Sketching as a tool for Algorithmic Design

Alex Andoni
(Columbia University)

Find similar pairs

Methodology ?

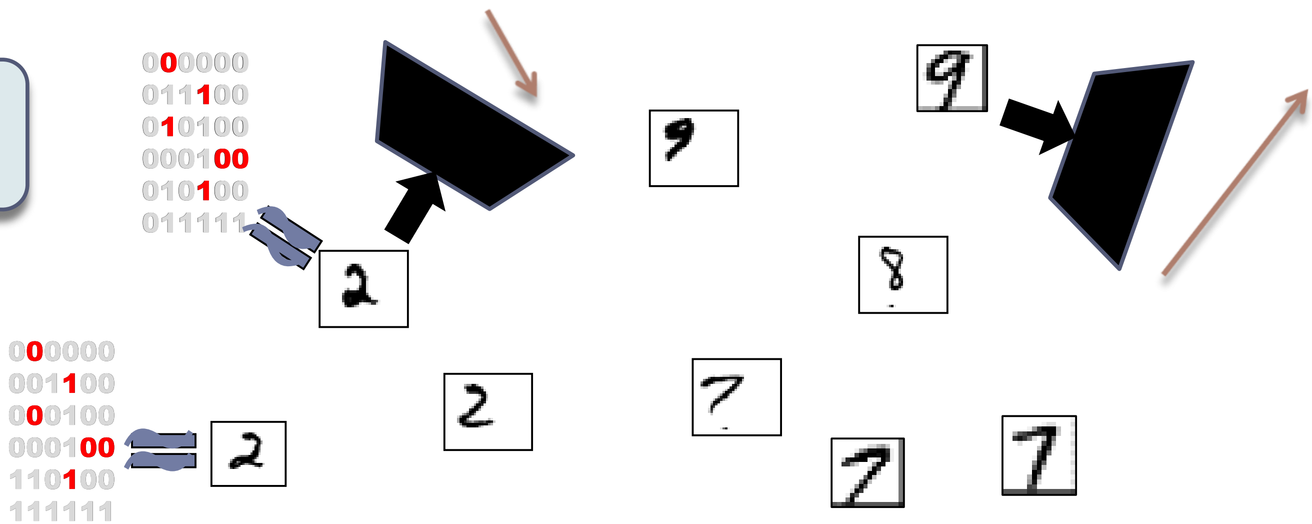
Small space algorithms

Fast algorithms

Sketching

dimension reduction

- compression
- good for specific task
- lossy



Dimension reduction: linear map $S: \mathbb{R}^n \to \mathbb{R}^k$ s.t:

- for any points $p, q \in \mathbb{R}^n$:

$$\Pr_S \left[\frac{||S(p)-S(q)||}{||p-q||} \in (1 \pm \epsilon) \right] \geq 1-\delta$$

[Johnson-Lindenstrauss'84]:

$S =$ Gaussian matrix

$$k = O\left(\frac{1}{\epsilon^2} \log \frac{1}{\delta}\right)$$

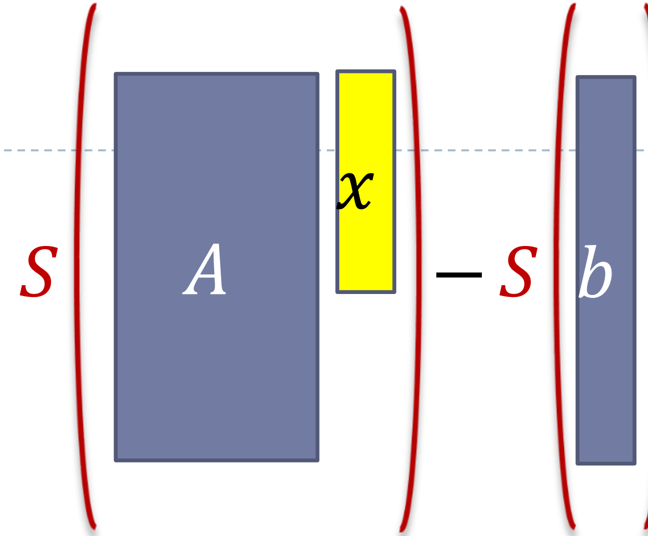
~~Plan~~ *a sketch of sketching applications…*

- Numerical Linear Algebra
- Nearest Neighbor Search
- Min-cost matching in plane

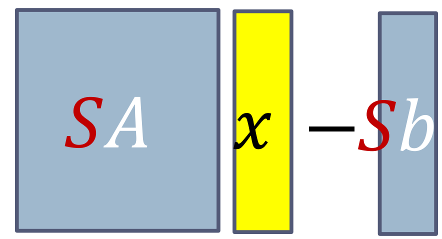
Plan

- Numerical Linear Algebra
  - the power of linear sketches
- Nearest Neighbor Search
- Min-cost matching in plane

Numerical Linear Algebra



- **Problem:** Least Square Regression
  - $x^* = argmin_x||Ax - b||$
  - where A is $n \times d$ matrix
  - $n \gg d$
  - $1 + \epsilon$ approximation



- **Idea:** Sketch-And-Solve
  - solve $x' = argmin_x||S \cdot (Ax - b)|| = argmin_x||SAx - Sb||$
    - where $S: \mathbb{R}^n \to \mathbb{R}^k$ is a dimension-reducing matrix
  - reduces to much smaller $k \times d$ problem
  - Hope: $||Ax' - b|| \leq (1 + \epsilon)||Ax^* - b||$

Sketch-And-Solve

[S'06, CW'13, NN'13, MM'13, C'16]

Oblivious Subspace Embedding: linear map $S: \mathbb{R}^n \rightarrow \mathbb{R}^k$ s.t.

- for any linear subspace $P \subset \mathbb{R}^n$ of dimension d:

$$\Pr_S\left[\forall p \in P : \frac{||S(p)||}{||p||} \in (1 \pm \epsilon)\right] \geq 1 - \delta$$

$k \sim d$

- Issue: time to compute sketch
  - When S=Gaussian ([JL]) $\Rightarrow$ computing SA takes $O(n \cdot d^2)$ time
  - Idea: structured S s.t. SA can be computed faster
- +structured S: $O\left(nnz(A) + \left(\frac{d}{\epsilon}\right)^{O(1)}\right)$ time
- +Preconditioner: $O\left(\left(nnz(A) + d^{O(1)}\right) \cdot \log\frac{1}{\epsilon}\right)$

slower than the original problem !

ℓ_1 regression

- No similar dimension reduction in ℓ_1 [BC'04,JN'09]

[I'00]

Weak DR: linear map $S: \mathbb{R}^n \to \mathbb{R}^k$, s.t.

- for any $p \in \mathbb{R}^n$: $\Pr_S\left[1 \le \frac{||S(p)||_1}{||p||_1} \le \frac{1}{\delta}\right] \ge 1 - O(\delta)$

$S_{ij} \sim$ Cauchy distribution, or 1/Exponential

[SW'11, MM'13, WZ'13, WW'18]

Weak(er) OSE: linear map $S: \mathbb{R}^n \to \mathbb{R}^k$ s.t.

- for any linear subspace $P \subset \mathbb{R}^n$ of dimension d:

$\Pr_S\left[\forall p \in P : 1 \le \frac{||S(p)||_1}{||p||_1} \le d^{O(1)}\right] \ge 0.9$

$k = O(d \cdot \log d)$

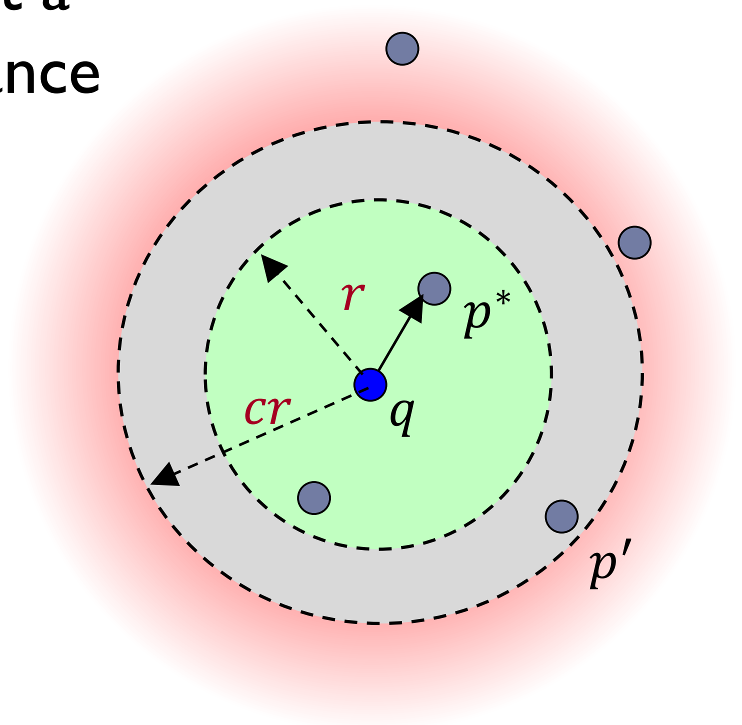
- +structured S, +preconditioner: $O\left(nnz(A) \cdot \log n + \left(\frac{d}{\epsilon}\right)^{O(1)}\right)$
- More: other norms (ℓ_p, M-estimator, Orlicz norms), low-rank approximation & optimization, matrix multiplication, see [Woodruff, FnTTCS'14,…]

Plan

- Numerical Linear Algebra
- Nearest Neighbor Search
  - ultra-small sketches
- Min-cost matching in plane

Approximate Near Neighbor Search

- Preprocess: a set of N point
  - approximation $c > 1$
- Query: given a query point q, report a point $p^* \in P$ with the smallest distance to q
  - up to factor c
- *Near* neighbor: threshold r
- Parameters: space & query time



Ultra-small sketches

Distance Estimation Sketch: for approx c, & all thresholds r
map $S: \mathbb{R}^d \to \{0,1\}^k$, estimator $\boldsymbol{E}(\cdot,\cdot)$, s.t. for any $p, q \in \mathbb{R}^d$:

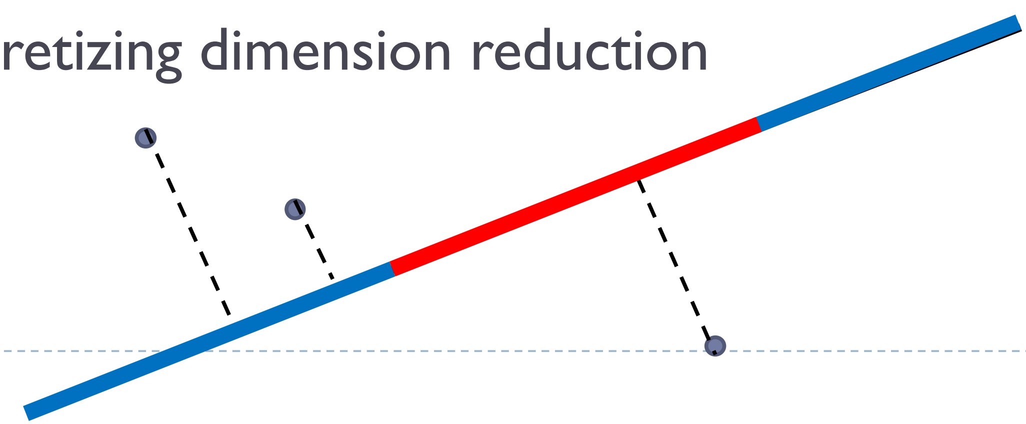
- $||p-q|| \le r$, then $\Pr_S\left[\boldsymbol{E}\big(S(p), S(q)\big) = \text{"}close\text{"}\right] \ge 1-\delta$
- $||p-q|| > cr$, then $\Pr_S\left[\boldsymbol{E}\big(S(p), S(q)\big) = \text{"}close\text{"}\right] \le \delta$

(c, δ, k)-DE sketch

const # of bits!

- [KOR'98,IM'98]: ℓ_2, ℓ_1 have $\left(1+\epsilon, 0.1, O\left(\frac{1}{\epsilon^2}\right)\right)$-DE sketches
  - Via: bit sampling (Hamming),
  - or discretizing dimension reduction

000000
011100
010100
000100
010100
011111



DE Sketch => NNS

[KOR'98,IM'98]: $(c, 1/3, k)$-DES imply c-approx NNS with space $N^{O(k)}$ and 1 memory probe per query

Proof: construct a sketch with failure probability $1/N$

- by concatenating $O(\log N)$ i.i.d. copies of the sketch, and taking majority vote
- Data structure: a look-up table for all possible sketches of a query: $2^{O(k \cdot \log N)} = N^{O(k)}$ possibilities only

Const size DES => NNS with polynomial space!

- Query time: computing the sketch, typically $\sim O(kd \log N)$ [see also AC'06]

[AK+ANNRW'18]: $(c, 0.1, k)$-DES implies NNS with $O(ck)$-approximation and $O(N^{1.1})$ space, $O(N^{0.1})$ memory probes per query

Beyond ℓ_1 and ℓ_2

α-embedding of metric X into ℓ_1: for distortion D, power $\alpha \geq 1$:
map $f: X \to \ell_1$, s.t. for any $p, q \in X$:

- $||f(p) - f(q)||^{\alpha} \leq dist_X(p,q) \leq D \cdot ||f(p) - f(q)||^{\alpha}$

Embedding with $D = c$ ⇒ $\big(O(c), 0.1, O(1)\big)$-DES ⇒ NNS

[AKR'15]: when X is a norm:

Embedding with $D = O(ck)$ ⇐ $(O(c), 0.1, k)$-DES

OPEN: if $\alpha = 1$ achievable

Not true for general X [KN]

NNS with smaller space?

- Space closer to linear in N ?

LSH Sketch: for approx c, & $\forall$ thresholds r
map $S: \mathbb{R}^d \to \{0,1\}^k$, estimator $\boldsymbol{E}(\cdot,\cdot)$, s.t. for any $p, q \in \mathbb{R}^d$:

- $||p-q|| \leq r$, then $\Pr_S[\boldsymbol{E}(S(p), S(q)) = \text{"close"}] \geq 2^{-\rho k}$
- $||p-q|| > cr$, then $\Pr_S[\boldsymbol{E}(S(p), S(q)) = \text{"close"}] \leq 2^{-k+1}$
- $E(\sigma, \tau) = \text{"close"}$ iff $\sigma = \tau$

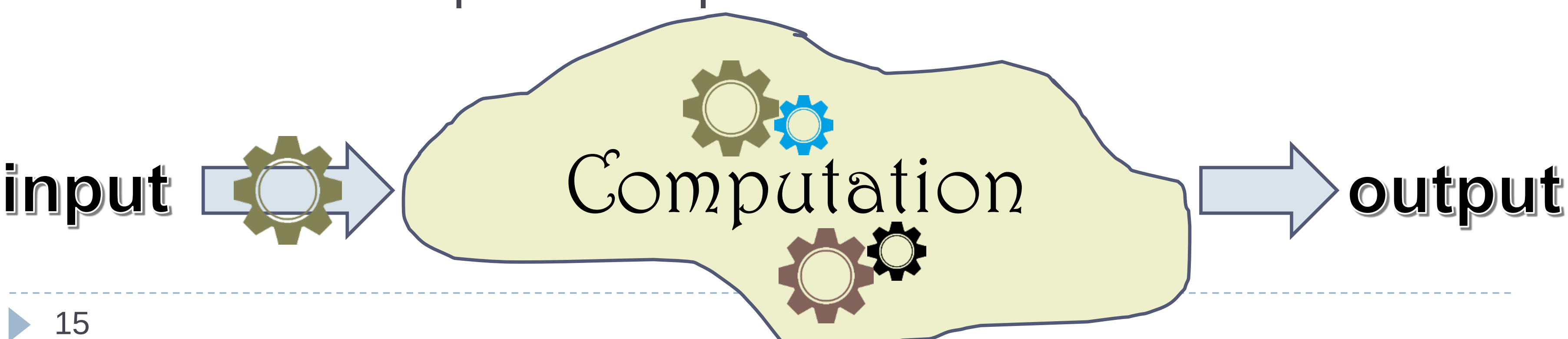
(c, ρ, k)-LSH

[IM'98]: (c, ρ, k)-LSH imply c-approx NNS with $O(N^{1+\rho})$ space and $O(N^{\rho})$ memory probes per query

[IM'98]: $\rho = 1/c$ for ℓ_1

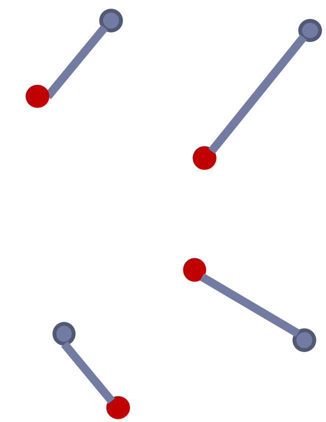
Plan

- Numerical Linear Algebra
- Nearest Neighbor Search
- Min-cost matching in plane
  - specialized sketches
- Exploit sketches for:
  - input
  - internal state / partial computations



LP for Geometric Matching

- **Problem:**
  - Given two sets A, B of points in $\mathbb{R}^2$,
  - Find min-cost matching ($1+\epsilon$ approx.)
  - a.k.a., Earth-Mover Distance, optimal transport, Wasserstein metric, etc



- Classically: LP with n^2 variables
  - General: $\tilde{O}(n^2/\epsilon^4)$ time [AWR'17]
  - In 2D: hope for $\approx n$ time [SA'12]

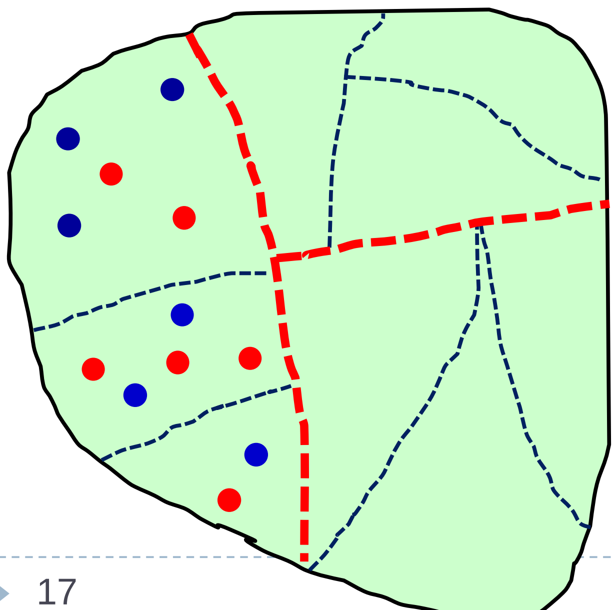
$$\min_{\pi \in \mathbb{R}_+^{n^2}} \sum_{ij} ||p_i - q_j|| \cdot \pi_{ij}$$

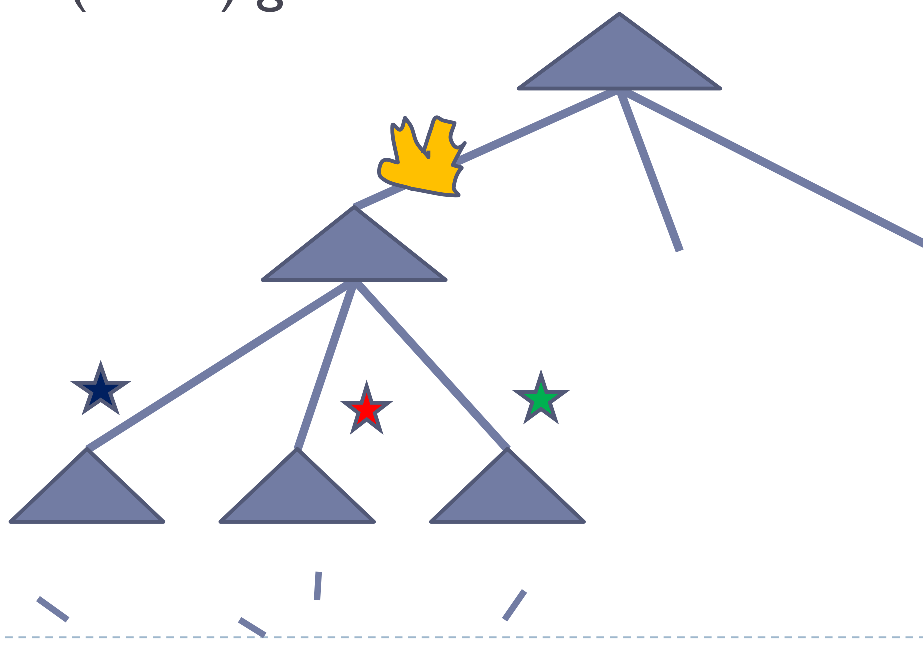
s.t. $\pi \mathbf{1} = \frac{1}{n}\mathbf{1}$ and $\pi^t \mathbf{1} = \frac{1}{n}\mathbf{1}$

[ANOY'14]: **Solve-And-Sketch** framework
Solves in $n^{1+o(1)}$ time (for fixed ϵ)

Solve-And-Sketch (=Divide & Conquer)

- Partition the space hierarchically in a "nice way"
- In each part
  - Compute a "solution" for the local view
  - Sketch the solution using small space
  - Combine local sketches into (more) global solution





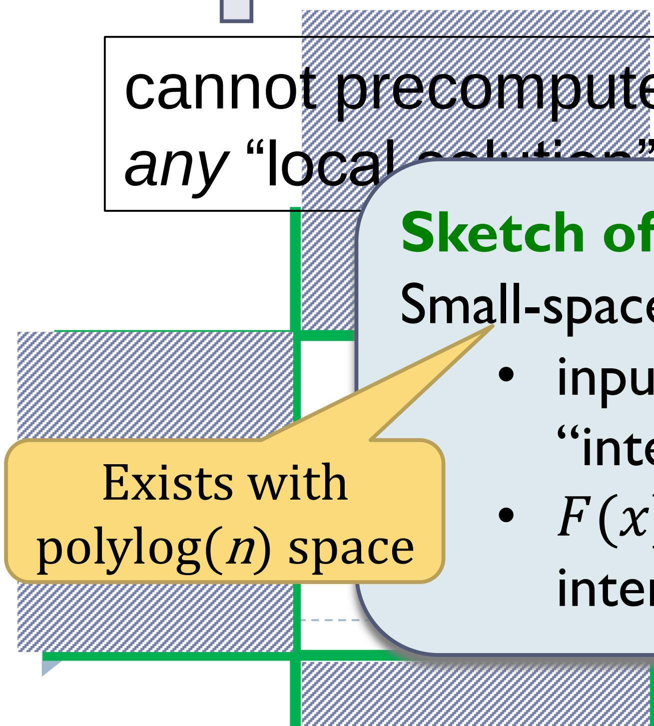
Solve-And-Sketch for 2D Matching

- Partition the space hierarchically in a "nice way"
- In each part
  - Compute a ~~"solution"~~ for the local view
  - Sketch ~~the solution using~~ small space
  - Combine local sketches into (more) global solution

all potential local solutions

quad-tree

cannot precompute *any* "local solution"



Sketch of all potential local solutions:

Small-space sketch of the "solution" function $F: \mathbb{R}^k \rightarrow \mathbb{R}_+$

- input $x \in \mathbb{R}^k$ defines the flow (matching) at the "interface" to the rest
- $F(x)$ is the min-cost matching assuming flow x at interface

Exists with polylog(n) space

Sketching → Fast algorithms

- Numerical Linear Algebra
  - linear sketching
- Nearest Neighbor Search
  - ultra-small sketches
- Min-cost matching in plane
  - specialized sketching
- Graph sketching
  - Linear sketch for graph => data structures for dynamic connectivity [AGM'12, KKM'13]
- Characterization of DE-sketch size for metrics:
  - For symmetric norms [BBCKY'17]
- Adaptive sketching: when we know we sketch set $A \subset \mathbb{R}^d$
  - Then $S(\cdot)$ may depend (weakly) on A
  - Non-oblivious subspace embeddings [DMM'06,…,Woodruff'14]
  - Data-dependent LSH [AINR'14, AR'15]

Bibliography 1

- Sarlos'06
- Clarkson-Woodruff'13,
- Nguyen-Nelson'13,
- Mahoney-Meng'13,
- Cohen'16
- Indyk'00
- Sohler-Woodruff'11
- Woodruff-Zhang'13
- Wang-Woodruff'18 (arxiv)

Bibliography 2

- Kushilevitz-Ostrovsky-Rabani'98
- Indyk-Motwani'98
- Ailon-Chazelle'06
- Khot-Naor (unpublished)
- A-Krauthgamer (unpublished)
- A-Naor-Nikolov-Razenshteyn-Weingarten'18
- Altschuler-Weed-Rigolet'17
- Sharathkumar-Agarwal'12
- A.-Nikolov-Onak-Yaroslavtsev'14
- Ahn-Guha-McGregor'12
- Kapron-King-Mountjoy'13
- Blasiok-Braverman-Chestnut-Krauthgamer-Yang'17
- Drineas-Mahoney-Muthukrishnan'06
- A-Indyk-Nguyen-Razenshteyn'14
- A-Razenshteyn'15