Report

on the Dagstuhl Seminar on

Computing with Faulty Inputs

May 29 - June 2, 1995

The seminary was concerned with questions arising in computing when the inputs to the computation can be corrupted. Such studies have implications for the manipulation, retrieval and transmission in the presence of errors and noise and they are related to many fields in Computer Science, Mathematics and Information Theory. Important technical tools come, e. g., from combinatorial search theory and the study of computational models (such as circuits, decision trees and branching programs) under the influence of random noise at the inputs.

The number of participants was rather small which was certainly due to the fact that the STOC meeting was held at the same time. As a consequence, almost all speakers gave 60 minutes talks (including discussion). These long lectures covered a wide range of topics and were appreciated very much by the participants.

There was also ample opportunity to discuss the topics of joint interest in the stimulating environment of Schloß Dagstuhl.

Reporter: Thorsten Prenzel

Contents

Eberhard Triesch	On Group Testing with a Bounded Number of Defectives
Joel Spencer	Paul and Carole Games
Rudolf Ahlswede	Some Ideas about a General Theory of Information Transfer
Prasad Tetali	Token Games on Graphs
Laurent Alonso	Determining the Majority
Miklós Ruszinkó	Intersection Theorems
Anna Gál	Fault Tolerant Circuits and Probabilistically Checkable Proofs
Ulrich Kulisch	Arithmetic Requirements and Language Support
Peter Damaschke	Bounds for several search problems
Gábor Tardos	Search in Multidimensional Grids
Hans Ulrich Simon	Bounds on the Number of Examples Needed for PAC-Learning; Concepts in the Presence of Noise
René Schott	Motion Planning with Uncertainty
Rüdiger Reischuk	Reliable Computations with Large Fanin Gates
Gyula O.M. Katona	Search with Small Sets
Ingo Schiermeyer	3-Satisfiability and Pure Literal Look-ahead
S. Muthukrishnan	Some Group Testing Issues in Experimental Molecular Biology
Martin Aigner	Finding the Maximum and Minimum

On Group Testing with a Bounded Number of Defectives

by E. Triesch (Universität Bonn, Germany)

Suppose we are given some hypergraph $H = (V, E)$ with a probability distribution $p : E \to \boldsymbol{R}_{>0}$ on the edges. We want to find some unknown edge $e^* \in E$ by asking questions of the form "Is $W \cap e^* \neq \emptyset$?" $(W \subseteq V)$. By some variant of alphabetic search, we show that the unknown edge can be found by asking at most

$$\lceil -\log p(e) \rceil + |e| \text{ questions (if } e^* = e, e \in E)$$

If $|e| \leq r$ for all $e \in E$, the number of questions is bounded by

$$-\log p(e) + r$$

and this time the inequality is strict. The average search length is thus bounded by $H(p) + r$, where $H(p)$ denotes the entropy of the distribution $p(\cdot)$.

As an application, we develop a-competitive and strongly a'-competitive group testing algorithms with $a = 1.58\ldots$ and $a' = 4$.

Paul and Carole Games

by Joel Spencer (Courant Institute, New York, USA)

In the (n, q, k)-Liar game Paul tries to find an unknown $x \in \{1, 2, \ldots, n\}$ with q Yes/No questions and Carole can lie at most k times. As Carole needn't really pick x in advance we may consider this a Perfect-Information game.

$$\textbf{Theorem: } \textit{If } n\frac{1 + q + \cdots + \binom{q}{k}}{2^q} > 1 \textit{ Then Carole Wins}$$

The proof involves analysis of a Random strategy for Carole, showing this works with positive probability. Using "Derandomization" we find an explicit strategy for Carole via maximization of a weight function w. This leads, by what we call "Antirandomization" to a strategy for Paul in which he asks queries so that the weight function is nearly independent of Carol's response. We show that for k fixed and $q \geq q(k)$ the converse of the Theorem is nearly (not exactly!) true and give precise conditions on n, q for Paul to win.

Paul = Paul Erdös, always asking questions.

Carole = ORACLE, whose answers need be wisely evaluated.

Some Ideas about a General Theory of Information Transfer

by R. Ahlswede (Universität Bielefeld, Germany)

We live in a world vibrating with "information" and in most cases we don't know how it is processed or even what it is at the semantic and pragmatic levels. A multitude of challenges to information theory comes from computer science. They, in particular, have stimulated us to reconsider the basic assumptions of Shannon's Theory and to investigate, whether its formulation is broad enough. This theory deals with "messages", which are elements of a prescribed set of objects, known to the communicators. The receiver wants to know the true message. This basic model, occurring in all engineering work on communication channels and networks addresses a very special communication situation. More generally they are characterized by

(I) The senders prior knowledge

(II) The prior knowledge of the receivers

(III) The questions of the receivers concerning the given “ensemble”, to be answered by the senders

We build up an understanding by considering first specific problems and then outline a general theory of information transfer. The classical transmission problem as formulated by Shannon and the identification problem are known special cases.

Token Games on Graphs

by Prasad Tetali (Georgia Tech., Atlanta, USA)

Consider n tokens placed on the n vertices of a (finite) undirected graph. There is an adversary, whose object is to keep the tokens apart for as long as possible, who decides which token moves at each step; then that token takes one step of a simple random walk. We study the expected time $M_G(n)$ till all the tokens coalesce under an optimal plan by the adversary. In particular, we show that $H(G) \leq M_G(n) \leq nH(G)$, where $H(G) = \max_{x,y \in G} E_x T_y$, with $E_x T_y$ being the (usual) expected first passage time to go from x to y. We believe the correct upper bound to be $M_G(n) = O(n^3)$, for all n-vertex graphs G.

We ask a weaker question here which we cannot resolve yet. Suppose you change the stopping time in the above situation to the first time by which every token either has moved or has been hit by some other token. Let us denote the expected time till this event happens by $MC_G(n)$. Clearly, $MC_G(n) \leq M_G(n)$.We believe the right bound to be $MC_G(n) \leq k' H(G)$, for some absolute constant $k' \geq 0$.

(Joint work with Peter Winkler)

Determining the Majority

by Laurent Alonso (CRIN, Nancy, France)

Given a set $\{x_1, x_2, \ldots, x_n\}$, each element of which is colored either red or blue, we must determine an element of the majority color by making equal/not-equal color comparisons $x_u : x_v$; when n is even, we must report that there is no majority if there are equal numbers of each color.

We show that in the worst case, exactly $n - \nu(n)$ questions are necessary and sufficient, where $\nu(n)$ is the number of 1-bits in the binary representation of n (new proof). Then we show that any algorithm that correctly determines the majority must on the average use at least

$$\frac{2n}{3} - \sqrt{\frac{8n}{9\pi}} + \Theta(1)$$

color comparisons, assuming all 2^n distinct colorings of the n elements are equally probable. Finally, we describe an algorithm that uses an average of

$$\frac{2n}{3} - \sqrt{\frac{8n}{9\pi}} + O(\log n)$$

color comparisons.

Intersection Theorems

by Miklós Ruszinkó (Hungarian Academy of Sciences, Budapest, Hungary)

A family $\mathbf{F}$ of subsets of an n-element set is r-cover-free if no set is covered by the union of r others. Let $T(r, n)$ denote the maximum size of such family. The aim is to get bounds on $T(r, n)$.

This question was posed by Kautz and Singleton in 1964 and later it was studied in Combinatorics (Erdös - Frankl - Füredi) and in group testing (T. Sós - Hwang) independently. Recent works of N. Linial; M. Szegedy and S. Vishwanathan apply this type of families in distributed graph coloring.

In the presented paper we improve the previously known upper bounds. Our main theorem says:

Theorem: $T(r, n) \leq 2^{cn\frac{\log \mathbf{r}}{\mathbf{r}^2}}$ *, where c is an absolute constant*

On the other hand, using algebraic geometric codes we give a lower bound for the $\frac{n}{r^2}$ uniform case which is tight. As a corollary, answering a question of V. Sós and Hwang, we get that for $\frac{n}{r^2}$ uniform case (up to a constant) the r-cover-free families are equivalent to the big distance codes.

Fault Tolerant Circuits and Probabilistically Checkable Proofs

by Anna Gál (University of Chicago, USA)

We study fault tolerant Boolean circuits. We consider the case when the faults are not random, but may be chosen by an adversary.

We show that every symmetric function has a synchronized circuit of size $O(n)$ and depth $O(\log n)$ that performs the "loose computation" of the function even if an adversary chooses a small constant fraction of the gates to be faulty at each level of the circuit. For the loose computation of a function f we require the output to be 1 whenever $f(x) = 1$, but the output has to be 0 only on inputs that have a large enough neighborhood where f is identically 0.

We also show a perhaps unexpected relation between our model and probabilistically checkable proofs. We show that from certain constructions of fault tolerant circuits the theorem of Arora and Safra, $NP = PCP(\log n, \log n)$, follows. Our results verify the existence of such constructions, using the stronger result of Arora, Lund, Motwani, Szedan, Szegedy stating that $NP = PCP(\log n, 1)$.

(Joint work with Mario Szegedy)

Arithmetic Requirements and Language Support

by Ulrich Kulisch (Universität Karlsruhe, Germany)

The speed of digital computers is ever increasing. While the emphasis in computing was traditionally on speed, more emphasis can and should now be put on accuracy and reliability of computed results. The arithmetic capability and repertoire, language and programming support of computers should be expanded to enhance the mathematical power of the digital computer and to turn it into a rigorous mathematical tool. The quality of the elementary floating-point operations should be extended to the usual product spaces of computation (vectors, matrices, complex numbers, intervals, etc.). This expanded capability is gained at modest cost and does not implicate a performance penalty. A vector arithmetic coprocessor with integrated PCI-interface has been developed. It performs dot products always with full accuracy or with a single rounding at the end. Language support is available on the basis of FORTRAN, PASCAL and C. ACRITH-XSC, PASCAL-XSC and C-XSC are very powerful programming environments which support the expanded arithmetic. Algorithms have been developed which deliver highly accurate and automatically verified results by applying mathematical fixed-point theorems. This means that such computations carry their own accuracy control. The computer itself can be used to appraise the quality and reliability of the computed results over a wide range of applications. Problem-solving routines with automatic result verification have been developed for many standard problems of Numerical Analysis as for linear or nonlinear systems of equations, for differential or integral equations etc. as well as for a large number of applications in the engineering and natural sciences.

Bounds for several search problems

by Peter Damaschke (Fernuniversität Hagen, Germany)

The talk addresses two combinatorial search problems:

(1) In a complete binary tree of depth k, an unknown subset of the leaves is "faulty". One may ask any node of the tree whether there is a faulty leaf in the subtree rooted at that node. We give a simple test strategy that finds all faulty leaves and is worst-case optimal, provided that the number of faults is equal (or approximately equal) to an estimated number r. The strategy needs $2r(k - \lceil \log_2 r \rceil - 1) + 2^{\lceil \log_2 r \rceil + 1}$ tests. The problem arouse in a practical situation in VLSI circuit testing.

(2) An unknown monotone function f with $|\mathrm{dom}(f)| \leq n, |\mathrm{range}(f)| \leq n$ shall be determined in parallel by queries of type "Is $f(x) \geq y$?" for suitable chosen pairs (x, y). We can solve this problem in $O(\log n)$ time by $O(n)$ queries, in such a way that simultaneous queries always refer to mutually distinct x and y. This is part of an optimal EREW PRAM algorithm for line segmentation of digital curves.

The results are to be presented at WG'95 (1) and ESA'95 (2).

Search in Multidimensional Grids

by Gábor Tardos (Math. Institute, Budapest, Hungary)

Let $\underline{x}$ be a positive integer vector of dimension d and $\underline{x} \leq \underline{a}$ upper bound is known.

The question considered is to find the minimal number $f(\underline{a})$ of queries of the form "$\underline{x} \leq b$?" in the adaptive worst case to find $\underline{x}$. Trivial bound for $\underline{a} = (a_1, \ldots, a_d)$:

$$\left\lceil \sum \log a_i \right\rceil \leq f(\underline{a}) \leq \sum \lceil \log a_i \rceil$$

Results:

- for fixed d and $a_i \to \infty$ the lower bound is almost always tight
  for $a_i \geq n_0(d) : f(\underline{a}) \leq \lceil \sum \log a_i \rceil + 1$
- $f(3, 3, \ldots, 3) = 2d$ thus $f(\underline{a}) - \lceil \sum \log a_i \rceil$ can be as big as $\lfloor (2 - \log 3)d \rfloor$

$$\textbf{Conjecture:} \ \limsup_{d} \frac{\max_{\underline{a}} \left(f(\underline{a}) - \lceil \sum \log a_i \rceil\right)}{d} = 2 - \log 3$$

- We know: $\forall \underline{a} : f(\underline{a}) \leq \lceil \sum \log a_i \rceil + 0.6d$

+ puzzles and results for the 2-dimensional case.
(Joint work with Miklós Ruszinkó)

Bounds on the Number of Examples Needed for PAC-Learning Concepts in the Presence of Noise

by Hans Ulrich Simon (Universität Dortmund, Germany)

We discuss the model of PAC-learning (PAC = Probably Approximately Correct) concepts under malicious noise. Tight upper and lower bounds on the sample complexity are derived (modulo logarithmic factors). More precisely: Let denote:

H the class of target concepts

d the VCdim of H

ϵ the accuracy parameter of the pac-model

β the noise rate, $\beta < \beta_0 := \frac{\epsilon}{1+\epsilon}$

Δ $:= \beta_0 - \beta$ the difference of β to the information-theoretic bound of Kearns and Li.

Then $\Omega\left(\frac{d\epsilon}{\Delta^2}\right)$ examples are needed by any learning algorithm which meets the requirements of the model. This matches the upper bound, obtained by using the minimum disagreement strategy, modulo a logarithmic factor.

Motion Planning with Uncertainty

by René Schott (CRIN, Université Henri Poincaré, Nancy, France)

In motion planning uncertainty arises from sensing errors, control errors and uncertainty in the geometric models of the environment.

Several models for representing and manipulating spatial uncertainty have been developed over the last decade: bayesian technique, Kalman filtering, Markov chains on colored graphs etc. We show how to model the motion of the robot in terms of random walks on the motion group or on some special graphs. Then we estimate the position of the robot after n steps as $n \to +\infty$.

Simulations based on these results are in progress.

Reliable Computations with Large Fanin Gates

by Rüdiger Reischuk (Med. Universität zu Lübeck, Germany)

For ordinary circuits with a fixed upper bound on the maximal fanin of gates it has been shown that logarithmic redundancy is necessary and sufficient to overcome random hardware faults. We review the main results and discuss the same question for circuits of sublogarithmic depth that make use of gates with unbounded fanin.

The answer depends on the type of gates that are used, and whether the error probabilities are known exactly or only an upper bound for the error probabilities is given. Also the fault model makes some difference, either only gates can deliver wrong values, or only wires can do so or both gates and wires can become faulty.

Reliable computation is basically impossible for $\wedge$-$\vee$-circuits even when restricted to faults in the wires. Gates with large fanin are of no use in case of noise. Restricted to a fixed depth only a small number of Boolean functions can be computed reliably even if the circuit size may be arbitrarily large. By a more elaborate analysis the same result can be obtained for threshold circuits if the error probabilities of the wires can vary between 0 and some upper bound ϵ_{max} and are chosen by an adversary.

Only in case of threshold circuits with fixed error probabilities redundancy is able to compensate faults. In this case a transformation from fault-free to fault-tolerant circuits can be performed that does not change the depth and increase the circuit size and fanin only moderately.

Search with Small Sets

by Gyula O.M. Katona (Math. Inst. Hungar. Acad. Sci., Budapest, Hungary)

Let X be a finite set of n elements, we want to find one unknown element $x_0 \in X$ by asking questions "Is $x_0 \in A$?" for subsets A of X. The search is non-adaptive, that is, the later questions do not depend on the answers of the previous ones. Suppose that the question sets satisfy the condition $|A| \leq k$. However, the results of the tests (answers of the question) can be faulty. At most l answers for the m questions can be faulty.

That is, using the characteristic vectors of the question sets, we obtain an $m \times n$ matrix ($0,1$-matrix) which contains at most k 1's in each row and the Hamming distance of any two columns is $\geq 2l + 1$. Given n, k, l, we need to minimize m.

Theorem: *m satisfies the inequality*

$$\log n + \log\left(1 + \ldots + \binom{m}{l}\right) \leq m\, h\left(\frac{k}{n}\right) \qquad \left(k \leq \frac{n}{2}\right)$$

where $h(x) = -x\log x - (1-x)\log(1-x)$.
If $\frac{k}{n} \to \kappa$, $\frac{l}{m} \to \lambda$ then this leads to

$$\frac{\log n}{h(\kappa) - h(\lambda)} \leq m \quad (\lambda < \kappa < \frac{1}{2})$$

We determined a weak upper bound for the case $\frac{k}{n} \longrightarrow \kappa, \frac{l}{m} \longrightarrow \lambda, \lambda < \frac{\kappa}{2} \leq \frac{1}{4}$.

3-Satisfiability and Pure Literal Look-ahead

by Ingo Schiermeyer (Technische Universität Cottbus, Germany)

In this talk we describe and analyze an improved algorithm for solving the 3-satisfiability problem. If F is a Boolean formula in conjunctive normal form with n variables and r clauses, then we will show that this algorithm solves the satisfiability problem for formulas with at most three literals per clause in time less than $O(1.498^n)$. The basic idea of our approach is to branch (if possible) in such a way that one (or more) of the new generated formulas contain pure literals.

Some Group Testing Issues in Experimental Molecular Biology

by S. Muthukrishnan (DIMACS, USA and University of Warwick, UK)

Here is the group testing scenario in molecular biology. A DNA is cut into several overlapping pieces called clones and libraries of such clones are tested for the occurences of "probes" which are unique sites. When we study the features, several problems can be abstracted:

(1) Consider the "trivial" questions in the second stage. Trivial $\Rightarrow$ Singleton objects tested. First stage is nonadaptive. What is the minimum number of tests needed? ($h = 0 \Rightarrow$ regular nonadaptive group testing.)

(2) Pool vs. tests tradeoff. # Pool = # of distinct sets in the decision tree. Fix v pools. What is the true min depth of any such decision tree?

(3) Sequencing by Hybridization. Determine a string $s \in \Sigma^n$ using substring queries "Is $s^\star$ a substring of S ?" The min. # of queries = $\Theta(n|\Sigma|)$. What if there are false positives / false negatives?

Many more problems can be formulated. (Reference: M. Krill (Los Alamos), S. Muthukrishnan: Technical report at Los Alamos LAUR-95-1503)

Finding the Maximum and Minimum

by Martin Aigner (Freie Universität Berlin, Germany)

We discuss three variants of the problem of finding the maximum and/or minimum out of a list of n elements with binary comparisons.

(1) Carole can lie a fixed number $l-1$ of times ($l \geq 1$). The common strategy for Paul is to repeat comparisons $x : y$ until he knows the true outcome, for Carole to reserve her lies to the endgame. If $M(n,l)$ resp. $MM(n,l)$ denote the worst-case costs, then

$$M(n,l) = ln - 1;\ 1 + \frac{\sqrt{2}}{l(1+\sqrt{2})^{l-1}} \leq \lim \frac{MM(n,l)}{ln} \leq 1 + \frac{\binom{2l}{l}}{2^{2l}}.$$

(2) Carole may lie a fixed proportion p at each stage of the game. Then $M(n,p) = \Theta\left(\frac{1}{1-p}\right)^n$ and $MM(n,p) \leq O\left(\frac{1}{1-p}\right)^{\frac{3n}{2}}$. The correct growth is not known.

(3) The nuts-and-bolts model. There are lists $x_1, \ldots, x_n$; $y_1, \ldots, y_n$ where (y_i) is a permutation of the x_j's, and only comparisons $x_i : y_j$ are permitted. Then $\overline{M}(n) = 2n-2$, $\overline{MM}(n) = 3n-4$. For l increasing $\frac{\overline{M}(n,l)}{M(n,l)}$ decreases, so lying does not always pay.