusting for (conditioning) the national HPI (Courtesy of Magnetar Capital).

the supervision of banking and insurance, estimates of the high quantiles of loss distributions are required. Such estimates are also needed for the reinsurance industry. Immediate applications exist in the field of alternative risk transfer between the insurance of natural causes and the financial markets (catastrophe bonds, mortality and longevity bonds).

In risk analysis, one needs to understand the impact of dimensionality on covariance matrix estimation, risk assessment, and portfolio allocation. One needs to select and identify important risk factors, a common theme in statistical model selection that has applications to many scientific disciplines. Modeling high-dimensional covariance matrices has broad significance for finance, longitudinal studies in economics and health sciences, and for biological, genomic and social networks. Different subject domains require different statistical modeling and understanding, but nevertheless share some common intellectual content. For example, credit risk analysis (time to default modeling) shares common statistical themes with health risk analysis (survival time analysis) and reliability (lifetime of products), but has also its distinguishing characteristics deriving from financial and economic theory. Disease classifications using microarrays and proteomic data share common statistical challenges with robust portfolio selection and risk management. In the former case, scientists wish to select tens of genes for disease classifications and to understand the underlying molecular mechanism, while in the latter, investors wish to select tens or hundred of stocks that achieve return-risk efficiency. In a similar vein, stochastic modeling for understanding the dynamics of market price co-movements also intersects with mathematical modeling of biological systems.

Macroeconomic policy making relies on the theory of inference for simultaneous equations models of high-dimensionality. Model uncertainty and estimation errors for parameters can have a large impact on decision making. Statistical model building and asymptotic inference play pivotal roles. Recently, economists at central banks have employed Bayesian MCMC methods that make possible inference with the more complex models that emerge from the rational expectations framework, and at a scale that is useful for the policy process. They produce measures of uncertainty that can in principle be fed, with due caution, into policy discussion that invokes judgemental beliefs of policy makers. A model estimated by these methods is now the central policy model at the Swedish Riksbank, and several other central banks have such models under development.

These are just a few examples of a methodology that combines ideas that have emerged from statistics and crossed over into very different fields, illustrating the role of the statistics

discipline in stimulating intellectual cross-fertilization.

3.6 Machine Learning

In the past few decades, the presence of increasingly large data sets in many fields of science and engineering, coupled with the widespread availability of powerful computers, has prompted a great deal of computationally-intensive statistical research, much of it carried out under the moniker of "machine learning." Machine learning is a set of activities that fall squarely within the general purview of statistical science, enriching statistics by enhancing the links to other areas of information science—including optimization theory, algorithms, signal processing, databases and information theory—and by strengthening the ties to applied users of statistical inference in areas such as genomics, information retrieval, speech recognition, remote detection and logistics.

Large data sets, characterized by many variables (large p) and/or many samples (large n), are now a common feature in many areas of science and engineering. They often arise from the use of high-throughput technologies, such as mass spectometry and gene expression arrays, and large-scale scientific studies, such as digital sky surveys and climate modeling.

Large, high-dimensional data sets present a number of statistical and computational challenges that are not present in more traditional studies involving small samples and relatively few variables. Machine learning, and the statistics field in general, has responded to these challenges in a timely fashion, through the development of core theory and methodology. Just as importantly, this statistical core has served as a hub through which key ideas in one discipline are developed, refined and transported to other disciplines.

Supervised and unsupervised learning. Machine learning has classically emphasized "supervised learning" problems (i.e., classification and regression problems), in which categorical labels or real-valued responses are attached to data points and the goal is to predict the response of a future data points. More recently many researchers have focused their attention on "unsupervised learning," a more diffusely-defined set of topics that includes clustering, pattern mining, dimension reduction and manifold learning. In these problems, only unlabeled data points are available, and the goal is to identify significant regularities among the data points, or between data points and variables. Both supervised and unsupervised learning are now highly developed areas of methodological and theoretical research and the ties to classical statistical theory are increasingly clear. Recent work has focused on so-called semi-supervised learning, in which information in labeled data points is combined with that from more readily accessible unlabeled data for the purposes of class prediction.

Much of the particular flavor of machine learning methodology has to do with its strong ties to optimization theory and numerical linear algebra. In particular, machine learning methods are often characterized by the use of surrogate or approximate loss functions to "convexify" or "relax" risk minimization problems. Examples include support vector machines, boosting, L_1 based methods for variable selection, and variational inference for graphical models.

Machine learning methodology often emphasizes simple, flexible statistical models that can be remarkably effective in the context of very large data sets; examples include boosting, bagging, random forests and hierarchical Bayesian models. More complex dependence structures can often be captured with graphical models, a formalism that merges graph theory and probability theory, and allows complex models to be built by combining simple modules.

Similarly, simple optimization methods (e.g., stochastic gradient descent) are often found to work very effectively on large data sets. Much effort has gone into understanding some of the reasons for this success, and the answers have involved statistical issues (e.g., control of overfitting) as well as numerical issues (e.g., regularization). This effort has also led to the development of new optimization methods, particularly constrained optimization methods associated with large-margin modeling.

Another theme that has emerged in the large-scale statistical setting is the important role of randomness. Examples include the random designs that are exploited in compressed sensing through L1-penalized minimization (Lasso) or Dantzig selectors, the use of random projections in feature selection algorithms, and the problem of finding largest submatrix of 1's in a 0-1 matrix, where a stochastic formulation yields quantitative results on the asymptotic size of such submatrices and has implications for the noise sensitivity of frequent itemset analysis [7]. Moreover, for massive data sets, randomness in the selection of training data often has two important positive side-effects: it can help mitigate overfitting issues and it can yield significant computational savings [14].

Statistical ideas also have an important role to play in the (unsupervised) problem of data mining, where the goal is to identify instances of one or more pre-defined patterns in a given data set. Researchers in computer science have developed sophisticated exact algorithms that can identify every pattern of a suitable type, but many of the patterns to which their methods apply are not robust to noise. Recent work [10] shows that elementary ideas from multiple testing can facilitate the noise-tolerant mining of patterns in high dimensional data. Just as importantly, measures of statistical significance resulting from a hypothesis-testing approach can provide an objective way to rank and select among the large number of potentially interesting patterns that are typically present in massive data sets. In short, statistical significance (under an appropriate null hypothesis) can act as a principled basis for pattern discovery in the exploratory analysis of large data sets. Statistically based data mining still presents computational challenges: some can be addressed with adaptations of existing techniques such as the EM algorithm; others require substantive collaborations between statisticians and computer scientists.

As machine learning matures for prediction, demand for interpretable models is increasing. Sparsity is a popular and useful proxy for interpretability, and is desirable for compression and transmission purposes. Sparse modeling tools, including L_1 based methods as Lasso and extensions to group and hierarchical sparsity, are intensively studied especially in the $p \gg n$ situation for regression and generalized linear models. This sparse modeling literature includes compressed sensing and covers classfication as well as regression and low rank matrix estimation. Moreover, it has a different angle to the problem from compressed sensing. Here, the design matrix is given by the particular application (e.g. gene expression levels) and its columns are often highly correlated. Theoretical results therefore assume more general and possibly dependent correlation structure for the design matrix than iid random entries as in the compressed sensing literature. The performance metrics are also broader than in compressed sensing and include L^2 error, L^2 prediction error, and to model selection (subset recovery). Recently, [8] provide a unified derivation of L_2 error bound in the $p >> n$ case for M-estimation (convex loss function) with a decomposable penalty. This general result covers both old and new results, and as special cases, Lasso, low-rank sparse approximation, and group-structured sparse matrix.

Finally, the new methodology developed by machine learning researchers has created challenges for statistical theory, many of which are being met by heightened activity in empirical process theory, where sharp concentration inequalities are a fundamental tool for

the analysis of nonparametric procedures, as well as in random matrix theory and convex geometry.

Bayesian nonparametric methods. Much of the focus in machine learning research has been on methods that allow the complexity of the underlying model to grow with the growth in the sample size; in statistical language these are nonparametric methods. Both the methods (e.g., the support vector machine and boosting) and their analysis have generally been developed within a frequentist framework in which one analyzes performance of a method based on repeated draws of the training data. There is also, however, a segment of the machine learning community interested in Bayesian methods, and most recently these researchers have begun to focus on Bayesian nonparametric methods. An existing literature in statistics dating back to the 1960's has provided an essential foundation for this effort; in particular, this literature provides a general framework for working with prior distributions that are general stochastic processes, and provides connections to other areas of mathematics (such as probability theory, functional analysis, and combinatorics) that are key to the manipulation of these stochastic processes.

Clustering is an important data analysis problem that has seen contributions from many applied communities (including machine learning). In Bayesian nonparametric statistics clustering problems can be attacked via the use of a prior known as the *Dirichlet process* (DP); the DP provides an appealing solution because it does not require the number of clusters to be known a priori. But the statistical framework provides something more; in Bayesian statistics it is natural to consider *hierarchical models*, in which multiple models are coupled probabilistically. Thus it is natural to define a *hierarchical Dirichlet process* (HDP) and thereby solve multiple related clustering problems [12]. Interestingly, while the need to solve the multiple clustering problem has been perceived within various applied communities, the problem was not faced head on until it was posed in a general statistical setting. Subsequently, the HDP solution has had significant impact on applied communities. Indeed, HDP-based models have yielded state-of-the-art solutions to problems such as object recognition in computational vision, natural language parsing, protein backbone modeling, haplotype inference in multiple subpopulations and image denoising.

Bayesian nonparametrics also provides methods for capturing and exploiting sparsity. In particular, the *beta process* is a stochastic process that yields a countably-infinite number of coins, which when tossed repeatedly yield a sparse binary matrix of exchangeable random variables [13]. Each row of this matrix can be viewed as a sparse featural representation of some object. Moreover, *hierarchical beta processes* can be defined to share sparsity patterns among collections of objects. There is currently significant activity in using the beta process for compression and sparse regression, and connections to the frequentist literature are beginning to be developed.

Challenges. The multiple clustering problem is a special case of the data integration or data fusion problem. Such problems arise frequently in the analysis of large data sets, in part because of the need to decompose large data sets into manageable pieces and in part because complex phenomena often provide many different views. For example, in understanding biological phenomena there is a pressing need to integrate across the great variety of genomic, genetic, proteomic and metabolomic data sets that are available. Climatology is another area in which data integration is paramount, and where issues of spatial and temporal scale make integration particularly challenging. Other challenging issues are also facing us:

- Could we as a community formalize canonical data processing operations that might

influence database design?

- What kind of a role can traditional exploratory data analysis play in the extraction of information from massive data sets?
- How should we go about trying to systematically visualize the outputs of modern statistical methods?
- How do we decide on the importance of variables?
- Streaming data requires considerations of compression and transmission in addition to computation. Can we develop a useful theory to encompass statistical estimation, computation, and data storage and transmission?

3.7 Network Modeling

What distinguishes network data from other examples of large scale data problems is the inherent dependencies among units. Indeed it is these dependencies or link that are a primary focus of analysis. Networks are usually characterized in terms of a set of n nodes, a set of N links among the nodes, and a set of r relations that characterize the links.

Examples and Models Galore. Examples of network datasets and problems involving relational data arise in diverse setting and areas. There are the early datasets from Stanley Milgram's 1960s small world experiments. Examples of other forms of network data include: (1) Social networks: Sampson's noviates in a monastery, Classroom friendship, My Space, Facebook, (2) Organization theory, (3) Homeland security, (4) Politics: Congressional voting behavior, bill co-sponsorship, (5) Public health: Needle sharing, Spread of AIDS, (6) Computer science: Email networks (Enron), Internet links, WWW routing systems, (7) Biology: Protein-protein interactions.

Often networks are embedded in policy problems such as those involving public health strategies, the design of economic markets, and alternative structures such as airline "hub and spoke" systems.

Researchers approach these with different analytical tools in different substantive areas: Erdos-Renyi random graph models and their generalizations, social network models such as p_1 and exponential random graph models, statistical physics approaches, and most recently latent variable models. Much of the substantive work in network modeling suffers from forms of "casual empiricism." Thus, there is an array of interesting network modeling problems to which statisticians can contribute. Most notable is the need to integrate across the different approaches to develop a common and reasonably flexible class of models that could then be adapted to the specifics of specific applications and problems of interest.

Two classes of problems have received limited statistical attention: (1) the design and analysis of studies involving dynamic networks in which nodes and links evolve over time. (2) integration of models based on attributes of nodes and models focused on the attributes of links. In addition, networks often contain external information about the nodes–jointly modeling node information with pairwise relationships is an important statistical issue. For recent descriptions of different classes of statistical models for networks see [5] and [4].

Some Overarching Statistical Issues. There are a number of major statistical modeling and inferential challenges in the analysis of network data that transcend individual models and classes of estimation methods. We mention six of these:

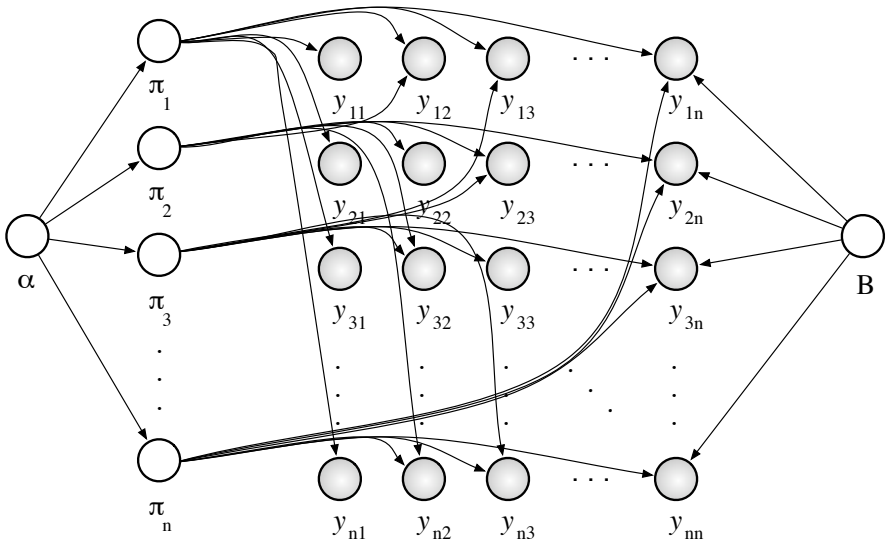


Figure 3: **The mixed membership stochastic blockmodel.** Nodes represent random variables; edges denote dependence between random variables. Each observed (shaded) node y_{ij} represents the observed link or observed lack of link between two elements of the network. This link is assumed drawn from a distribution that depends on the per-element hidden (unshaded) mixed membership vectors π_i and π_j. Note that each element's mixed membership vector plays a role in the distribution of all of its relationships to other nodes.

Computability. Can we do statistical estimation computations and model fitting exactly for large networks, or do we need to resort to approximations such as those involved in the variational approximations?

Asymptotics. There is no standard large sample asymptotics for networks, e.g., as the number of nodes n goes to infinity, which can be used to assess the goodness-of-fit of models.

Sampling. Do our data represent the entire network or are they based on only a subnetwork or subgraph? Should we take a random sample of nodes and their links, or look at the links to nodes outside the subgraph? When the data come from a subgraph, even one selected at random, we need to worry about boundary effects and the attendant biases they induce. This problem can be considered from both a sample designed based or a model based perspective.

Embeddability. Underlying most dynamic network models is a continuous time stochastic process even though the data used to study the models and their implications may come in the form of repeated snapshots at discrete time points (epochs)—a form of time sampling as opposed to node sampling referred to above—or cumulative network links. Can we represent and estimate the continuous-time parameters in the actual data realizations used to fit models? This is the *embeddability* problem and was studied for Markov processes in the 1970s, and more recently in the context of econometric models and by others in the computational finance literature.

Prediction. In dynamic network settings, data generated over time there are a series of forecasting problems. How should we evaluate alternative predictions from different models?

Privacy. As social networks on the WWW expand, concerns about the privacy of network data, recorded and shared, increase as well. The literature on privacy protection of

traditional statistical databases has burgeoned over the past decade, but a fundamental assumption regarding most disclosure limitation methodologies is the independence of data for different individuals or units of analysis. The depends among units that are the focus of network analysis make the privacy protection of network data a major challenge.

What statisticians can bring to the table here is not only the full set of statistical tools and methods used in other settings, but also the ability to abstract key elements from different modeling traditions to create a general theory which can in turn be carried back to the applications and to new statistical problems.

4 Acknowledgements

We are grateful for financial and planning support from NSF/DMS through Peter March, Deborah Lockhart, Grace Yang, Gabor Szekely, and Yazhen Wang. Organizational support was provided by Jianqing Fan and Alexis Kelley at Princeton.

The report was prepared through the efforts of many of the participants (listed below): session summaries were drafted and edited by the session chairs and/or the speakers. The overall document was shepherded by Peter Bickel, Iain Johnstone and Bin Yu.

5 Workshop Presenters

The workshop was held October 16-17, 2007 at the Mathematical Association of America (MAA) Conference Center in Washington D.C. Session organizers and presenters are listed below. In addition, there were opening remarks by Peter March, Director, NSF-DMS, Deborah Lockhart, Executive Officer NSF-DMS, and Iain Johnstone, Stanford U. There was a panel discussion on infrastructure needs led by Mary Ellen Bock, Purdue, Alan Karr (NISS), and Bruce Lindsay (Penn State).

1. Statistics in Biological Sciences [Peter Bickel, UC Berkeley]
    - Wing Wong, Stanford University
    - Xihong Lin, Harvard University
    - Samuel Kou, Harvard University
2. Sparsity [Iain Johnstone, Stanford University]
    - David Donoho, Stanford University
3. Computer Models [Jim Berger, Duke University and SAMSI]
    - Douglas Nychka, NCAR
    - Chris Jones, UNC-Chapel Hill
4. Statistics in Particle & Astrophysics [Tom Kurtz, Wisconsin]
    - Harrison Prospper, Florida State
    - Larry Wasserman, Carnegie Mellon
5. Economics/Finance/Risk Management [Jianqing Fan, Princeton]
    - Chris Sims, Princeton

- Paul Embrechts, ETH Zürich

6. Machine Learning [Bin Yu, UC Berkeley]
    - Andrew Nobel, UNC-Chapel Hill
    - Mike Jordan, UC Berkeley
7. Network Modeling [Steve Fienberg, Carnegie Mellon]
    - Mark Handcock, U. Washington
    - David Blei, Princeton

References

[1] Bickel, P. J., Brown, J. B., Huang, H. and Li, Q. [2009], 'An overview of recent developments in genomics and associated statistical methods', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **367**(1906), 4313–4337.

[2] Candès, E. J. [2006], Compressive sampling, *in* 'International Congress of Mathematicians. Vol. III', Eur. Math. Soc., Zürich, pp. 1433–1452.

[3] Donoho, D. L. [2006], 'Compressed sensing', *IEEE Trans. Inform. Theory* **52**(4), 1289–1306. **URL:** *http://dx.doi.org/10.1109/TIT.2006.871582*

[4] Goldenberg, A., Zheng, A., Fienberg, S. and Airoldi, E. [2010], 'A survey of statistical network models', *Foundations and Trends in Machine Learning* **2**(2), 129–233.

[5] Kolaczyk, E. D. [2010], *Statistical Analysis of Network Data*, Springer, New York.

[6] Kou, S. C., Cherayil, B. J., Min, W., English, B. P. and Xie, X. S. [2005], 'Single-molecule michaelismenten equations', *The Journal of Physical Chemistry B* **109**(41), 19068–19081. PMID: 16853459.

[7] Liu, J., Paulsen, S., Sun, X., Wang, W., Nobel, A. B. and Prins, J. [2006], Mining approximate frequent itemsets in the presence of noise: Algorithms and analysis., *in* 'Proceedings of the 2006 SIAM Conference on Data Mining (SDM)', Bethesda, MD.

[8] Negahban, S., Ravikumar, P., Wainwright, M. and Yu, B. [2009], A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers, *in* 'Proc. NIPS'.

[9] on Mathematical Sciences Research for DOEs Computational Biology, N. C. [2005], *Mathematics and 21st Century Biology*, National Academies Press. M. Olson, chair.

[10] Shabalin, A. A., Weigman, V. J., Perou, C. M. and Nobel, A. B. [2009], 'Finding large average submatrices in high dimensional data', *The Annals of Applied Statistics* **3**(3), 985–1012.

[11] Szalay, A. and Gray, J. [2006], 'Science in an exponential world', *Nature* **440**, 413–414.

[12] Teh, Y. W., Jordan, M. I., Beal, M. J. and Blei, D. M. [2006], 'Hierarchical Dirichlet processes', *Journal of the American Statistical Association* **101**(476), 1566–1581.

[13] Thibaux, R. and Jordan, M. I. [2009], Hierarchical beta processes and the indian buffet process., *in* M. Meila and X. S. (Eds.), eds, 'Proceedings of the Eleventh Conference on Artificial Intelligence and Statistics (AISTATS)', Puerto Rico.

[14] Yu, B. [2007], 'Embracing Statistical Challenges in the Information Technology Age', *Technometrics* **49**(3), 237–248.