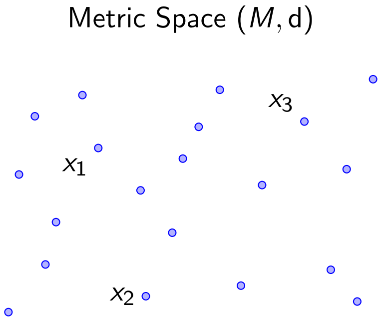
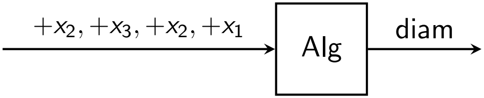
A Polynomial Space Lower Bound for Diameter Estimation in Dynamic Streams

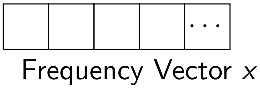
Joint work with Sanjeev Khanna, Ashwin Padaki, and Erik Waingarten

December 6, 2025

Streaming Diameter

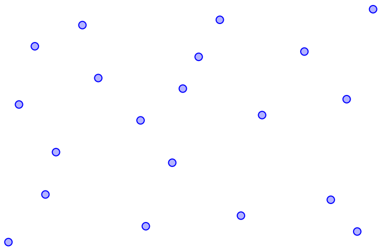


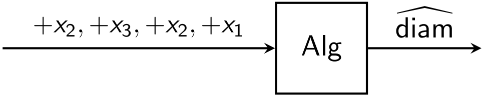




Streaming Diameter

Metric Space (M, d)

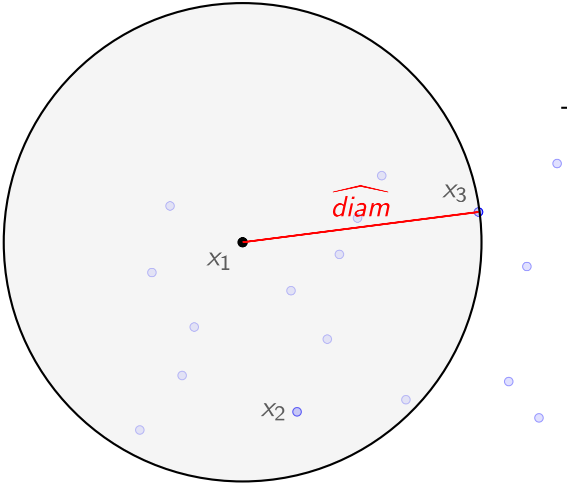


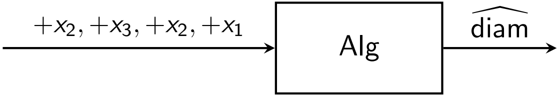


Frequency Vector x

Insertion-Only Streams

Metric Space (M, d)

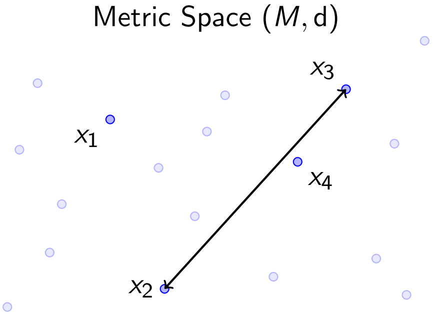


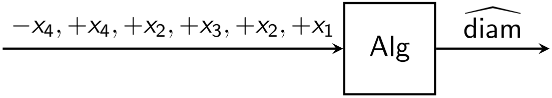


1	2	1	0	$\cdots$

Frequency Vector x

Dynamic Streams





1	2	1	0	$\cdots$

Frequency Vector x

Dynamic Streams

But what about dynamic streams?

Dynamic Streams

But what about dynamic streams?

- Insertion-only algorithm finds a surviving point and computes the radius around it
- But it's unclear how to identify such a surviving point a priori

Streaming Diameter in ℓ_∞

When can we say something non-trivial?

Streaming Diameter in ℓ_∞

When can we say something non-trivial?

There exists a deterministic and exact dynamic streaming algorithm for $([\Delta]^d, \ell_\infty)$ using space $O(d \cdot \Delta)$

1. For each coordinate, maintain a histogram for each value
2. Diameter is the maximum difference along a coordinate

Streaming Diameter in General Metrics

Theorem (Matousek)

Any metric space (M, d) *can be embedded into* ℓ_∞^k *with distortion* $2c - 1$ *and* $k = \tilde{O}(n^{1/c})$.

This immediately implies a dynamic streaming algorithm for c-approximate diameter in general metric spaces using $O(\Delta \cdot n^{1/c})$ space.

Theorem

There exists a dynamic streaming algorithm for c*-approximate diameter in general metric spaces using* $n^{O(1/c)}$ *space*

Streaming Diameter in General Metrics

What about lower bounds?

Streaming Diameter in General Metrics

What about lower bounds?

- Euclidean MST [CJLW21, CCAJ$^+$23]: $n^{O(1/\sqrt{c})}$ space
- Euclidean Facility Location [CJK$^+$22]: $n^{O(1/c)}$ space
- EMD over $[\Delta]^2$: $\Delta^{O(1/c)}$
- Euclidean Diameter [Ind03]: $n^{O(1/c^2)}$

Streaming Diameter in General Metrics

What about lower bounds?

- Euclidean MST [CJLW21, CCAJ+23]: $n^{O(1/\sqrt{c})}$ space
- Euclidean Facility Location [CJK+22]: $n^{O(1/c)}$ space
- EMD over $[\Delta]^2$: $\Delta^{O(1/c)}$
- Euclidean Diameter [Ind03]: $n^{O(1/c^2)}$

Motivating Question: Can we show a polynomial space lower bound for Euclidean diameter?

Streaming Diameter in General Metrics

What about lower bounds?

- Euclidean MST [CJLW21, CCAJ$^+$23]: $n^{O(1/\sqrt{c})}$ space
- Euclidean Facility Location [CJK$^+$22]: $n^{O(1/c)}$ space
- EMD over $[\Delta]^2$: $\Delta^{O(1/c)}$
- Euclidean Diameter [Ind03]: $n^{O(1/c^2)}$

Motivating Question: Can we show a polynomial space lower bound for Euclidean diameter?

Answer: Not yet :(But we can show the existence of a metric on which diameter is hard

Streaming Diameter in General Metrics

Theorem (Informal Main Result)

There exists a metric space (M, d) where any dynamic streaming algorithm c-approximating diameter requires $n^{\Omega(1/c)}$ space.

Towards a Lower Bound via Communication Complexity

Indexing Problem

Alice	Bob
$x \in \{0,1\}^n$	$i \in [n]$

Bob must output x_i

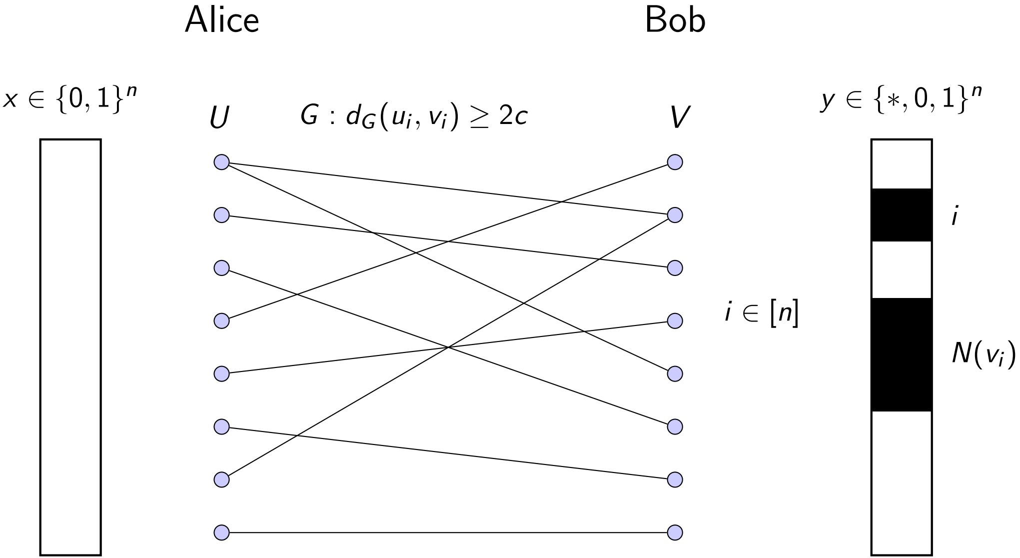
Towards a Lower Bound via Communication Complexity

Indexing with a Twist

Alice	Bob
$x \in \{0,1\}^n$	$i \in [n]$ $+$ knowledge about x

Bob must output x_i

Towards a Lower Bound via Communication Complexity



Towards a Lower Bound via Communication Complexity

Distinguishing whether $\mathrm{diam}(N(v_{\mathbf{i}}) \cup \{u_{\mathbf{i}}\}) \geq 2c$ or ≤ 2 allows us to distinguish whether $x_i = 1$ or 0.

Towards a Lower Bound via Communication Complexity

Distinguishing whether $\text{diam}(N(v_{\mathbf{i}}) \cup \{u_{\mathbf{i}}\}) \geq 2c$ or ≤ 2 allows us to distinguish whether $x_i = 1$ or 0.

Given a diameter estimation algorithm:

- Alice inserts all u_j for which $x_j = 1$, hands the memory contents of the algorithm over to Bob
- Bob deletes all v_j for which $y_j = 1$, returns the output of the algorithm

Towards a Lower Bound via Communication Complexity

- This communication game looks a lot like Indexing but with a *twist*
- Studied as *Index Coding with Side Information*
- Its communication complexity remains open, but progress has been made when restricting to linear protocols over finite fields

Towards a Lower Bound via Communication Complexity

- This communication game looks a lot like Indexing but with a *twist*
- Studied as *Index Coding with Side Information*
- Its communication complexity remains open, but progress has been made when restricting to linear protocols over finite fields

Theorem (Informal [BYBJK06])

For linear protocols over $\mathbb{F}_2$, $\Theta(\mathrm{MinRank}(G))$ *bits of communication are needed*

Towards a Lower Bound via Communication Complexity

- This communication game looks a lot like Indexing but with a *twist*
- Studied as *Index Coding with Side Information*
- Its communication complexity remains open, but progress has been made when restricting to linear protocols over finite fields

Theorem (Informal [BYBJK06])

For linear protocols over $\mathbb{F}_2$*,* $\Theta(\mathrm{MinRank}(G))$ *bits of communication are needed*

Theorem (Informal [ABG$^+$20])

For $\mathbf{G} \sim \mathcal{G}(n, n, p)$*,* $\mathrm{MinRank}(\mathbf{G}) \geq \tilde{\Omega}(np)$ *with high probability*

Towards a Lower Bound via Communication Complexity

- This communication game looks a lot like Indexing but with a *twist*
- Studied as *Index Coding with Side Information*
- Its communication complexity remains open, but progress has been made when restricting to linear protocols over finite fields

Theorem (Informal [BYBJK06])

For linear protocols over $\mathbb{F}_2$, $\Theta(\mathrm{MinRank}(G))$ *bits of communication are needed*

Theorem (Informal [ABG+20])

For $\mathbf{G} \sim \mathcal{G}(n, n, p)$, $\mathrm{MinRank}(\mathbf{G}) \geq \tilde{\Omega}(np)$ *with high probability*

Results generalize when domain is $\mathbb{F}$ and linear protocols are over $\mathbb{F}$, but will need a different approach to show *general* streaming lower bounds

Min Rank

Consider the following incomplete matrix induced by a graph $G = (U, V, E)$

1. 1's on the diagonal
2. $M_{ij} = 0$ if $(u_i, v_j) \in E$

$$M = \begin{bmatrix} 1 & 0 & & & 0 & & & \\ & 1 & 0 & & & & 0 & \\ & & 1 & & 0 & & & 0 \\ 0 & & & 1 & & & 0 & \\ & 0 & & & 1 & & & 0 \\ 0 & & & & & 1 & & 0 \\ & & 0 & & & & 1 & 0 \\ & 0 & & 0 & & & & 1 \end{bmatrix}$$

$$\text{Minrank(G)} := \min_{\text{Real completions } M' \text{ of } M} \text{rank(M')}$$

Streaming Algorithms are Equivalent to Linear Sketches [Gan08, LNW14]

For a distribution $\mathcal{D}$ over frequency vectors $x \in \{0, \ldots, m\}^n$ and a decision problem P.

Dynamic Streaming Algorithm $\mathcal{A}$ $\longrightarrow$

Linear Sketch

$\mathbf{T} \in \mathbb{Z}^{s \times n}$

$\mathrm{Dec} : \mathbb{Z}^s \to \{0, 1\}$

$\mathrm{Dec}(Tx) = P(x)$ w.h.p. over $\mathcal{D}$

A Naive Plan

For a distribution $\mathcal{D}$ over frequency vectors $x \in \{0, \ldots, m\}^n$.

Dynamic Streaming Algorithm $\mathcal{A}$ $\rightarrow$

Linear Sketch

$\mathbf{T} \in \mathbb{Z}^{s \times n}$

$\mathrm{Dec} : \mathbb{Z}^s \rightarrow \{0, 1\}$

$\mathrm{Dec}(Tx) = \mathrm{diam}(x)$ w.h.p. over $\mathcal{D}$

A Naive Plan

For a distribution $\mathcal{D}$ over frequency vectors $x \in \{0, \ldots, m\}^n$.

Dynamic Streaming Algorithm $\mathcal{A}$ $\mapsto$

Linear Sketch

$\mathbf{T} \in \mathbb{Z}^{s \times n}$

$\mathrm{Dec} : \mathbb{Z}^s \to \{0, 1\}$

$\mathrm{Dec}(Tx) = \mathrm{diam}(x)$ w.h.p. over $\mathcal{D}$

Find a matrix $H \in \mathbb{R}^{n \times s}$ such that HT is a minrank completion of some $\mathbf{G} \sim \mathcal{G}(n, n, p)$.

$$s = \mathrm{rank}(T) \geq \mathrm{rank}(HT) \geq \tilde{\Omega}(np)$$

The Problem

But there's a problem!

- When constructing our matrix H, our argument requires $m \gg \exp(s)$
- In the reduction of [LNW14], however, $s \approx \text{Space}(\mathcal{A}) \cdot \log m$

$$\text{Space}(\mathcal{A}) \cdot \log m \approx s = \text{rank}(T) \geq \tilde{\Omega}(np)$$

- This will give a vacuous lower bound on Space($\mathcal{A}$)!

The $\log m$ factor

Is the $\log(m)$ necessary?

The $\log m$ factor

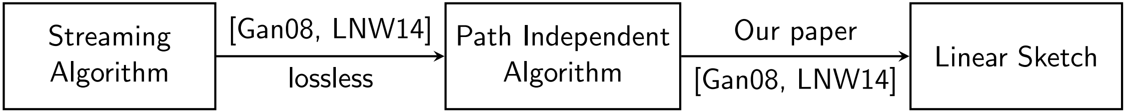
Is the $\log(m)$ necessary?

Consider the function $f(x) = x_1 \mod 2$.

- Optimal streaming algorithm stores one bit
- For any correct linear sketch, $\mathbf{T} \cdot e_1 \neq 0$ with constant probability. Thus $\mathbf{T} \cdot i \cdot e_1$ results in m unique states

Two Reductions

(scale-invariant functions)
$f(ax) = f(x)$ for $a \in \mathbb{Z}^+$



(general functions,
$\log m$)

The Plan

1. A new connection between streaming algorithms and linear sketches for scale-invariant functions (rids dependence on $\log m$)
2. Lower bound on linear sketches for diameter
3. A tight upper bound for diameter in general metrics
4. A new characterization of metrics that are "hard" to embed into ℓ_∞

2. Lower bound on Linear Sketches for Diameter

Starting point: We assume a lossless reduction from streaming algorithms to linear sketches.

Goal: Come up with a *hard* distribution over metrics for linear sketches

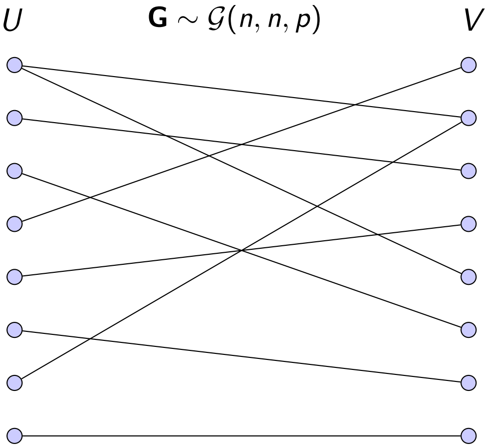
Theorem (Informal)

Let $\mathbf{G} \sim (n, n, p)$ *for suitable* p*. There exists a distribution* $\mathcal{D}(\mathbf{G})$ *such that any linear sketch* (T, Dec) *satisfying*

$$\Pr_{\mathbf{x} \sim \mathcal{D}(\mathbf{G})}[\mathrm{Dec}(T\mathbf{x}) \text{ } c\text{-approximates diameter of } \mathbf{x}] \geq 1 - o(1)$$

must have $\dim(T) \geq n^{\Omega(1/c)}$.

The Hard Metric



The Hard Metric

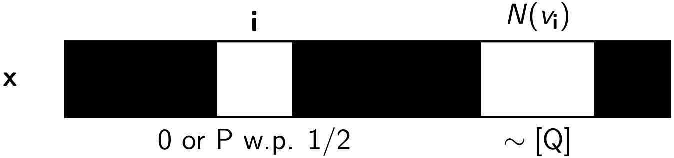
Setting $p \approx n^{-1+1/c}$, a draw $\mathbf{G} \sim \mathcal{G}(n, n, p)$ will satisfy

1. $\Omega(n)$ indices $i \in [n]$ will have $d_{\mathbf{G}}(u_i, v_i) \geq 2c$
2. $\text{Minrank}(\mathbf{G}) \geq \tilde{\Omega}(np)$

with high probability.

Designing the Hard Distribution

1. Choose $\mathbf{i} \in [n]$ and let $\mathbf{x_i} = 0$ or P each w.p. $1/2$
2. Choose $\mathbf{x}_j \sim [\mathrm{Q}]$ when $j \in N(v_\mathbf{i})$ and $\mathbf{x}_j = 0$ otherwise

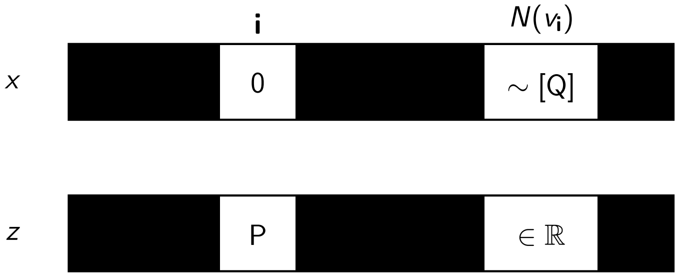


- When $\mathbf{x_i} = 0$, the subset is contained within $N(v_\mathbf{i})$ and thus has diameter ≤ 2
- When $\mathbf{x_i} = 1$, the diameter is $\geq 2c$

Any linear sketch c-approximating diameter must distinguish between $\mathbf{x_i} = 1$ or 0!

Fooling the Linear Sketch

Under this distribution, when does our linear sketch get fooled?



Suppose $Tz = 0$,

- $Tx = T(x + z)$, but the diameters of x and z are different!
- Can shift any $x \in \text{support}(\mathcal{D})$ to find many mistakes

Fooling the Linear Sketch

Otherwise, if $Tz \neq 0$ for all z, then it must be the case

$$0 \neq Tz = T^i z_i + \sum_{j \in N(v_i)} T^j z_j$$

This is a convex set, so by the separating hyperplane theorem there exists a hyperplane $h \in \mathbb{R}^s$ such that $\langle h, T^i z_i + \sum_{j \in N(v_i)} T^j z_j \rangle > 0$

$$\langle h, T^i \rangle \neq 0 \text{ and } \langle h, T^j \rangle = 0$$

Stacking these hyperplanes for each $i \in [n]$ gives us H such that HT is a minrank completion of $\mathbf{G}$!

Two Problems

1. We crucially relied on z being real-valued... But T does not need to be correct on real-valued vectors! The shifting argument won't work

Fixing Problem 1

- Suppose $z \in \mathbb{R}^n$ such that $Tz = 0$
- T obtained from reduction has bounded entries $\|T\|_\infty \leq \exp(n \log n)$
- This allows us to find an *integer* fooling vector w with entries $\gg \exp(sn)$

Two Problems

2. Even if w is integer valued, we must ensure that $\mathbf{x} + w$ has non-negligible mass under $\mathcal{D}$ to find significant mass of mistakes

Fixing Problem 2

- If we knew w_i exactly, we could set $\mathrm{P} = w_i$. This would result in a $1/2n$ error probability for the linear sketch
- But w_i depends on T, which in turn, depends on $\mathcal{D}$ (circularity)
- Set $\mathrm{P} \approx (\exp(sn))!$ so that $x + \frac{\mathrm{P}}{z_i} \cdot z$

This is where the $m \gg \exp(s)$ occurs! Makes it necessary to remove the $O(\log m)$ loss in space complexity

Conclusion

Open Questions

1. Can we get a similar result for Euclidean space?
2. Where else can we apply the *lossless* reduction from linear sketches to streaming algorithms for scale-invariant functions?

Thanks :)

References

Noga Alon, Igor Balla, Lior Gishboliner, Adva Mond, and Frank Mousset.
The minrank of random graphs over arbitrary fields.
Israel Journal of Mathematics, 235(1):63–77, 2020.

Ziv Bar-Yossef, Yitzhak Birk, T. S. Jayram, and Tomer Kol.
Index coding with side information.
In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, USA, 2006. IEEE Computer Society.

Xi Chen, Vincent Cohen-Addad, Rajesh Jayaram, Amit Levi, and Erik Waingarten.
Streaming euclidean mst to a constant factor.
In *Proceedings of the 55th ACM Symposium on the Theory of Computing (STOC '2023)*, pages 156–169, 2023.

Artur Czumaj, Shaofeng H.-C. Jiang, Robert Krauthgamer, Pavel Veselý, and Mingwei Yang.
Streaming facility location in high dimension via geometric hashing.