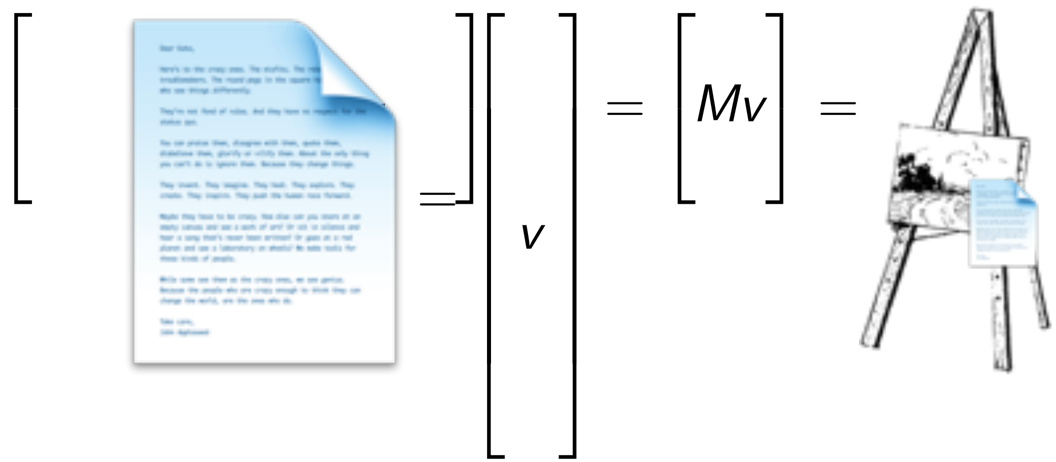
Homomorphic Sketches

Shrinking Big Data without Sacrificing Structure

Andrew McGregor
University of Massachusetts



Sketches: Encode data as vector; use *linear projections* to compress the data while preserving properties.

Extensive theory with connections to compressed sensing, metric embeddings; *widely applicable* since parallelizable and suitable for stream processing.

Many positive results such as distinct elements, entropy, frequency moments, quantiles, histograms...

Previously at the workshop...

Regression Problem → *Problem in Sketch Space*

$$\begin{pmatrix} \\ A \\ \\ \end{pmatrix} \in \mathbb{R}^{n \times d} \quad \begin{pmatrix} \\ b \\ \\ \end{pmatrix} \in \mathbb{R}^{n}$$

$$\begin{pmatrix} \tilde{A} \end{pmatrix} \in \mathbb{R}^{s \times d} \quad \begin{pmatrix} \tilde{b} \end{pmatrix} \in \mathbb{R}^{s}$$

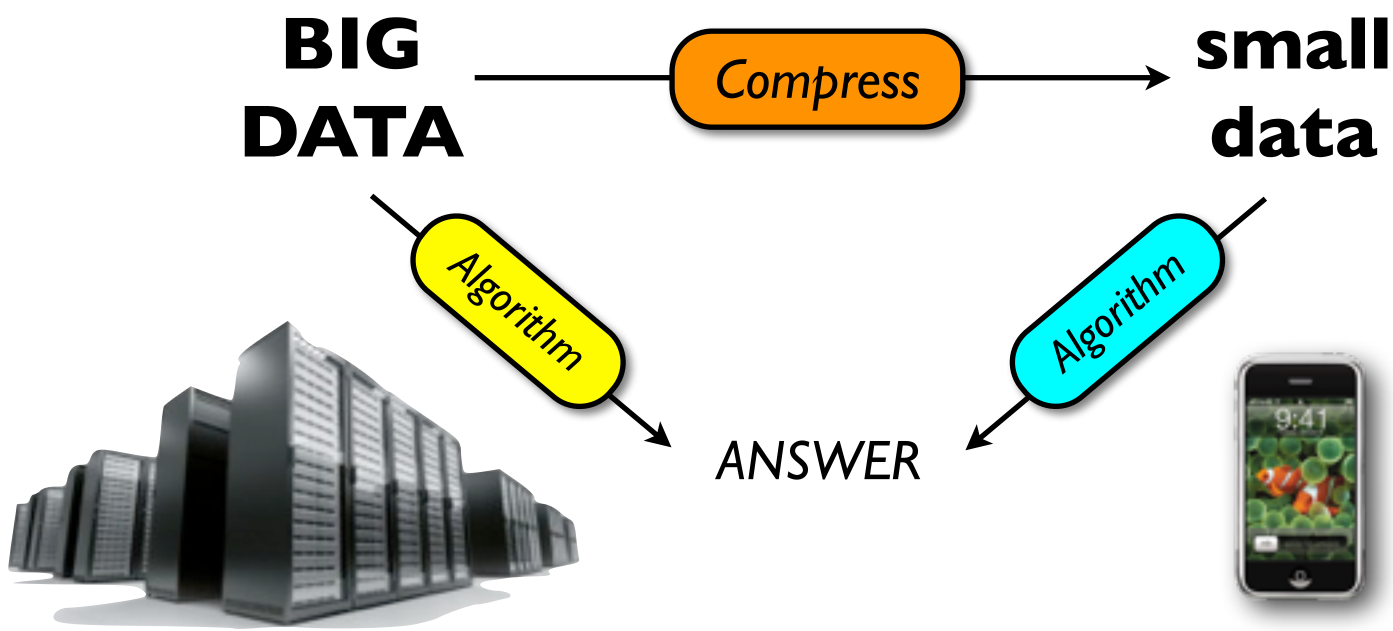
Find $x \in \mathbb{R}^d$ such that $Ax \approx b$

Find $x \in \mathbb{R}^d$ such that $\tilde{A}x \approx \tilde{b}$

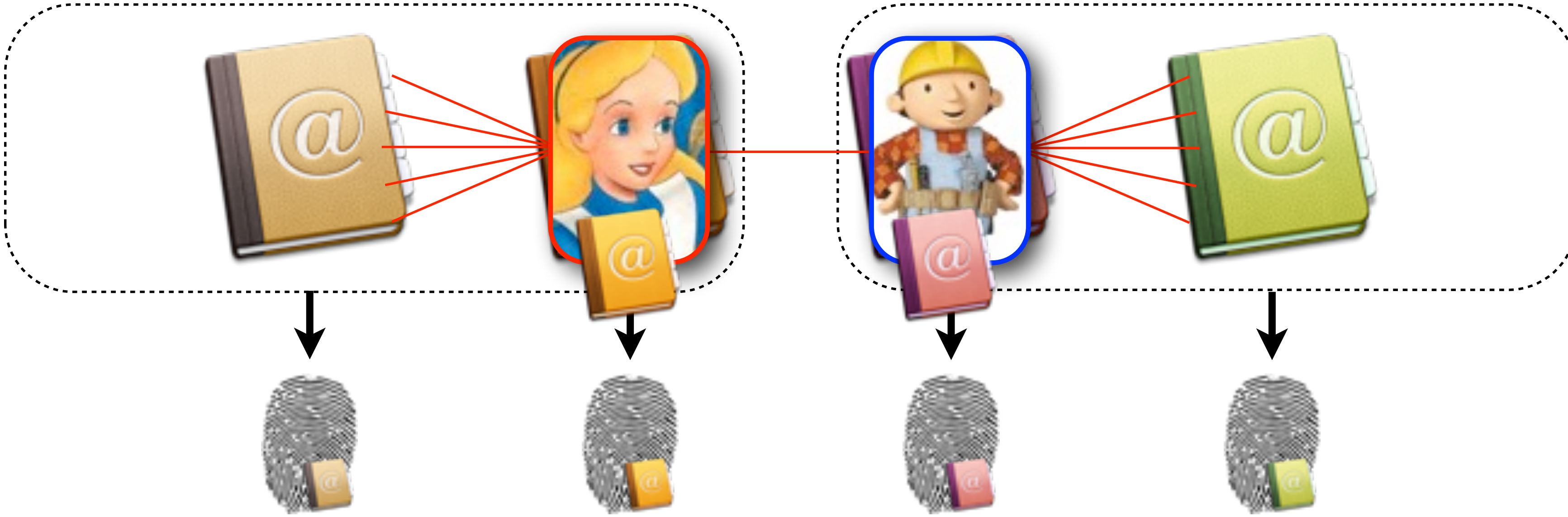
Underlying Idea: Reduce large instance to small instance and solve in *sketch space*. Sketches are natural fit for linear algebra problems since linear operations on sketches correspond to operations on original data.

Is it possible to analyze richer combinatorial and group-theoretic structure via linear sketches?

Can we make compression "homomorphic" in a more general sense and run algorithms on sketched data?



First Result...

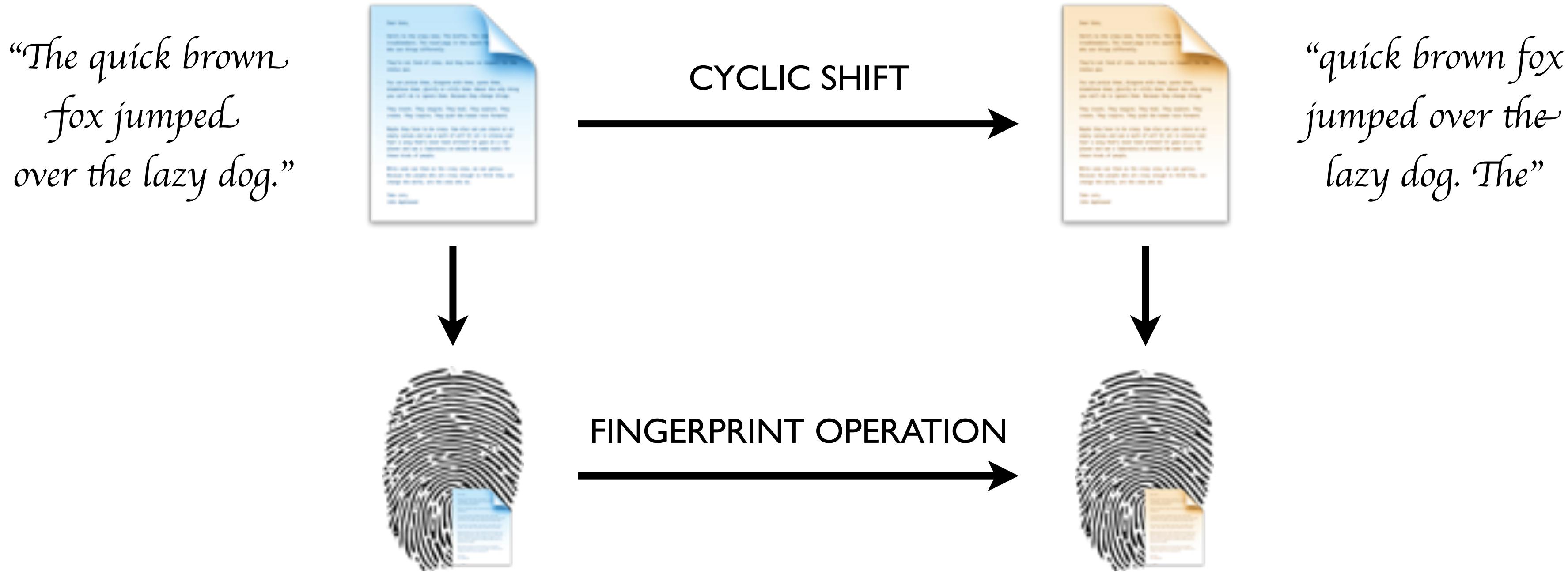


Problem: Sketch each row of nxn adjacency matrix such that we can check connectivity using sketches.

Theorem: Sketches of size O(polylog n) bit suffice!

Surprising? Seems impossible if there are bridge edges.

Second Result...



Problem: Sketch files such that we can test if files are close under some cyclic rotation.

Theorem: Sketches of size ≈ no. of divisors of n suffice.

Surprising? Sketch size isn't monotonic in file size!

I. **Connectivity**

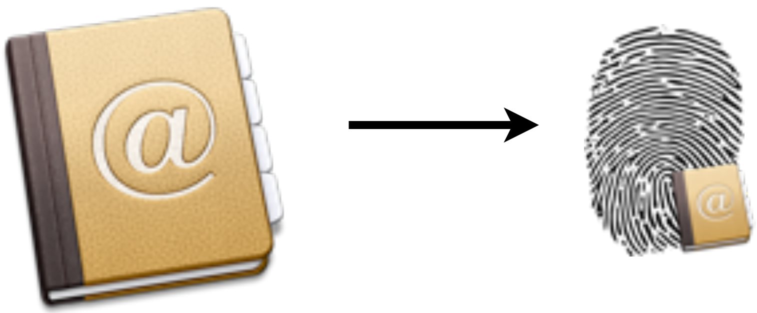
II. **Misalignment**

a) Connectivity via O(polylog n) bit Fingerprints

b) Extension to Estimating Cuts and Eigenvalues

Joint work with Kook Jin Ahn and Sudipto Guha
with Michael Crouch and Daniel Stubbs

Sketches for Connectivity



- *Theorem:* Can check connectivity with high probability using a O(polylog n) bit fingerprint of each adjacency list.
- *Corollary:* Can monitor connectivity of dynamic graph streams where edges both inserted and deleted.

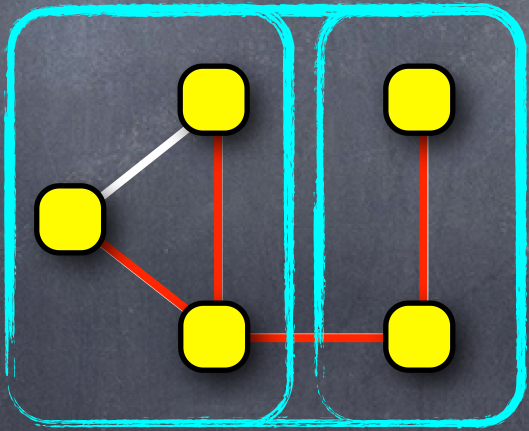
e.g., [Feigenbaum, Kannan, McGregor, Suri, Zhang 2004, 2005], [McGregor 2005] [Jowhari, Ghodsi 2005], [Zelke 2008], [Sarma, Gollapudi, Panigrahy 2008, 2009] [Ahn, Guha 2009, 2011], [Konrad, Magniez, Mathieu 2012], [Goel, Kapralov, Khanna 2012]

- *More recently:* Estimating cut sizes and spectral properties from short linear sketches and processing sliding windows.

[Crouch, McGregor, Stubbs 2013], [Ahn, Guha, McGregor 2012, 2013]

Basic Algorithm

- Plan: Sketch data and emulate connectivity algorithm in sketch space. What algorithm should we emulate?
- Algorithm (Spanning Forest):
  1. For each node: pick incident edge
  2. For each connected comp: pick incident edge
  3. Repeat until no edges between connected comp.

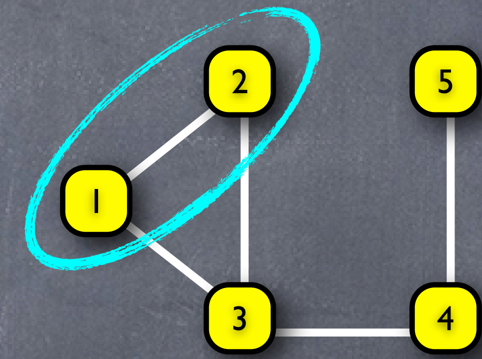


- Lemma: Find a spanning forest after O(log n) rounds.

Emulating Algorithm via Sketches

- **Defn:** Let a_i be i^{th} row of signed vertex-edge matrix

	{1,2}	{1,3}	{1,4}	{1,5}	{2,3}	{2,4}	{2,5}	{3,4}	{3,5}	{4,5}
$\mathbf{a}_1 =$	1	1	0	0	0	0	0	0	0	0
$\mathbf{a}_2 =$	−1	0	0	0	1	0	0	0	0	0
$\mathbf{a}_1 + \mathbf{a}_2 =$	0	1	0	0	1	0	0	0	0	0



For $S \subset V$, non-zero entries of $\sum_{i \in S} a_i$ equals $E(S, V \setminus S)$

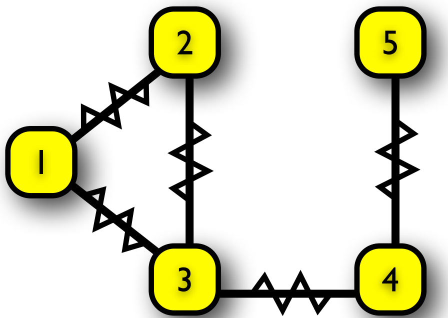
- **Sketch:** Ma_i where M is a random projection to $\mathbb{R}^{\text{polylog } N}$ such that $\forall x$, can recover a non-zero entry from Mx.

- **Utility:** Can find an edge across any cut S.

$$\sum_{j \in S} M\mathbf{a}_j = M\Big(\sum_{j \in S} \mathbf{a}_j\Big) \underset{process}{\longrightarrow} \Big(\text{non-zero entry of } \textstyle\sum_{j \in S} a_j\Big) = \text{cut edge}$$

Sparsification

- *Algorithm:* Sample each edge uv with probability p_{uv} and weight sampled edges by $1/p_{uv}$.
- *Theorem (Fung et al.)* If $p_{uv} \geq 1/c_{uv}$ then we $(1\pm\varepsilon)$ approx. all cuts where c_{uv} is size of min uv cut.
- *Theorem (Spielman-Srivastava)* If $p_{uv} \geq r_{uv}$ then we get $(1\pm\varepsilon)$ spectral sparsifier where r_{uv} is the effective resistance.



r_{uv} is potential difference when unit of flow injected at u and extracted at v

- *Lemma:* If uv is an edge then $1 \leq r_{uv}/c_{uv} \leq O(n^{2/3})$
- *Theorem:* Can sample w/probability t/c_{uv} with $\tilde{O}(t)$ sketches.

Sampling edges via k-Skeletons

- Goal: Sample edge e with probability t/c_e.
- Connectivity Result: Given $\tilde{O}(k)$ bit sketches, can find all edges in a cut of $\leq k$ edges. Call this a "k-skeleton".
- Algorithm (Edge sampling via k-skeletons)
  - Let G_i be graph with edges sampled w/p 2^{-i}.
  - Return k-skeleton H_i for each G_i where $k= 2t$
- Thm: $e=(u,v)$ is in some H_i with probability at least t/c_e
- Proof: Let C be edges in min u-v cut in G.
  - For $i= \log c_e/t$, then $E[|C\cap G_i|]=t$ and whp $|C\cap G_i|\leq 2t$.
  - Hence $e\in H_i$ iff $e\in G_i$ which happens w/p t/c_e

I. **Connectivity** II. **Misalignment**

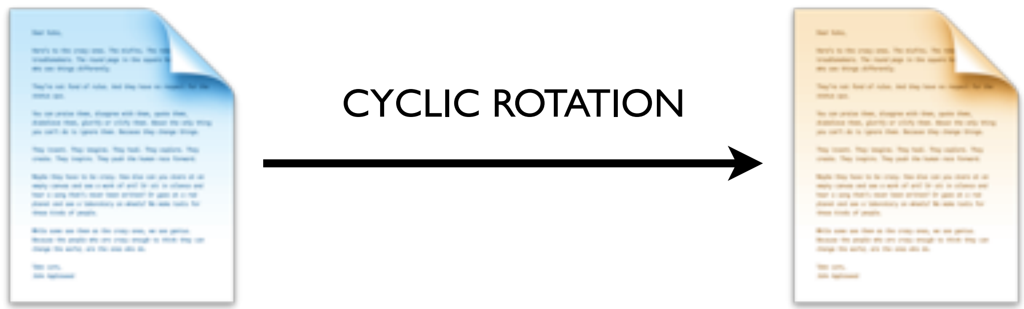
a) Testing Equality with Rotation

b) Matching Lower Bound

Joint work with Alexandr Andoni, Assaf Goldberger, Ely Porat

Fingerprints for Rotation

"The quick brown fox jumps over the lazy dog."



"quick brown fox jumps over the lazy dog. The"

- *Theorem:* There's a D(n) polylog n bit fingerprint F that is:
  - *Useful:* F(a) and F(b) determine if a, b$\in\mathbb{Z}^n$ are rotations w.h.p.
  - *Homomorphic:* From F(a) can construct F(any rotation of a)
  - *Linear:* From F(a) and F(b) can compute F(a+b).
- *Theorem:* Fingerprints with above properties need D(n) bits.
- *Extension:* Extends to case where files aren't perfect rotations.

*D(n) is the number of divisors of n

False Start: Fermat's Little Theorem

- Karp-Rabin: For some p and r, encode $a=a_0a_1a_2 \ldots a_{n-1}$ as

$$f(r,a) = a_0 + a_1 r + a_2 r^2 + \ldots a_{n-1} r^{n-1} \bmod p$$

- Fermat's Little Thm: If p=n+1 prime, $r^n=1$ mod p and so,

$$\begin{aligned} rf(r, a_0 a_1 \ldots a_{n-1}) &= a_0 r + a_1 r^2 + a_2 r^3 + \ldots + a_{n-1} r^n \\ &= a_{n-1} + a_0 r + a_1 r^2 + \ldots + a_{n-2} r^{n-1} \\ &= f(r, a_{n-1} a_0 \ldots a_{n-2}) \end{aligned}$$

- So, if b is k-shift of a then $g(r) = r^k f(r,a) - f(r,b) = 0$
- Schwartz-Zippel: If r is random and g non-zero:

$$P[g(r)=0] \leq (n-1)/p = 1 - O(1/n)$$

- Conclusion: No false negatives but likely false positives.

Beyond Schwartz-Zippel

- Evaluate g on roots of x^n-1 but work in larger field
- x^n-1 factorizes as D(n) irreducible polys over rationals:

$$\begin{aligned} x^{10} - 1 &= \Phi_1(x)\Phi_2(x)\Phi_5(x)\Phi_{10}(x) \\ &= (x-1)(1+x)(1-x+x^2-x^3+x^4)(1+x+x^2+x^3+x^4) \end{aligned}$$

- At least one ϕ_i has no shared roots with g:
  - If ϕ_i shares one root, ϕ_i divides g (Abel's Irred. Thm)
  - Can't all divide g because g has degree ≤ n-1
- Suffices to test g on an arbitrary root of each ϕ_i
- Bad News: Can't guarantee g(r) has finite precision.
- Good News: Work modulo a random p. Can show ϕ_i still doesn't share roots with g whp by analyzing resultant.

Lower Bound: Basic Idea

- Can recover D(n) bits about a from F(a): sum the fingerprints of various rotations
- To deduce $\alpha = \sum a_i$ from $F(a_0 a_1 a_2 a_3 a_4 a_5)$

$$F(a_0 a_1 a_2 a_3 a_4 a_5) + F(a_1 a_2 a_3 a_4 a_5 a_0) + \ldots + F(a_5 a_0 a_1 a_2 a_3 a_4) = F(\alpha\alpha\alpha\alpha\alpha\alpha)$$

  and compare $F(gggggg)$ for all g until matches.

- To deduce $\beta = a_1 + a_3 + a_5$

$$F(a_0 a_1 a_2 a_3 a_4 a_5) + F(a_2 a_3 a_4 a_5 a_0 a_1) + F(a_4 a_5 a_0 a_1 a_2 a_3) = F(\beta\gamma\beta\gamma\beta\gamma)$$

  and compare $F(gg'gg'gg')$ for all g, g'=α-g until matches.

- And so on for other divisors of n...

Thanks!

- *Homomorphic Sketches:* Compress using sketches such that we can run algorithms on compressed data directly. Resulting algorithms are *parallelizable* + *streamable*.
- *Graphs:* Dimensionality reduction for preserving structural properties. Enables dynamic graph streaming.
- *Fingerprinting with Misalignments:* Tight bounds on size of fingerprint necessary for testing equality up to rotations.