The Physics of Algorithms

1. PI: Michael Chertkov (T-13) #149600
   co-PI: Allon Percus (CCS-3) #146353

2. Additional Investigators:

- Frank Alexander (CCS-3) #110216
- Eli Ben-Naim (T-13) #119658
- Brent Daniel (D-3) #180380
- Anders Hansson (CCS-5) #191558
- Matthew Hastings (T-13) #175388
- David Izraelevitz (D-6) #185962
- Gabriel Istrate (CCS-5) #169454
- Charles Reichhardt (T-13) #170260
- Mikhail Stepanov (T-13) #188972

3. Primary DOE Mission addressed

- **Reducing the Threats From Weapons of Mass Destruction and Advancing National Defense Capabilities**

4. Additional Missions Supported by Work

- **Enhancing the Nation's Fundamental Scientific Capabilities**
- **Ensuring the Safety and Reliability of the U.S. Nuclear Deterrent**
- **Improving the Nation's Energy Security**

5. Support for either or both of the Capability Thrusts

- **Predictive Science**
- **Materials Science**

6. Budget request for FY07, FY08, and FY09, in $K

- FY07: $1300K
- FY08: $1350K
- FY09: $1400K

Institutional and Scientific Motivation

Three key challenges at Los Alamos are analyzing large data sets, inferring information from incomplete data, and verifying and validating large-scale computer simulations. Central to all of these tasks is the need for effective algorithms that work in reasonable time spans, as well as methods that quantify the reliability of the answers. Owing to the scale of the data sets and the importance of the results, algorithms for national security applications must meet a much higher standard than that used in industrial and commercial settings.

This proposal describes a new method of algorithmic development. Our approach, based on advanced techniques of statistical physics, will be used to address a suite of problems in networks and computer science that are vital to the mission of the Laboratory. We aim to establish a long-term fundamental research capability in analysis and design of algorithms at LANL, oriented towards specific national security needs: pattern detection in adversary social networks; routing, coding, and optimization in wireless sensor networks; and formal verification for the validation of scientific computing codes. The new techniques to be developed in the proposed research program will advance the rapidly emerging field of statistical physics of algorithms, which is already impacting a wide range of computational methods and, more broadly, scientific disciplines.

In describing algorithmic performance, two metrics count: (1) the resources (computing time and memory) required to find the solution and (2) the quality of the solution found. Conventionally, the tools of discrete mathematics have been applied to *complete* algorithms, which guarantee the best solution to a problem. In many cases, however, these algorithms require an exponentially long time to run and are thus impractical. In contrast, *incomplete* algorithms achieve computational feasibility at the expense of solution quality. Such algorithms are frequently heuristic in nature, which severely limits our ability to quantify their performance as well as to scale them to larger systems. Our project will involve the development of both types of algorithms, exploiting powerful theoretical physics methods developed over the past few years. In this way, we will jointly address the issues of typical performance and of worst-case performance in extreme scenarios where high reliability is required. Our aim is to develop, quantify and improve the performance of these algorithms in the context of large-scale computing and data processing.

Our overall approach is to map hard computational problems onto physical systems of interacting spins or particles and then apply the techniques of statistical physics. These techniques, developed to study macroscopic systems of 10^{23} particles, are naturally suited to deal with the large data sets that arise in the problems of interest. This approach has already proven successful in industrial applications involving massive data sets. The PageRank algorithm used by Google to sort through the World Wide Web, one of the largest data sets in existence, is at its core an efficient algorithm for solving the diffusion equation on heterogeneous networks. Microsoft recently established a research group utilizing the theory of phase transitions to improve the reliability of their software. Likewise, developments in academia are leading to novel algorithms, such as the survey propagation algorithm discussed below, as well as to improved coding schemes.

Tasks and Probable Accomplishments

The tasks detailed below have been chosen carefully to satisfy three criteria: 1) they address specific needs arising in national security applications, 2) they deal with fundamental, unsolved questions in computer and information science, and 3) they are timely in that a breakthrough is possible using recent developments in the physics of algorithms.

Algorithms in Network Science.

Community detection involves decomposing a network or relational database into communities such that nodes are highly connected within a community and weakly connected between communities. The problem is closely related to network partitioning, and is of strong interest for detecting cliques in adversary social networks and motifs in biological interaction networks, the subjects of intense current research. Existing algorithms, while often successful at revealing hidden structures in networks, use largely ad hoc definitions of community. We approach the challenge of community detection differently, defining the network structure using a Hamiltonian formulation where the energy unambiguously quantifies the community assignment. Starting from a set of hypothetical community assignments, a maximal likelihood inference procedure utilizing techniques such as energy minimization will be used to select the most probable community assignment. We will also implement a heuristic search algorithm, “belief propagation,” inspired by an approximation by Hans Bethe in 1935 to explain melting. This allows for a search on graphs in linear rather than exponential time. To test the efficiency and reliability of our algorithms, we will analyze large-scale data sets with known community structure such as the collaboration network between scientists defined by co-authorship relations. The data for this network are extensive, highly accessible, and trustworthy. This builds upon our ongoing research in statistical analysis of bibliometric databases.

Routing, coding, and optimization problems in sensor networks can be similarly expressed as inference problems where algorithms can be applied using physics-based techniques. Ad-hoc wireless networks are large-scale collections of miniaturized sensors which can be used for the detection of biological agents and toxic chemicals, environmental measurement of radiation, or real-time video surveillance. Transmission of data across this large-scale, interacting, and noisy network requires knowing how to route, compress, and protect data against errors. We will develop a generalization of Shannon’s canonical information theory for communication across a noisy network. We will also build upon the work developed at LANL of using an optimal fluctuation technique for a single channel. This will be extended to networks and used to characterize the performance of network coding algorithms. A significant challenge in optimizing these networks is that a single sensor only has a limited energy available to process and transmit information, such as roughly 1 Joule in a cubic millimeter sensor. Thus, we will address the problem of optimizing sensor arrangements subject to these constraints. To accomplish this combined task of routing, coding, and optimization for a large network we will utilize generalized belief propagation algorithms.

Algorithms in Computer Science.

Partitioning and verification are two core combinatorial challenges in distributed computing at the Laboratory. Partitioning involves the allocation of tasks to different processors, so as to minimize inter-processor communication and conflict. Rapid and effective partitioning strategies are vital in simulation codes that use methods such as adaptive mesh refinement, where the processor load distribution evolves significantly over time. Based on our results in the graph partitioning problem, we expect that *phase transition* behavior will serve as a powerful predictor of efficient partitioning in these distributed codes. We intend to employ our framework of *extremal optimization*, explicitly making use of phase structure to motivate improved partitioning algorithms for scientific computing models such as the Parallel Ocean Code. At the same time, methods of formal verification promise vast improvements in these applications. Verification is used to prove or disprove the correctness of software with respect to a set of specifications. An approach that has been used successfully for debugging commercial software is to apply logical inference to a distributed system, modeling it as an instance of the canonical problem of *satisfiability*. Recent breakthroughs in understanding satisfiability, both at LANL and elsewhere, are leading to the design of survey propagation algorithms that perform well on even the hardest instances. These advances make it possible to develop formal verification for more complex scientific codes arising in the Laboratory applications mentioned here. We will provide a systematic approach to verification that is far more comprehensive than existing techniques.

Data storage and retrieval is, likewise, a central challenge of information science. Handling massive amounts of data requires compression, decompression, storage, and then retrieval of data. Many data storage and sorting algorithms use tree architectures. The novelty of our approach is mapping the growth of these trees to a collision process in a gas and then utilizing techniques in kinetic theory, in particular nonlinear dynamics analysis using the traveling waves method for data retrieval and zipping that we pioneered recently. We will determine how the retrieval procedure scales with the size of the data set and characterize how data ranking protocols, used in data storage and retrieval, perform both in normal and worst-case scenarios. We will use these scaling properties to develop more efficient storage and retrieval schemes. Computationally, we will test these theoretical ideas on tree structures generated using existing data compression algorithms, including the Lempel-Ziv algorithm, for which the traveling wave technique applies.

Overcoming data corruption requires solving an inference problem (as above) that is especially difficult in modern architectures such as a two-dimensional lattice, as opposed to the traditional co-centric architecture. The lack of efficient algorithms is the current bottleneck in high-density data restoration and related areas of two- and three-dimensional (holographic) image processing. We will address the challenge using methods developed in the physics of disordered magnets, specifically, mapping the multi-dimensional inference problem to a spin system. We will then design an approximate algorithm where nearest-neighbor bits are considered at the first level, next nearest-neighbors at the second level, etc. This mapping will be used to develop efficient algorithmic solutions for data restoration. Additionally, we will characterize statistical properties of data corruption using spectral analysis of distorted images, with high-speed video from fluid experiments at LANL providing datasets. This noise correlation analysis will enable us to characterize and improve existing data and image restoration algorithms.

Institutional Goals and Objectives

The development of efficient algorithmic solutions for computationally hard tasks is fundamental to the Laboratory's mission. It impacts modeling and simulation, network science, nonproliferation, and homeland and infrastructure security. We aim to develop a long-term capability at Los Alamos in the new field of statistical physics of algorithms. Our proposed research fits into a long-standing tradition at the Laboratory. From MANIAC to the introduction of the Monte Carlo method to the creation of the scholarly information archive (xxx.lanl.gov), LANL has leveraged strengths in the physical sciences to secure leadership in information and computer science.

Our team consists of physicists, experts in network science, computer scientists and engineers already working in close collaboration. Key accomplishments include the development of fidelity analysis in optics communications and coding theory, inventing the extremal optimization algorithm, and establishing LANL as the leader in the statistical physics of infrastructure networks. (For references and additional information, see http://cnls.lanl.gov/˜chertkov/alg.htm.)