08341 Abstracts Collection
Sublinear Algorithms
— Dagstuhl Seminar —

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

[1] University of Warwick, GB
czumaj@dcs.warwick.ac.uk
[2] Google Inc - New York, USA
[3] MIT - Cambridge, USA
ronitt@theory.lcs.mit.edu
[4] Universität Paderborn, D
sohler@informatik.uni-bonn.de

Abstract. From August 17 to August 22, 2008, the Dagstuhl Seminar 08341 "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.

Keywords. Sublinear algorithms, property testing, data streaming, graph algorithms, approximation algorithms

08341 Executive Summary – Sublinear Algorithms

With the increasing role of information technologies we are often confronted with a huge amount of information that is generated without pace by distributed sources or by large scale complex information systems.

In many scenarios, it is not possible to entirely store this information on standard storage devices.

Examples include the World Wide Web, data accumulated in network traffic monitoring, or sensor network data. One of the key challenges in this context is to efficiently process these massive data sets and to extract knowledge by summarizing and aggregating their major features. Most of the time, it is impossible to use traditional algorithms for this purpose. Even linear time algorithms are typically too slow because they require random access to the input data. We require algorithm that either look only at a small random sample of the input or process the data as it arrives extracting a small summary.

Dagstuhl Seminar Proceedings 08341
Sublinear Algorithms
http://drops.dagstuhl.de/opus/volltexte/2008/1698

Algorithms of this type are called *sublinear algorithms*.

The purpose of this workshop was to bring together leading researchers in the are of sublinear algorithms to discuss recent advances in the area, identify new research directions, and discuss open problems.

The area of sublinear algorithms can be split into three subareas: *p*roperty testing, *sublinear time approximation*, 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. Researchers from all three areas came to attend this workshop and we believe that it 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.

Keywords: Sublinear algorithms, property testing, data streaming, graph algorithms, approximation algorithms

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

Extended Abstract: http://drops.dagstuhl.de/opus/volltexte/2008/1696

Overcoming the L_1 non-embeddability barrier: or Choose your host space wisely

Alexandr Andoni (MIT - Cambridge)

A common approach for solving computational problems over a "difficult" metric space is to embed the hard metric into L_1, which admits efficient algorithms and is thus considered an "easy" metric. Over years, this approach has proved successful or partially successful for important spaces such as the edit distance, but it also has inherent limitations: it is provably impossible to go below certain approximation for some metrics.

We propose a new approach, of embedding the difficult space into richer host spaces, namely iterated products of standard spaces like L_1 and L_∞. We show that this class is rich since it contains useful metric spaces with only a constant distortion, and, at the same time, it is tractable and admits efficient algorithms.

Using this approach, we obtain for example the first nearest neighbor data structure with $O(\log \log d)$ approximation for edit distance in non-repetitive strings (the Ulam metric). This approximation is exponentially better than the lower bound for embedding into L_1. Furthermore, we give constant factor approximation algorithms for two other computational problems. Along the way, we answer positively some questions left open in [Ajtai-Jayram-Kumar-Sivakumar, STOC'02]. One of our algorithms has already found applications for smoothed edit distance over 0-1 strings [Andoni-Krauthgamer, ICALP'08].

Keywords: Embedding, normed space, edit distance, nearest neighbor search, sketching, streaming

Joint work of: Andoni, Alexandr; Indyk, Piotr; Krauthgamer, Robert

A Sublinear-Time Approximation Scheme for Bin Packing

Tugkan Batu (London School of Economics)

The bin packing problem is defined as follows: given a set of n items with sizes $0 < w_1, w_2, \ldots, w_n \leq 1$, find a packing of these items into minimum number of unit-size bins possible.

We present a sublinear-time asymptotic approximation scheme for the bin packing problem; that is, for any $\epsilon > 0$, we present an algorithm A_ϵ that has sampling access to the input instance and outputs a value k such that $C_{opt} \leq k \leq (1+\epsilon) \cdot C_{opt} + 1$, where C_{opt} is the cost of an optimal solution. It is clear that uniform sampling by itself will not allow a sublinear-time algorithm in this setting; a small number of items might constitute most of the total weight and uniform samples will not hit them. In this work we use weighted samples, where item i is sampled with probability proportional to its weight: that is, with probability $w_i / \sum_i w_i$. In the presence of weighted samples, the approximation algorithm runs in $\tilde{O}(\sqrt{n} \cdot \mathrm{poly}(1/\epsilon)) + g(1/\epsilon)$ time, where $g(x)$ is an exponential function of x. When both weighted and uniform sampling are allowed, $\tilde{O}(n^{1/3} \cdot \mathrm{poly}(1/\epsilon)) + g(1/\epsilon)$ time suffices. In addition to an approximate value to C_{opt}, our algorithm can also output a constant-size "template" of a packing that can later be used to find a near-optimal packing in linear time.

Joint work of: Batu, Tugkan; Berenbrink, Petra; Sohler, Christian

Testing Linear-Invariant Non-Linear Properties

Arnab Bhattacharyya (MIT - Cambridge)

We consider the task of testing properties of Boolean functions that are invariant under linear transformations of the Boolean cube. Previous work in property testing, including the linearity test and the test for Reed-Muller codes, has mostly focussed on such tasks for linear properties. The one exception is a test due to Green for "triangle freeness": A function $f : \mathcal{F}_2^n \to \mathcal{F}_2$ satisfies this property if $f(x), f(y), f(x+y)$ do not all equal 1, for any pair $x, y \in \mathcal{F}_2^n$.

Here we extend this test to a more systematic study of testing for linear-invariant non-linear properties. We consider properties that are described by a single forbidden pattern (and its linear transformations), i.e., a property is given by k points $v_1, \ldots, v_k \in \mathcal{F}_2^k$ and $f : \mathcal{F}_2^n \to \mathcal{F}_2$ satisfies the property that if for all linear maps $L : \mathcal{F}_2^k \to \mathcal{F}_2^n$ it is the case that $f(L(v_1)), \ldots, f(L(v_k))$ do not all equal 1. We show that this property is testable if the underlying matroid

specified by $v_1, \ldots, v_k$ is a graphic matroid. This extends Green's result to an infinite class of new properties.

Our techniques extend those of Green and in particular we establish a link between the notion of "1-complexity linear systems" of Green and Tao, and graphic matroids, to derive the results.

Keywords: Property Testing, Regularity Lemma, Finite fields

Joint work of: Bhattacharyya, Arnab; Chen, Victor; Sudan, Madhu; Xie, Ning

"Robust" Communication Complexity and Random-Order Data Streams

Amit Chakrabarti (Dartmouth College - Hanover)

Communication complexity typically uses the following setup: Alice gets *these* specific bits of the input and Bob gets *those* bits. Do the lower bounds we know crucially depend on this precise allocation? Or can we prove *robust* lower bounds that hold w.h.p. for random allocations?

We study the communication complexity of evaluating functions when the input data is randomly allocated (according to some known distribution) amongst two or more players, possibly with information overlap. This model naturally extends previously studied variable partition models such as the best-case and worst-case partition models. We call a communication lower bound "robust" if it holds in this new model.

A key application is to the heavily studied data stream model. Our communication results imply space lower bounds in the data stream model with the order of the items in the stream being chosen not adversarially but rather uniformly from the set of all permutations.

Our results include the first robust lower bounds for (two- and multi-party) set disjointness and gap-Hamming-distance (both tight), and for tree pointer jumping. Collectively, these yield lower bounds for a variety of problems in the random-order data stream model, including estimating the number of distinct elements, approximating frequency moments, median finding, quantile estimation and certain graph streaming problems.

Keywords: Communication Complexity, Data Streams, Lower Bounds, Information Complexity, Random Order

Joint work of: Chakrabarti, Amit; Cormode, Graham; McGregor, Andrew

Full Paper:
http://www.cs.dartmouth.edu/∼ac/Pubs/stoc08-rcc.pdf

Algorithms for distributed functional monitoring

Graham Cormode (AT&T Research - Florham Park)

We study what we call *functional monitoring* problems.

We have k players each receiving a stream of items, and communicating with a central coordinator. Let the multiset of items received by player i up until time t be $A_i(t)$. The coordinator's task is to monitor a given function f computed over the union of the inputs $\cup_i A_i(t)$, *continuously* at all times t. The goal is to minimize the number of bits communicated between the players and the coordinator. Of interest is the approximate version where the coordinator outputs 1 if $f \geq \tau$ and 0 if $f \leq (1-\epsilon)\tau$. This defines the (k, f, τ, ϵ) distributed, functional monitoring problem.

Functional monitoring problems are fundamental in distributed systems, in particular sensor networks, where we must minimize communication; they also connect to problems in streaming algorithms, communication complexity, communication theory, and signal processing. Yet few formal bounds are known for functional monitoring.

We give upper and lower bounds for the (k, f, τ, ϵ) problem for some of the basic f's. In particular, we study frequency moments (F_0, F_1, F_2). For F_0 and F_1, we obtain continuously monitoring algorithms with costs almost the same as their one-shot computation algorithms. However, for F_2 the monitoring problem seems much harder.

We give a carefully constructed multi-round algorithm that uses "sketch summaries" at multiple levels of detail and solves the (k, F_2, τ, ϵ) problem with communication $\tilde{O}(k^2/\epsilon + k^{3/2}/\epsilon^3)$.

Joint work of: Cormode, Graham; Muthukrishnan, S.; Yi, Ke

On Distance to Monotonicity and Longest Increasing Subsequence of a Data Stream

Ayse Funda Ergun (Simon Fraser University - Burnaby)

In this talk we consider problems related to the sortedness of a data stream. First we investigate the problem of estimating the distance to monotonicity; given a sequence of length n, we give a deterministic $(2+\epsilon)$-approximation algorithm for estimating its distance to monotonicity in space $O(1/\epsilon^2 log^2(\epsilon n))$. This improves over the randomized $(4+\epsilon)$-approximation algorithm of Gopalan et al. We then consider the problem of approximating the length of the longest increasing subsequence of an input stream of length n. We use techniques from multi-party communication complexity combined with a fooling set approach to prove that any $O(1)$-pass deterministic streaming algorithm that approximates the length of the longest increasing subsequence within $1+\epsilon$ requires $\Omega(\sqrt{n})$ space. This proves the conjecture in Gopalan et al. and matches the current upper bound.

Joint work of: Ergun, Ayse Funda; Jowhari, Hossein

See also: SODA 2008

Testing orientations for being Eulerian

Eldar Fischer (Technion - Haifa)

The graph orientation model, initiated by Halevy, Lachish, Newman and Tsur, is the following: An undirected graph is given to the algorithm in advance (making it in essence a parameter of the testing problem), and the input to be queried is a directed graph whose edges are orientations of the edges of the given undirected graph. Distances are measured with respect to the number of edges in the undirected graph, which determines how dense or sparse the problem is.

One of the natural questions of this model is testing whether the directed graph is Eulerian, a property that is equivalent to every vertex having an incoming degree identical to its outgoing degree.

As it turns out, testing this property is surprisingly involved, because graphs can be far from being Eulerian in hard to detect ways. In essence the only way to detect a violation of this property is to show that a cut is unbalanced in the directions of its edges, while even a small imbalance can trigger a domino effect of many required alterations.

Our work shows that it is still possible to test every graph with m edges for being Eulerian using m^a for $a < 1$ queries, but the only currently known way to do so is adaptive and computationally inefficient. We have a very strong worst case lower bound for 1-sided algorithms, while for general algorithms only weaker non-constant lower bounds are known (which however hold even for simple tori graphs).

The talk presents the essence of the difficulties and known solutions to this problem.

Keywords: Property testing, graph orientation, massively parameterized properties

Joint work of: Fischer, Eldar; Lachish, Oded; Matsliah, Arie; Newman, Ilan; Yahalom, Orly

Lower bound for estimating frequency for update data streams

Sumit Ganguly (Indian Inst. of Technology - Kanpur)

We consider general update streams, where, the stream is a sequence of updates of the form $(index, i, v)$, where, $i \in \{1, 2 \ldots, n\}$ and $v \in \{-1, +1\}$, signifying deletion or insertion, respectively of an instance of i. The frequency of $i \in \{1, 2, \ldots, n\}$ is given as the sum of the updates to i, that is, $f_i(\sigma) = \sum_{(index,i,v)\in\sigma} v$. The n-dimensional vector $f(\sigma)$ with ith coordinate $f_i(\sigma)$ is called the frequency vector of the stream σ. We consider the problem of finding an n-dimensional integer vector $\hat{f}(\sigma)$ that estimates the frequency vector $f(\sigma)$ of an input stream σ in the following sense:

$$||\hat{f}(\sigma) - f(\sigma)||_\infty \leq \epsilon ||f(\sigma)||_p$$

For $p = 1$ and 2, there are randomized algorithms known with space bound $\tilde{O}(\epsilon^{-p})$. A space lower bound of $\Omega(\epsilon^{-1} \log(n\epsilon))$ is also known. However, the deterministic space upper bound is $\tilde{O}(\epsilon^{-2})$.

In this work, we present a deterministic space lower bound of $\Omega(n^{2-2/p}\epsilon^{-2} \log |\sigma|)$, for $1 \leq p < 2$ and $1/4 \leq \epsilon = \Omega(n^{1/2-1/p})$. For $p \geq 2$, we show an $\Omega(n)$ space lower bound for all $\epsilon < 1/4$.

The results are obtained using a new characterization of data stream computations, that show that any uniform computation over a data stream may be viewed as an appropriate linear map.

Keywords: Data stream, lower bound, frequency estimation, stream automata, linear map

Full Paper: http://drops.dagstuhl.de/opus/volltexte/2008/1695

Full Paper:
http://www.cse.iitk.ac.in/users/sganguly/csr-full.pdf

Whither sublinear algorithms in engineering?

Anna C. Gilbert (University of Michigan)

I will talk about recent development of hardware devices in electrical engineering which employ sublinear measurement schemes of signals, images, and video and sometimes (but not often!) sublinear algorithms for analysis of those measurements. I will discuss what role, if any, sublinear algorithms play in these devices and where both the algorithms and the engineering is headed.

Keywords: Sublinear algorithms, compressed sensing

On Proximity Oblivious Testing

Oded Goldreich (Weizmann Inst. - Rehovot)

We initiate a systematic study of a special type of property testers.

These testers consist of repeating a basic test for a number of times that depends on the proximity parameters, whereas the basic test is oblivious of the proximity parameter.

We refer to such basic tests by the term proximity-oblivious testers.

While proximity-oblivious testers were studied before - most notably in the algebraic setting - the current study seems to be the first one to focus on graph properties.

We provide a mix of positive and negative results, and in particular characterizations of the graph properties that have constant-query proximity-oblivious

testers in the two standard models (i.e., the adjacency matrix and the bounded-degree models).

Furthermore, we show that constant-query proximity-oblivious testers do not exist for many easily testable properties, and that even when proximity-oblivious testers exist repeating them does not necessarily yield the best standard testers for the corresponding property.

Keywords: Property Testing, Graph Properties

Joint work of: Goldreich, Oded; Ron, Dana

Lower bounds on streaming algorithms for approximating the length of the Longest Increasing Subsequence

Anna Gál (Univ. of Texas at Austin)

We show that any deterministic data-stream algorithm that makes a constant number of passes over the input and gives a constant factor approximation of the length of the longest increasing subsequence in a sequence of length n must use space $\Omega(\sqrt{n})$.

This proves a conjecture made by Gopalan, Jayram, Krauthgamer and Kumar [GJKK07] who proved a matching upper bound. Our results yield asymptotically tight lower bounds for all approximation factors.

This is joint work with Parikshit Gopalan.

Joint work of: Gál, Anna; Gopalan, Parikshit

Longest Increasing Subsequence and Distance to Monotonicity in Data Stream Model

Hossein Jowhari (Simon Fraser University - Burnaby)

We consider problems related to the sortedness of a data stream. First we investigate the problem of estimating the distance to monotonicity; given a sequence of length n, we give a deterministic $(2+\epsilon)$-approximation algorithm for estimating its distance to monotonicity in space $O(\frac{1}{\epsilon^2}\log^2(\epsilon n))$. This improves over the randomized $(4+\epsilon)$-approximation algorithm of Gopalan et al SODA 2007.

Then we consider the problem of approximating the length of the longest increasing subsequence of an input stream of length n. We use techniques from multi-party communication complexity combined with a fooling set approach to prove that any $O(1)$-pass deterministic streaming algorithm that approximates the length of the longest increasing subsequence within $1+\epsilon$ requires $\Omega(\sqrt{n})$ space. This proves the conjecture in Gopalan et al, and matches the current upper bound.

Keywords: Longest increasing subsequence, Distance to Monotonicity, Data Stream, Communication Complexity

Joint work of: Ergun, Funda; Jowhari, Hossein

Full Paper:
http://portal.acm.org/citation.cfm?id=1347162&jmp=cit&coll=GUIDE&dl=GUIDE

Trading Tensors for Combinatorial Cloning: Approximation Schemes for Metric CSP

Marek Karpinski (Universität Bonn)

We design constant time approximation schemes (CTASs) for the metric and quasi-metric CSP problems in a preprocessed model of computation. They entail the first sublinear time approximation schemes for the above classes of optimization problems.

Keywords: Approximation Algorithms, Metric and Quasi-Metric CSP, Tensors, Combinatorial Cloning

Joint work of: Karpinski, Marek; de la Vega, W. Fernandez

Algebraic Property Testing: The Role of Invariance

Tali Kaufman (MIT)

We argue that the symmetries of a property being tested play a central role in property testing. We support this assertion in the context of algebraic functions, by examining properties of functions mapping a vector space K^n over a field K to a subfield F. We consider (F-)linear properties that are invariant under linear transformations of the domain and prove that an $O(1)$-local "characterization" is a necessary and sufficient condition for $O(1)$-local testability, when $|K| = O(1)$. (A local characterization of a property is a definition of a property in terms of local constraints satisfied by functions exhibiting a property.) For the subclass of properties that are invariant under affine transformations of the domain, we prove that the existence of a *single* $O(1)$-local constraint implies $O(1)$-local testability. These results generalize and extend the class of algebraic properties, most notably linearity and low-degree-ness, that were previously known to be testable. In particular, the extensions include properties satisfied by functions of degree linear in n that turn out to be $O(1)$-locally testable.

Joint work of: Kaufman, Tali; Sudan, Madhu

Sound 3-query PCPPs are Long

Oded Lachish (University of Warwick)

We initiate the study of the tradeoff between the length of a probabilistically checkable proof of proximity (PCPP) and the maximal soundness that can be guaranteed by a 3-query verifier with oracle access to the proof.

Our main observation is that a verifier limited to querying a short proof cannot obtain the same soundness as that obtained by a verifier querying a long proof.

Moreover, we quantify the soundness deficiency as a function of the proof-length and show that any verifier obtaining "best possible" soundness must query an exponentially long proof.

In terms of techniques, we focus on the special class of inspective verifiers that read at most 2 proof-bits per invocation. For such verifiers we prove exponential length-soundness trade-offs that are later on used to imply our main results for the case of general (i.e., not necessarily inspective) verifiers. To prove the exponential tradeoff for inspective verifiers we show a connection between PCPP proof length and property-testing query complexity, that may be of independent interest. The connection is that any linear property that can be verified with proofs of length m by linear inspective verifiers must be testable with query complexity approximately $\log(m)$.

Keywords: PCPP, Property testing

Joint work of: Ben-Sasson, Eli; Harsha, Prahladh; Lachish, Oded; Matsliah, Arie

Facility Location in Dynamic Geometric Data Streams

Christiane Lammersen (Universität Bonn)

In the uniform facility location problem, we are given a set of clients that have to be served by a set of facilities. It is possible to open a facility at any client for a given cost of f. The cost of serving a client is proportional to its distance to the nearest facility.

We present a randomized algorithm that maintains a constant factor approximation for the cost of the facility location problem in the dynamic geometric data stream model. In this model, the input is a sequence of insert and delete operations of points from a discrete space $\{1 \dots \Delta\}^d$, where d is a constant.

The algorithm needs $\log^{\mathcal{O}(1)} \Delta$ time to process an insertion or deletion of a point, uses $\log^{\mathcal{O}(1)} \Delta$ bits of storage, and has a failure probability of $1/\Delta^{\Theta(1)}$.

Keywords: Facility location, dynamic data streams, approximation

Joint work of: Lammersen, Christiane; Sohler, Christian

See also: D. Halperin and K. Mehlhorn (Eds.): ESA 2008, LNCS 5193, pp. 660-671, Springer-Verlag Berlin Heidelberg, 2008

Approximate Hypergraph Partitioning and Applications

Arie Matsliah (Technion - Haifa)

We show that many partition-problems of dense hypergraphs have an $O(n)$ time (sublinear-time) approximate partitioning algorithm and an efficient property tester. This extends the results of Goldreich, Goldwasser and Ron who obtained similar algorithms for the case of graph partition problems in their seminal paper [GGR98].

The partitioning algorithm is used to obtain the following results:

A surprisingly simple $O(n)$ time algorithmic version of Szemerédi's regularity lemma. Unlike the previous approaches for this problem, which only guaranteed to find partitions of towersize, our algorithm will find a small regular partition in the case that one exists;

For any $r \geq 3$, we give an $O(n)$ time randomized algorithm for constructing regular partitions of r-uniform hypergraphs, thus improving the previous $O(n^{2r-1})$ time (deterministic) algorithms;

The property tester is used to unify several previous results, and to obtain the partition densities for the above problems (rather than the partitions themselves) by making only a constant number of queries.

Keywords: Hypergraph, testing, regularity

Joint work of: Matsliah, Arie; Fischer, Eldar; Shapira, Assaf

Testing Halfspaces

Kevin Matulef (MIT - Cambridge)

We address the problem of testing whether a Boolean-valued function f is a halfspace, i.e. a function of the form $f(x) = \mathrm{sgn}(w * x - t)$. We consider functions over the continuous domain R^n endowed with the standard multivariate Gaussian distribution, as well as halfspaces over the boolean cube $\{-1, 1\}^n$ endowed with the uniform distribution. In both cases we give an algorithm that distinguishes halfspaces from functions that are ϵ-far from any halfspace using only $\mathrm{poly}(1/\epsilon)$ queries, independent of the dimension n.

Keywords: Property testing, halfspaces, fourier analysis

Joint work of: Matulef, Kevin; O'Donnell, Ryan; Rubinfeld, Ronitt; Servedio, Rocco

Lower Bounds for Stream Computation via Pass Elimination (feat. More LIS Bounds!)

Andrew McGregor (University of California)

There is a natural relationship between lower bounds in the multi-pass stream model and lower bounds in multi-round communication.

However, this connection is less understood than the connection between single-pass stream computation and one-way communication. In this paper, we consider data-stream problems for which reductions from natural multi-round communication problems do not yield tight bounds or do not apply. While lower bounds are known for some of these data-stream problems, many of these only apply to deterministic or comparison-based algorithms, whereas the lower bounds we present apply to any (possibly randomized) algorithms. Our results are particularly relevant to evaluating functions that are dependent on the ordering of the stream, such as the longest increasing subsequence and a variant of tree pointer jumping in which pointers are revealed according to a post-order traversal.

Our approach is based on establishing "pass-elimination" type results that are analogous to the round-elimination results of Miltersen et al. [MNSW98] and Sen [Sen03]. We demonstrate our approach by proving tight bounds for a range of data-stream problems including finding the longest increasing sequences (a problem that has recently become very popular [LVZ06, GJKK07, SW07, GalG07, ErgunJ08] and we resolve an open question of [SW07]), constructing convex hulls and fixed-dimensional linear programming (generalizing results of [ChanC07] to randomized algorithms), and the "greater-than" problem (improving results of [CK06]). These results will also clarify one of the main messages of our work: sometimes it is necessary to prove lower bounds directly for stream computation rather than proving a lower bound for a communication problem and then constructing a reduction to a data-stream problem.

Keywords: Data streams, Communication Complexity, Longest Increasing Subsequence

Joint work of: Guha, Sudipto; McGregor, Andrew

Full Paper:
http://talk.ucsd.edu/andrewm/papers/08-lis.pdf

See also: ICALP 2008

On Trade-Offs in External-Memory Diameter-Approximation

Ulrich Carsten Meyer (J.W. Goethe Universität Frankfurt)

Computing diameters of huge graphs is a key challenge in complex network analysis. However, since exact diameter computation is computationally too costly, one typically relies on approximations.

In fact, already a single BFS run rooted at an arbitrary vertex yields a factor two approximation. Unfortunately, in external-memory, even a simple graph traversal like BFS may cause an unacceptable amount of I/O-operations. Therefore, we investigate alternative approaches with worst-case guarantees on both I/O-complexity and approximation factor.

Keywords: External-Memory, Graph Algorithms, Diameter, Approximation

Full Paper:
http://dx.doi.org/10.1007/978-3-540-69903-3_38

See also: J. Gudmundsson (Ed.): SWAT 2008, LNCS 5124, pp. 426-436, Springer-Verlag Berlin Heidelberg, 2008

High Dimensional Clustering Problems

Morteza Monemizadeh (Universität Bonn)

Given a point set P in d-dimensional space and an integer $j > 0$, the j-flat mean clustering problem is to find a flat OPT of dimension j such that the sum of squared distances between a point $q \in P$ and OPT is minimized. In this paper we show that every point set P has a weak coreset of size $O(2^{j^2})$ for this problem. A weak coreset, S, is a weighted set not necessarily a subset of P together with a set T such that T contains a $(1+\epsilon)$-approximation of OPT and for every j-dimensional flat $F \in T$, the cost of F for S is a $(1+\epsilon)$-approximation of the cost of F for P.

We apply our weak coreset to obtain a PTAS for the one j-flat mean problem with running time $O(ndj^2 + d2^{j^6})$. We also show this problem can be solved in the one-pass data stream scenario and in the space $O(d \log^6 n 2^{j^2})$.

Joint work of: Feldman, Dan; Monemizadeh, Morteza; Sohler, Christian

Property Testing in the subgraph/orientation model - a survey

Ilan Newman (Haifa University)

Property Testing considers the following relaxation of standard decision problems: Given a property $\mathcal{P}$ of some combinatorial structures, one wants to decide whether a given instance (structure) S has the property $\mathcal{P}$ or is ϵ-far from having the property. By ϵ-far we mean that at least an ϵ-fraction of the representation of S should be modified in order to make S satisfy $\mathcal{P}$. The goal in property testing is to design randomized algorithms, called *testers*, which read a very small portion of the input and distinguish between the two cases. The complexity of the tester is the number of queries that it asks from the input.

We consider here the problem of property testing in a model that was introduced by Halevi et. al (2005) and is referred to as the *underlying graph model*. In this model the tester has full knowledge of an underlying undirected fixed graph $G = (V, E)$. A property is then a collection of assignments on the edges (vertices). Such an assignment can be interpreted in several ways. One such interpretation is considering the assignment as a characteristic function of the

edges of a subgraph of G. Hence, a property will be a collection of G-subgraphs, e.g the subgraphs that are bipartite, 3-colorable and so on. Another is to interpret the Boolean assignment on the edges as an orientation of the edges (relative to some fixed predefined orientation). In this case, a property is a property of G-orientations e.g being strongly connected, Eulerian etc. A third, which sounds less natural, but has interesting connections to other areas in theoretical CS (such as PCP's) is an interpretation of the assignment as an assignment to a collection of edge variables. In this case, we also assume some fixed predefined collection of vertex formulae; $\{\phi_v,\ v \in V\}$ where ϕ_v is defined on the edge-variables that are associated with the edges that are adjacent to v. A property in this case contains all the assignments that satisfy every vertex formula (e.g: the formula in each vertex asserts that the number of 1-edges is equal to the number of 0-edges, which is equivalent to Eulerianity, in this case).

I will survey several positive results (namely, a construction of testers) for several natural properties, several impossibility results (lower bounds on the complexity of any tester) and some connections to other models. Among these we will discuss certain types of constraint graph formulae (as explained above, in the third interpretation), the properties of being connected, having st-path and being Eulerian, in the orientation interpretation, and properties such as being bipartite, planar and others in the subgraph model.

Keywords: Property testing, orientation model

Joint work of: Chakraborty, Sourav; Fischer, Eldar; Halevi, Shirley; Lachish, Oded; Matsliah, Arie; Newman, Ilan; Rozenberg, Eyal; Tsur, Dekel; Yahalom, Orly

Constant-Time Approximation Algorithms via Local Improvements

Krzysztof Onak (MIT - Cambridge)

We present a technique for transforming classical approximation algorithms into constant-time algorithms that approximate the size of the optimal solution. Our technique is applicable to a certain subclass of algorithms that compute a solution in a constant number of phases. The technique is based on greedily considering local improvements in random order.

The problems amenable to our technique include Vertex Cover, Maximum Matching, Maximum Weight Matching, Set Cover, and Minimum Dominating Set. For example, for Maximum Matching, we give the first constant-time algorithm that for the class of graphs of degree bounded by d, computes the maximum matching size to within ϵn, for any $\epsilon > 0$, where n is the number of nodes in the graph. The running time of the algorithm is independent of n, and only depends on d and ϵ.

Keywords: Approximation algorithms, sublinear-time algorithms, matchings

Joint work of: Nguyen, Huy N.; Onak, Krzysztof

Transitive-Closure Spanners

Sofya Raskhodnikova (Pennsylvania State University)

We define the notion of a transitive-closure spanner of a directed graph. Given a directed graph $G = (V, E)$ and an integer $k >= 1$, a k-transitive-closure-spanner (k-TC-spanner) of G is a directed graph $H = (V, E_H)$ that has (1) the same transitive-closure as G and (2) diameter at most k. These spanners were studied implicitly in property testing, access control, and data structures, and properties of these spanners have been rediscovered over the span of 20 years. We bring these areas under the unifying framework of TC-spanners. We abstract the common task implicitly tackled in these diverse applications as the problem of constructing sparse TC-spanners.

We study the size of the sparsest k-TC-spanners for specific graph families, as well as the computational task of finding the sparsest k-TC-spanner for a given input graph. We present new positive and negative results on both fronts. In the talk, we will explain some ramifications of these results for testing monotonicity of functions.

Keywords: Spanners, property testing, access control, data structures, monotonicity of functions

Joint work of: Bhattacharyya, Arnab; Grigorescu, Elena; Jung, Kyomin; Raskhodnikova, Sofya; Woodruff, David

On the benefits of adaptivity in property testing of dense graphs

Dana Ron (Tel Aviv University)

We consider the question of whether adaptivity can improve the complexity of property testing algorithms in the dense graphs model. It is well known that there can be at most a quadratic gap between adaptive and non-adaptive testers in this model, but it was not known whether any gap indeed exists. In this this talk I will discuss such gaps for several properties, amongst them bipartiteness.

In addition to demonstrating that adaptivity can play a role in the dense-graphs model, these results show that testing in the dense-graphs model is not only a question of combinatorics. Rather, in some cases, in order to obtain better bounds on the query complexity, algorithmic aspects should come into play.

Keywords: Property testing, Dense-graphs model, Adaptivity

Joint work of: Goldreich, Oded; Gonen, Mira; Ron, Dana

Lower bounds for multi-pass processing of multiple streams

Nicole Schweikardt (Universität Frankfurt)

In this talk I consider the following machine model:

Two streams S and T are processed by k cursors (i.e. "heads"). Each cursor either processes S or T, and each cursor can move one-way only (either forward or backward). Note, however, that cursors are allowed to move asynchronously. Throughout the computation, internal memory that consists of $\log m$ bits may be used. During each computation step, the machine sees the elements of S and T at the current cursor positions, and the current content of internal memory. Depending on these pieces of information, a deterministic transition function tells the machine (a) the new content of internal memory, and (b) which of the k cursors should be advanced to the next position. The main result presented in this talk is a lower bound for solving the set-disjointness problem DISJ_n. The precise definition of DISJ_n is as follows: Let U be a set of size at least $2n$. The input of DISJ_n consists of two subsets S and T of U with $|S| = |T| = n$. These two sets are represented by two streams that enumerate the elements in an arbitrary order. The goal is to decide whether S and T are disjoint.

THEOREM: If $(k^5 \log m + k^6 \log n) < n/2$, then no machine of the above kind can solve the problem DISJ_n.

[Based on joint work with M. Grohe, Y. Gurevich, D. Leinders, J. Tyszkiewicz, J. Van den Bussche. "Database Query Processing using Finite Cursor Machines". Proc. ICDT 2007. Journal version to appear in Theory of Computing Systems, 2008. The result presented in this talk can be found as Theorem 5.11 in the journal version, resp. as Theorem 12 in the conference proceedings.]

Distributed Monotonicity Reconstruction

C. Seshadhri (Princeton University)

We investigate the problem of monotonicity reconstruction, and introduce the model of distributed reconstruction.

We have oracle access to a nonnegative real-valued function f defined on domain $[n]^d = \{1, \ldots, n\}^d$. We would like to closely approximate f by a monotone function g. This should be done by a procedure (a filter) that given as input a point $x \in [n]^d$ outputs the value of $g(x)$, and runs in time that is highly sublinear in n. The procedure can (indeed must) be randomized, but we require that all of the randomness be specified in advance by a single short random seed. We construct such an implementation where the time and space per query is $(\log n)^{O(1)}$ and the size of the seed is polynomial in $(\log n)$ and d. Furthermore the distance of the approximating function g from f is at most a constant multiple of the minimum distance of any monotone function from f.

This allows for a distributed implementation: one can initialize many copies of the filter with the same short random seed, and they can autonomously handle queries, while producing outputs that are consistent with the same approximating function g.

Joint work of: Saks, Michael; Seshadhri, C.

Sublinear Algorithms in Private Data Anlaysis

Adam D. Smith (Penn State University)

Consider an agency holding a large database of sensitive personal information (perhaps medical records, census answers, or web search records). The agency would like to discover and publicly release global characteristics of the data (say, to inform policy and business decisions) while protecting the privacy of individuals whose data it collected. This problem has been studied in statistics and data mining, and recently received attention in theoretical computer science.

I will describe a recent notion, "differential privacy", which formalizes the privacy requirement for statistical databases. Informally, an algorithm is differentially private if its output does not depend too heavily on individual inputs. I will discuss how techniques from sublinear algorithms can be used to design differentially private algorithms and outline a few directions for future research.

We will present several techniques for designing differentially private algorithms, and demonstrate them on computational problems ranging from finding the average and the median of individuals' salaries to general learning tasks.

Revisiting Norm Estimation in a Data Stream

David P. Woodruff (IBM Almaden Center - San José)

There are three independent components of the talk concerning norm estimation in a data stream.

In the first part, I will describe a $\log m$-pass $m^{1-2/k}\text{polylog}(mM/\epsilon)$ space algorithm for $(1+\epsilon)$-approximating the L_k norm, $k > 2$, in a data stream with n arbitrary updates to a vector of m coordinates, each coordinate bounded by M. It is already known (Indyk-Woodruff, STOC 2005 and Bhuvanagiri et al, SODA 2006) that $m^{1-2/k}\text{polylog}(mM/\epsilon)$ space is optimal in 1-pass. The current algorithm is much simpler than either of those. It directly follows an approach suggested earlier by Kumar of building a low-space L_p sampler for any $0 \leq p \leq 2$, which may be of independent interest.

In the second part, I will describe a way to improve the $\Omega(1/\epsilon^2)$ bound (Indyk-Woodruff, FOCS 2003 and Woodruff, SODA 2004) for $(1+\epsilon)$-approximating L_p for any $p \geq 0$ to achieve an $\Omega(\log n)/(\epsilon^2 \log 1/\epsilon)$ bound when there are deletions. It is known how to achieve less space for estimating F_0 in an insertion-only stream, so this lower bound separates the deletion-model from the insertion-only

model for distinct elements. The technique establishes tight lower bounds for L_0 and L_2 estimation, as well as stronger lower bounds for many problems, such as additive entropy approximation. The tight lower bound for L_0 follows from a new space and time-optimal algorithm that I will give an overview of.

In the last part, I will describe extensions to norms of product spaces. Given an $n \times n$ matrix A presented as an arbitrary sequence of insertions and deletions, we want to compute $F_k(F_p)(A)$, meaning that we first compute F_p of each of the rows of A, and then F_k on the resulting column of F_p values. I will describe an algorithm which is space-optimal (up to polylog(mMn/ϵ) factors) for any k and any $p \geq 2$. This resolves several open questions of Cormode and Muthukrishnan (PODS, 2005).

Keywords: Data stream, norm estimation

Joint work of: Jayram, T.S.; Nelson, Jelani; Woodruff, David P.

Breaking the ϵ-Soundness Bound of the Linearity Test over GF(2)

Ning Xie (MIT - Cambridge)

For Boolean functions that are ϵ-far from the set of linear functions, we study the lower bound on the rejection probability (denoted by $\text{REJ}(\epsilon)$) of the linearity test suggested by Blum, Luby and Rubinfeld.

This problem is arguably the most fundamental and extensively studied problem in property testing of Boolean functions.

The previously best bounds for $\text{REJ}(\epsilon)$ were obtained by Bellare, Coppersmith, Håstad, Kiwi and Sudan. They used Fourier analysis to show that $\text{REJ}(\epsilon) \geq \epsilon$ for every $0 \leq \epsilon \leq \frac{1}{2}$. They also conjectured that this bound might not be tight for ϵ's that are close to 1/2. In this paper we show that this indeed is the case.

Specifically, we improve the lower bound of $\text{REJ}(\epsilon) \geq \epsilon$ by an additive term that depends only on ϵ: $\text{REJ}(\epsilon) \geq \epsilon + \min\{1376\epsilon^3(1-2\epsilon)^{12}, \frac{1}{4}\epsilon(1-2\epsilon)^4\}$, for every $0 \leq \epsilon \leq \frac{1}{2}$.

Our analysis is based on a relationship between $\text{REJ}(\epsilon)$ and the weight distribution of a coset of the Hadamard code. We use both Fourier analysis and coding theory tools to estimate this weight distribution.

Joint work with Tali Kaufman and Simon Litsyn.

Keywords: Linearity test, Fourier analysis, coding theory

Joint work of: Kaufman, Tali; Litsyn, Simon; Xie, Ning

Full Paper: http://drops.dagstuhl.de/opus/volltexte/2008/1697

Full Paper:
http://www.springerlink.com/content/0ph6155h41704hl4/

See also: APPROX-RANDOM 2008: 498-511

Algorithms for Streaming Graphs

Mariano Zelke (HU Berlin)

A semi-streaming algorithm has no random access to the input graph and uses a working memory of restricted size, which is sublinear for dense graphs.

We will see how to obtain semi-streaming algorithms with optimal running times for testing graph connectivity, bipartiteness and the computation of a minimum spanning tree. Interestingly, these running times match the corresponding ones in the traditional RAM model. For the problem of finding a maximum weighted matching, which is intractable in the semi-streaming sodel, we will discuss the best know approximation algorithm. Moreover, we will see why the computation of a minimum cut in a graph is not possible for a streaming algorithm and how such a cut can be approximated in a randomized fashion.

Keywords: Streaming algorithm, graph algorithm, semi-streaming model