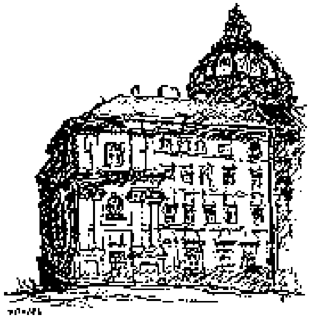
Computing edit distance

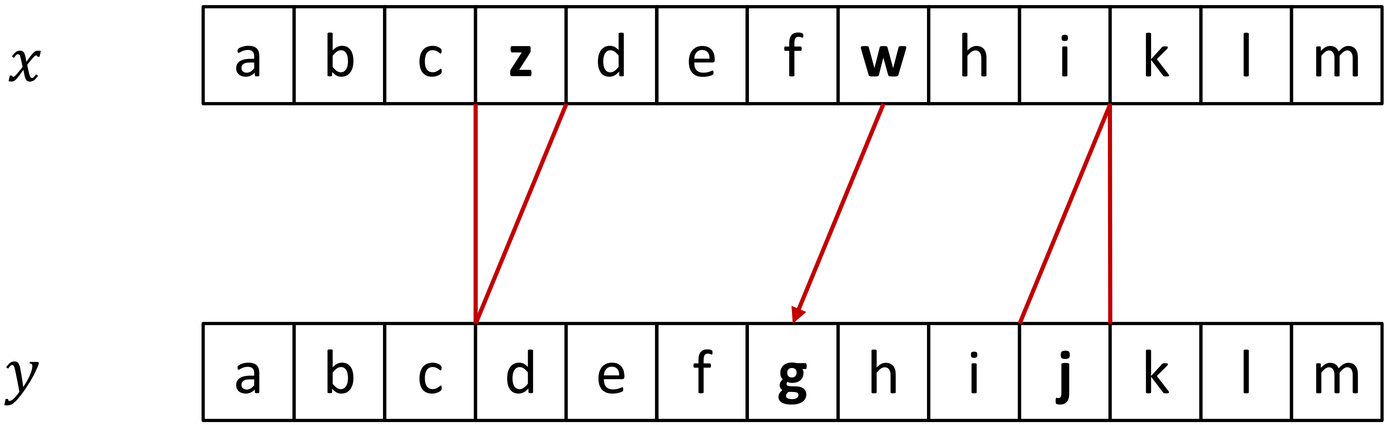
Michal Koucký

Charles U., Prague



Wrocław – CPM 2021

Edit distance



Edit distance $\mathrm{ED}(x, y)$:

the number of
1) bit flips/symbol changes
2) insertions, and
3) deletions

that transform x into y.

Variants of edit distance

- *Levenshtein distance:* vanilla edit distance.
- *Longest Common Subsequence:* dual measure.
- *Ulam distance:* large alphabet, each symbol appears at most once.
- *Edit distance with moves:* additional operation – block move.
- *Hamming distance.*

Main questions

How do you compute edit distance efficiently:

- Exact algorithms.
- Approximate algorithms.

Scenarios:

- Full access to x and y.
- Sketches of x and/or y.

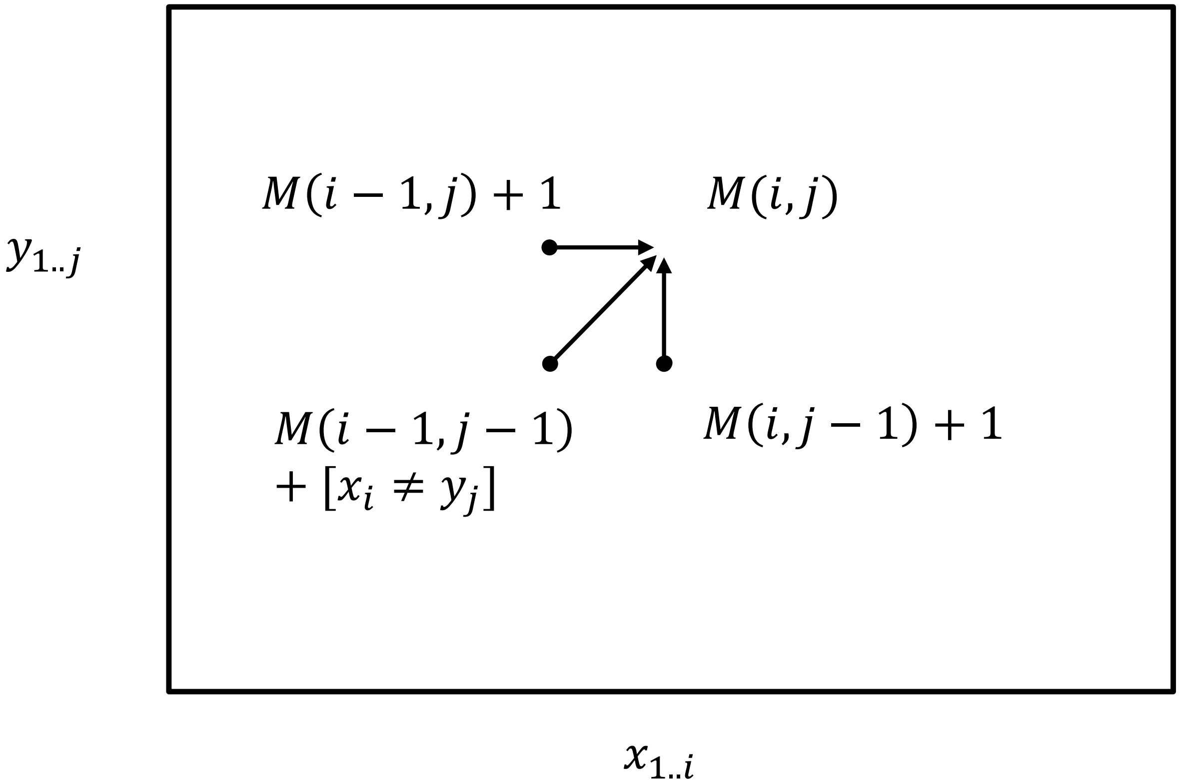
Computing edit distance

- Wagner-Fischer'74, Masek-Paterson'80, …
  Grabowski'16 $O(n^2 / \log^2 n)$
- Ukkonen'85 $O(kn)$
- Myers'86, Landau-Vishkin'88,
  Landau-Myers-Schmidt'98 $O(n + k^2)$

… and many others

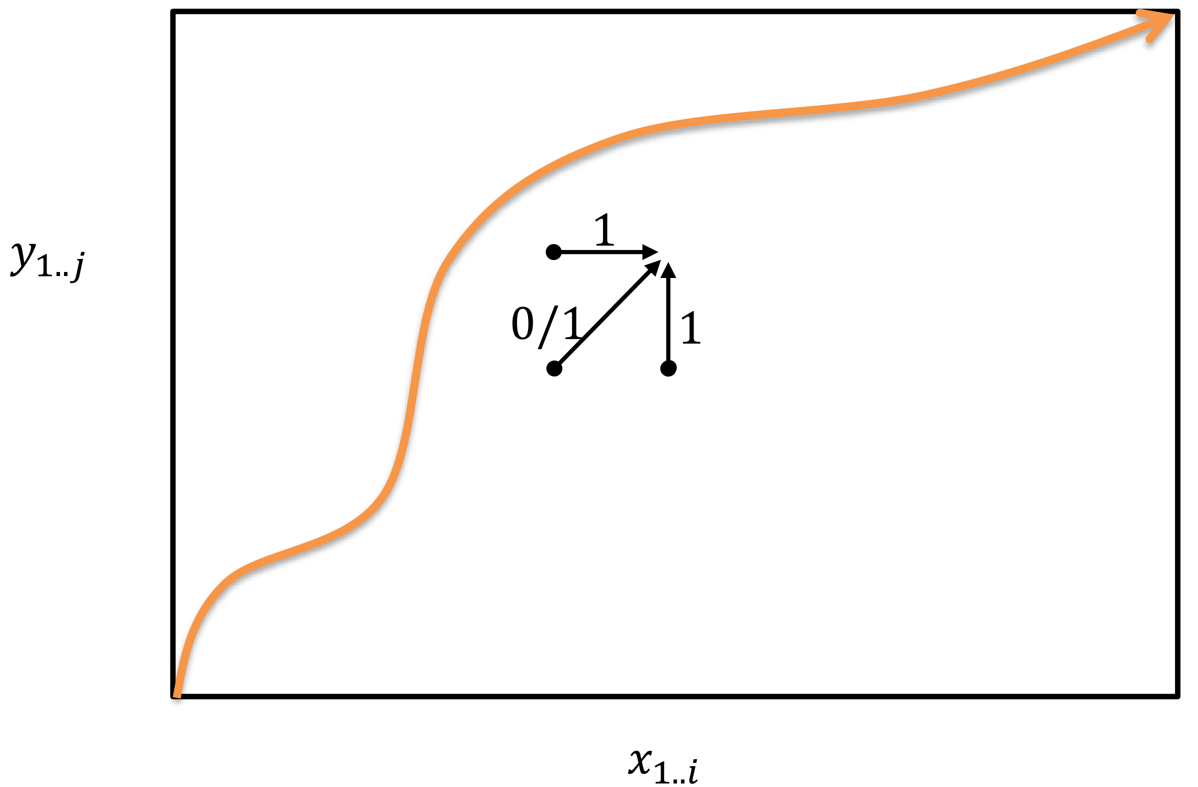
$k = \mathrm{ED}(x, y)$

Computing ED: Dynamic programming



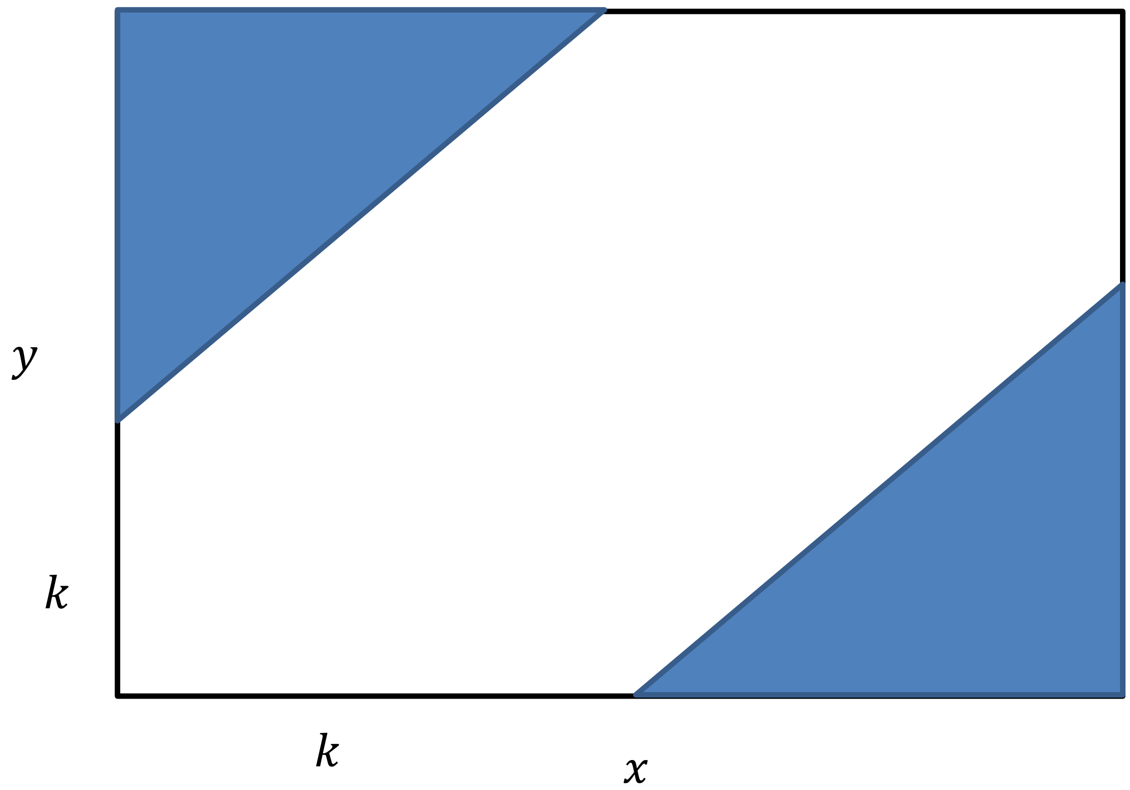
$M(i,j) = \mathrm{ED}\,(x_{1\dots i}, y_{1\dots j})$ → $O(n^2)$ time algorithm

Computing ED: Dynamic programming



$M(i,j) = \mathrm{ED}\,(x_{1\ldots i}, y_{1\ldots j})$ $\rightarrow O(n^2)$ time algorithm

Computing ED: Dynamic programming



Ukkonen’95: $O(kn)$ time algorithm

Computing edit distance

- Wagner-Fischer'74, Masek-Paterson'80, …
  Grabowski'16 $O(n^2 \ / \ \log^2 n)$
- Ukkonen'85 $\mathrm{O}(kn)$
- Myers'86, Landau-Vishkin'88,
  Landau-Myers-Schmidt'98 $\mathrm{O}(n + k^2)$

… and many others

$k = \mathrm{ED}(x, y)$

Fine-grained complexity

Backurs-Indyk'15:

An algorithm for edit distance in time $O(n^{2-\epsilon})$

implies

an algorithm for SAT in time $2^{(1-\delta)n}$.

(contradicting Strong Exponential Time Hypothesis (SETH).)

formula φ	→	instance of edit distance (x, y)
n variables		$2^{n/2}$ length

Fine-grained complexity

Backurs-Indyk’15:

An algorithm for edit distance in time $O(n^{2-\epsilon})$

implies

an algorithm for SAT in time $2^{(1-\delta)n}$.

(contradicting Strong Exponential Time Hypothesis (SETH).)

Abboud-Hansen-Vassilevska Williams-Williams'16, Abboud-Backurs-Vassilevska Williams'15, Bringmann-Künnemann'15:

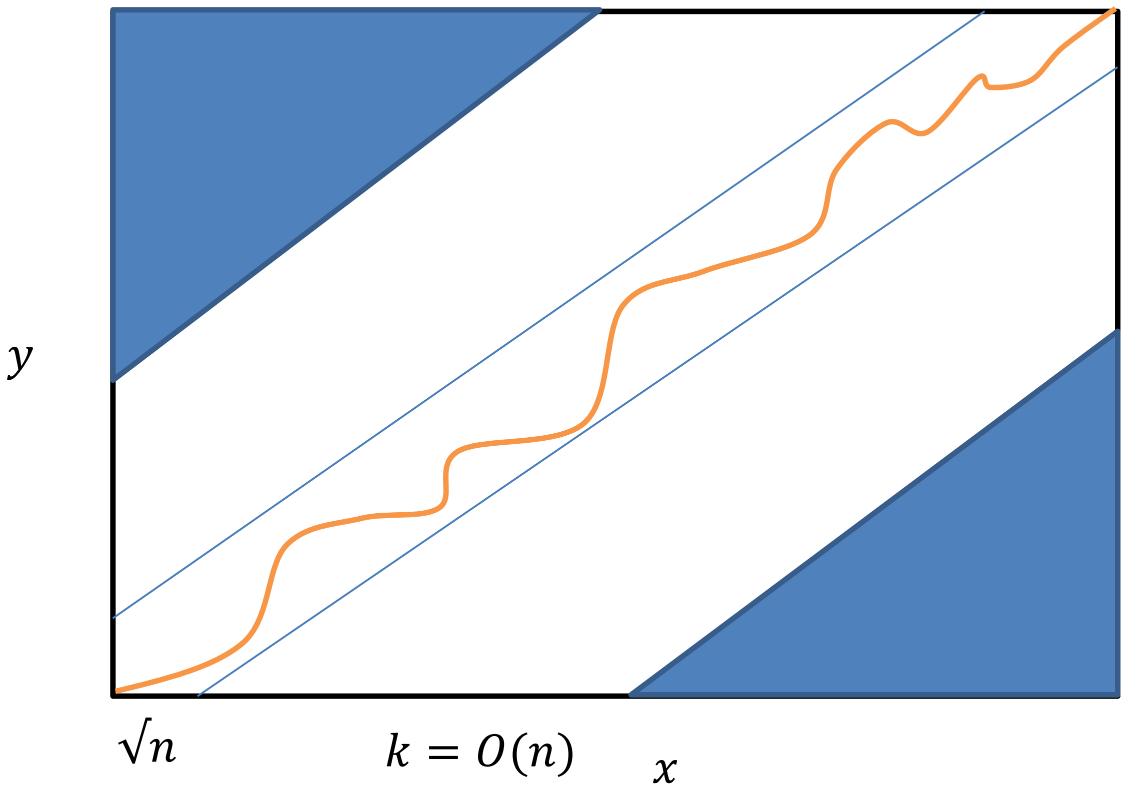
$o(n^2)$ algorithms for edit distance imply circuit lower bounds.

Question

- Algorithm running in time $O(n^{2-\epsilon})$ which for most pairs of strings x and y computes edit distance correctly?

[Goldenberg-Karthik'19]

Approach?



Approximating edit distance

	approximation	time
Landau-Myers-Schmidt'98	$\sqrt{n}$	$O(n)$
B.Yossef-Jayram-Krauthgamer-Kumar'04	$n^{3/7}$	$\tilde{O}(n)$
Batu-Ergun-Sahinalp'06	$n^{1/3+o(1)}$	$\tilde{O}(n)$
Andoni-Onak'09	$2^{\sqrt{\log n}}$	$O(n^{1+o(1)})$
Andoni-Krauthgamer-Onak'10	$\log^{O(1/\varepsilon)} n$	$O(n^{1+\varepsilon})$

Abboud-Backurs'17: $(1 + 1/poly \log)$-inapprox. in time $n^{2-\epsilon}$

Approximating edit distance

	approximation	time
Boroujeni-Ehsani-Ghodsi-Hajiaghayi-Seddighin'18		
quantum	$O(1)$	$O(n^{1.708\ldots})$
Chakraborty-Das-Goldenberg-K.-Saks'18	$O(1)$	$O(n^{1.647\ldots})$
Andoni'18	$O(1)$	$O(n^{3/2})$
Goldenberg-Rubinstein-Saha'20	$3+\epsilon$	$O(n^{1.6})$
Brakensiek-Rubinstein'20, K.-Saks'20		
far inputs	$O(1)$	$O(n^{1+\epsilon})$
Andoni-Nosatzki'20	$O(1)$	$O(n^{1+\epsilon})$

Abboud-Backurs'17: $(1+1/poly\log)$-inapprox. in time $n^{2-\epsilon}$

Approximating edit distance

Chakraborty-Das-Goldenberg-K.-Saks'18:

$O(1)$-approximation algorithm for edit distance in time $O(n^{2-2/7})$

$12/7 = 1.714\ldots$

Gap Edit Distance

Fixed constant $C > 1$.

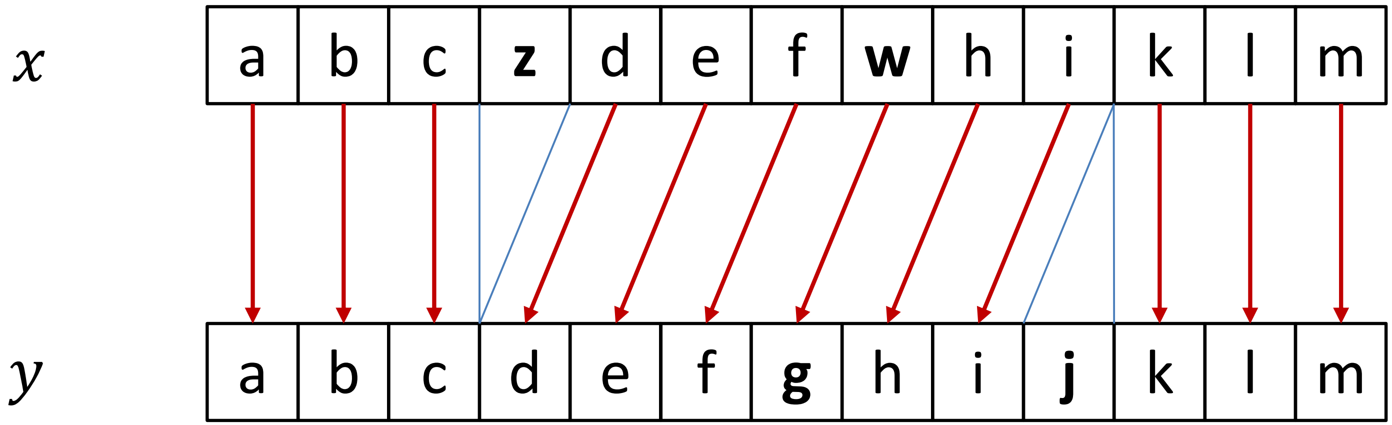
Input: $x, y \in \Sigma^n$, $\theta \in (0,1]$.

Output:

YES	if	$\mathrm{ED}(x, y) \leq \theta n$.
NO	if	$\mathrm{ED}(x, y) > C\theta n$.

CDGKS'18: Algorithm for some C running in time $O(n^{12/7})$.

Edit distance



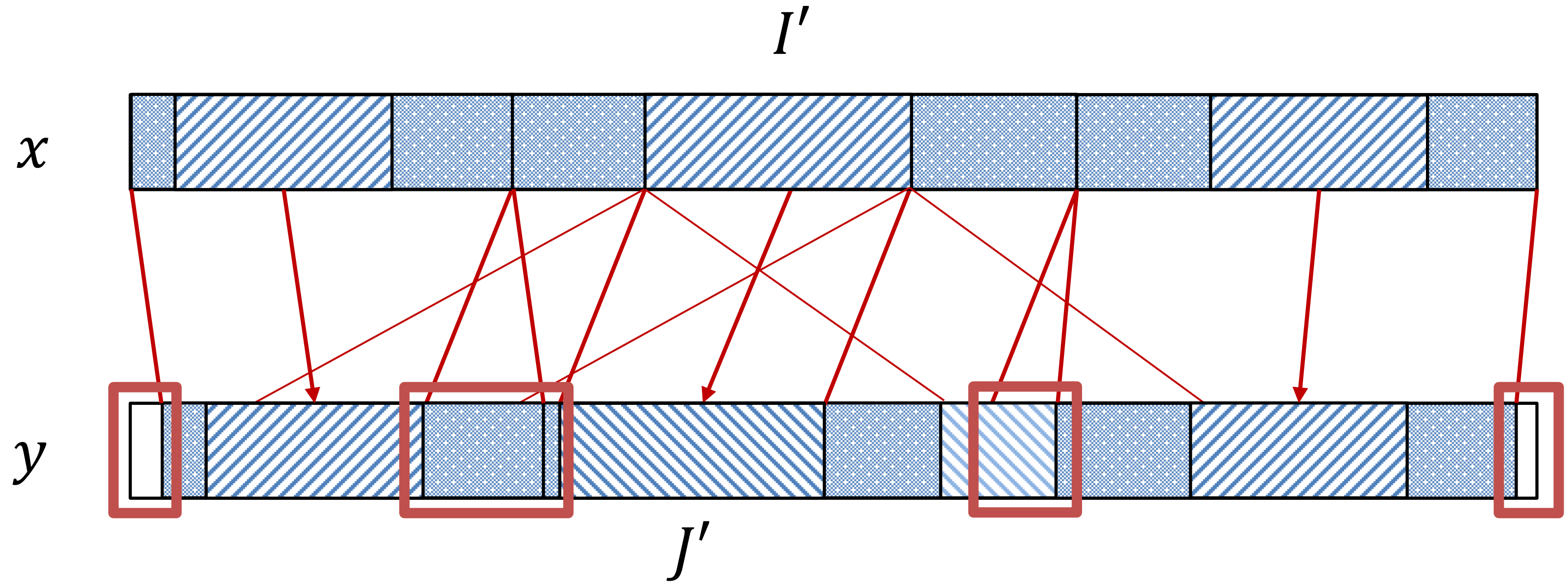
Edit distance $\mathrm{ED}(x, y)$:

the number of

1) bit flips/symbol changes
2) insertions, and
3) deletions

that transform x into y.

Main ideas of CDGKS

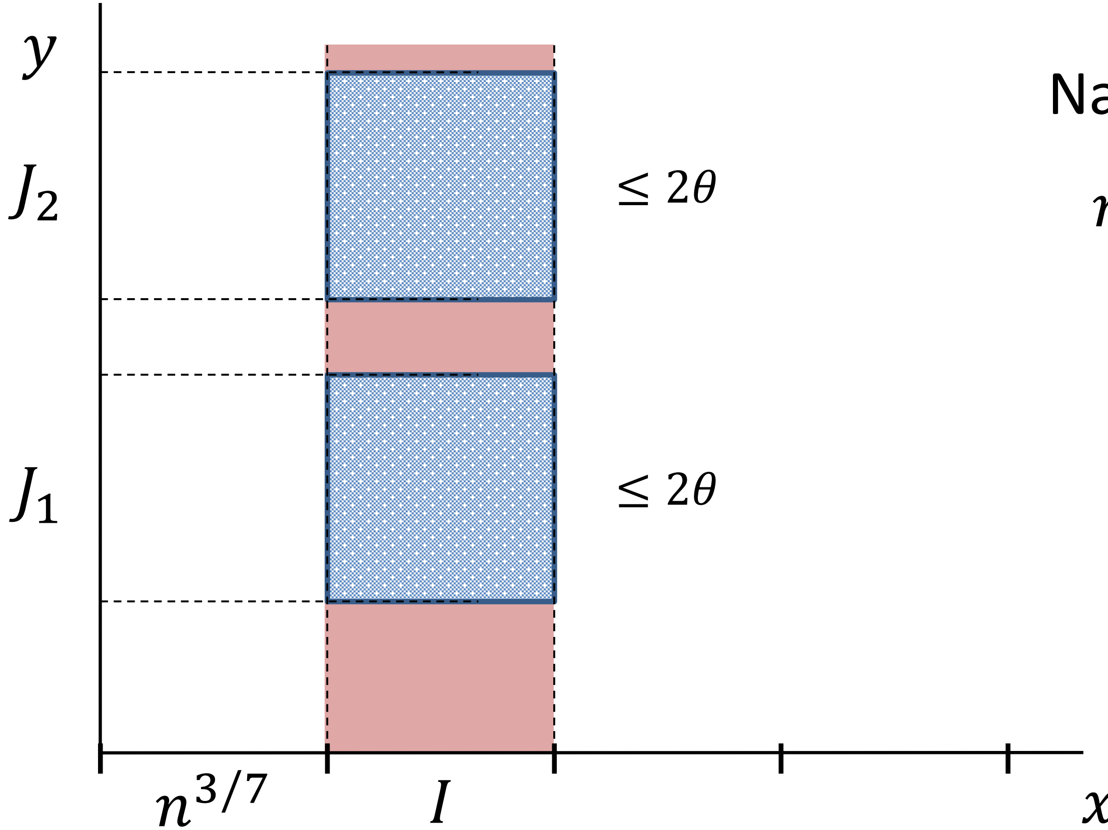


Assume $\mathrm{ED}(x, y) \leq \theta n$:

For most I' of size ℓ, $\mathrm{ED}(x_{I'}, y_{J'}) \leq 2\theta\ell$

$\ell = n^{\kappa}$

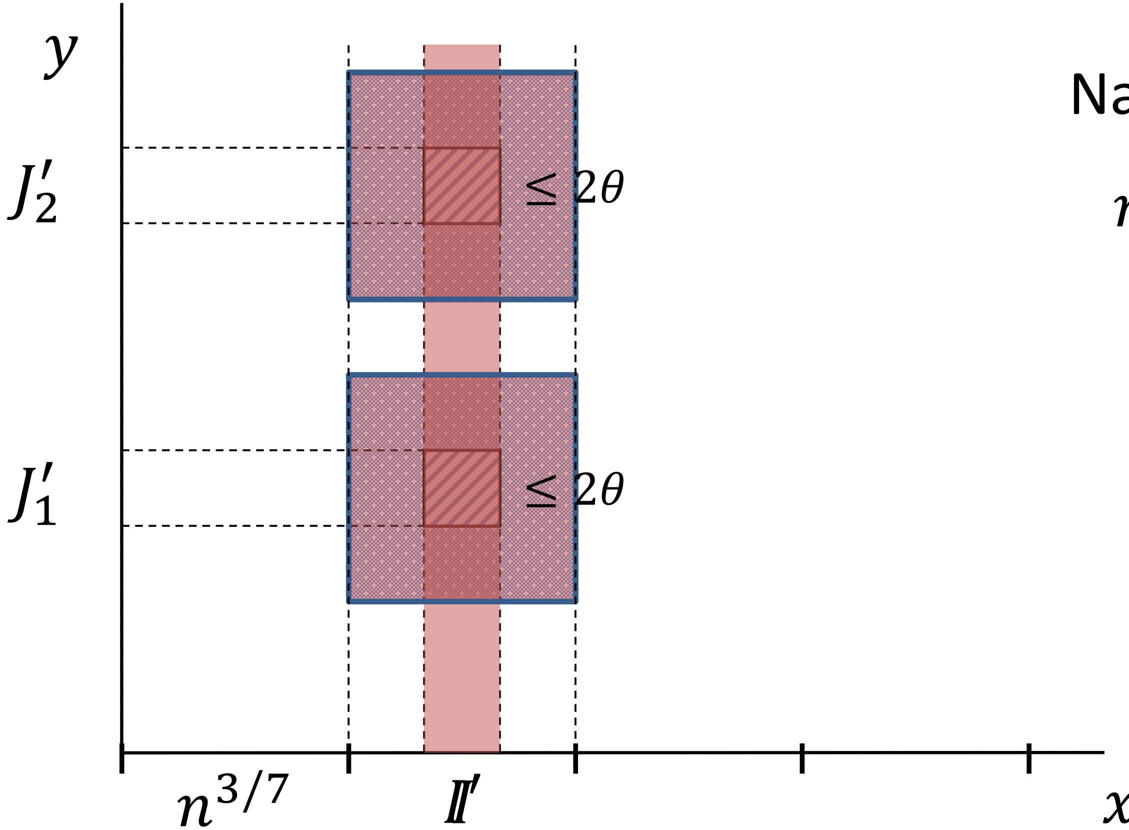
Searching for matches



Naïve cost:

$n^{3/7} \cdot n = n^{10/7}$

Sparse case of CDGKS



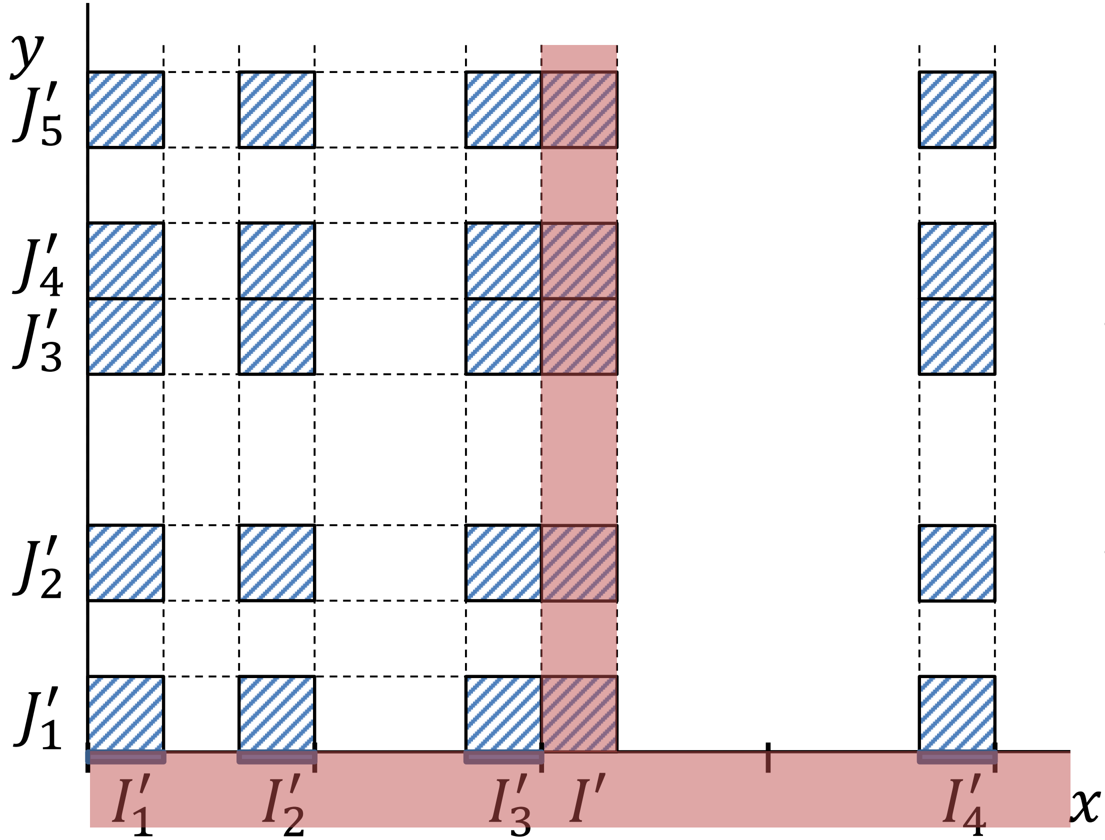
Naïve cost:

$n^{3/7} \cdot n = n^{10/7}$

$|I'| = n^{1/7}$ threshold $d = n^{2/7}$

New cost: $n^{1/7} \cdot n + (n^{3/7})^2 \cdot d = n^{8/7}$.

Dense case of CDGKS



$\Delta_{ed}\left(x_{I'}, x_{I_i'}\right) \leq 2\epsilon$

$\Delta_{ed}(x_{I'}, y_{J_j'}) \leq 3\epsilon$

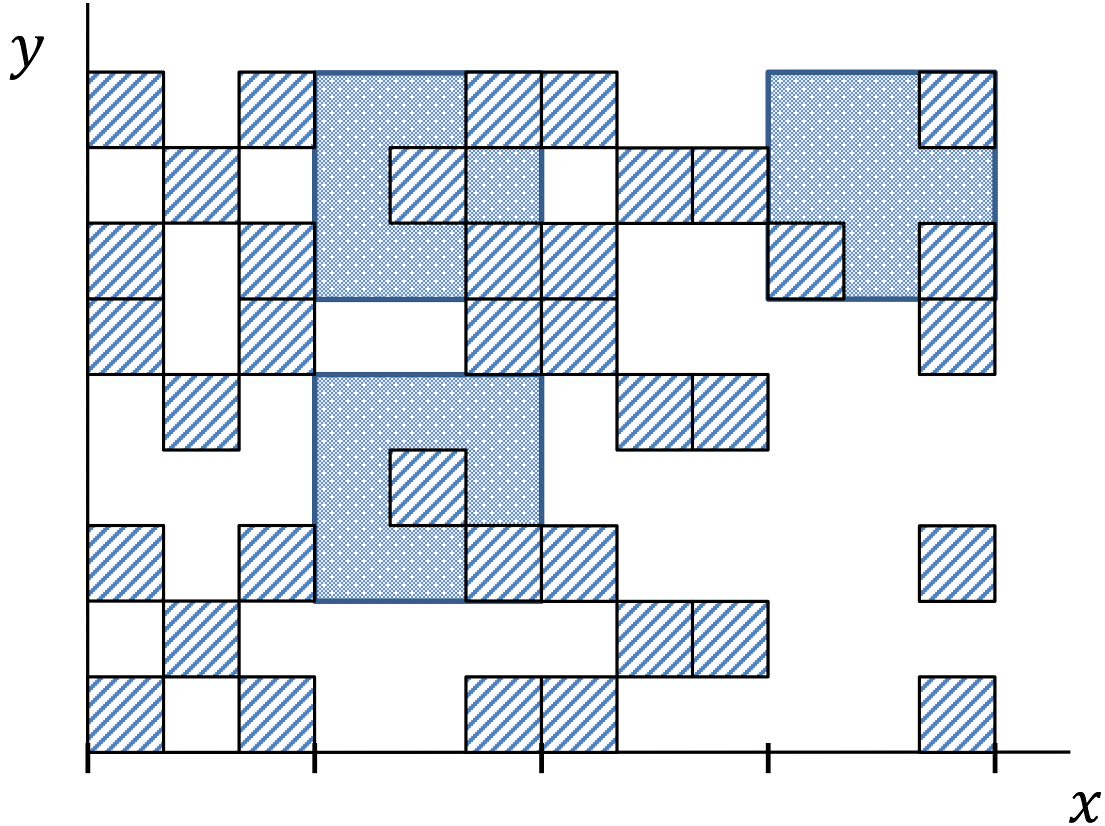
$\Rightarrow$

$\Delta_{ed}(x_{I_i'}, y_{J_j'}) \leq 5\epsilon$

$|I'| = n^{1/7}$ threshold $d = n^{2/7}$

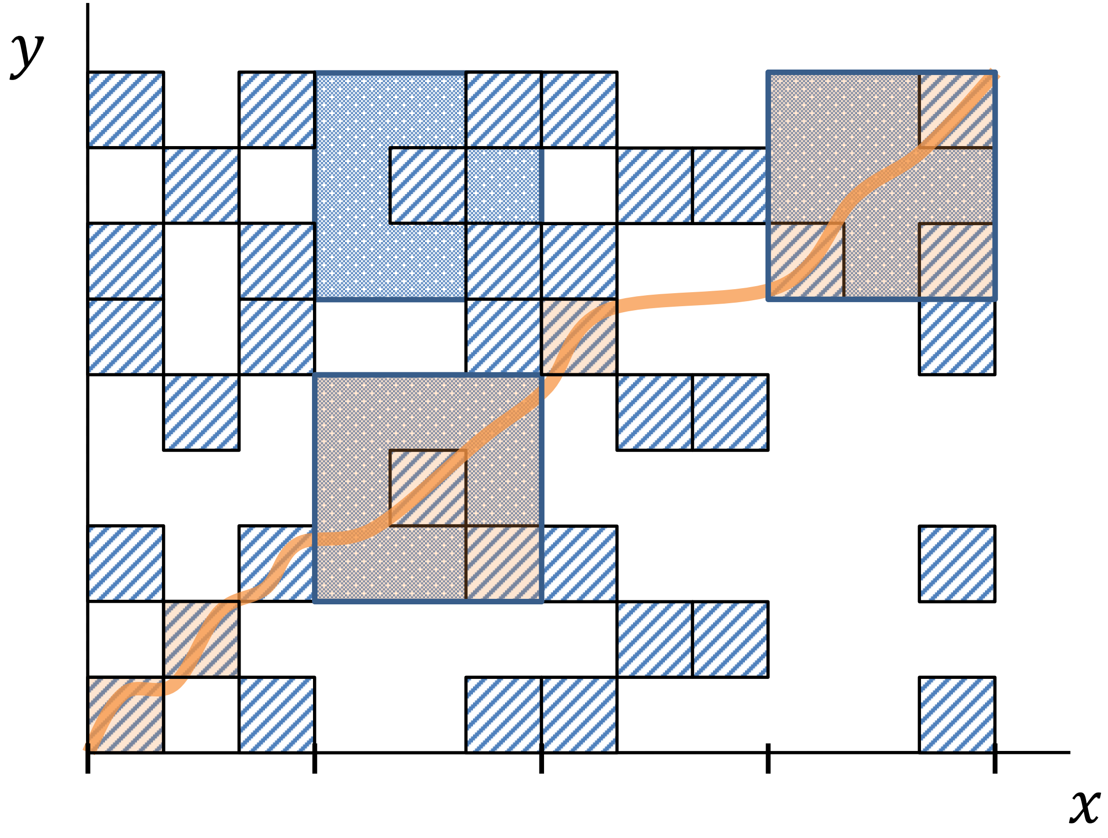
Total cost: $n^{8/7} \cdot n/(|I'| \cdot d) = n^2/d = n^{12/7}$.

Combining the two cases



1) Test each narrow column and process dense ones.
2) In each wide column, sample a sparse column and extend.
3) Repeat 1-2 for closeness parameters $\epsilon \in [\theta, 1], \epsilon = 2^{-i}$.

Recovering a good match



4) Find boxes that well approximate the best match.

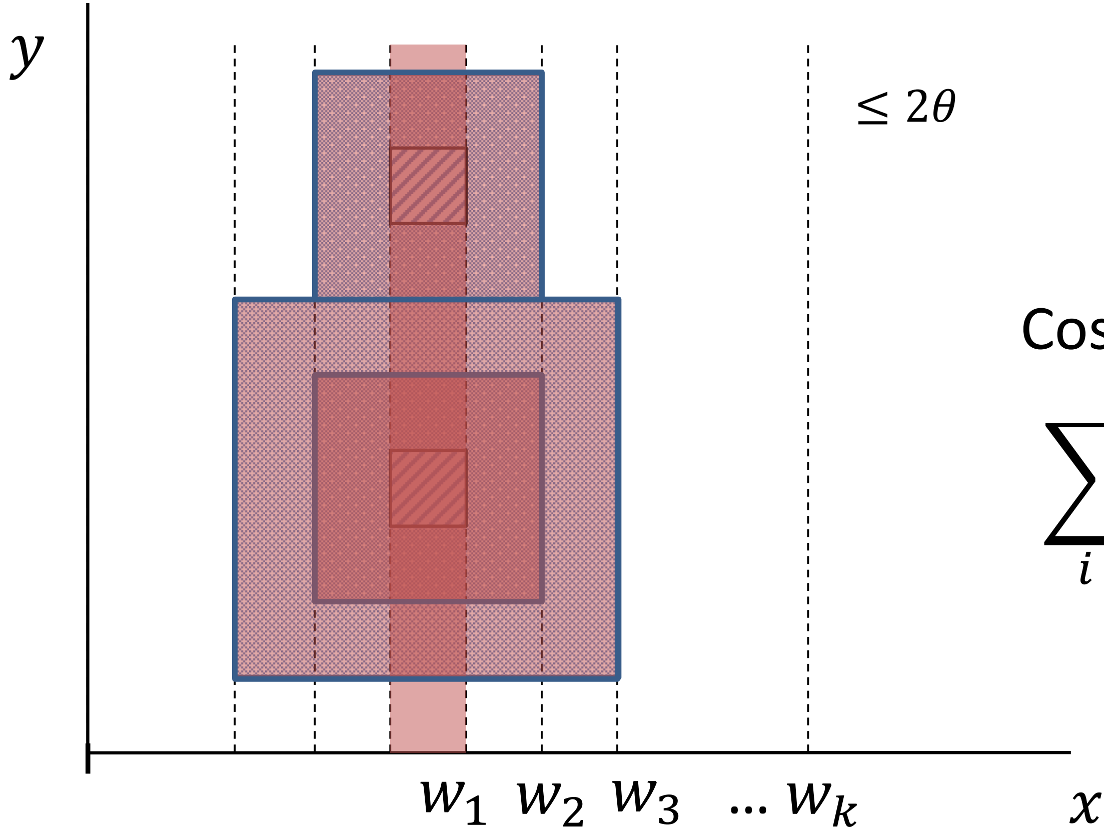
$n^{1+\epsilon}$ time algorithms

*Brakensiek-Rubinstein’20, *K.-Saks’20, Andoni-Nosatzki’20

- Build on Chakraborty-Das-Goldenberg-K.-Saks algorithm.
- Refine the algorithm by dual recursion.
- Data structure-like approach.

* Works on inputs of edit distance $\geq n^{1-\delta}$.

Multiple levels – sparse case



Cost:

$$\sum_i (w_i)^{1+\epsilon} \cdot d_{i-1}$$

$$\sqrt{n} = w_1 < w_2 \cdots < w_k < n$$

$$\sqrt{n} = d_0 > d_1 \cdots > d_k = 1$$

Questions

- $1+\epsilon$ approximation in time $O(n^{2-\epsilon})$?
- $O(1)$ approximation in time $n\log^{O(1)} n$?

Sub-linear algorithms

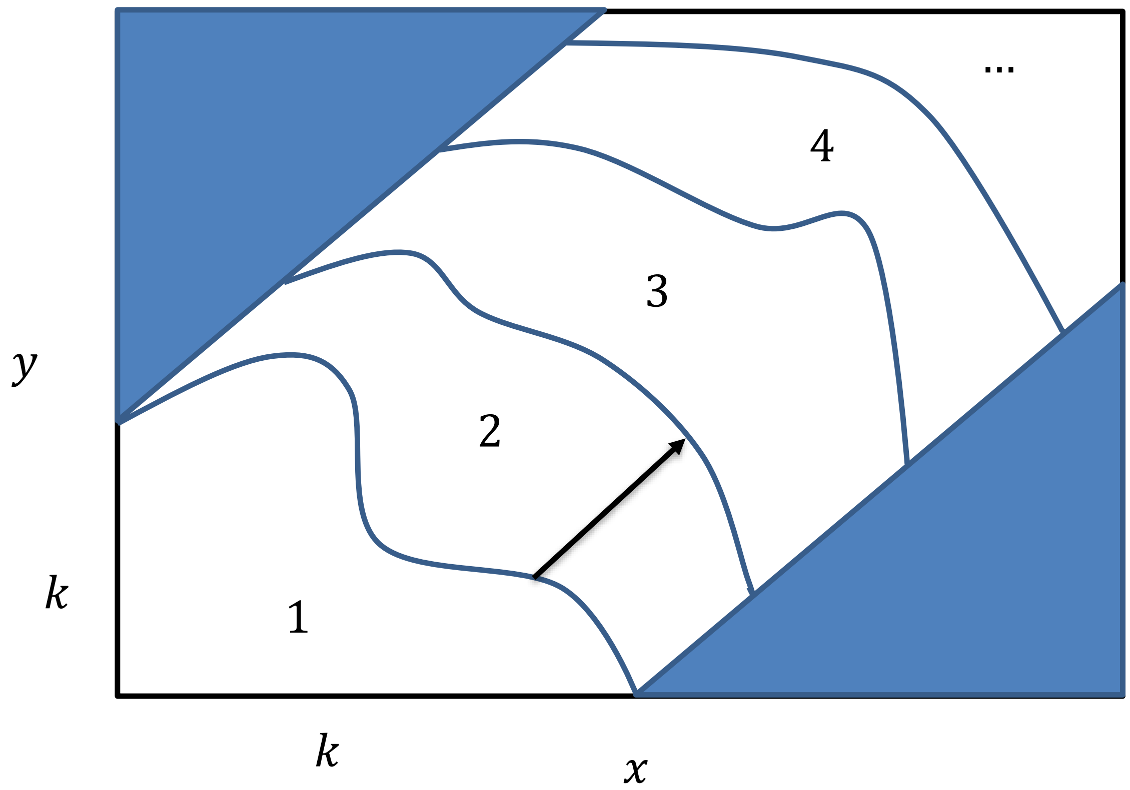
Input: $x, y \in \Sigma^n$, integer k.

Output:

YES	if	$\mathrm{ED}(x,y) \le k$.
NO	if	$\mathrm{ED}(x,y) > k^2$.

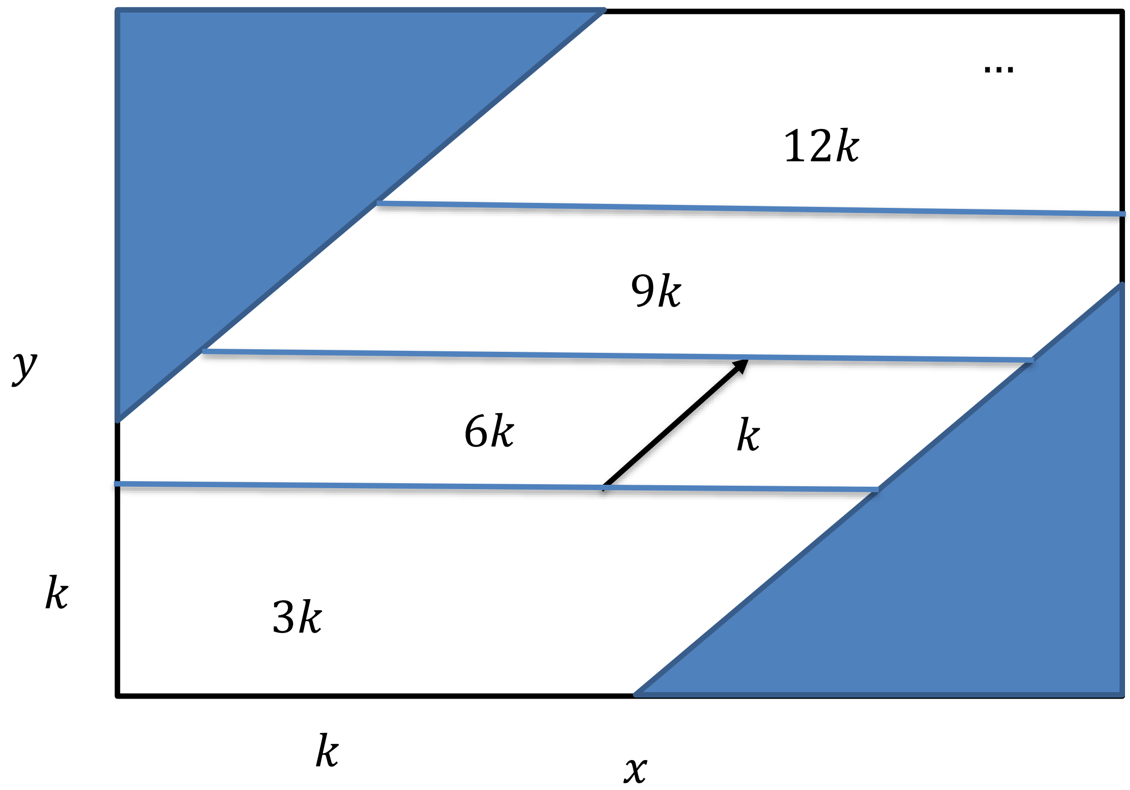
Batu-Ergun-Kilian-Magen-Raschodnikova-Rubinfeld-Sami'03	$O(n^{\alpha})$ vs $\Omega(n)$	$O(n^{\alpha/2})$
Andoni-Onak'09		$O(\frac{n^{2+o(1)}}{k^3})$
Goldenberg-Krauthgamer-Saha'19		$O(\frac{n}{k} + k^3)$
Kociumaka-Saha'20		$O(\frac{n}{k} + k^2)$
Brakensiek-Charikar-Rubinstein'20		$O(\frac{n}{\sqrt{k}})$

Edit distance waves



Landau-Myers-Schmidt'98: $O(n + k^2)$ time algorithm

Sublinear "waves": k vs k^2

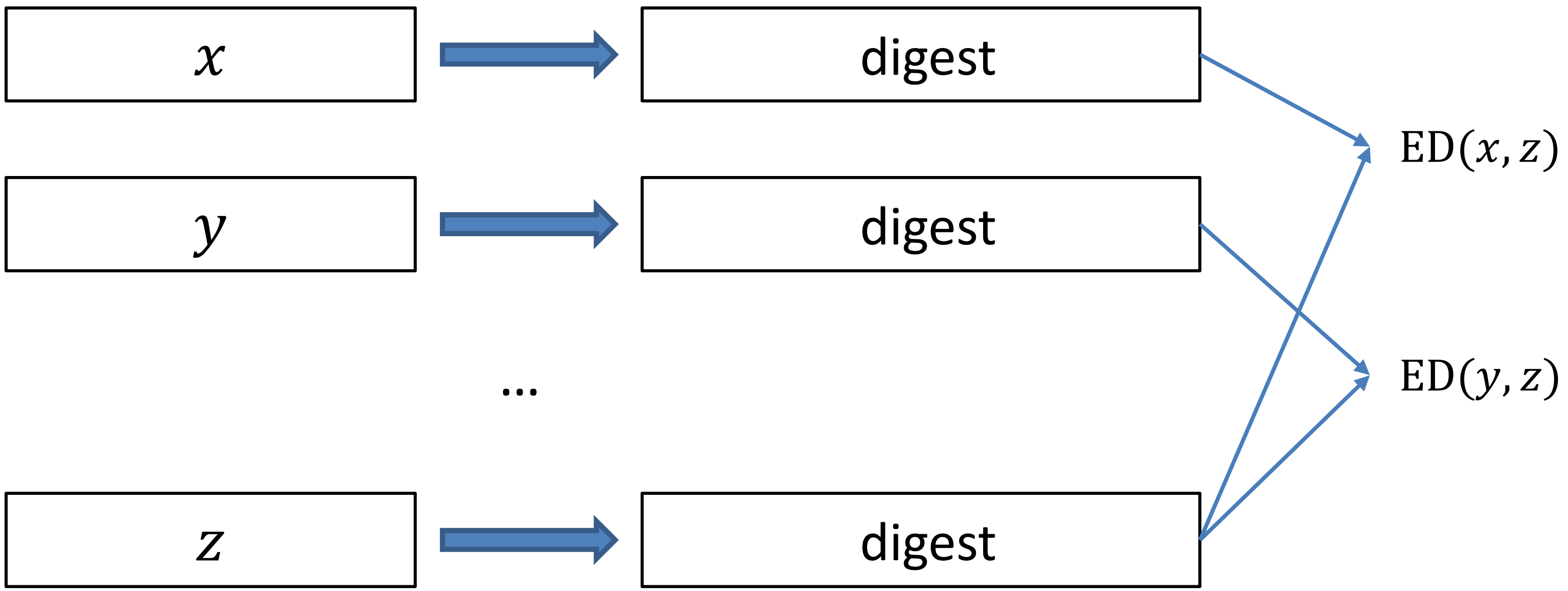


Goldenberg-Krauthgamer-Saha'19, Kociumaka-Saha'20, Brakensiek-Charikar-Rubinstein'20

Question

- $O(1)$ approximation in time $O(n/k)$?

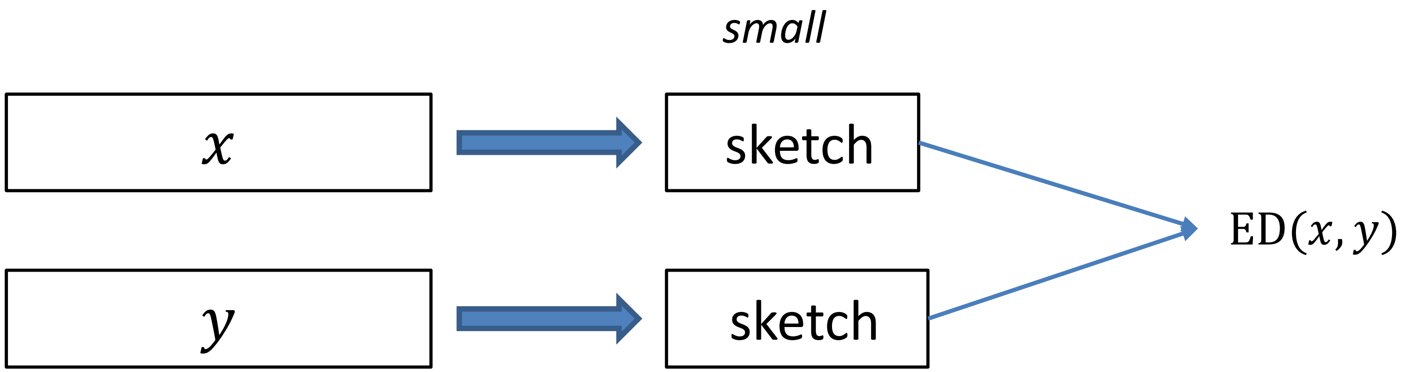
Preprocessing



Goldenberg-Rubinstein-Saha'20:

- Preprocessing time $O(n)$
- Exact edit computation $\tilde{O}(k^2)$

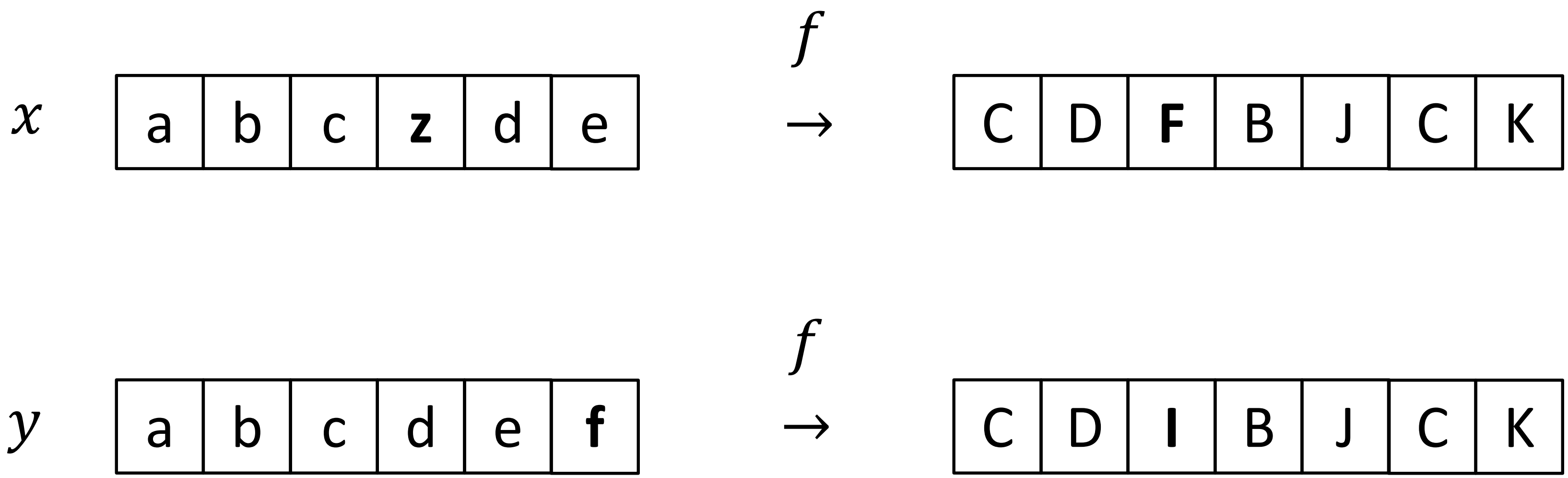
Sketching



Belazzougui, Zhang'16, Jin, Nelson, Wu'21 :

- Preprocessing time $O\left(n\, k^{O(1)}\right)$
- Exact edit computation $\tilde{O}\left(k^{O(1)}\right)$
- Sketch size $\tilde{O}\left(k^{3}\right)$

Embedding edit distance into Hamming distance



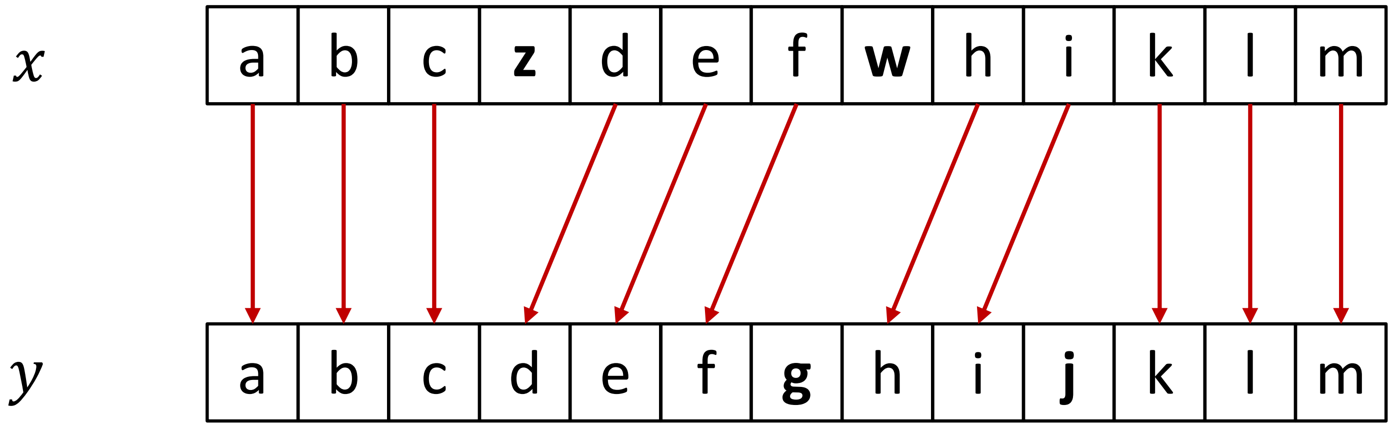
$$\mathrm{ED}(x, y) \approx \mathrm{Ham}(f(x), f(y))$$

Embedding edit into ℓ_1 distance

Embedding		distortion
Bar-Yossef-Jayram-Krauthgamer-Kumar'04		$O(n^{2/3})$
Ostrovsky-Rabani'07		$2^{O(\sqrt{\log n \ \log \ \log n})}$
Cormode-Muthukrishnan'02	(with moves)	$O(\log n \log^* n)$
Chakraborty-Goldenberg-K.'16	(random)	$O(k)$

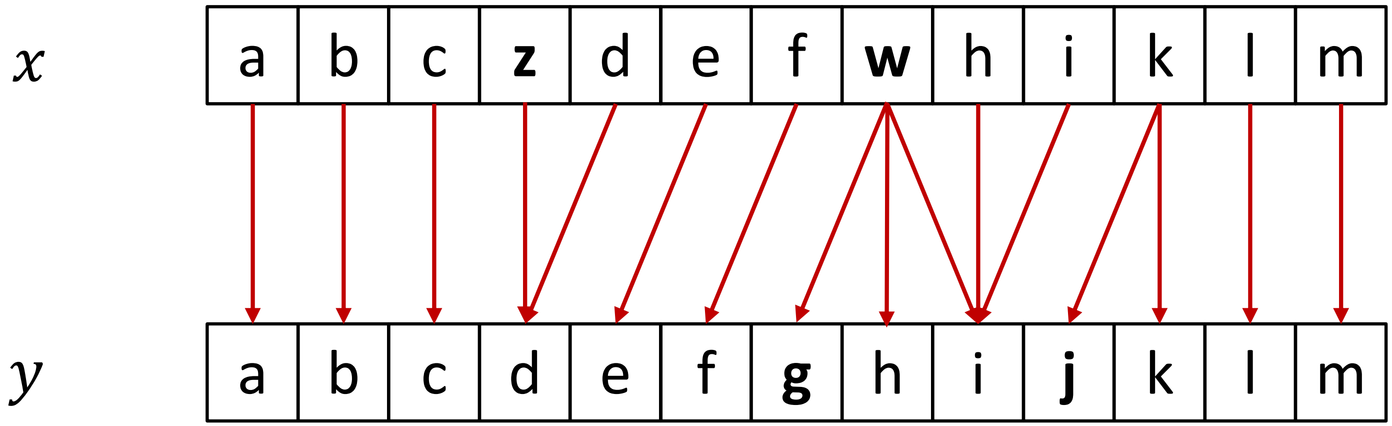
Lower bounds	
Andoni-Deza-Gupta-Indyk-Raskhodnikova'03	$\geq 3/2$
Knot-Naor'05	$\Omega((\log n)^{1/2-o(1)})$
Krauthgamer-Rabani'09	$\Omega(\log n)$

Optimal alignment



- Optimal alignment of size $\geq n - k$.

Large alignment



Saha'14:

- W.h.p. alignment of size $n - 20k^2$.
- Time $O(n)$.

Randomized embedding of *edit distance → Hamming distance*

Chakraborty-Goldenberg-K.'16:

$$f: \{0,1\}^n \times \{0,1\}^l \to \{0,1\}^{3n}$$

for any x and $y \in \{0,1\}^n$

$$\frac{1}{2}\mathrm{ED}(x,y) \;\leq\; \mathrm{Ham}(f(x,r), f(y,r)) \;\leq\; \mathrm{O}(\mathrm{ED}(x,y)^2)$$

with probability $\geq 2/3$ over a random choice of r.

Algorithm for embedding f

Chakraborty-Goldenberg-K.'16:

Input: $x \in \{0,1\}^n$ and random bits $r \in \{0,1\}^l$.

Interpret r as hash functions $h_1, h_2, \ldots h_{3n}: \{0,1\} \to \{0,1\}$.

$i := 1$

For $j := 1$ to $3n$ do

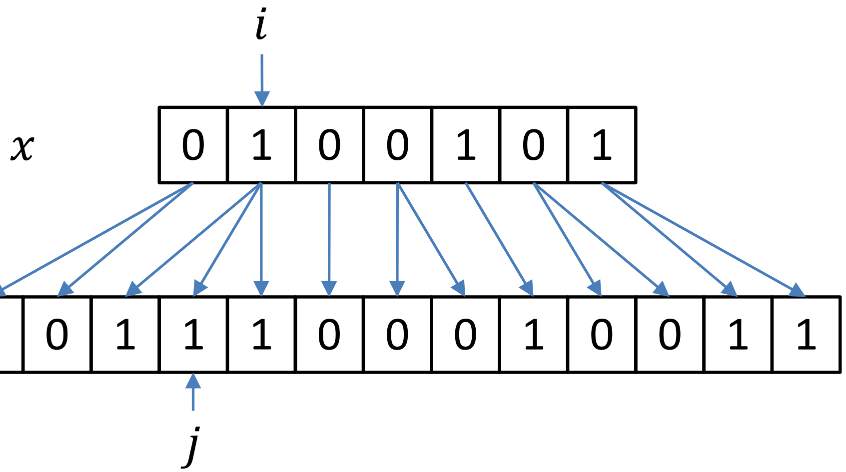
1. If $i \leq n$ then

   Output x_i

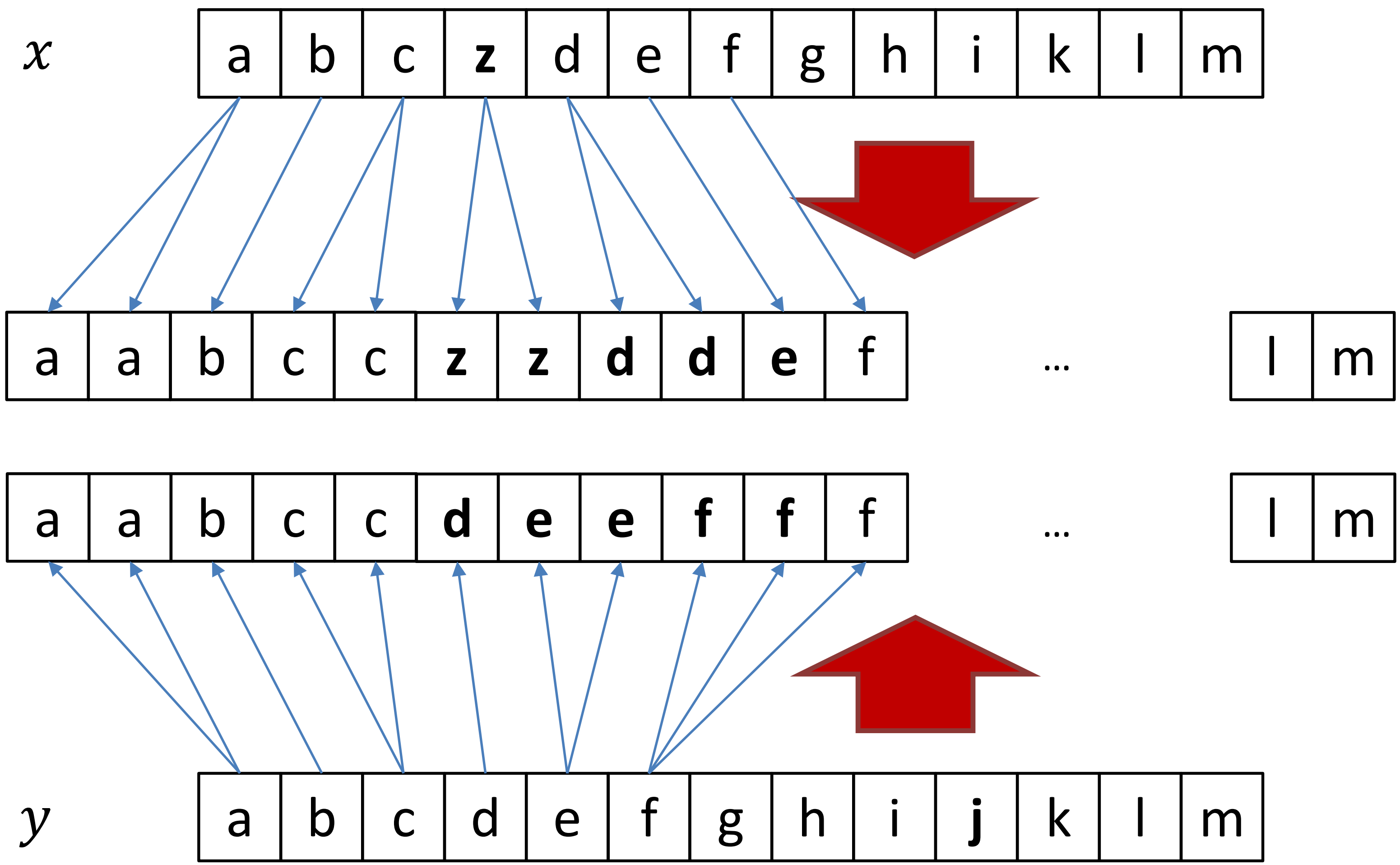
   $i := i + h_j(x_i)$

2. Else

   Output 0



Why it works



x
a
b
c
z
d
e
f
g
h
i
k
l
m
a
a
b
c
c
z
z
d
d
e
f
…
l
m
a
a
b
c
c
d
e
e
f
f
f
…
l
m
y
a
b
c
d
e
f
g
h
i
j
k
l
m

Synchronization

- The two pointers into x and y behave like a random walk on a line.

➢ With probability $\geq 2/3$ they synchronize in $\mathrm{O}(k^2)$ steps.

➢ But, the expected number of steps to synchronize is $\mathrm{O}(n)$.

Randomized embedding of *edit distance → Hamming distance*

Kociumaka-Saha'20

$$f\colon \{0,1\}^n \times \{0,1\}^l \to \{0,1\}^{6n/p}$$

computable in time $O(n/p)$ for chosen parameter $p < k$.

Allows to distinguish edit distance k vs pk^2.

Algorithm for embedding f

Kociumaka-Saha'20:

Input: $x \in \{0,1\}^n$, bits $r \in_R \{0,1\}^l$ and $s_1, \dots, s_{6n/p} \in_R \{1, \dots, p\}$.

Interpret r as hash functions $h_1, h_2, \dots h_{6n/p}: \{0,1\} \to \{0,1\}$.

$i \coloneqq 1$

For $j \coloneqq 1$ to $6n/p$ do

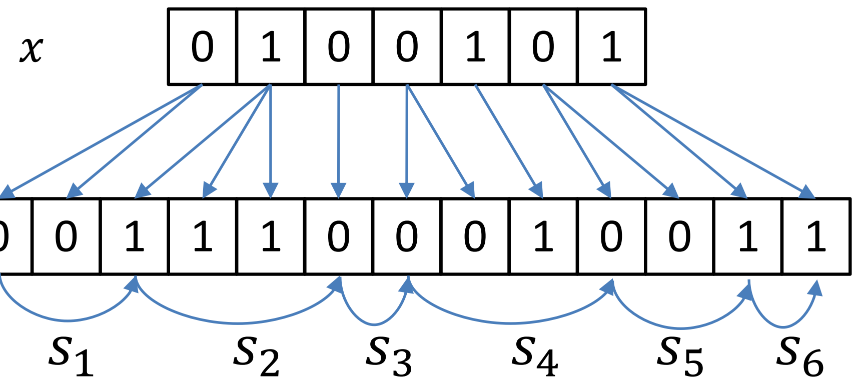
1. If $i \leq n$ then

   Output x_i

   $i \coloneqq i + s_j + h_j(x_i)$

2. Else

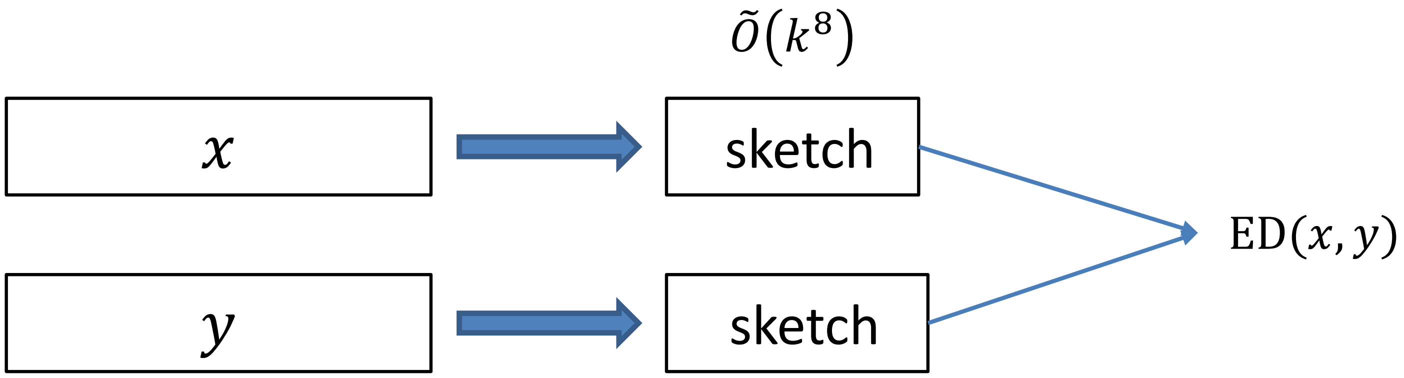
   Output 0



Question

- Embedding with better distortion?

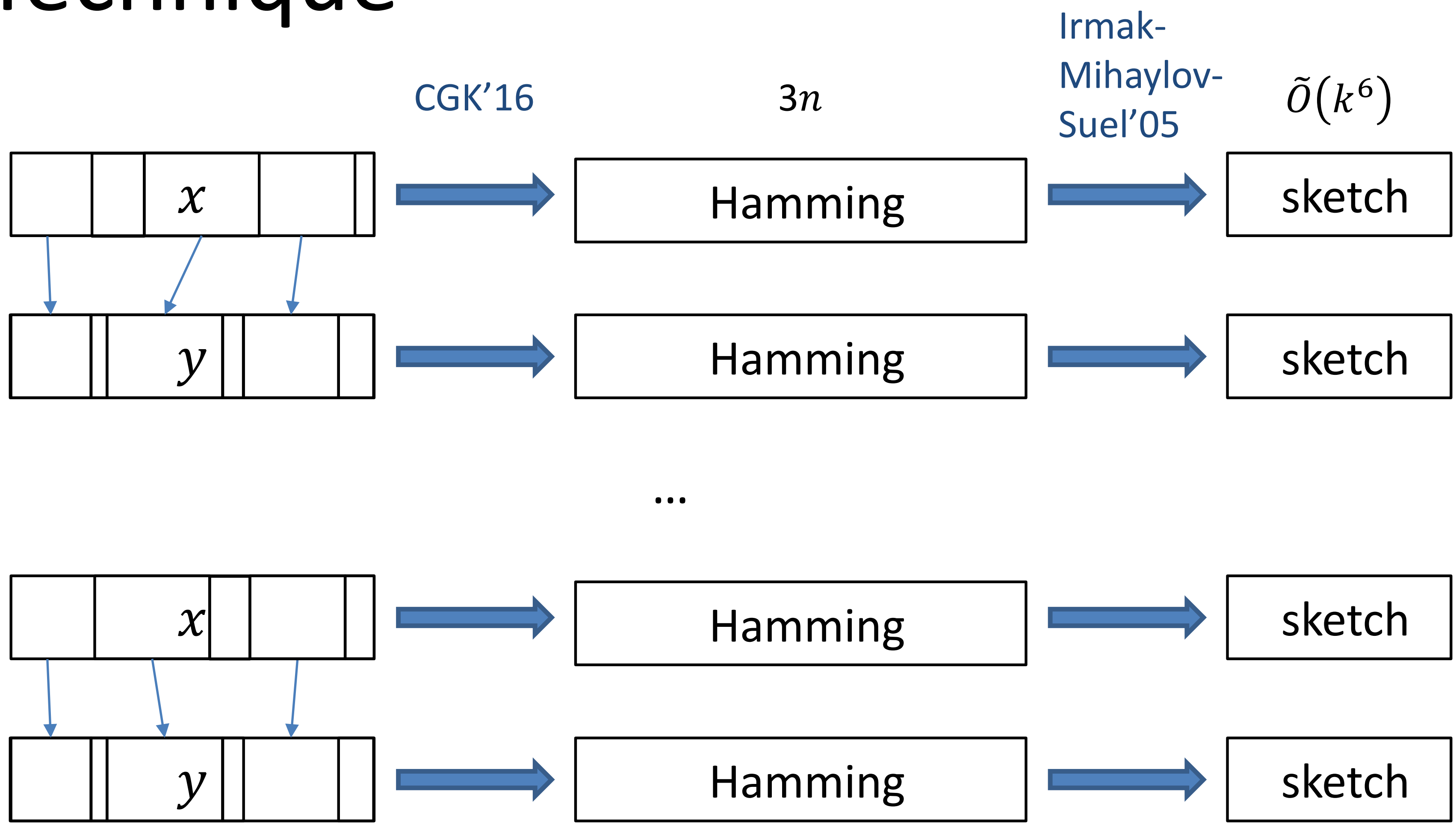
Sketching



Belazzougui-Zhang'16 :

- Preprocessing time $O\left(n\, k^{O(1)}\right)$
- Exact edit computation $\tilde{O}\left(k^{O(1)}\right)$
- Sketch size $\tilde{O}\left(k^8\right)$

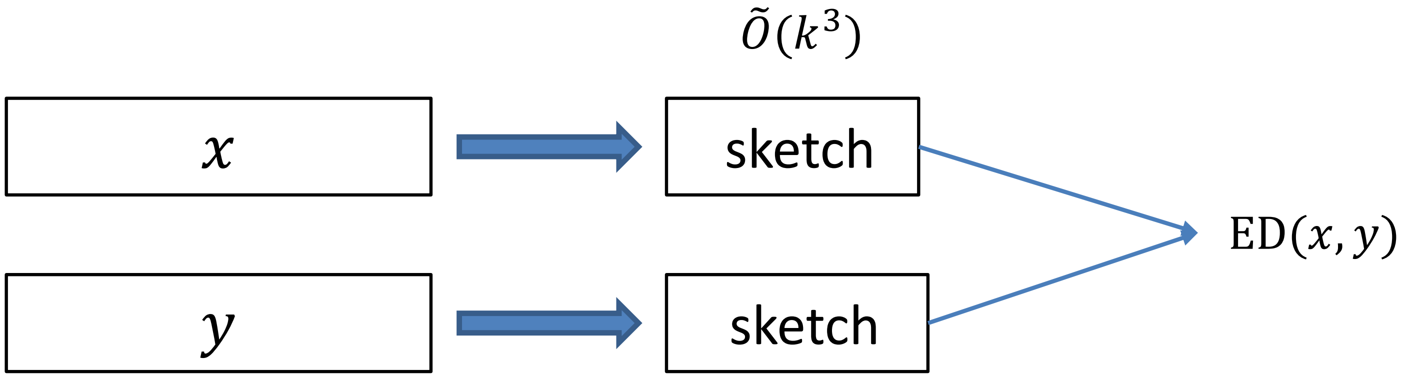
Technique



common edges → optimal matching

Belazzougui-Zhang'16

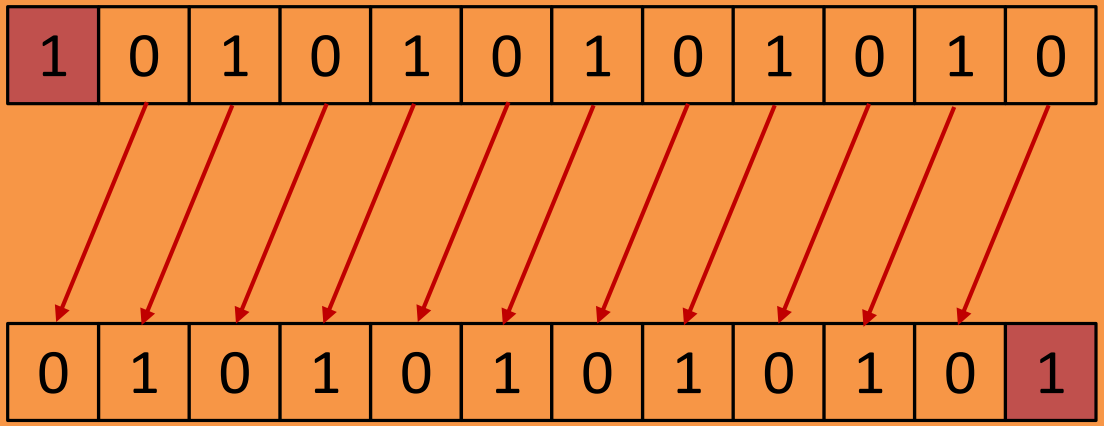
Sketching



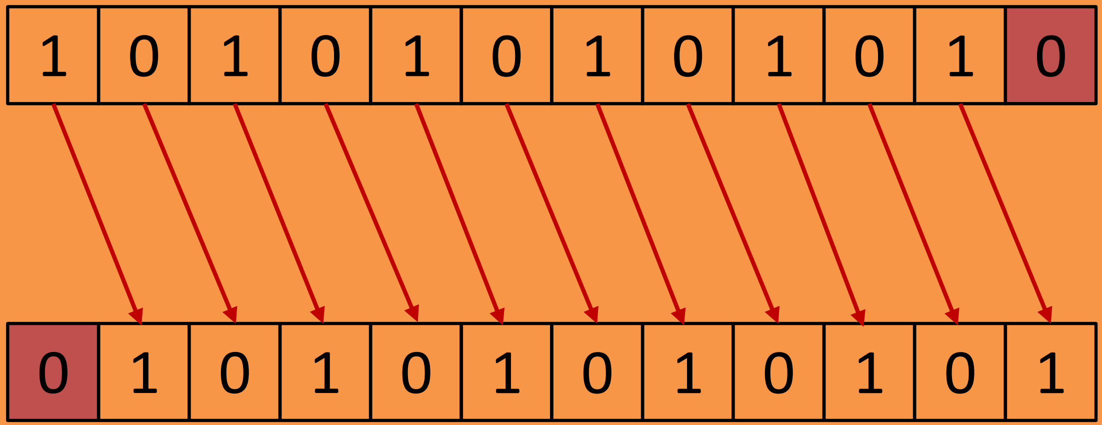
Jin-Nelson-Wu'21 :

- Preprocessing time $O\left(n\, k^{O(1)}\right)$
- Exact edit computation $\tilde{O}\left(k^{O(1)}\right)$
- Sketch size $\tilde{O}\left(k^3\right)$

Question



Question



Question

1	0	1	0	1	0	1	0	1	0	1	0

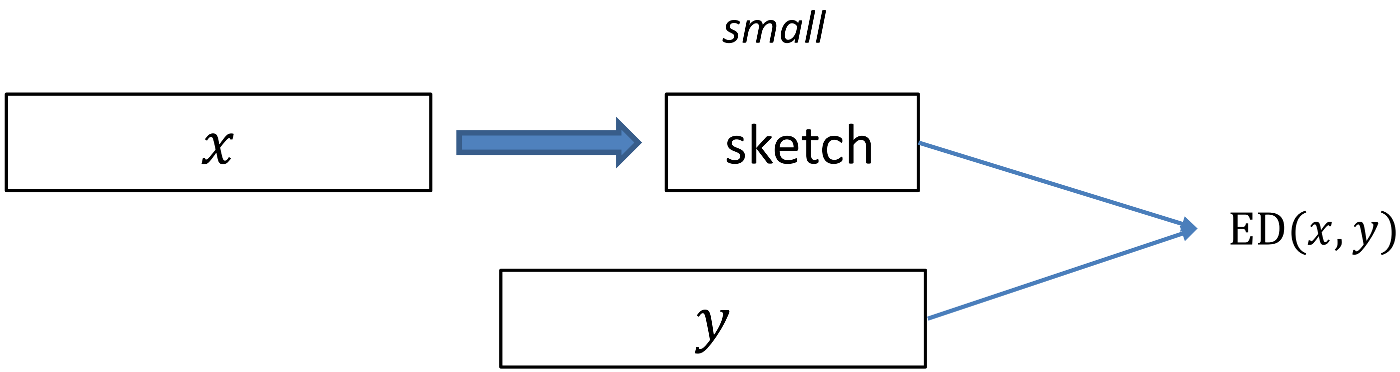
0	1	0	1	0	1	0	1	0	1	0	1

- The total number of unmatched symbols at most $O(k)$?

Questions

- Preprocessing x and y into approximate sketches of size $\log^{O(1)}(n+k)$?
- Preprocessing x and y so that query in time $\log^{O(1)}(n+k)$?

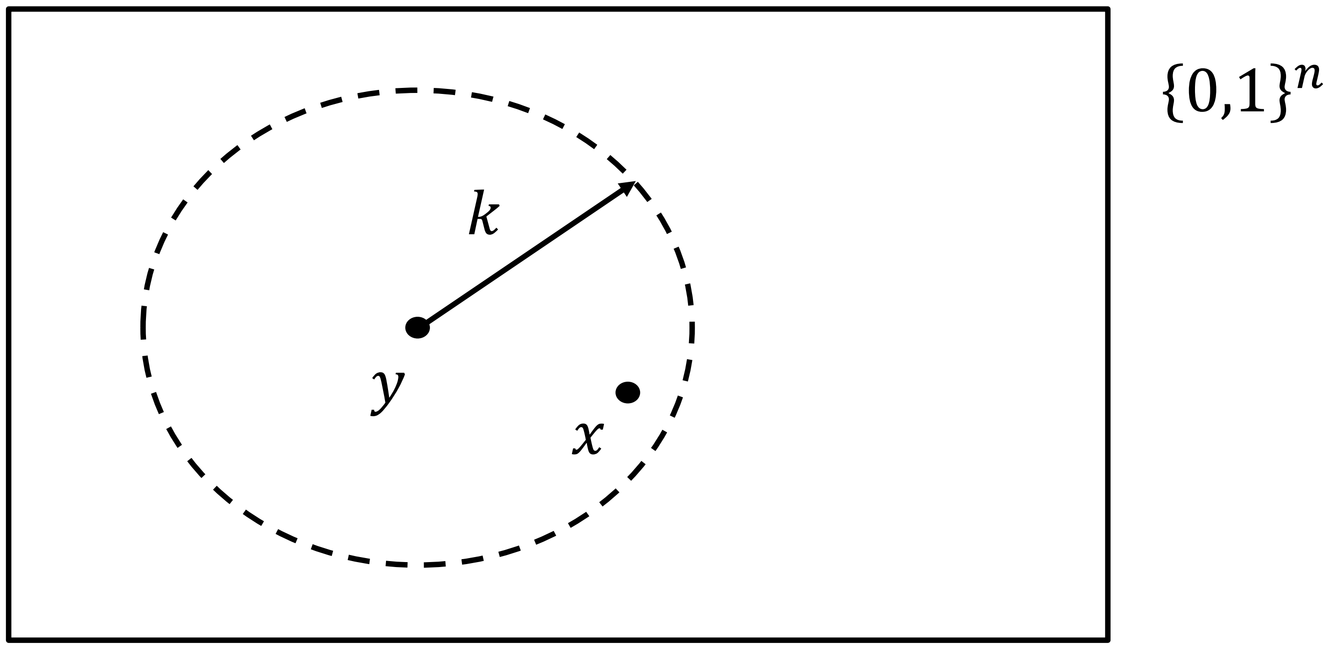
Document exchange problem



Cheng-Jin-Li-Wu'18, Haeupler'19:

	sketch size	
deterministic	$k \log^2 n/k$	
randomized	$k \log n/k$	… optimal

Document exchange problem



$$Ball(y,k) = \{z \in \{0,1\}^n, ED(y,z) \leq k\}$$

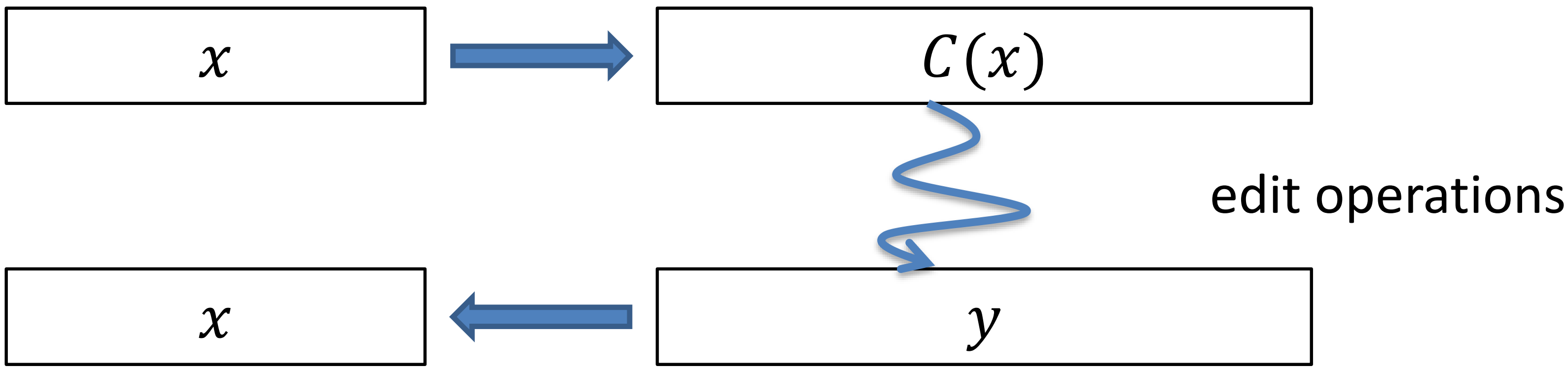
$$|Ball(y,k)| \approx 2^{3k} \cdot \binom{n}{2k}$$

$$\log|Ball(y,k)| \approx k \log n/k$$

Document exchange problem

		sketch size	time
Orlitsky'91	(det.)	$k \log n/k$	$n^{O(k)}$
Irmak-Mihaylov-Suel'05		$k \log(n/k) \log n$	$\tilde{O}(n)$
Jowhari'12		$k \log^2 n \log^* n$	$\tilde{O}(n)$
Belazzougui'12	(det.)	$k^2 + k \log^2 n$	$\tilde{O}(n)$
Chakraborty-Goldenberg-K.'16		$k^2 \log n$	$\tilde{O}(n)$
Belazzougui-Zhang'16		$k(\log^2 k + \log n)$	$\tilde{O}(n)$
Cheng-Jin-Li-Wu'18,			
Haeupler'19	(det.)	$k \ \log^2 n/k$	$\tilde{O}(n)$
		$k \ \log n/k$	$\tilde{O}(n)$

Error correcting codes



Cheng-Jin-Li-Wu'18

redundancy $O(k \log n)$

Haeupler'19:

redundancy $O(k \; \log^2 n/k)$ … systematic

END

Bonus Question

- Can you reduce the low-regime edit distance into high-regime edit distance?