THE FIELDS INSTITUTE
FOR RESEARCH IN MATHEMATICAL SCIENCES

ABSTRACTS 1.2

SHAI BEN-DAVID
Technion

Computational Complexity vs. Statistical Generalization in Learning – A survey of current knowledge and major questions

In the past decade computational learning witnessed a fascinating interplay between theory and practice. Several ideas that originated from theoretical research were translated into highly successful practical algorithms. Furthermore, the experimental success of these algorithms exceeded the most optimistic theoretical predictions, thus creating a new challenge for theoreticians - to justify the unexpected success of their own algorithmic ideas. In some cases, further research results in even more pessimistic theoretical predictions.

In this talk I shall survey these developments from the point of view of the interplay between the two major measures of learning complexity; The computational complexity and the statistical generalization ability (or information complexity). It seems that algorithms that succeeds with respect to one of these measures are bound to behave poorly with respect to the other.

I shall discuss some recent insights and research directions concerning two of the most successful algorithmic ideas; Boosting and Support Vector Machines. Finally, I wish to present a new notion of approximation that requires optimal solutions on common "regular" input instances, while relaxing the success requirements on some exceptional inputs. Such approximations can be achieved by learning algorithms that are efficient with respect to both the computational and the information complexity measures.

JEFF EDMONDS
York University

TCP has Competitive Flow Times

By viewing the popular TCP (Transport Control Protocol) as non-clairvoyant scheduling algorithm, we are able to prove that on a general network with $O(1)$ times as much bandwidth TCP is $O(1)$ competitive against the optimal global algorithm in minimizing the average transmission times of jobs. In addition, we extend this result in four ways: We consider a congestion control mechanism, RED (Random Early Detection), that achieves smoother transmissions and higher throughput by randomly dropping packets. We allow a server to use broadcast to simultaneously satisfy a number of independent requests for the same page. We allow the network to change dynamically. Finally, we consider jobs with fully parallelizable, sequential, and other nondecreasing sublinear speedup functions.

The transmission network is modeled as follows. There are a number of routers that act as bottlenecks. Each has a maximum bandwidth B_k and drops random packets beyond this maximum. Each job J_i to be scheduled is defined by its arrival time a_i, its file size l_i, and the subset of the bottlenecks $B(i)$ that it passes through. A scheduling algorithm allocates bandwidth $b_{i,t}$ to job J_i at time t, subject to the bottleneck constraints. The

average flow time of the algorithm is $Avg_i[c_i^A - a_i]$, where c_i^A is the completion time of the job under the algorithm.

TCP is a simple distributed online scheduler. In the simplified version of TCP, each sender of a job is completely in the dark, knowing nothing of the other jobs. It increases its own transition rate $b_{i,t}$ linearly until it detects that one of its packets has been dropped. We assume that a sender learns a packet has been dropped after a delta time delay. At this point, the sender cuts his own rate $b_{i,t}$ in half and resumes increasing the rate linearly.

In contrast, the optimal scheduler knows of all jobs past and future in the system and is able to completely direct how much each sender transmits. With a single bottleneck, we know that Shortest-Remaining-Work-First is optimal and that Equal Partition with 2+epsilon times the speed is competitive. However, with many bottlenecks, we had no idea what either the optimal or "EQUI" would do. Though some papers [F] prove that TCP is unfair to jobs that pass through many bottlenecks, we prove that if we charge each job for its bandwidth for each bottleneck that it passes through an amount proportional to the demand on that bottleneck, then TCP in its "steady-state" in fact proportions these resources out evenly amongst all jobs. We go on to prove that this algorithm TCP-EQUI with 2+epsilon times the bandwidth is competitive.

TCP, however, without global knowledge, takes some time to self adjust to this steady state TCP-EQUI. For example, if n jobs increase their bandwidth linearly at a rate of A, it takes B_k/nA time until the maximum bandwidth is reached. We call this an adjustment period. Define $D(J)$ to be the sum over all jobs of $O(1)$ adjustment periods at the beginning and at the end of its life. (We assume that this will be a small fraction of the job's total life.) Then we prove that for TCP with extra bandwidth s in $O(1)$, $TCP_s(J)/(OPT_1(J) + D(J)) = O(1)$.

RAN EL-YANIV
Technion

On online learning of expert advice

In the "expert advice" problem we attempt to sequenially form a prediction of the future based on the current opinions of a pool of experts. The goal is to combine these opinions as effectively as possible, without knowing anything about the underlying mechanism generating the observation. In this talk we will survey some known and new expert advice algorithms and their analyses within a "competitive" or "regret" framework.

C.C. GOTLIEB
University of Toronto

The New Real Estate

For a very long time wealth flowed from land, and after that from capital. There is a growing acceptance that today wealth flows from informationm and this in turn depends on access to information which is achieved through communication. The new channels of communication aare the Internet and the electromagnetic spectrum. We are seeing fierce battles being waged for control of these channels, and this is why ICANN, media convergence, and issues around them are so important.

DAVID KIRKPATRICK
University of British Columbia

Restructuring Ordered Binary Trees

We consider the problem of restructuring an ordered binary tree T, preserving the in-order sequence of its nodes, so as to reduce its height to some target value h. Such a restructuring necessarily involves the downward displacement of some of the nodes of T. Our results, focusing both on the maximum displacement over all nodes and on the maximum displacement over leaves only, provide (i) an explicit tradeoff between the worst-case displacement and the height restriction (including a family of trees that exhibit the worst case displacements) and (ii) efficient algorithms to achieve height-restricted restructuring while minimizing the maximum node displacement. (joint work with Will Evans)

JON KLEINBERG
Cornell University

Small-World Phenomena and the Dynamics of Information

Navigating in an unfamiliar environment is a problem that fits naturally within the paradigm of on-line algorithms — with the subtle difference that the "unknown future" characterizing most on-line problems is replaced here with an unknown setting that is gradually revealed as it is explored. While most of the initial work on these types of problems considered a robot maneuvering through the physical world, the rise of the World Wide Web has led to a related but different situation in which Internet users and software agents navigate through unfamiliar virtual terrain.

These virtual environments are perhaps best viewed as a type of social network, and they generally exhibit the "small-world phenomenon" — the average distance between nodes is very small relative to the size of the whole network. The study of this phenomenon was inaugurated as an area of experimental research in the social sciences through the

pioneering work of Stanley Milgram in the 1960's, and his basic discovery concerned a type of navigation; he found that individuals using local information are collectively very effective at actually constructing short paths between two points in a social network.

We describe a model for generating networks in which this type of local navigation is feasible, and contrast it with a large family of models in which it is not. The feasibility of navigation turns out to be connected to a natural structural property of the underlying network: it should have a sufficient density of links at all "distance scales." We then describe an application of these ideas to a related problem, the design of "gossip protocols" for spreading information in a large distributed networks. This latter result is joint work with David Kempe and Al Demers.

IAN MUNRO
University of Waterloo

A Worst Case Constant Time Priority Queue

The $O(\lg \lg m)$ priority queue of Van Emde Boas remains, after more than 25 years, one of the most intriguing data structures. It side steps the $\Omega(\lg n)$ lower bound of the comparison model by considering a bounded universe of size m. Mehlhorn et al proved the technique is optimal under a cell probe model. Paul Beame and Faith Fich returned to considering the number of elements actually present to prove matching upper and lower bounds of $\Theta(\sqrt{\lg n / \lg \lg n})$. Here we switch models again, this time to the Rambo (random access machine with byte overlap) model of Fredman and Willard, in which an individual bit may be in several words. Under this model, we give a constant time algorithm for the problem, which we have implemented in hardware.

RAFAIL OSTROVSKY
Telcordia Technologies

Non-Interactive and Non-Malleable Commitment and Zero-Knowledge

How do you maintain the security of Zero-Knowledge and Commitment protocols on the Internet? What does it mean? In this talk, I'll describe two recent results on this topic. In particular, I'll show that if all users have access to a common random string, such protocols can be achieved without the need of interaction and with very strong security guarantees. The talk will be self-contained.

THE FIELDS INSTITUTE
FOR RESEARCH IN MATHEMATICAL SCIENCES

ABSTRACTS 1.2

NICHOLAS PIPPENGER
University of British Columbia

Random Boolean Functions

We consider random Boolean functions of n arguments, $f : \mathbf{B}^n \to \mathbf{B}$ (where $\mathbf{B} = \{0, 1\}$), for which each value $f(x_1, \ldots, x_n)$ is independently 1 with probability p (and thus 0 with probability $1 - p$). We study the expected length $\overline{\ell}(n)$ of the shortest representation of such a function in disjunctive normal form (that is, as the disjunction of zero or more terms, each of which is the conjunction of zero or more literals, each of which is either an argument or its complement). The bounds we obtain for this problem involve correlation inequalities, cluster expansions, and other tools of statistical physics.

YUVAL RABANI
Technion

Geometric Search Structures in High Dimensional Spaces

We show that certain random linear transformations can be used to estimate the Hamming distance between points of the binary cube. We use these transformations to solve approximation versions of several proximity problems in the cube as well as in geometric settings. We analyze in particular the nearest neighbor search problem and clustering problems. In contrast, we show that similarly efficient results cannot be obtained for exact nearest neighbor search.

The talk is based on joint papers with Omer Barkol, Allan Borodin, Eyal Kushilevitz, and Rafail Ostrovsky.

A. A. RAZBOROV
Steklov Mathematics Institute/IAS

Proof Complexity of Pigeonhole Principles

Pigeonhole principles (differing from each other by the number of pigeons and optional constraints put on their behaviour) are probably the most extensively studied tautologies in Proof Complexity. They are amazingly simple, capture one of the most basic primitives in mathematics and Theoretical Computer Science (counting) and at the same time possess a clean combinatorial structure which make them akin to other important classes of tautologies. Respectively, beginning with the classical paper by Haken (1985), much effort has been put in understanding their proof complexity. Several results obtained during the last couple of years make valuable additions to the overall picture but some important problems still remain open.

In this talk I will try to summarize what is known about the proof complexity of pigeonhole principles, and what we still would like to prove.

ADI ROSEN
Technion

Tight Bounds for the Performance of Longest-in- System on DAGs

A growing amount of work has been invested in recent years in analyzing packet-switching networks under worst-case scenarios rather than under probabilistic assumption. Most of this work makes use of the model of "adversarial queuing theory" proposed by Borodin et al. [**?**], under which an adversary is allowed to inject into the network any sequence of packets as long as — roughly speaking — it does not overload the network.

We show that the protocol Longest In System, when applied to DAGs, uses buffers of only linear size (in the length of the longest path in the network). Furthermore, we show that any packet incurs only linear delay as well. These are, to the best of our knowledge, the first deterministic polynomial bounds on queue sizes and packet delays in the adversarial queuing theory model (other than on the line and the cycle). Our upper bounds are complemented by matching lower bounds on buffer sizes and packet delays.

Joint work with Micah Adler.

STEVEN RUDICH
Carnegie Mellon University

Formal Code Obfuscation

Informally, an *obfuscator* $\mathcal{O}$ is an (efficient, probabilistic) "compiler" that takes as input a program (or circuit) P and produces a new program $\mathcal{O}(P)$ that has the same functionality as P yet is "unreadable" in some sense. Obfuscators, if they exist, would have a wide variety of cryptographic and complexity-theoretic applications, ranging from software protection to homomorphic encryption to complexity-theoretic analogues of Rice's theorem. Most of these applications are based on an interpretation of the "unreadability" condition in obfuscation as meaning that $\mathcal{O}(P)$ is a "virtual black box," in the sense that anything one can efficiently compute given $\mathcal{O}(P)$, one could also efficiently compute given oracle access to P.

In this work, we initiate a theoretical investigation of obfuscation. Our main result is that, even under very weak formalizations of the above intuition, obfuscation is impossible. We prove this by constructing a family of functions $\mathcal{F}$ that are *inherently unobfuscatable* in the following sense: there is a property $\pi : \mathcal{F} \to \{0,1\}$ such that (a) given *any* program that computes a function $f \in \mathcal{F}$, the value $\pi(f)$ can be efficiently computed, yet (b) given oracle access to a (randomly selected) function $f \in \mathcal{F}$, no efficient algorithm can compute $\pi(f)$ much better than random guessing.

We extend our impossibility result in a number of ways, including even obfuscators that (a) are not necessarily computable in polynomial time, (b) only *approximately* preserve the functionality, and (c) only need to work for very restricted models of computation ($\mathcal{T}C0$). We also rule out several potential applications of obfuscators, by constructing

"unobfuscatable" signature schemes, encryption schemes, and pseudorandom function families.

BARUCH M. SCHIEBER
IBM T.J. Watson Research Center

Online Server Allocation in a Server Farm via Benefit Task Systems

A web content hosting service provider needs to dynamically allocate servers in a server farm to its customers' web sites. Ideally, the allocation to a site should always suffice to handle its load. However, due to a limited number of servers and the overhead incurred in changing the allocation of a server from one site to another, the system may become overloaded. The problem faced by the web hosting service provider is how to allocate the available servers in the most profitable way. Adding to the complexity of this problem is the fact that future loads of the sites are either unknown or known only for the very near future.

In this talk we model this server allocation problem, and consider both its offline and online versions. We give a polynomial time algorithm for computing the optimal offline allocation. In the online setting, we show almost optimal algorithms (both deterministic and randomized) for any positive lookahead. The quality of the solution improves as the lookahead increases. We also consider several special cases of practical interest. Finally, we present some experimental results using actual trace data that show that one of our online algorithm performs very close to optimal.

Interestingly, the online server allocation problem can be cast as a more general benefit task system that we define. Our results extend to this task system, which captures also the benefit maximization variants of the k-server problem and the metrical task system problem. It follows that the benefit maximization variants of these problems are more tractable than their cost minimization variants.

MADHU SUDAN
Massachusetts Institute of Technology

Random walks with Back Buttons

A "Backoff process" is an idealized stochastic model of browsing on the world-wide web. This model incorporates both hyperlink traversals and use of the "back button." With some probability the next state of the process is generated by a distribution over out-edges from the current state, as in a traditional Markov chain. With the remaining probability, however, the next state is generated by clicking on the back button, and returning to the state from which the current state was entered. Repeated clicks on the back button require access to increasingly distant history.

Backoff processes have fascinating similarities to and differences from Markov chains. In particular, they include the class of finite Markov chains as a special case, but are

included within the class of denumerable Markov chains. Backoff processes always have a limit distribution, like (irreducible, aperiodic) Markov chains. Unlike Markov chains, the limit distribution may depend on the start state. Further this distribution can be computed by an efficient algorithm. In this talk, we will describe some of the mathematical questions and algorithmic answers associated with backoff processes.

Joint work with Ron Fagin, Anna Karlin, Jon Kleinberg, Prabhakar Raghavan, Sridhar Rajagopalan, Ronitt Rubinfeld and Andrew Tomkins.

HISAO TAMAKI
Meiji University

Heuristic algorithms for Euclidean TSP based on Arora's dynamic programming scheme

Arora constructed a PTAS for Euclidean TSP using a dynamic programming approach. We show, through implementation and experiments, that the same approach combined with several heuristics leads to practically efficient algorithms that compete with the currently prevailing local search heuristics.

MARTIN TOMPA
University of Washington

Identifying Motifs in Orthologous DNA Sequences from Multiple Species

The identification of conserved patterns in DNA sequences is a fundamental method for suggesting good candidates for biologically functional regions such as regulatory sequences, splice sites, protein binding sites, etc. We will discuss the following approach for identifying motifs: given a collection of orthologous (i.e., corresponding) sequences from multiple species that are related by a known evolutionary tree, search for motifs that are well conserved (according to a parsimony measure) in the species. We present an exact algorithm for solving this problem. We then discuss experimental results on finding regulatory sequences for some sample genes in sample families of species.

This is joint work with Mathieu Blanchette and Benno Schwikowski. No prior knowledge of molecular biology will be assumed.

PANAYIOTIS TSAPARAS
University of Toronto

Link Analysis Ranking Algorithms on the World Wide Web

Recently, there have been a number of algorithms proposed for analyzing hypertext link structure so as to determine the best "authorities" for a given topic or query. While such analysis is usually combined with content analysis, there is a sense in which some algorithms are deemed to be "more balanced" and others "more focused". We undertake a comparative study of hypertext link analysis algorithms. Guided by some experimental queries, we propose some formal criteria for evaluating and comparing link analysis algorithms.

ELI UPFAL
Brown University

Can Entropy Characterize Performance of Online Algorithms?

Viewing online problems with stochastic input as iterative gambling games, we explore the relation between the entropy of the input sequence and the performance of the best online algorithm for that problem. We present both positive and negative results, showing that entropy is a good performance characterizer for list accessing and prefetching problems, but a poor characterizer for online caching. The motivation for this work are advanced system and architecture designs which allow the operating system to dynamically allocate resources to online protocols such as prefetching and caching. To utilize these features the operating system needs to identify data streams that can benefit from more resources. This question is not addressed by the standard online competitive analysis.

LESLIE G. VALIANT
Harvard University

Quantum Computations that Can be Simulated Classically in Polynomial Time

The universality of computational models depends on physical suppositions, most notably the polynomial time version of the Turing Hypothesis. Little systematic study of these suppositions has been attempted until relatively recently. However, the quantum computational model does now offer an avenue to these foundational questions. In this talk we shall describe some recent results that show that significant subclasses of quantum circuits can be simulated with only polynomial slowdown by classical computation. In particular, we define a class of circuits, composed of unitary gates that generate and process highly entangled states, that is insufficient in giving superpolynomial speedups over classical computers. The result says, in brief, that physics magic does not necessarily

imply computational magic. We shall also discuss the broader context for such results, and some new problems that are raised.

AVI WIGDERSON
IAS Princeton and The Hebrew University

Expander graphs - where Combinatorics and Albegra compete and cooperate

Expansion of graphs can be given equivalent definitions in combinatorial and algebraic terms. This is the most basic connection between combinatorics and algebra illuminated by expanders and the quest to construct them. The talk will survey how fertile this connection has been to both fields, focusing on recent results.