OPEN PROBLEMS IN DATA STREAMS, PROPERTY TESTING, AND RELATED TOPICS
June 14, 2011

ABSTRACT. This document contains a list of open problems and research directions that have been suggested by participants at the Bertinoro Workshop on Sublinear Algorithms (May 2011) and IITK Workshop on Algorithms for Processing Massive Data Sets (December 2009). Many of the questions were discussed at the workshop or were posed during presentations. Further details can be found at

`www.dcs.warwick.ac.uk/~czumaj/Bertinoro_2011`
`www2.cse.iitk.ac.in/~fsttcs/2009/wapmds`

Lists compiled by Piotr Indyk (indyk@mit.edu), Andrew McGregor (mcgregor@cs.umass.edu), Ilan Newman (ilan@cs.haifa.ac.il), and Krzysztof Onak (konak@cs.cmu.edu).

BERTINORO WORKSHOP PARTICIPANTS:
Nir Ailon, Noga Alon, Alexandr Andoni, Arnab Bhattacharyya, Vladimir Braverman, Amit Chakrabarti, Graham Cormode, Artur Czumaj, Pierre Fraigniaud, Oded Goldreich, Nir Halman, Sariel Har-Peled, Piotr Indyk, Tali Kaufman, Robert Krauthgamer, Oded Lachish, Michael Mahoney, Andrew McGregor, Morteza Monemizadeh, Jelani Nelson, Ilan Newman, Krzysztof Onak, Ely Porat, Sofya Raskhodnikova, Ronitt Rubinfeld, Rocco Servedio, Madhu Sudan, Ben Recht, Justin Romberg, Dana Ron, C. Seshadhri, Asaf Shapira, Christian Sohler, Gilad Tsur, Paul Valiant, Roger Wattenhofer, David Woodruff, Ning Xie, and Yuichi Yoshida.

KANPUR WORKSHOP PARTICIPANTS:
Pankaj K. Agarwal, Kook Jin Ahn, Paul Beame, Amit Chakrabarti, Inderjit Dhillon, Dan Feldman, Sumit Ganguly, Sudipto Guha, Piotr Indyk, T. S. Jayram, Christiane Lammersen, Michael Mahoney, Andrew McGregor, Jelani Nelson, Krzysztof Onak, Rina Panigrahy, Ely Porat, Jaikumar Radhakrishnan, Christian Sohler, Joel Tropp, Matthias Westermann, and David Woodruff.

CONTENTS

THE BERTINORO LIST

QUESTION 1: LEARNING AN f-TRANSFORMED PRODUCT DISTRIBUTION (ROCCO A. SERVEDIO)

In this learning setting there are n independent Bernoulli random variables $X_1, ..., X_n$ with *unknown* $E[X_i] = p_i$. There is a *known* transformation function $f : \{0,1\}^n \mapsto R$, where R is some range. The learner has access to independent draws from $f(X_1, ..., X_n)$; i.e. each example for the learner is obtained by independently drawing $X_1, ..., X_n$, applying f, and giving the result to the learner. Call this distribution D_f. The learner's job is to construct a hypothesis distribution D' over the range set such that the variation distance between D_f and D' is at most ϵ, with high probability.

Question: Give some necessary or sufficient conditions on f that make the "learn an f-transformed product distribution" problem solvable using $O_\epsilon(1)$ queries, independent of n.

Background: The following is known [DDS11]:

(1) For $f(X) = X_1 + ... + X_n$, there's a learning algorithm using poly$(1/\epsilon)$ queries independent of n.
(2) For $f(X) = \sum_{i=1}^n i \cdot X_i$, any algorithm for learning to constant accuracy must make $\Omega(n)$ queries.

QUESTION 2: TESTING SUBMODULARITY (C. SESHADHRI)

A function $f : \{0,1\}^n \mapsto R$ is *submodular* if for every $i \in [n]$ and every $S \subset T$, such that $i \notin T$,

$$f(T \cup \{i\}) - f(T) \leq f(S \cup \{i\}) - f(S) .$$

Question: How efficient can we test that f is submodular (in terms of number of queries to f). In particular, can one do it in $\mathrm{poly}(n/\epsilon)$? Special cases of interest that are open:

(1) f is monotone and for every S and $i \in [n]$, $f(S \cup \{i\}) - f(S)$ is either 0 or 1. In this case f is the rank function of a matroid.
(2) A more special case (suggested by Noam Nisan): f is said to be a *coverage valuation* if every $i \in [n]$ is associated with a set V_i from some measurable space with a measure μ (we might want to think of V_i as discrete, in which case the measure is just the cardinality). Then f is defined by $f(S) = \mu(\bigcup_{i \in S} V_i)$. Observe that such f is a submodular function.

Background: The problem is interesting in algorithmic game theory. The best known upper bound on the number of queries is $O(\epsilon^{-\sqrt{n}\log n})$ [SV11]. We do not know the answer even for constant size R, although for $R = \{0,1\}$ it is easy.

QUESTION 3: QUERY COMPLEXITY OF LOCAL PARTITIONING ORACLES (KRZYSZTOF ONAK)

A local partitioning oracle is defined in the paper of Hassidim, Kelner, Nguyen, and Onak [HKNO09], and an implicit construction of a partitioning oracle is shown in the earlier paper of Benjamini, Schramm, and Shapira [BSS08]. Partitioning oracles are a useful abstraction for approximation and testing algorithms in the bounded degree model.

The best known oracle for bounded-degree planar graphs makes at most $d^{\mathrm{poly}(1/\epsilon)}$ queries to the underlying graph to answer each query about the resulting partition, where d is the bound on the maximum vertex degree in the graph. See [Ona10] for a description of the method.

Question: Can one design an oracle that makes only $\mathrm{poly}(d/\epsilon)$ queries? If so, then among other things, this would lead to a tester for planarity in the bounded-degree model that makes only $\mathrm{poly}(1/\epsilon)$ queries, resolving an open question of Benjamini et al. [BSS08].

Question 4: Approximating Maximum Matching Size (Krzysztof Onak)

Consider graphs with maximum degree bounded by d. It is possible to approximate the size of the maximum matching up to an additive ϵn in time that is a function of only ϵ and d [NO08, YYI09]. The fastest currently known algorithm runs in $d^{O(1/\epsilon^2)}$ time [YYI09].

Question: Is there an algorithm that runs in $\mathrm{poly}(d/\epsilon)$ time?

Question 5: Testing Monotonicity and the Lipschitz Property (Sofya Raskhodnikova)

Positive answers to the conjectures below would imply better testers for monotonicity and the Lipschitz property. Consider a function $f : \{0,1\}^d \to \mathbb{R}$. It corresponds to a d-dimensional hypercube with the vertex set $\{0,1\}^d$ and (directed or undirected, depending on the problem) edges (x, y) for all x and y, where y can be obtained from x by increasing one bit. Each node x is labeled by a real number $f(x)$.

(1) A directed edge (x, y) of the hypercube is *violated* if $f(x) > f(y)$. Function f is *monotone* if no edges are violated.

Question: Suppose v edges are violated. Give an upper bound on the number of node labels that have to be changed to make f monotone.

Background: The best known bound is vd [DGL$^+$99] but the conjecture is v.

(2) An undirected edge (x, y) of the hypercube is *violated* if $|f(x) - f(y)| > 1$. Function f is *Lipschitz* if no edges are violated.

Question: Suppose v edges are violated. Give an upper bound on the number of node labels that have to be changed to make function f Lipschitz in terms of v and d.

Background: Nothing nontrivial is known for real labels. The conjecture is $O(v)$. For integer labels, the best known bound is $2v \cdot \mathrm{ImageDiameter}(f)$, where $\mathrm{ImageDiameter}(f) = \max_x f(x) - \min_x f(x)$ [JR11].

Question 6: Testing Acyclicity (Dana Ron)

Consider the problem of testing acyclicity in *directed* bounded-degree graphs (in the incidence list model, where one can query both outgoing and incoming edges).

Question: What is the best algorithm for this problem?

Background: There is a lower bound of $\Omega(n^{1/3})$ for adaptive, two-sided error algorithms, where n is the number of vertices [BR02]. No sublinear upper bound is known. (For dense graphs, in the adjacency matrix model, one can test the property using $\mathrm{poly}(1/\epsilon)$ queries.) The best known lower bound for 1-sided error testing is only $\Omega(\sqrt{n})$.

Question 7: Graph Frequency Vectors (Noga Alon)

For a graph G, a k-disc around a vertex v is the subgraph induced by the vertices that are at distance at most k from v. The frequency vector of k-discs of G is a vector indexed by all isomorphism types of k-discs of vertices in G which counts, for each such isomorphism type K, the fraction of k-discs of vertices of G that are isomorphic to K. The following is a known fact observed in a discussion with Lovász. It is proved by a simple compactness argument.

Fact: For every $\epsilon > 0$, there is an $M = M(\epsilon)$ such that for every 3-regular graph G, there exists a 3-regular graph H on at most $M(\epsilon)$ vertices (independent on $|V(G)|$), such that variation distance between the frequency vector of the 100-discs in G and the frequency vector of the 100-discs in H is at most ϵ.

Question: Find *any* explicit estimate on $M(\epsilon)$. Nothing is currently known.

QUESTION 8: RANK LOWER BOUND (MADHU SUDAN)

We want to prove that the following tall matrix has full column rank. The columns are indexed by $a_1, \ldots, a_k$ from the field F_{2^n} where n is prime; the rows are indexed by degrees $d_1 \ldots d_r$. The entry in the ith column and jth row is equal to $a_i^{d_j}$.

Question: Is it true that for all k there exists an r such that for all $d_1, ..., d_r$ that are powers of 2 and for all $a_1, ..., a_k$ that are linearly independent over F_2, the rank of the matrix is k?

Background: Note that if $d_i = i$ and $r \geq k$, then the matrix is Vandermonde and so has full rank. If $d_i = 2^i$, then also the matrix has full rank [GKS08, Lemma 19]. The general case, when d_i's are arbitrary, and not successive powers of two remains open [BGM$^+$11, Conjecture 5.9].

QUESTION 9: APPROXIMATING LIS LENGTH IN THE STREAMING MODEL (AMIT CHAKRABARTI)

The goal of LIS is to compute a 2-approximation of the length of the longest increasing subsequence in a given stream of elements.

Question: What is the randomized streaming space complexity of LIS, for one pass or possibly a constant number of passes?

Background: Gopalan et al. [GJKK07] gave an $O(n^{1/2} \operatorname{polylog} n)$-space *deterministic* streaming algorithm, using one pass, that achieves c-approximation for any fixed $c > 0$. For deterministic algorithms [EJ08, GG07] showed an $\Omega(n^{1/2})$ space lower bound, for a constant number of passes. The latter arguments proceed by proving a lower bound for related communication complexity problems. However, it is known that the randomized communication complexity of those problem is $O(\log n)$ [Cha10] so a different approach is needed.

QUESTION 10: STREAMING MAX-CUT/MAX-CSP (ROBERT KRAUTHGAMER)

The problem is defined as follows: given a stream of edges of an n-node graph G, estimate the value of the maximum cut in G.

Question: Is there an algorithm with an approximation factor strictly better than $1/2$ that uses $o(n)$ space?

Background: Note that $1/2$ is achievable using random assignment argument. Moreover, using sparsification arguments [Tre09, AG09], one can obtain a better approximation ratio using $O(n \operatorname{polylog} n)$ space. Woodruff and Bhattacharyya (private communication) noted that subsampling $O(n/\epsilon^2)$ edges gives, with high probability, an ϵ-additive approximation for all cuts, and thus $1 + \epsilon$ multiplicative approximation for MAX-CUT.

Question: What about general constraint satisfaction problems with fixed clause-length and alphabet-size? In this case it is even not known how to obtain $O(n \operatorname{polylog} n)$ space bound.

QUESTION 11: FAST JL TRANSFORM FOR SPARSE VECTORS (JELANI NELSON)

Consider a distribution over linear mappings A from R^d to R^k, $k = O(\log(1/P)/\epsilon^2)$. We say that it has an (ϵ, P)-JL property if for any vector $x \in R^d$ we have

$$\|Ax\|_2 = (1 \pm \epsilon)\|x\|_2$$

with probability $1 - P$.

Question: Can we construct a distribution with this property such that the matrix-vector product Ax can be evaluated in time $(s + k) \cdot \mathrm{polylog}(d)$ time given an s-sparse x?

Background: Such an algorithm is not known even for $s = d$ (unless k is larger [AL11]).

Question: Provide an explicit construction of a distribution with the (ϵ, P)-JL property such that the random variable A can be generated using $O(\log(d/P))$ bits.

QUESTION 12: ANNOTATED STREAMING (GRAHAM CORMODE)

In the annotated stream model [CCM09], a stream is augmented with 'annotation', which takes the form of a proof of a property of the stream. In its simplest form, the annotation may just be a reordering of the stream to make it easy to compute a function of interest. The key parameters in this model are H, the size of the annotation, and V, the space needed by the streaming party to view the stream and verify the proof. The annotation proposed should be such that an honest annotation will always be accepted, while a mistaken annotation will be detected and rejected with high probability.

We consider the problem of counting the number of triangles in a graph described by a stream of edges (where each edge is promised to occur at most once). Partial results from the above reference are that $H = O(n^2)$ and $V = \tilde{O}(1)$ is possible, as is $H = O(n^{3/2}), V = O(n^{3/2})$.

Question: Can one achieve $H = V = \tilde{O}(n)$?

QUESTION 13: SKETCHING SHIFT METRICS (ALEX ANDONI)

For any $x, y \in \{0,1\}^n$, define the *shift metric*

$$\mathrm{sh}(x, y) = \min_{\sigma} H(x, \sigma(y)),$$

where σ ranges over all n cyclic permutations of $\{1 \dots n\}$, and $H()$ is the hamming distance.

For any $c > 20$, the promise problem P_c is to distinguish whether $\mathrm{sh}(x, y) > n/10$ or $\mathrm{sh}(x, y) < n/c$. Consider probabilistic mappings $L_c : \{0,1\}^n \to \{0,1\}^s$. We say that L_c is a sketching scheme for P_c if there is an algorithm that, for any $x, y \in \{0,1\}^n$ satisfying the promise of P_c, given $L_c(x)$ and $L_c(y)$, solves P_c with probability at least 0.9.

Question: Is there a sketching scheme for P_c where $c = O(1)$ and $s = O(1)$?

Background: If the shift metric is replaced by Hamming metric, one can achieve $s = O(1)$ using random sampling [KOR00]. The actual problem can be solved for $c = O(\log^2 n)$ and $s = O(1)$ [AIK08]. The algorithm proceeds by embedding the shift metric into Hamming metrics, and it is known that this step must induce $\Omega(\log n)$ distortion [KN06].

Question 14: Sketching Earth Mover Distance (Piotr Indyk)

For any two subsets A, B of the planar grid $[n]^2$, $|A| = |B|$, define

$$\mathrm{EMD}(A, B) = \min_{\pi: A \to B} \sum_{a \in A} \|a - \pi(a)\|_1$$

where π ranges over one-to-one mapping from A to B.

Question: What is the sketching complexity of c-approximating EMD? That is, consider a distribution over mappings L_c that maps subset of $[n]^2$ to $\{0,1\}^s$, such that for any sets A, B with $|A| = |B|$, given $L_c(A), L_c(B)$, one can estimate $\mathrm{EMD}(A, B)$ up to a factor of c, with probability $\geq 2/3$. Is it possible to construct such a distribution for constant c and $s = \mathrm{polylog}\, n$?

Background: It is known that one can achieve $s = O(\log n)$ for $c = O(\log n)$ by embedding EMD into ℓ_1 [IT03, Cha02], and $s = n^{O(1/c)} \mathrm{polylog}\, n$ for any $c \geq 1$ [ABIW09].

Question 15: Sparse Recovery for Tree Models (Piotr Indyk)

For any $n = 2^h - 1$, we can identify the coordinates of a vector $v \in \mathbb{R}^n$ with the nodes of a full binary tree B_h of height h and root 1. We define a *k-sparse tree model* $\mathcal{T}_k$ to be a set of all $T \subset [n]$ of size k which form a connected subtree in B_h rooted at 1.

We want to design an $m \times n$ matrix A such that for any $x \in \mathbb{R}^n$, one can recover from Ax a vector $x^* \in \mathbb{R}^n$ such that:

$$\|x^* - x\|_1 \leq \min_{x' \in \mathbb{R}^n, \mathrm{supp}(x') \subset T \text{ for some } T \in \mathcal{T}_k} C \left\|x' - x\right\|_1,$$

where $\mathrm{supp}(x')$ is the set of non-zero coefficients of x', and $C > 0$ is a constant.

Question: Is it possible to achieve $m = O(k)$ for some constant $C > 0$?

Background: It is known that a weaker bound of $m = O(k \log(n/k))$ is possible to achieve even if $\mathcal{T}_k$ is replaced by the set of all k-subsets of $[n]$ [CRT06]. However, since $|\mathcal{T}_k| = \exp(O(k))$, one can expect a better bound for $\mathcal{T}_k$. By using *model-based compressive sensing* framework of [BCDH10] (cf. [IP11]), one can achieve the desired bound of $m = O(k)$ but with *superconstant* C.

THE KANPUR LIST

Question 16: Random Walks (Rina Panigrahy)

The paper of Das Sarma, Gollapudi, and Panigrahy [DGP08] shows a method for performing random walks in the streaming model. In particular, a random walk of length l can be simulated using $O(n)$ space and $O(\sqrt{l})$ passes over the input stream. Is it possible to simulate such a random walk using $\tilde{O}(n)$ space and a much smaller number of passes, say, at most polylogarithmic in n and l? The goal is either to show an algorithm or prove a lower bound.

Das Sarma et al. [DGP08] simulate random walks to approximate the probability distribution on the vertices of the graph after a random walk of length l. What is the streaming complexity of approximating this distribution? What is the streaming complexity of finding the k (approximately) most likely vertices after a walk of length l?

QUESTION 17: APPROXIMATE 2D WIDTH (PANKAJ AGARWAL AND PIOTR INDYK)

The width of a set P of points in the plane is defined as

$$\mathrm{width}(P) = \min_{\|a\|_2=1} \left(\max_{p \in P} a \cdot p - \min_{p \in P} a \cdot p \right).$$

For a stream of insertions and deletions of points from a $[\Delta] \times [\Delta]$ grid, we would like to maintain an approximate width of the point set. We conjecture that there is an algorithm for this problem that achieves a constant approximation factor and uses $\mathrm{polylog}(\Delta + n)$ space.

Progress: The conjecture has been resolved (in the positive direction) by Andoni and Nguyen, 2011 (personal communication).

QUESTION 18: "ULTIMATE" DETERMINISTIC SPARSE RECOVERY (PIOTR INDYK)

We say that a vector $v \in \mathbb{R}^n$ is k*-sparse* for some $k \in \{0, \ldots, n\}$ if there are no more than k non-zero coordinates in v. The goal in the problem being considered is to design an $m \times n$ matrix A such that for any $x \in R^n$, one can recover from Ax a vector $x^* \in \mathbb{R}^n$ that satisfies the following "L_2/L_1" approximation guarantee:

$$\|x^* - x\|_2 \leq \min_{k\text{-sparse } x' \in \mathbb{R}^n} \frac{C}{\sqrt{k}} \left\|x' - x\right\|_1,$$

where $C > 0$ is a constant.

We conjecture that there is a solution that (a) uses $m = O(k \log(n/k))$ *and* (b) supports recovery algorithms running in time $O(n \,\mathrm{polylog}\, n)$.

Background: It is known that one can achieve *either* (a) *or* (b) (see, e.g., [NT10]). It is also possible to achieve both (a) and (b), but with a different "L_1/L_1" approximation guarantee, where $\|x^* - x\|_1 \leq \min_{k\text{-sparse } x'} C\|x' - x\|_1$ [IR08, BIR08].

QUESTION 19: COMMUNICATION COMPLEXITY AND METRIC SPACES (T. S. JAYRAM)

POINCARÉ INEQUALITIES. Alice and Bob are given two points x and y, respectively, from a specific metric space $\mathcal{M}$. We are interested in deciding whether $d_{\mathcal{M}}(x, y) \leq R$ or $d_{\mathcal{M}}(x, y) \geq \alpha R$, where $d_{\mathcal{M}}$ is the distance function of $\mathcal{M}$, $R > 0$, and $\alpha > 1$. What amount of information must be exchanged in order to solve this problem? Answering this question is interesting in any standard communication model: unrestricted communication between the players, one-way communication, sketching, etc.

The above question can partially be answered if the metric satisfies a specific "gap" Poincaré inequality [AJP10]. It is known that another kind of Poincaré inequality is equivalent to non-embeddability into ℓ_2^2 [Mat02], but it is not known if non-embeddability into ℓ_2^2 implies lower bounds for communication complexity. Can one show a formal connection between the communication complexity for approximating the distance between two points and non-embeddability into ℓ_2^2?

PRODUCT METRICS. We are also interested in the following general class of metrics. Let each $\mathcal{M}_i = \langle S_i, d_i \rangle$, $1 \leq i \leq k$, be a metric space on a set S_i with a distance function d_i. A *product metric space* $\bigoplus_{i=1}^k \mathcal{M}_i$ is defined on the product $S_1 \times \ldots \times S_k$ with the distance function

$$d\left((x_1, \ldots, x_k), (y_1, \ldots, y_k)\right) = \mathit{op}\left(d_1(x_1, y_1), \ldots, d_k(x_k, y_k)\right),$$

where *op* is a symmetric operator. For instance, $\bigoplus_{i=1}^k \mathcal{M}_i$ is a proper metric space if *op* is the maximum operator or the p-th norm for any $p \in [1, \infty)$. The case when $\bigoplus_{i=1}^k \mathcal{M}_i$ is not necessarily a metric space also finds applications.

Applications of product metric spaces include a nearest neighbor data structure for Ulam distance [AIK09], and a near-linear time subpolynomial-approximation algorithm for edit distance [AO09].

The following questions arise in the context of product spaces:

(1) Can one design efficient communication protocols for computing the distance between a pair of points? Suppose that there is an efficient communication protocol for each $\mathcal{M}_i$. What is the communication complexity for computing the distance between two points in $\bigoplus_{i=1}^{k} \mathcal{M}_i$? Andoni, Jayram, and Pătraşcu [AJP10] prove lower bounds for some product metrics. Jayram and Woodruff [JW09] show streaming algorithms which yield communication protocols.

(2) Can one design efficient streaming algorithms and data structures for product metric spaces? In particular, can one efficiently compute the distance between a pair of points? Jayram and Woodruff [JW09] consider the related question of computing *cascaded norms*.

QUESTION 20: EQUIVALENCE OF TWO MAPREDUCE MODELS (PAUL BEAME)

The original MapReduce paper [DG04] gives two distributed models. First it only says that intermediate key/value pairs with the same key are combined and sent as batch jobs to workers. Then in Section 4.2, it additionally guarantees that the batch jobs received by a single worker are sorted according to the corresponding key values. There are algorithms that rely on this additional feature of MapReduce. Are these two models equivalent? For decision problems in the complexity world, we know strong time-space trade-offs for sorting, but no similar lower bounds are known for distinctness.

QUESTION 21: MODELING OF DISTRIBUTED COMPUTATION (PAUL BEAME)

MapReduce has recently inspired two distributed models of computation in the theory community. One is the MUD model of Feldman et al. [FMS$^+$10]. In this model they assume that every worker has at most a polylogarithmic amount of space available, which in total gives at most $\tilde{O}(n)$ space, where n is the input size (in the order of at least terabytes). The other model of computation, due to Karloff et al. [KSV10], assumes that each of $n^{1-\varepsilon}$ workers has at most $n^{1-\varepsilon}$ space, where ε is a fixed positive constant. This totals to $n^{2-2\varepsilon}$ space in the entire system. Can one design an interesting and practical model that only uses $n^{1+o(1)}$ space/resources?

QUESTION 22: RANDOMNESS OF PARTIALLY RANDOM STREAMS (SUDIPTO GUHA)

Many streaming algorithms are designed for worst-case inputs and the first step of analysis is conducted using truly random hash functions, which in the second step are replaced by hash functions that can be described using a small number of truly random bits. In practice, the input stream is often a result of some random process. Mitzenmacher and Vadhan [MV08] show that as long as it has sufficiently large entropy, hash functions from a restricted family are essentially as good as truly hash functions. On a related note, Gabizon and Hassidim [GH10] show that algorithms for random-order streams need essentially no additional entropy apart from what can be extracted from the input.

In these two cases, the input can be seen as a source of randomness for the algorithm. How can one quantify the randomness of the stream in a natural way? For instance, Mitzenmacher and Vadhan set a lower bound for the Renyi entropy of each element of the stream, conditioned on the previous elements of the stream. Are there measures that are likely to be useful in practice and that are possible to verify?

Once we fix a measure of randomness, how much randomness according to this measure is necessary to derandomize or simplify specific streaming algorithms?

QUESTION 23: STRONG LOWER BOUNDS FOR GRAPH PROBLEMS (KRZYSZTOF ONAK)

A large number of streaming papers consider graph problems. Typically, the input stream is an arbitrarily-ordered sequence of edges. For many problems, one can show that solving the problem, even approximately, requires $\Omega(n)$ bits of space. For instance, one can relatively easily prove that finding a constant-factor approximation to the maximum matching problem requires $\Omega(n)$ bits of space. Therefore, in many cases, the desired space complexity is of the form $\tilde{O}(n)$. Despite this relaxation, it is plausible that for some popular problems, there are barriers that cannot be overcome by (approximate) algorithms that use $n^{1+o(1)}$ space and a small number of passes.

For example, let $M(G)$ be the maximum matching size in the input graph G. McGregor [McG05] shows that there is an algorithm that finds a matching of size $(1-\varepsilon) \cdot M(G)$ in a number of passes that is a function of only ε. It is plausible that for any constant k, there is no k-pass $\tilde{O}(n)$-space algorithm that finds a matching of size greater than $(1-\varepsilon_k) \cdot M(G)$ times the optimum, where ε_k is a positive constant. In particular, to the best of my knowledge, no one-pass $\tilde{O}(n)$-space algorithm that finds a $(1-\varepsilon)$-approximation for any constant $\varepsilon \in (0, 1/2)$ is known. Can one prove lower bounds as suggested above? The question generalizes to other problems. For instance, the best known $\tilde{O}(n)$-space algorithms for simulating random walks require a large number of passes (see [DGP08] and Rina Panigrahy's question). Can one prove for these problems that a small number of passes requires $n^{1+\Omega(1)}$ space?

To the best of my knowledge, the only problem for which this kind of lower bound is known is approximating graph distances. Feigenbaum et al. [FKM$^+$08] show that obtaining a t-approximation for the distance between two nodes in a single pass requires $\Omega(n^{1+1/t})$ space.

QUESTION 24: UNIVERSAL SKETCHING (JELANI NELSON)

Rather than designing different sketching algorithms for every problem, it would be desirable to have algorithms that where *universal*, in some sense, for a variety of problems. Specifically, let $\mathcal{F}$ be a family of functions mapping frequency vector $[-M, M]^n$ to $\mathbb{R}$. We say could say a sketching algorithm A is (ϵ, δ) universal for $\mathcal{F}$ if for all $x \in [-M, M]^n$, A can recover a $(1+\epsilon)$ approximation each $f(x)$ for any $f \in \mathcal{F}$ with probability $1-\delta$.

An example would be when $\mathcal{F}$ is $\{F_p : 0 \leq p \leq 2\}$. A simple approach would be to discretize p and to utilize the fact that $\ell_p(x) \approx \ell_{p'}(x)$ if p and p' are sufficiently close. Better yet would be to interpolate through a small set of values, using ideas from Harvey, Nelson, and Onak [HNO08]. Consequently it should be possible to be universal for $\mathcal{F} = \{F_p : 0 \leq p \leq 2\}$ while using only slightly more space than that required to estimate a specific F_p. For what other families are there efficient universal algorithms? It seems that the Indyk-Woodruff [IW05] technique would be useful here, and that the work of Braverman and Ostrovsky [BO10] is also highly relevant.

QUESTION 25: GAP-HAMMING INFORMATION COST (AMIT CHAKRABARTI)

In the Gap-Hamming problem, two players Alice and Bob have vectors $x, y \in \{0,1\}^n$ respectively and wish to compute the function f

$$f(x,y) = \begin{cases} 0 & \text{if } \Delta(x,y) \leq n/2 - \sqrt{n} \\ 1 & \text{if } \Delta(x,y) \geq n/2 + \sqrt{n} \end{cases}$$

where $\Delta(x,y) = |\{i : x_i \neq y_i\}|$ is the Hamming distance between the vectors and we are promised that $|\Delta(x,y) - n/2| \geq \sqrt{n}$. The problem became interesting in the streaming community because a lower bound on the communication complexity of evaluating f yields a lower bound on the space required by a streaming algorithm to estimate the number of distinct elements or the entropy of a stream. After a series of papers, it is know that evaluating f requires $\Omega(n)$ communication [IW03, Woo04, BC09, BCR$^+$10, CR11] even if an unlimited number of rounds of communication are used.

An increasingly popular technique in communication complexity is to prove bounds by bounding the information cost [CSWY01, BYJKS04]. Here we consider random input (X, Y) and consider the mutual information between the input and the random transcript of the protocol $\Pi(X, Y)$:

$$I(XY; \Pi(X, Y)) = H(XY) - H(XY|\Pi(X, Y)) .$$

It would be interesting to prove a lower bound on the information cost for the Gap-Hamming problem for some natural input distribution.

QUESTION 26: THE VALUE OF A REVERSE PASS (ANDREW MCGREGOR)

Multi-pass stream algorithms have been designed for a range of problems including longest increasing subsequences [LNVZ06, GM08], graph matchings [McG05], and various geometric problems [CC07]. However, the existing literature almost exclusively considers the case when the multiple passes are in the same direction. One exception is recent work by Magniez et al. [MMN10] on the DYCK_2 problem: given a length n string in the alphabet "$(,), [,]$", determine whether it is well-parenthesized, i.e., it can be generated by the grammar $S \to (S) \mid [S] \mid SS \mid \epsilon$? For this problem it can be shown that with one forward and one reverse pass over the input, the problem can be solved with $O(\log^2 n)$ space. On the other hand, any algorithm using $O(1)$ forward passes and no reverse passes, requires $\Omega(\sqrt{n})$ space [CCKM10, JN10]. For what other natural problems is there such a large separation?

QUESTION 27: GROUP TESTING (ELY PORAT)

Given a set $S \subset [n]$ of size at most k, we want to identify S by the following 2-stage process.

(1) We choose a set of subsets $T_1, \ldots, T_m \subset [n]$. For each T_i we learn whether or not $T_i \cap S = \emptyset$.
(2) Based on the outcomes of the first m tests, we may choose $j_1, \ldots, j_{O(k)} \in [n]$. For each j_i we learn whether or not $j_i \in S$.

The goal is to minimize m, the number of tests performed in the first stage. Without any further restrictions it has been shown that $m = O(k \log n/k)$ suffices [BGV05]. However for various pattern matching applications we have the constraint that each T_i needs to be an arithmetic progression, e.g., $T_i = \{2, 8, 14, 20, \ldots\}$. In this case, $m = O(k \log^2 n)$ suffices. Is it possible to decrease this to $m = O(k \log n)$?

QUESTION 28: LINEAR ALGEBRA COMPUTATION (MICHAEL MAHONEY)

It is often not the case that the entire data sits on a single machine and that we are allowed to make one or more passes over it. Instead the data is often distributed across multiple systems. This is one of the reasons why the streaming model does not have more impact in practice for linear algebra computation. It would be great to design new models that address this shortcoming.

Consider also the following problem. Let A be an $m \times n$ matrix and let k be a rank parameter. Let $P_{A,k}$ be the projection matrix on the best rank-k left (or right) singular subspace. The goal is to compute the diagonal of $P_{A,k}$ exactly or approximately in a small number of passes in the streaming model, or even better, in a new model that addresses the aforementioned shortcoming.

Question 29: Maximal Complex Equiangular Tight Frames (Joel Tropp)

Consider a system of unit vectors $\{x_k : k = 1, 2, \ldots, N\}$ in $\mathbb{C}^d$. It can be shown that the maximum inner product among these vectors satisfies the Welch bound

$$\max_{i \neq j} |\langle x_i, x_j \rangle| \geq \sqrt{\frac{N - d}{d(N-1)}}.$$

Miraculously, when this bound is attained, the modulus of the inner product between every pair of vectors is identical. Such a configuration is referred to as an equiangular tight frame (ETF).

It can be shown that the cardinality N of an ETF is $\mathbb{C}^d$ must satisfy the bound $N \leq d^2$. When this bound is attained, the ETF is referred to as a maximal ETF. In other words, a maximal ETF is a system of d^2 unit vectors in $\mathbb{C}^d$ whose pairwise inner products share the modulus $(d+1)^{-1/2}$.

A striking geometric fact about maximal ETFs is that each one corresponds with a regular simplex consisting of d^2 points embedded in the set of rank-one, trace-one, complex, Hermitian matrices with dimension d. This correspondence is achieved by mapping each vector x in the ETF to the matrix xx^*. Researchers believe that there is a maximal ETF for every dimension d. This question, so far, has resisted all efforts at solution.

References

[ABIW09] Alexandr Andoni, Khanh Do Ba, Piotr Indyk, and David P. Woodruff. Efficient sketches for earth-mover distance, with applications. In *IEEE Symposium on Foundations of Computer Science*, pages 324–330, 2009.

[AG09] Kook Jin Ahn and Sudipto Guha. Graph sparsification in the semi-streaming model. In *International Colloquium on Automata, Languages and Programming*, pages 328–338, 2009.

[AIK08] Alexandr Andoni, Piotr Indyk, and Robert Krauthgamer. Earth mover distance over high-dimensional spaces. In *SODA*, pages 343–352, 2008.

[AIK09] Alexandr Andoni, Piotr Indyk, and Robert Krauthgamer. Overcoming the ℓ_1 non-embeddability barrier: algorithms for product metrics. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 865–874, 2009.

[AJP10] Alexandr Andoni, T. S. Jayram, and Mihai Patrascu. Lower bounds for edit distance and product metrics via poincaré-type inequalities. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 184–192, 2010.

[AL11] Nir Ailon and Edo Liberty. An almost optimal unrestricted fast johnson-lindenstrauss transform. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 185–191, 2011.

[AO09] Alexandr Andoni and Krzysztof Onak. Approximating edit distance in near-linear time. In *ACM Symposium on Theory of Computing*, pages 199–204, 2009.

[BC09] Joshua Brody and Amit Chakrabarti. A multi-round communication lower bound for gap hamming and some consequences. In *IEEE Conference on Computational Complexity*, pages 358–368, 2009.

[BCDH10] Richard G. Baraniuk, Volkan Cevher, Marco F. Duarte, and Chinmay Hegde. Model-based compressive sensing. *IEEE Transactions on Information Theory*, 56(4):1982–2001, 2010.

[BCR$^+$10] Joshua Brody, Amit Chakrabarti, Oded Regev, Thomas Vidick, and Ronald de Wolf. Better gap-hamming lower bounds via better round elimination. In *APPROX-RANDOM*, pages 476–489, 2010.

[BGM$^+$11] Eli Ben-Sasson, Elena Grigorescu, Ghid Maatouk, Amir Shpilka, and Madhu Sudan. On sums of locally testable affine invariant properties. *Electronic Colloquium on Computational Complexity (ECCC)*, 18:79, 2011.

[BGV05] Annalisa De Bonis, Leszek Gasieniec, and Ugo Vaccaro. Optimal two-stage algorithms for group testing problems. *SIAM J. Comput.*, 34(5):1253–1270, 2005.

[BIR08] Radu Berinde, Piotr Indyk, and Milan Ruzic. Practical near-optimal sparse recovery in the l_1 norm. *Allerton*, 2008.

[BO10] Vladimir Braverman and Rafail Ostrovsky. Zero-one frequency laws. In *ACM Symposium on Theory of Computing*, pages 281–290, 2010.

[BR02] Michael A. Bender and Dana Ron. Testing properties of directed graphs: acyclicity and connectivity. *Random Struct. Algorithms*, 20(2):184–205, 2002.

[BSS08] Itai Benjamini, Oded Schramm, and Asaf Shapira. Every minor-closed property of sparse graphs is testable. In *ACM Symposium on Theory of Computing*, pages 393–402, 2008.

[BYJKS04] Ziv Bar-Yossef, T. S. Jayram, Ravi Kumar, and D. Sivakumar. An information statistics approach to data stream and communication complexity. *J. Comput. Syst. Sci.*, 68(4):702–732, 2004.

[CC07] Timothy M. Chan and Eric Y. Chen. Multi-pass geometric algorithms. *Discrete & Computational Geometry*, 37(1):79–102, 2007.

[CCKM10] Amit Chakrabarti, Graham Cormode, Ranganath Kondapally, and Andrew McGregor. Information cost tradeoffs for augmented index and streaming language recognition. In *IEEE Symposium on Foundations of Computer Science*, pages 387–396, 2010.

[CCM09] Amit Chakrabarti, Graham Cormode, and Andrew McGregor. Annotations in data streams. In *International Colloquium on Automata, Languages and Programming*, pages 222–234, 2009.

[Cha02] Moses Charikar. Similarity estimation techniques from rounding algorithms. In *ACM Symposium on Theory of Computing*, pages 380–388, 2002.

[Cha10] Amit Chakrabarti. A note on randomized streaming space bounds for the longest increasing subsequence problem. *Electronic Colloquium on Computational Complexity (ECCC)*, 10(10), 2010.

[CR11] Amit Chakrabarti and Oded Regev. An optimal lower bound on the communication complexity of gap-hamming-distance. In *ACM Symposium on Theory of Computing*, pages 51–60, 2011.

[CRT06] Emmanuel J. Candès, Justin Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.*, 59(8):1208–1223, 2006.

[CSWY01] Amit Chakrabarti, Yaoyun Shi, Anthony Wirth, and Andrew C. Yao. Informational complexity and the direct sum problem for simultaneous message complexity. In *IEEE Symposium on Foundations of Computer Science*, pages 270–278, 2001.

[DDS11] Constantinos Daskalakis, Ilias Diakonikolas, and Rocco A. Servedio. Learning transformed product distributions. *CoRR*, abs/1103.0598, 2011.

[DG04] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: Simplified data processing on large clusters. In *OSDI*, pages 137–150, 2004.

[DGL$^+$99] Yevgeniy Dodis, Oded Goldreich, Eric Lehman, Sofya Raskhodnikova, Dana Ron, and Alex Samorodnitsky. Improved testing algorithms for monotonicity. In *RANDOM-APPROX*, pages 97–108, 1999.

[DGP08] Atish Das Sarma, Sreenivas Gollapudi, and Rina Panigrahy. Estimating PageRank on graph streams. In *PODS*, pages 69–78, 2008.

[EJ08] Funda Ergün and Hossein Jowhari. On distance to monotonicity and longest increasing subsequence of a data stream. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 730–736, 2008.

[FKM$^+$08] Joan Feigenbaum, Sampath Kannan, Andrew McGregor, Siddharth Suri, and Jian Zhang. Graph distances in the data-stream model. *SIAM J. Comput.*, 38(5):1709–1727, 2008.

[FMS$^+$10] Jon Feldman, S. Muthukrishnan, Anastasios Sidiropoulos, Clifford Stein, and Zoya Svitkina. On distributing symmetric streaming computations. *ACM Transactions on Algorithms*, 6(4), 2010.

[GG07] Anna Gál and Parikshit Gopalan. Lower bounds on streaming algorithms for approximating the length of the longest increasing subsequence. In *IEEE Symposium on Foundations of Computer Science*, pages 294–304, 2007.

[GH10] Ariel Gabizon and Avinatan Hassidim. Derandomizing algorithms on product distributions and other applications of order-based extraction. In *ICS*, pages 397–405, 2010.

[GJKK07] Parikshit Gopalan, T. S. Jayram, Robert Krauthgamer, and Ravi Kumar. Estimating the sortedness of a data stream. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 318–327, 2007.

[GKS08] Elena Grigorescu, Tali Kaufman, and Madhu Sudan. 2-transitivity is insufficient for local testability. In *IEEE Conference on Computational Complexity*, pages 259–267, 2008.

[GM08] Sudipto Guha and Andrew McGregor. Tight lower bounds for multi-pass stream computation via pass elimination. In *International Colloquium on Automata, Languages and Programming*, pages 760–772, 2008.

[HKNO09] Avinatan Hassidim, Jonathan A. Kelner, Huy N. Nguyen, and Krzysztof Onak. Local graph partitions for approximation and testing. In *IEEE Symposium on Foundations of Computer Science*, pages 22–31, 2009.

[HNO08] Nicholas J. A. Harvey, Jelani Nelson, and Krzysztof Onak. Sketching and streaming entropy via approximation theory. In *IEEE Symposium on Foundations of Computer Science*, pages 489–498, 2008.

[IP11] Piotr Indyk and Eric Price. K-median clustering, model-based compressive sensing, and sparse recovery for earth mover distance. In *ACM Symposium on Theory of Computing*, pages 627–636, 2011.

[IR08] Piotr Indyk and Milan Ruzic. Near-optimal sparse recovery in the ℓ_1 norm. In *IEEE Symposium on Foundations of Computer Science*, pages 199–207, 2008.

[IT03] Piotr Indyk and Nitin Thaper. Fast color image retrieval via embeddings. *Workshop on Statistical and Computational Theories of Vision (at ICCV)*, 2003.

[IW03] Piotr Indyk and David P. Woodruff. Tight lower bounds for the distinct elements problem. *IEEE Symposium on Foundations of Computer Science*, pages 283–288, 2003.

[IW05] Piotr Indyk and David P. Woodruff. Optimal approximations of the frequency moments of data streams. In *ACM Symposium on Theory of Computing*, pages 202–208, 2005.

[JN10] Rahul Jain and Ashwin Nayak. The space complexity of recognizing well-parenthesized expressions. *Electronic Colloquium on Computational Complexity (ECCC)*, 17:71, 2010.

[JR11] Madhav Jha and Sofya Raskhodnikova. Testing and reconstruction of Lipschitz functions with applications to data privacy. *Electronic Colloquium on Computational Complexity (ECCC)*, 18:57, 2011.

[JW09] T. S. Jayram and David P. Woodruff. The data stream space complexity of cascaded norms. In *IEEE Symposium on Foundations of Computer Science*, 2009.

[KN06] Subhash Khot and Assaf Naor. Nonembeddability theorems via Fourier analysis. *Math. Ann.*, 334(4):821–852, 2006.

[KOR00] Eyal Kushilevitz, Rafail Ostrovsky, and Yuval Rabani. Efficient search for approximate nearest neighbor in high dimensional spaces. *SIAM J. Comput.*, 30(2):457–474, 2000.

[KSV10] Howard J. Karloff, Siddharth Suri, and Sergei Vassilvitskii. A model of computation for mapreduce. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 938–948, 2010.

[LNVZ06] David Liben-Nowell, Erik Vee, and An Zhu. Finding longest increasing and common subsequences in streaming data. *J. Comb. Optim.*, 11(2):155–175, 2006.

[Mat02] Jiří Matoušek. *Lectures on Discrete Geometry*. Springer, 2002.

[McG05] Andrew McGregor. Finding graph matchings in data streams. In *APPROX-RANDOM*, pages 170–181, 2005.

[MMN10] Frédéric Magniez, Claire Mathieu, and Ashwin Nayak. Recognizing well-parenthesized expressions in the streaming model. In *ACM Symposium on Theory of Computing*, pages 261–270, 2010.

[MV08] Michael Mitzenmacher and Salil P. Vadhan. Why simple hash functions work: exploiting the entropy in a data stream. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 746–755, 2008.

[NO08] Huy N. Nguyen and Krzysztof Onak. Constant-time approximation algorithms via local improvements. In *IEEE Symposium on Foundations of Computer Science*, pages 327–336, 2008.

[NT10] Deanna Needell and Joel A. Tropp. Cosamp: iterative signal recovery from incomplete and inaccurate samples. *Commun. ACM*, 53(12):93–100, 2010.

[Ona10] Krzysztof Onak. *New Sublinear Methods in the Struggle Against Classical Problems*. PhD thesis, Massachusetts Institute of Technology, 2010.

[SV11] C. Seshadhri and Jan Vondrák. Is submodularity testable? In *ICS*, 2011.

[Tre09] Luca Trevisan. Max cut and the smallest eigenvalue. In *ACM Symposium on Theory of Computing*, pages 263–272, 2009.

[Woo04] David P. Woodruff. Optimal space lower bounds for all frequency moments. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 167–175, 2004.

[YYI09] Yuichi Yoshida, Masaki Yamamoto, and Hiro Ito. An improved constant-time approximation algorithm for maximum matchings. In *ACM Symposium on Theory of Computing*, pages 225–234, 2009.