Title:
The Physics of Algorithms: a new approach to Information Science

Principal Investigator:
Michael Chertkov (T-13, Complex Systems) chertkov@lanl.gov, Z149600

Co-Principal Investigator:
Gabriel Istrate (CCS-DSS, Dynamic Simulation Science) istrate@lanl.gov, Z169454

Co-Investigators:
Eli Ben-Naim (T-13, Z119658), Leonid Gurvits (CCS-3, Z174471), Anders Hansson (CCS-DSS, Z191558), Matthew Hastings (T-13, Z175388), Allon Percus (CCS-3, Z146353), Zoltan Toroczkai (CNLS, Z149541)

PD-Investigators:
Julien Barre (CNLS/T-11, Z194178), Zohar Nussinov (T-11, Z186507), Erzsebet Ravasz (CNLS/T-13, Z201863), Mikhail Stepanov (CNLS/T-13, Z188972)

External Collaborators: Ralf Koetter (UIUC), Paul Krapivsky (BU), Marc Mezard (Orsay), Andrea Montanari (ENS), Cristopher Moore (UNM), Bane Vasic (UA)

Primary institutional Goal addressed:
1. Science-Based Prediction (1.3 Computer Science and 1.4 Computational Science)

Additional Goals Supported by Work:
3. Preferred laboratory for defense, intelligence and homeland security (3.2 Monitoring wide areas and 3.3 Nuclear and Radiological Threats) and
4. Energy Security (6.2 Enhance Infrastructure for Utilization)

Proposed program duration: 3 years

Proposed Funding:
FY2006   1.2M
FY2007   1.26M
FY2008   1.32M
Title:
The Physics of Algorithms: a new approach to Information Science

Abstract:

Many problems of national importance in data transmission/storage/inference and optimization are computationally complex at a fundamental level: the difficulty of solving the problem grows exponentially with the size, so that even a small increase in problem size can make it completely intractable. Advances in algorithms in these cases can far outweigh advances in hardware. Recently, many of these computationally hard problems have been studied using techniques that physicists develop to analyze nature. This is leading to an intriguing convergence of engineering and computer-science themes with physics. On the information science side most advances have been based on heuristic approximations while the problems considered by the computer science community thus far have been mostly theoretical. This constitutes a unique opportunity for us to combine theory and practice. We will develop a comprehensive analysis based on statistical physics and apply this to three problems that are crucial to LANL's missions. First, the classical inference problems of error-correction and data storage/restoration are central to secure/reliable communication and inference from large data sets. Second, combinatorial optimization offers methods with generic applicability in infrastructure planning and software/hardware certification. Third, coding and clustering on complex graphs are intimately related to high-performance computing, efficient routing and genetic-circuits. Historically, LANL has played a leading role in applying the tools of statistical physics to algorithms, dating back over a half a century. Recent successes of our team highlight innovative approaches by which LANL may regain leadership in this critical area.

Michael Chertkov
Institutional Goals and Objectives: The development of efficient algorithmic solutions for computationally hard tasks is an important ingredient in many problems of Laboratory interest, from high-performance computing and data mining/inference to secure/reliable communication. These problems require “seamless extraction of knowledge from computational datasets that are orders of magnitude larger than generated and exploited currently” and “develop algorithms and hardware cooperatively”, that are at the core of Laboratory efforts in science-based prediction of complex systems (Goal 1.3 and Goal 1.4 of the LDRD DR strategic priorities white paper). Our research will also contribute to Goal 3.2 in the context of communication approaches, information flow and network reliability for large sensor systems, Goal 3.3 related to defensive capabilities in “efficient data communication/operation protocols, and complementary information” and Goal 6.2 to develop modern communication architectures that “require new approaches such as smart power delivery systems with integrated energy and communication”. This research will also have impact on other topics of importance to LANL such as modeling of gene-regulatory circuits and development of novel cryptographic principles.

Science and Technology Objectives: Our proposal consists of three inter-related parts. Understanding the similarities and differences of algorithmic solutions for these three problems will lead to a deeper understanding, and will enable us to apply our results with more confidence than would be possible by studying a single problem. The three problems share common themes. They can be described in statistical physics terms as discrete systems on large graphs. Furthermore, existing methods for these problems present a dilemma. The methods consist either of exact algorithms that cannot be implemented efficiently, or of efficient algorithms that work in practice but yield approximate solutions with uncontrolled errors. In this context, the single most important unifying idea of this project is to relate efficient but suboptimal algorithms to an approximation (e.g., cavity or mean-field approximation) that can be systematically improved upon. We aim to provide a detailed science-based understanding of existing heuristics and use this to develop novel algorithms. The algorithmic emphasis makes our approach unique; its special focus on individual instances compliments more traditional approaches such as the worst-case scenario approach of computer science and the ensemble average approach of information theory. We next describe details of the specific problems.

Error-correction, data storage/reconstruction: We address sets of problems in coding and data storage under noisy conditions and inter-symbol interference. One key recent result is that the efficient approximate decoding algorithms developed for modern error-correcting codes are exact on tree-like structures but are only approximate on graphs with loops representing real codes. When operating at moderate values of the noise these so-called belief propagation algorithms show an unprecedented ability to correct errors. However, belief propagation is unable to match the performance of exact (but impractical) decoding in reducing errors below the so-called error-floor threshold. The error-floor is a serious impediment for high-performance systems, such as modern fiber optics links that require less than a $10^{-12}$ probability of a bit error, while the error floor starts already at roughly $10^{-7}$ even for the best known codes. Aiming at reducing the error-floor, we will develop nonlinear dynamics and extreme statistics tools to characterize performance and subsequently improve error-correction schemes/algorithms for a variety of LANL relevant applications including communication and inference of

Michael Chertkov
large data sets, cooperative development of software-hardware algorithms and cryptography.

**Combinatorial optimization** is the task of finding the “optimum” in a large space of solutions. If the problem complexity is exponential, enumeration of all possible solutions is not possible, but powerful approximate techniques have been developed such as survey propagation. Recently, combining ideas from statistical physics and computer science, it has been found that the difficult domains in combinatorial optimization where these approximate techniques break down are associated with phase transitions. We will apply phase transition techniques to describe the typical case complexity, and use extreme statistics tools to analyze the “hardest” instances. We will improve analysis of existing algorithms and use the insight provided by statistical physics to develop and analyze novel algorithms for practical tasks such as infrastructure planning and software/hardware certification.

**Clustering (community detection, data mining) and Coding on complex graphs:** A system of a large number of nodes typically involves communities such that nodes within a community are strongly coupled. Thus, detecting communities is an optimization problem of finding the most likely set of communities. It is also an inference problem: if an intruder rewire some fraction of the nodes then we aim to find the most likely initial distribution of nodes among the different communities. Our task is to develop efficient algorithmic solutions for optimization/inference in the clustering problem. One important application is to genetic circuits, where communities correspond to groups of genes that regulate specific functions. Another problem we will study is coding on a highly structured graphs. Shannon's classical theory is confined to point-to-point communication between two parties. In contrast, ad hoc complexes of transceivers are less vulnerable to damage, because the information is divided into packets that are dynamically routed via multiple switching nodes. From an information theoretical point of view, a node can function as an encoder by combining information from all its input links and sending information to all output links. CCS-DSS is currently developing a scalable simulator for packet-switched communications. The simulator needs to compute efficient routes, but standard algorithms for solving this problem do not scale efficiently. Our approach in cracking this outstanding problem is to develop new efficient alternatives to standard routing algorithms capitalizing on recently suggested randomized coding ideas. Development of the new multi-cast algorithms will also boost threat-reduction capabilities in design of efficient monitoring protocols and will help improve communication architecture for smart power delivery systems.

**Expected Payoff/Impact:** Our research will provide new fundamental advances in algorithmic tools for mining, inference, manipulations and security of large information flows. The improvements promise a qualitative, orders of magnitude, increase in efficiency. Currently, the center of gravity for these “physics of algorithms” developments is in Europe. USA cannot afford to be second class in these areas vital to national security. We see a strong opportunity to establish LANL as the world center for this field. This is a chance to build on unique capabilities in our team that combines statistical physics techniques in extreme statistics, kinetic theory, renormalization group, and cavity method, with expertise in electrical engineering, information technology and computer science. (For references and additional information, see [http://cnls.lanl.gov/~chertkov/alg.htm](http://cnls.lanl.gov/~chertkov/alg.htm).)