05291 Abstracts Collection
Sublinear Algorithms
— Dagstuhl Seminar —

Artur Czumaj[1], S. Muthu Muthukrishnan[2], Ronitt Rubinfeld[3] and Christian Sohler[4]

[1] New Jersey Inst. of Technology, US
czumaj@oak.njit.edu
[2] Rutgers Univ. Piscataway, US
muthu@cs.rutgers.edu
[3] MIT - Cambridge, US
ronitt@theory.lcs.mit.edu
[4] Universität Paderborn, DE
csohler@upb.de

Abstract. From 17.07.05 to 22.07.05, the Dagstuhl Seminar 05291 "Sublinear Algorithms" was held in the International Conference and Research Center (IBFI), Schloss Dagstuhl. During the seminar, several participants presented their current research, and ongoing work and open problems were discussed. Abstracts of the presentations given during the seminar as well as abstracts of seminar results and ideas are put together in this paper. The first section describes the seminar topics and goals in general. Links to extended abstracts or full papers are provided, if available.

05291 Executive Summary - Sublinear Algorithms

This paper summarizes the content and structure of the Dagstuhl seminar Sublinear Algorithms, which was held from 17.7.2005 to 22.7.2005 in Schloss Dagstuhl, Germany.

The purpose of the Dagstuhl seminar 'Sublinear Algorithms' was to bring together researchers working on the development of algorithms for very large data sets. Over the last few years data sets have become increasingly massive and the need to design special algorithms and data structures that deal with such amounts of data has emerged. For example, the set of all credit card transactions in the world for a month would have been considered a massive data set some time ago. That is comparable to the number of packet transactions a single router processes in *one* hour on an interface and we are now facing problems of analyzing the traffic at a large network of such routers, each with many interfaces! Internet traffic logs, clickstreams, web data are all examples of modern data sets that show unprecedented scale. Managing and analyzing such data sets forces us to revisit the traditional notions of efficient algorithms. The long-held golden

Dagstuhl Seminar Proceedings 05291
Sublinear Algorithms
http://drops.dagstuhl.de/opus/volltexte/2006/681

standard of “linear algorithms”—algorithms that take time proportional to the input and store no more space than it takes to archive the input—is no longer as efficient as one needs or can afford. Thus, there is now a need for *sublinear* algorithms, that is algorithms that use resources (time and space) significantly less than the input size.

The main areas addressed in the workshop were *property testing*, *sublinear time approximation algorithms*, and *data streaming algorithms*. These areas are not only connected by the fact that they require algorithms with sublinear resources but also that they heavily rely on randomization and random sampling. Therefore, we hoped that this workshop helped to exchange ideas between these different areas.

During the seminar one could obtain a good overview of the current state of sublinear algorithms. In many interesting talks new algorithms and models as well as solutions to well-known open problems were presented. To give an idea about the topics of the seminar we present a few examples of topics that were discussed in a number of talks at the seminar. These examples are not meant to be exhaustive.

Testing graph properties

Property testing deals with a certain notion of approximation of decision problems. One tries to distinguish the case that an object has a certain property from the case that it is far from the property. One topic of the workshop was the classification of testable graph properties. In this context we say that a property is testable, if it can be tested in time/query complexity independent of the graph. This classification problem was recently solved and a talk about these recent achievements was given at the seminar.

Although we know property testers for many graph properties their running is often extraordinary high, i.e. a tower of towers of $1/\epsilon$. It has been suggested during the seminar that existing results should be further refined to be able to distinguish between properties that are, say, testable with query complexity polynomial in $1/\epsilon$ and those testable in superpolynomial time.

Testing and approximating of distributions

Another interesting topic of the workshop was the question how to approximate and/or determine properties of an unknown distributions. Typically, an unknown distribution, for example over the numbers $1, ..., n$, is given as a black box. The only access to the distribution is by randomly sampling elements from it. A typical question considered for this model would be to determine whether the unknown distribution is (close to) uniform or differs significanly from the uniform distribution. A number of other approximation and testing algorithms for fundamental properties of distributions were presented in a number of talks.

Sublinear geometric algorithms

Geometric problems have been considered in the context of sublinear approximation algorithms and data streaming. A survey talk about the recent developments in geometric data streams has been given. Other contributions included algorithms for clustering problems, sublinear time intersection detection of convex polygons and polyhedra and online data reconstruction.

Connections between the areas

One focus of the workshop was to find connections between the different areas of sublinear algorithms. One such connection is the notion of tolerant property testing and distance approximation. In tolerant property testing one wants to accept all objects that have a property or that are very close to it and reject all objects that are far from the property. This is in contrast to classical property testing where one must only accept objects that have the property. One may think of tolerant testing as a hybrid between testing and sublinear time approximation. In distance approximation we are asked to approximate the distance of an object to a given property, which can be interpreted as a sublinear approximation algorithm. Another result discussed at the workshop showed that certain property testers can be turned into streaming algorithms.

Also a connection between data streaming and the area of *compressed sensing*, which recently received some amount of attention in mathematics, was presented at the workshop.

Open problem session

On Wednesday evening an open problem session was held, where many researchers presented interesting open problems.

Concluding remarks

The seminar was attended by 52 researchers from eight countries (19 USA, 13 Israel, 10 Germany, 4 Canada, 2 France, 2 United Kingdom, 1 Switzerland, 1 Hungary). From our own experience and the feedback from the participants we believe that the workshop was very successful. Interesting talks, fruitful discussions between researchers working on different areas of sublinear algorithms, and the wonderful working and living environment of Schloss Dagstuhl contributed to the success of the workshop.

Joint work of: Czumaj, Artur; Muthukrishnan, S. Muthu; Rubinfeld, Ronitt; Sohler, Christian

Testing Monotone and Unimodal Distributions

Tugkan Batu (Simon Fraser University, CA)

The complexity of testing properties of monotone and unimodal distributions, when given access only to samples of the distribution, is investigated. Two kinds of sublinear-time algorithms—those for testing monotonicity and those that take advantage of monotonicity—are provided.

The first algorithm tests if a given distribution on $[n]$ is monotone or far away from any monotone distribution in L_1-norm; this algorithm uses $O(\sqrt{n})$ samples and is shown to be nearly optimal.

The next algorithm, given a joint distribution on $[n] \times [n]$, tests if it is monotone or is far away from any monotone distribution in L_1-norm; this algorithm uses $O(n^{3/2})$ samples.

The problems of testing if two monotone distributions are close in L_1-norm and if two random variables with a monotone joint distribution are close to being independent in L_1-norm are also considered. Algorithms for these problems that use only $\text{poly}(\log n)$ samples are presented. The closeness and independence testing algorithms for monotone distributions are significantly more efficient than the corresponding algorithms as well as the lower bounds for arbitrary distributions.

Some of the above results are also extended to unimodal distributions.

Joint work of: Batu, Tugkan; Kumar, Ravi; Rubinfeld, Ronitt

Full Paper:
http://www.cs.sfu.ca/ batu/personal/papers/monodist.pdf

Online Data Reconstruction

Bernard Chazelle (Princeton University, USA)

Consider a dataset that is expected to satisfy various structural properties (eg, monotonicity, convexity, k-coloring, angular constraints).

Because of noise and errors, however, an unknown fraction of the data might be violating some of these properties.

Can one "reconstruct" the dataset in an online (sublinear) setting in a manner that strictly enforces all of the properties?

Keywords: Online reconstruction, sublinear algorithms

Joint work of: Chazelle, Bernard; Ailon, Nir; Comandur, Seshadhri; Liu, Ding

Towards an Algorithmic Theory of Compressed Sensing

Graham Cormode (Bell Labs - Murray Hill, USA)

In Approximation Theory, the fundamental problem is to reconstruct a signal $\in IR^n$ from linear measurements with respect to a dictionary Ψ for IR^n.

Recently, there has been tremendous excitement about the novel direction of *Compressed Sensing* [**?**] where the reconstruction can be done with very few—$\tilde{O}(k)$—linear measurements over a modified dictionary Ψ' if the information of the signal is concentrated in k coefficients over an orthonormal basis Ψ.

These results have reconstruction error on any given signal that is optimal with respect to a broad class of signals. In a series of papers and meetings over the past year, a theory of Compressed Sensing has been developed by mathematicians. We develop an algorithmic perspective for the Compressed Sensing problem, showing that Compressed Sensing results resonate with prior work in Group Testing, Learning theory and Streaming algorithms.

Our main contributions are new algorithms that present the most general results for Compressed Sensing with $1 + \epsilon$ approximation on *every* signal, faster algorithms for the reconstruction, as well as succinct transformations of Ψ to Ψ'.

Joint work of: Cormode, Graham; Muthukrishnan, S. Muthu

Sampling Algorithms for ℓ_2 Regression and Applications

Petros Drineas (Rensselaer Polytechnic, USA)

We present and analyze the first-known sampling algorithm for the basic linear-algebraic problem of ℓ_2 regression. The ℓ_2 regression (least-squares fit) problem takes as input a matrix $A \in \mathbb{R}^{n \times d}$ (where $n \gg d$) and a target vector $b \in \mathbb{R}^n$, and computes $\mathcal{Z} = \min_{x \in \mathbb{R}^d} \|b - Ax\|_2$. Also of interest is $x_{opt} = A^+ b$, which is the minimum length vector among those achieving the minimum. Our first algorithm randomly samples r constraints (and thus rows) from the matrix A and vector b to construct an induced ℓ_2 problem with many fewer constraints (but with the same number of variables). A crucial feature of the algorithm is the nonuniform sampling probabilities used to construct the induced ℓ_2 problem. These probabilities depend in a sophisticated manner on the lengths of the rows of the left singular vectors of A and the manner in which b lies in the left nullspace of A. Under appropriate assumptions, we show relative error approximations for both $\mathcal{Z}$ and x_{opt}.

We also present and analyze a sampling algorithm for a problem of interest in Matrix Approximation Theory; this algorithm uses our algorithm for approximate ℓ_2 regression in an essential manner. This second algorithm takes as input a matrix A and a subset of its columns. It performs an approximate least squares fit for every column of A using the input columns as a basis and it leads to an approximation to original matrix A of the form $\tilde{A} = CUR$, where C is a matrix consisting of the input columns of A, R is a matrix consisting of the chosen rows of A, and $U = (DW)^{+}D$, where W is the matrix consisting of the intersection between those columns and those rows, and D is a diagonal rescaling matrix. It is shown that under appropriate assumptions $\|A - \tilde{A}\|_F \leq (1 + \epsilon)\, \|A - P_C A\|_F$, where P_C is the projection onto the full column space of C. When combined with recent results in matrix approximation theory, this leads to improved bounds for matrix approximations of the form CUR.

Joint work of: Drineas, Petros; Mahoney, Michael; Muthukrishnan, S. Muthu

Testing for frequent patterns in a string

Ayse Funda Ergun (Simon Fraser University, CA)

In this talk we discuss how to test whether a given string S of length n contains a large number of occurrences of the same substring in sublinear time within the property testing framework. Since each occurrence of the substring can start at any arbitrary point within S, there are a superlinear number of pattern pairs to compare and store as candidates. To overcome this difficulty, we resort to a highly structured sampling technique and use a sparse representation of the data points which maintain the sublinearity of the running time.

Using these techniques we first present a method for testing whether a large fraction of S is covered by a repeating pattern of a given size k.

We then discuss how to use this technique to look for a pattern of any size. Since this problem is a generalization of periodicity testing, we also compare and contrast the two problems and their respective solutions.

Testing and estimation of dense graph properties

Eldar Fischer (Technion - Haifa, IL)

A topic in property testing that is recently receiving attention is that of tolerant testing, first defined by Parnas, Ron and Rubinfeld, in which the testing algorithm is not only required to reject inputs that are far from satisfying the property to be tested, but is also guaranteed to accept (with high probability) all inputs that are close enough to satisfying the property (and not only the inputs that satisfy it).

The best one could hope for is to distinguish for any constants $\epsilon > \delta$ between inputs that are ϵ-far from the property and inputs that are δ-close to it, with a number of queries that depends only on ϵ and δ. For graph properties in the dense model, we describe a proof that shows a dichotomy, in the sense that all properties that admit a constant query size testing algorithm, also admit a constant query size estimation algorithm (i.e. a constant query size tolerant testing algorithm for any ϵ and δ).

The proof requires a framework extending Szemeredi's Regularity Lemma.

This new framework is used both to define the graph parameters we need to estimate its distance from the property and to formulate the way to obtain them.

This result has appeared in the proceedings of the 2005 ACM STOC.

Keywords: Property testing, distance approximation, dense graph model, regularity lemma

Joint work of: Fischer, Eldar; Newman, Ilan

See also: Proc. 37th STOC (2005), pp.138-146.

Coresets in Dynamic Geometric Data Streams

Gereon Frahling (Universität Paderborn, D)

A dynamic geometric data stream consists of a sequence of m insert/delete operations of points from the discrete space $\{1, \ldots, \Delta\}^d$ [Indyk 04]. We develop streaming $(1+\epsilon)$-approximation algorithms for k-median, k-means, Max-Cut, maximum weighted matching (MaxWM), maximum travelling salesperson (MaxTSP), maximum spanning tree (MaxST), and average distance over dynamic geometric data streams. Our algorithms maintain a small weighted set of points (a coreset) that approximates with probability 2/3 the current point set with respect to the considered problem during the m insert/delete operations of the data stream. They use $poly(\epsilon^{-1}, \log m, \log \Delta)$ space and update time per insert/delete operation for constant k and dimension d.

Having a coreset one only needs a fast approximation algorithm for the weighted problem to compute a solution quickly. In fact, even an exponential algorithm is sometimes feasible as its running time may still be polynomial in n.

For example one can compute in $poly(\log n, \exp(O((1+\log(1/\epsilon)/\epsilon)^{d-1})))$ time a solution to k-median and k-means where n is the size of the current point set and k and d are constants. Finding an implicit solution to MaxCut can be done in $poly(\log n, \exp((1/\epsilon)^{O(1)}))$ time. For MaxST and average distance we require $poly(\log n, \epsilon^{-1})$ time and for MaxWM we require $O(n^3)$ time to do this.

We furthermore present an implementation of the coreset algorithm for k-means. Running popular iterative algorithms on smaller coresets and transfering the solutions to larger coresets can save important computation time needed by former iterative algorithms to converge to a good solution.

Keywords: Coresets, Streaming Algorithms, k-means, k-median, maxcut, approximation algorithms

Joint work of: Frahling, Gereon; Sohler, Christian

See also: Gereon Frahling and Christian Sohler, Coresets in dynamic geometric data streams, in STOC '05: Proceedings of the thirty-seventh annual ACM symposium on Theory of computing, 2005, ISBN 1-58113-960-8, pp 209–217, http://doi.acm.org/10.1145/1060590.1060622, ACM Press, New York, NY, USA

Contemplations on Testing Graph Properties

Oded Goldreich (Weizmann Inst. - Rehovot, IL)

This note documents two programmatic comments regarding testing graph properties, which I made during the workshop. The first comment advocates paying more attention to the dependence of the tester's complexity on the proximity parameter.

The second comment advocates paying more attention to the question of testing general graphs (rather than dense or bounded-degree ones). In addition, this note includes a suggestion to view property testing within the framework of promise problems.

Keywords: Property testing, graph properties

Full Paper: http://drops.dagstuhl.de/opus/volltexte/2006/555

Distribution-Free Property-Testing

Shirley Halevy (Technion - Haifa, IL)

We consider the problem of distribution-free property-testing of functions. In this setting of property testing, the distance between functions is measured with respect to a *fixed but unknown* distribution D on the domain, and the testing algorithms have an oracle access to random sampling from the domain according to this distribution D. This notion of distribution-free testing was previously defined, but no distribution-free property-testing algorithm was known for any (non-trivial) property prior to our work. We present the first such distribution-free algorithms for some of the central problems in this field, such as:

– A distribution-free testing algorithm for low-degree multivariate polynomials with query complexity $O(d^2 + d \cdot \epsilon^{-1})$, where d is the total degree of the polynomial.
– A distribution-free monotonicity testing algorithm for functions $f : [n]^d \to A$ for low dimensions (e.g., when d is a constant) with query complexity $O(\frac{\log^d n \cdot 2^d}{\epsilon})$. By this, showing that for the low-dimensional case, distribution-free testing can be done using similar query complexity to the complexity required in the uniform setting of property-testing.
– A distribution-free testing algorithm for connectivity of sparse graphs.

In addition, we show that, though in the uniform setting monotonicity testing of boolean functions defined over the boolean hypercube can be done using query complexity that is polynomial in $\frac{1}{\epsilon}$ and in the dimension d, in the distribution-free setting such testing requires a number of queries that is exponential in the dimension d. Therefore, in the high-dimensional case (in oppose to the low-dimensional case), the gap between the query complexity for the uniform and the distribution-free settings is exponential.

Keywords: Distribution-free Property testing

Joint work of: Halevy, Shirley; Kushilevitz, Eyal

Short PCPs verifiable in polylogarithmic time

Prahladh Harsha (Microsoft - Mountain View, USA)

In this paper, we revisit the study of Probabilistically Checkable Proofs (PCPs) in the context of efficient (i.e., sublinear-time) proof verification.

The study of PCPs was initiated in the seminal works of Babai et. al. [BFLS] and Feige at. al. [FGLSS] with very different motivation and emphases. The work of Babai et. al. considered the direct motivation of verifying proofs highly efficiently. Their work led them to study the length of the PCP and efficiency of the PCP verifier with respect to running time.

In contrast, Feige et. al. established a dramatic connection PCPs and the inapproximability of optimization problems. Almost all succeeding works have focused on the latter. On the other hand, very few works have focused on the length of the PCP and in fact no later work considered the extreme efficiency of the verifier. This is unfortunate because the latter efficiency parameters are significant in the context of proof-verification.

Motivated by the recent progress in the length of PCP due to Ben-Sasson et. al [BGHSV] and Ben-Sasson and Sudan [BS], in this work we revisit the study of efficient PCPs. We show that every language in NP has a probabilistically checkable proof of proximity [PCPP] (i.e., proofs asserting that an instance is "close" to a member of the language), where the verifier's running time is POLYLOGARITHMIC in the input size and the length of the probabilistically checkable proof is only polylogarithmically larger that the length of the classical proof. Almost all previous PCP constructions, with the exception of Babai et. al., required the verifier to run in time POLYNOMIAL in the input size even if it queried just a constant number bits of the proof. Our results thus give efficient (in the sense of running time) versions of the shortest known PCPs.

Keywords: Proof verification, sublinear-time, Probabilistically checkable proofs

Joint work of: Ben-Sasson, Eli; Goldreich, Oded; Harsha, Prahladh; Sudan, Madhu; Vadhan, Salil

Full Paper:
http://ttic.uchicago.edu/ prahladh/papers/

See also: Preliminary Version in Proc. 20*th IEEE Conference on Computational Complexity*, pages 120–134, San Jose, California, 12-15 June 2005.

Streaming Algorithms for Geometric Problems

Piotr Indyk (MIT - Cambridge, USA)

I will present an overview of recent developments in the design of streaming algorithms for geometric problems.

Keywords: Geometric data streams, core sets, random projections, diameter, clustering, spanning trees, matching

Full Paper:
http://theory.lcs.mit.edu/ indyk/GEOSTREAM/geostream-bib.ps

Algorithms and Lower Bounds for Streamed Graphs

Sampath Kannan (University of Pennsylvania, USA)

The streaming model of computation is relevant to situations where the amount of input data far exceeds the storage capacity of the computer. The model typically assumes that the space available is polylogarithmic in the size of the input and that the input is streamed in a read once fashion.

In this talk we consider the situation where the input is a graph with n vertices and m edges, whose edges are streamed (in adversarial order). After arguing that o(n) space is insufficient even for the "simplest" tasks, we focus on what can be done with space O(n polylog n). Here too the news is not very good when we restrict our attention to one-pass algorithms.

After showing some one pass algorithms for approximating the distance, approximating the size and weight of matchings etc., we turn our attention to multipass algorithms. We show pass-space trade-offs for matching and lower bounds for a pointer-chasing-like problem.

Joint work of: Feigenbaum, Joan; Kannan, Sampath; McGregor, Andrew; Suri, Sid; Zhang, Jian

Almost Orthogonal Linear Codes are Locally Testable

Tali Kaufman (Tel Aviv University, IL)

A code is said to be locally testable if an algorithm can distinguish between a codeword and a vector being essentially far from the code using a number of queries that is independent of the code's length.

The question of characterizing codes that are locally testable is highly complex. In this work we provide a sufficient condition for linear codes to be locally testable. Our condition is based on the weight distribution (spectrum) of the code and of its dual.

Codes of length n and minimum distance $n/2 - \Theta(\sqrt{n})$ have size which is at most polynomial in n. We call such codes almost orthogonal.

We use our condition to show that almost orthogonal codes are locally testable, and, moreover, their dual codes can be spanned by words of constant weights (weight of a codeword refers to the number of its non-zero coordinates).

Dual-BCH codes are generalizations of the well studied Hadamard codes, and are known to be almost orthogonal. Alon et al. raised the question whether these codes are locally testable. Our result provides positive answer to this question. Moreover, it shows that the BCH codes are spanned by their almost shortest words.

Joint work of: Kaufman, Tali; Litsyn, Simon

Property testing in graphs of general density

Michael Krivelevich (Tel Aviv University, IL)

Traditionally, two models of testing graph properties have been considered. In the first one, introduced by Goldreich, Goldwasser and Ron in 1996, an input graph G is represented by its adjacency matrix, and an algorithm queries whether a pair of vertices u,v forms an edge of G. The second one, due to Goldreich and Ron (1999), assumes that G is represented by the incidence lists of its vertices, and the algorithm queries for the i-th neighbor of vertex v. The former model is appropriate for dense graphs, i.e. graphs with a quadratic number of edges, while the latter is tailored for testing bounded degree graphs.

Recently, a hybrid model, combining the previous two models, has emerged. In this model, an input graph is thought of to be represented by both adjacency matrix and incidence lists, and an algorithm is allowed to ask both types of queries, i.e., vertex-pair queries of the dense model and neighbor queries of the bounded degree model. This approach allows to circumvent the inherent drawbacks of the more traditional models and is thus suitable in principle for testing graphs of arbitrary density.

I will discuss this model and recent results pertaining to it, such as testing bipartiteness and H-freeness, for a fixed graph H.

Sublinear Geometric Algorithms

Avner Magen (University of Toronto, CA)

We present sublinear algorithms to such problems as Detecting of Polytope intersection, Shortest Path on 3D convex Polytopes and volume approximation.

Keywords: Sublinear algorithms, computational geometry

Joint work of: Chazelle, Bernard; Liu, Ding; Magen, Avner

Full Paper: http://drops.dagstuhl.de/opus/volltexte/2006/554

See also: SIAM SICOMP, to appear

Streaming and Sublinear Approximation of Entropy and Information Distances

Andrew McGregor (Univ. of Pennsylvania, USA)

In many problems in data mining and machine learning, data items that need to be clustered or classified are not points in a high-dimensional space, but are distributions (points on a high dimensional simplex). For distributions, natural measures of distance are not the ℓ_p norms and variants, but information-theoretic measures like the Kullback-Leibler distance, the Hellinger distance, and others.

Efficient estimation of these distances is a key component in algorithms for manipulating distributions. Thus, sublinear resource constraints, either in time (property testing) or space (streaming) are crucial.

In this talk we design streaming and sublinear time property testing algorithms for entropy and various information theoretic distances. We start by resolving two open questions regarding property testing of distributions. Firstly, we show a tight bound for estimating bounded, symmetric *f-divergences* between distributions in a general property testing (sublinear time) framework (the so-called *combined oracle model*). This yields optimal algorithms for estimating such well known distances as the Jensen-Shannon divergence and the Hellinger distance. Secondly, we close a $(\log n)/H$ gap between upper and lower bounds for estimating entropy H in this model. We provide an optimal algorithm over all values of the entropy, and for small entropy the improvement is significant. In a stream setting (sublinear space), we give the first algorithm for estimating the entropy of a distribution. Our algorithm runs in polylogarithmic space and yields an asymptotic constant factor approximation scheme. An integral part of the algorithm is an interesting use of F_0 (the number of distinct elements) estimation algorithms; we also provide other results along the space/time/approximation tradeoff curve.

Our results have interesting structural implications that connect sublinear time- and space-constrained algorithms. The mediating model is the random

order streaming model, which assumes the input is a random permutation of a multiset and was first considered by Munro and Patterson. We show that any property testing algorithm in the combined oracle model for permutation invariant functions can be simulated in the random order model in a single pass. This addresses a question raised by Feigenbaum et al. regarding the relationship between property testing and stream algorithms. Further we give a polylog-space PTAS for the estimating the entropy of a one pass random order stream. This bound cannot be achieved in the combined oracle model.

Joint work of: Guha, Sudipto; Venkatasubramanian, Suresh

Tolerant Property Testing and Distance Approximation

Michal Parnas (Academic College of Tel Aviv Yaffo, IL)

A standard property testing algorithm is required to determine with high probability whether a given object has property P or is ϵ-*far* from having P, for a given distance parameter $\epsilon > 0$.

In this paper we study a generalization of standard property testing where the algorithms are required to be more *tolerant* with respect to objects that do not have, but are close to having, the property. Specifically, a *tolerant property testing algorithm* is required to accept objects that are ϵ_1-close to having a given property P and reject objects that are ϵ_2-far from having P, for some parameters $0 \leq \epsilon_1 < \epsilon_2 \leq 1$. Another related natural extension of standard property testing that we study, is *distance approximation*. Here the algorithm should output an estimate $\hat{\epsilon}$ of the distance of the object to P, where this estimate is sufficiently close to the true distance of the object to P.

We first formalize the notions of tolerant property testing and distance approximation and discuss the relationship between the two tasks, as well as their relationship to standard property testing. We then apply these new notions to the study of two problems: tolerant testing of clustering, and distance approximation for monotonicity. We present and analyze algorithms whose query complexity is either polylogarithmic or independent of the size of the input.

Our tolerant testing algorithm for clustering exploits a general framework, which is an extension of the abstract combinatorial programs of Czumaj and Sohler to the tolerant setting. This framework may be applicable for tolerant testing of other cost measures for clustering, as well as for tolerant testing of other properties.

Keywords: Tolerant Property Testing

Joint work of: Parnas, Michal; Ron, Dana; Rubinfeld, Ronitt

Approximating string compressibility and the distribution support size

Sofya Raskhodnikova (Weizmann Inst. - Rehovot, IL)

Imagine having to choose between a few compression schemes to compress a very long file. Before deciding on the scheme, you might want to obtain a rough estimate on how well each scheme performs on your file. We consider the question of approximating compressibility of a string with respect to a fixed compression scheme, in sublinear time.

In the talk, we will concentrate on the run-length encoding and a version of Lempel-Ziv as our example compression schemes. We present algorithms and lower bounds for approximating compressibility with respect to both schemes. We show that compressibility with respect to Lempel-Ziv is related to approximating the support size of a distribution. This problem has been considered under different guises in the literature. We prove a lower bound for it, at the heart of which is a construction of two positive integer random variables, X and Y, with very different expectations and the following condition on the moments up to k:

$$E[X]/E[Y] = E[X^2]/E[Y^2] = ... = E[X^k]/E[Y^k].$$

Keywords: Compression, sublinear approximation algorithms, lower bound

Joint work of: Raskhodnikova, Sofya; Ron, Dana; Rubinfeld, Ronitt; Smith, Adam; Shpilka, Amir

Approximating Average Parameters of Graphs

Dana Ron (Tel Aviv University, IL)

Inspired by Feige (36th STOC, 2004), we initiate a study of sublinear randomized algorithms for approximating average parameters of a graph.

Specifically, we consider the average degree of a graph and the average distance between pairs of vertices in a graph. Since our focus is on sublinear algorithms, these algorithms access the input graph via queries to an adequate oracle. We consider two types of queries. The first type is standard neighborhood queries (i.e., what is the i'th neighbor of vertex v?), whereas the second type are queries regarding the quantities that we need to find the average of (i.e., what is the degree of vertex v? and what is the distance between u and v, respectively).

Loosely speaking, our results indicate a difference between the two problems: For approximating the average degree, the standard neighbor queries suffice and in fact are preferable to degree queries. In contrast, for approximating average distances, the standard neighbor queries are of little help whereas distance queries are crucial.

Keywords: Graph parameters, degree, distance

Joint work of: Goldreich, Oded; Ron, Dana

Full Paper: http://drops.dagstuhl.de/opus/volltexte/2006/553

Strong Linearity and Quadraticity Tests for Boolean Functions

Alex Samorodnitsky (The Hebrew University of Jerusalem, IL)

Let f be a function from ${Z_2}^n$ to Z_2.

We show that if an appropriately defined iterated difference function (a "second derivative") of f is zero with probability bounded away from 1/2, then f is somewhat close to a quadratic polynomial.

A similar question for Z_n (with n large) has been considered by Gowers in his proof of Szemeredi's theorem for arithmetic progressions of length four. For Z_n the above claim does not hold, but Gowers's approach can be adopted to give (with some additional technicalities) the proof for $(Z_2)^n$.

This result is used to construct strong linearity and quadraticity tests for Boolean functions. The test will query a function in a small number of places, and distinguish with very high probability between linear (or quadratic) functions and functions which are far from any quadratic polynomial over ${Z_2}^n$.

We will discuss possible generalizations and applications of such tests.

Testing monotone high-dimensional distributions

Rocco Servedio (Columbia University, USA)

A *monotone distribution* P over a (partially) ordered domain assigns higher probability to y than to x if $y \geq x$ in the order.

We study several natural problems concerning testing properties of monotone distributions over the n-dimensional Boolean cube, given access to random draws from the distribution being tested. We give a poly(n)-time algorithm for testing whether a monotone distribution is equivalent to or ϵ-far (in the L_1 norm) from the uniform distribution. A key ingredient of the algorithm is a generalization of a known isoperimetric inequality for the Boolean cube. We also introduce a method for proving lower bounds on various problems of testing monotone distributions over the n-dimensional Boolean cube, based on a new decomposition technique for monotone distributions.

We use this method to show that our uniformity testing algorithm is optimal up to polylog(n) factors, and also to give exponential lower bounds on the complexity of several other problems, including testing whether a monotone distribution is identical to or ϵ-far from a fixed known monotone product distribution and approximating the entropy of an unknown monotone distribution.

Keywords: Sublinear algorithms, property testing, distribution testing, monotone distributions

Joint work of: Rubinfeld, Ronitt; Servedio, Rocco

Full Paper:
http://www1.cs.columbia.edu/ rocco/papers/papers.html

See also: Preliminary version in STOC 2005, pp. 147-156.

Recent results on graph property testing

Asaf Shapira (Tel Aviv University, IL)

I will survey some recent results on graph property testing. These include:

1) A proof that any monotone graph property is testable.

2) A result showing that the "natural" graph properties that can be tested with 1-sided error are precisely those that are "close" to being hereditary. This result implies in particular that the properties of being Perfect, Chordal, Interval, Comparability and more are all testable with 1-sided error.

3) A separation-type result showing that there is a coNP language that can be tested if the error parameter ϵ is known in advance, but cannot be tested if ϵ should be accepted as part of the input.

Joint work of: Shapira, Asaf; Alon, Noga

An Online Spanning Tree Problem in Randomly Weighted Graphs

Angelika Steger (ETH Zürich, CH)

In this talk we consider an online variant of the minimum spanning tree problem in randomly weighted graphs. We assume that the input graph is complete and the edge weights are uniform distributed over $[0, 1]$.

An algorithm receives the edges one by one and has to decide immediately whether to include the current edge into the spanning tree or to reject it. The corresponding edge sequence is determined by some adversary. We propose an algorithm which achieves $\mathbb{E}[Alg]/\mathbb{E}[Opt] = O(1)$ against a fair adaptive adversary, i.e., an adversary which determines the edge order online and is fair in a sense that he does not know more about the edge weights than the algorithm. Furthermore, we prove that no online algorithm performs better than $\mathbb{E}[Alg]/\mathbb{E}[Opt] = \Omega(\log n)$ if the adversary knows the edge weights in advance. This lower bound is tight, since we also exhibit an algorithm which achieves $\mathbb{E}[Alg]/\mathbb{E}[Opt] = O(\log n)$ against the strongest imaginable adversary.

Keywords: Online algorithms, random graphs, minimum spanning tree

Joint work of: Steger, Angelika; Remy, J.; Souza, A.

Projective Clustering With Outliers

Sergei Vassilvitskii (Stanford University, USA)

Given a set of n points in R^d, a family of shapes S and a number of clusters k, the projective clustering problem is to find a collection of k shapes in S such that the maximum distance from a point to its nearest shape is minimized. Some special cases of the problem include the k-line center problem where the goal is to cover the points with minimum radius hypercylinders and the k-hyperplane center problem where the goal is to cover the points with minimum width slabs.

For the k-line center problem, we give an $O(\frac{dk^4}{\epsilon})$ time algorithm that identifies a collection $O(k \log \frac{kd}{\epsilon})$ cylinders of radius at most twice the optimum that cover $(1-\epsilon)n$ of the points.

Since real data often contains outliers, we strengthen our algorithms to handle the case when the optimum solution covers only a $(1-\gamma)$ fraction of the points.

Our algorithms with high probability find a collection of $O(k \log \frac{kd}{\epsilon})$ cylinders that cover $(1-\epsilon-\gamma)n$ of the points in time $O(\frac{dk^4}{\epsilon^2(1-\gamma)^2})$.

We then present a general framework in which sublinear projective clustering results may be obtained for any shape. In this framework, we prove that the k q-dimensional hyperplane center problem can be solved in time $O(d(\frac{kq}{\epsilon})^{q+1}$ where the algorithm finds a collection of $O(k \log \frac{kdq}{\epsilon})$ slabs of width at most 2^q times the optimum and cover all but an ϵ fraction of the points.

Keywords: Projective clustering, dimension reduction

Algorithms for Distances between Distributions

Suresh Venkatasubramanian (AT & T Research - Florham Park, USA)

In many data analysis settings, whether they be in data cleaning, network analysis, machine learning, computer vision, or even neuroscience, the basic primitive is a distribution, or a histogram. The analysis task involves estimating distances between such primitives, exactly or approximately.

Distributions can be viewed as points in a d-dimensional space, and "standard" distances (like those derived from l_p norms) can be employed in tasks like classification and clustering. It turns out though that other distances (especially measures like relative entropy that are derived from information-theoretic considerations) are more meaningful in many application settings, both formally and in practice.

This presents an interesting algorithmic challenge: Are there good algorithms for estimating such distances, preferably in sublinear space/time ? Can these be used effectively in various application domains, and at scale ?

In this talk, I will describe current research directions in the "algorithmics" of distributions, and sketch out recent work that I've been doing on both the applied and theoretical ends of this area.

Property and Equivalence Testing on Strings under the Edit Distance with Moves

Michel de Rougemont (Université Paris Sud, F)

We investigate property testing and related questions, where instead of the usual Hamming and edit distances between input strings, we consider the edit distance with moves. Using a statistical embedding of words (into ℓ_1) which has similarities with the Parikh mapping, several problems related to property testing and characterization of languages can be solved. We first construct a tolerant tester for the equality of two words, whose complexity is independent of the string size, from which we derive an approximation algorithm for the normalized edit distance with moves.

We then consider the question of testing if a string is a member of a given language. We develop a method to compute, in polynomial time in the representation, a geometric approximate description of a regular language by a finite union of polytopes. As an application, we have a new tester for regular languages given by their nondeterministic finite automaton (or regular expressions), whose complexity does not depend on the automaton, except for a polynomial time preprocessing step. Furthermore, this embedding method allows us to compare whole languages, and validates the new notion of equivalent testing that we introduce. Using the geometrical embedding we can distinguish between a pair of automata thatb compute the same language, and a pair of automata whose languages are not ϵ-equivalent in an appropriate sense. Our equivalence tester is deterministic and has polynomial time complexity, whereas the non-approximated version is PSPACE-complete. Last, we extend the geometric embedding, and hence the tester algorithms, to infinite regular languages and to context-free grammars as well.

For context-free grammars the equivalence test has now exponential time complexity, but in comparison, the non-approximated version is not even recursively decidable.

Keywords: Property testing, Equivalence testing, Edit distance with Moves

Joint work of: Fischer, Eldar; Magniez, Frédéric; de Rougemont, Michel

Full Paper:
http://eccc.uni-trier.de/eccc-reports/2004/TR04-096/Paper.pdf