Fast Protocols for Edit Distance Through Locally Consistent Parsing

Hossein Jowhari

MADALGO
Aarhus University

Definitions

Edit Distance

$x, y \in \Sigma^n$

$ed(x,y)$: Minimum number of character substitutions, insertions, deletions for Converting x to y

Hamming

H(x,y):
Minimum number of
Substitutions only

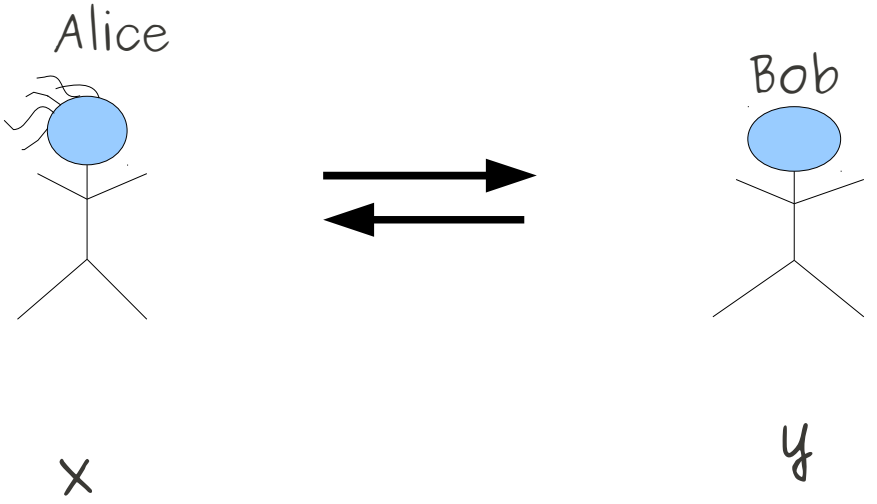
Ulam

Edit distance
Over Non-repetitive
Strings (permutations)

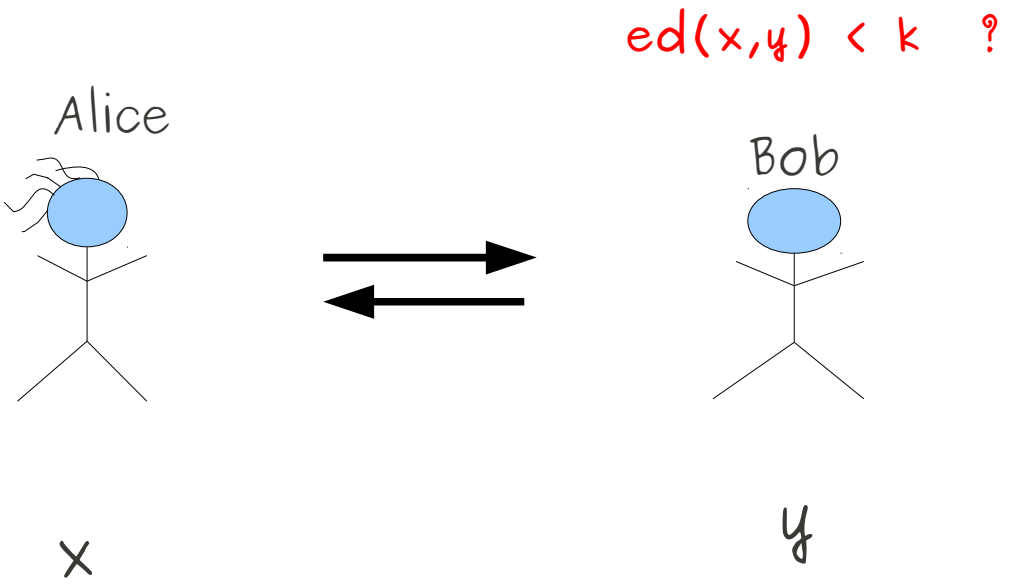
Definitions : two-player model

Definitions : two-player model

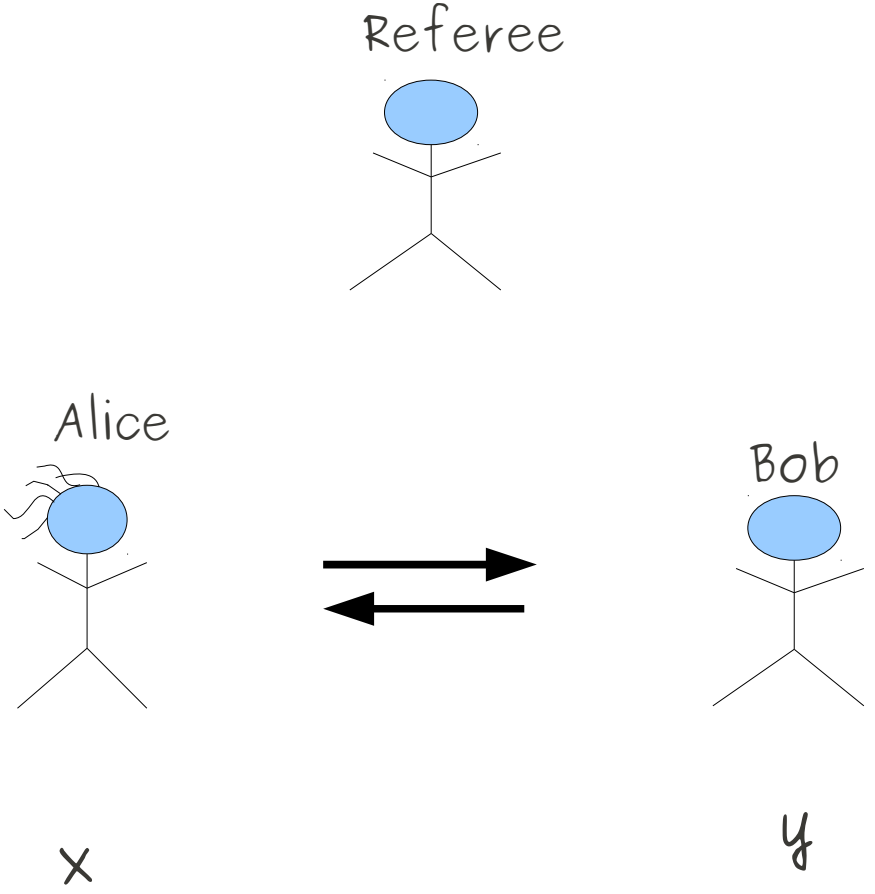
Definitions : two-player model



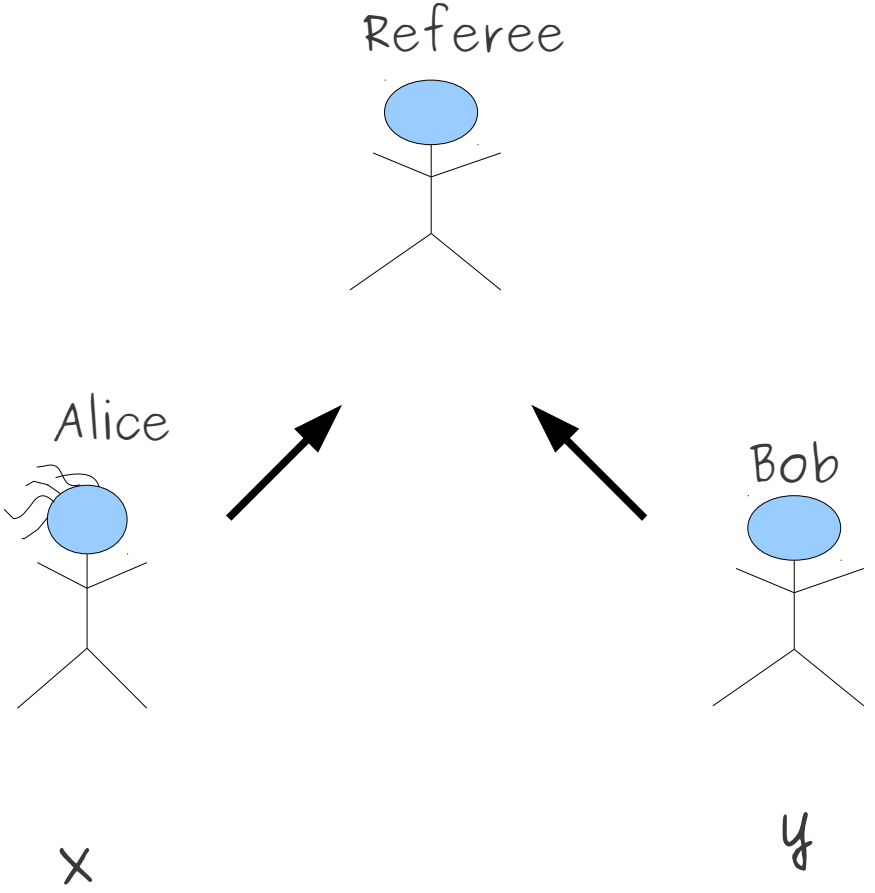
Definitions : two-player model



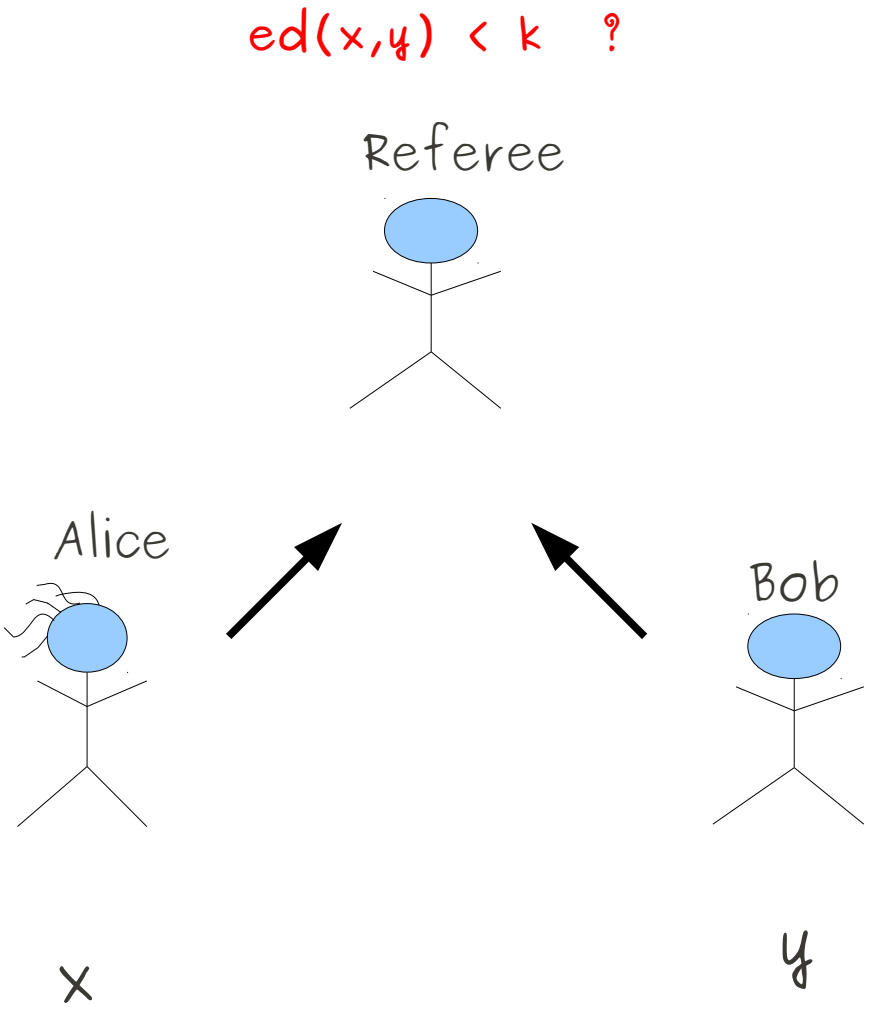
Definitions : simultaneous model



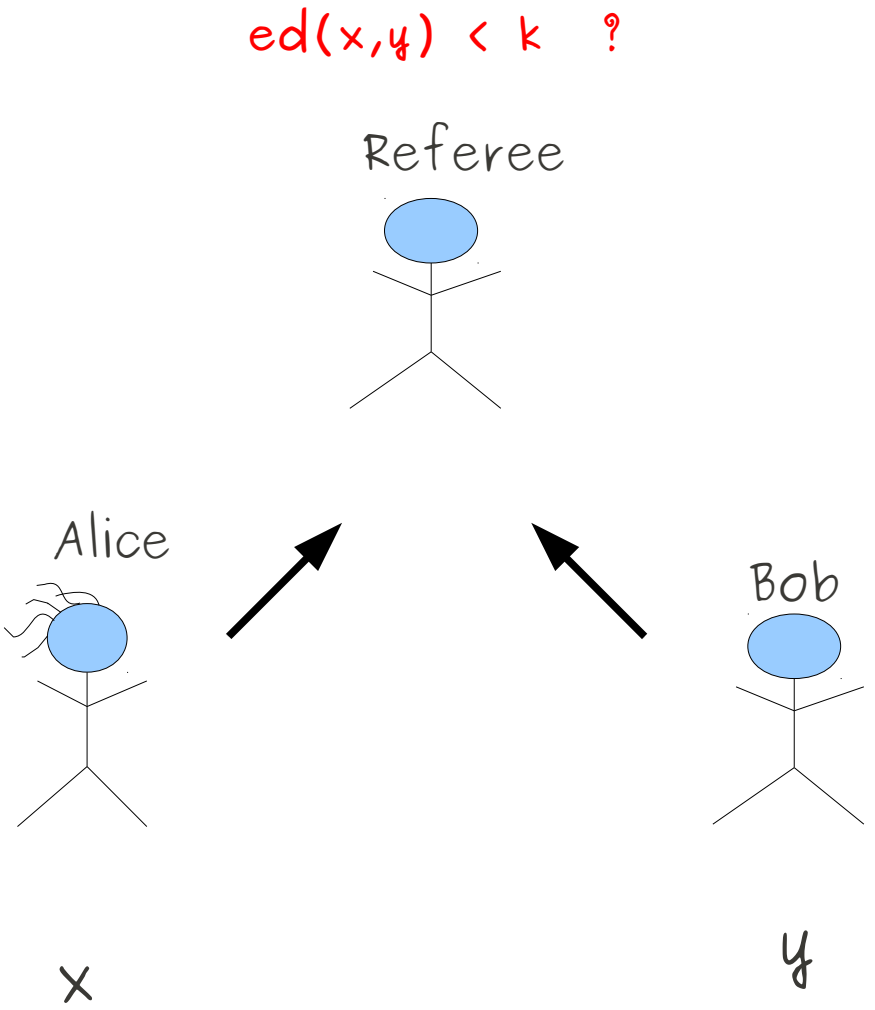
Definitions : simultaneous model



Definitions : simultaneous model



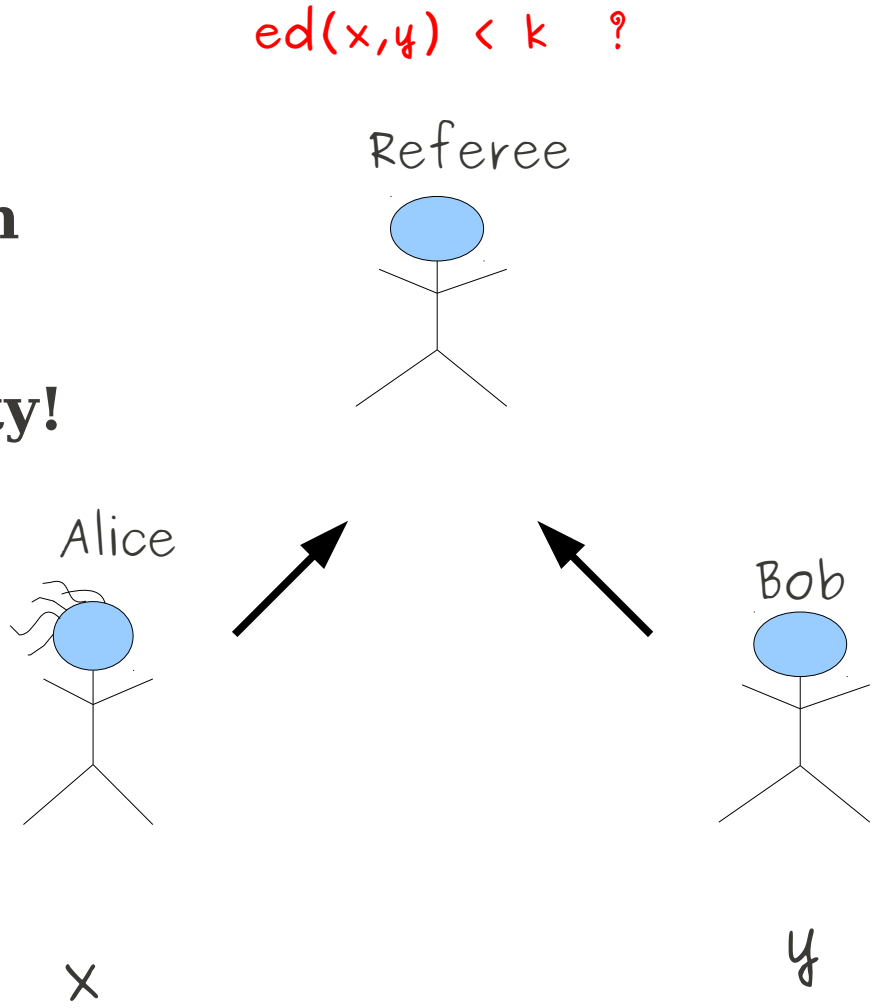
Definitions : simultaneous model



Definitions : simultaneous model

Communication complexity

Time complexity!



Applications

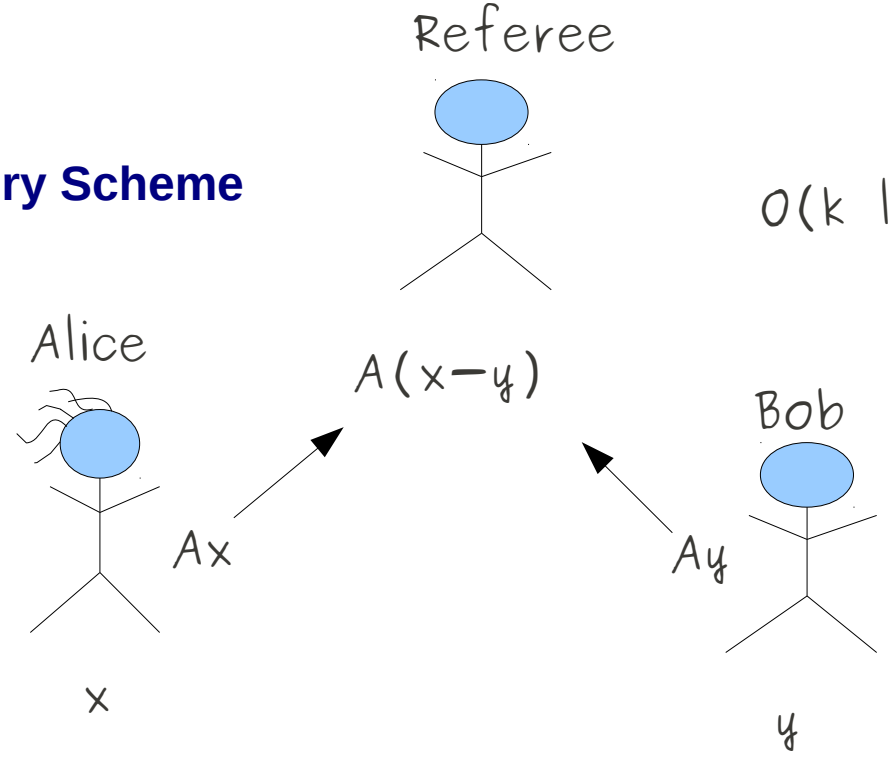
Transmission of data over errornous channels

Remote File Synchronization

Summarization of rankings for comparisons

Previous Results: Hamming

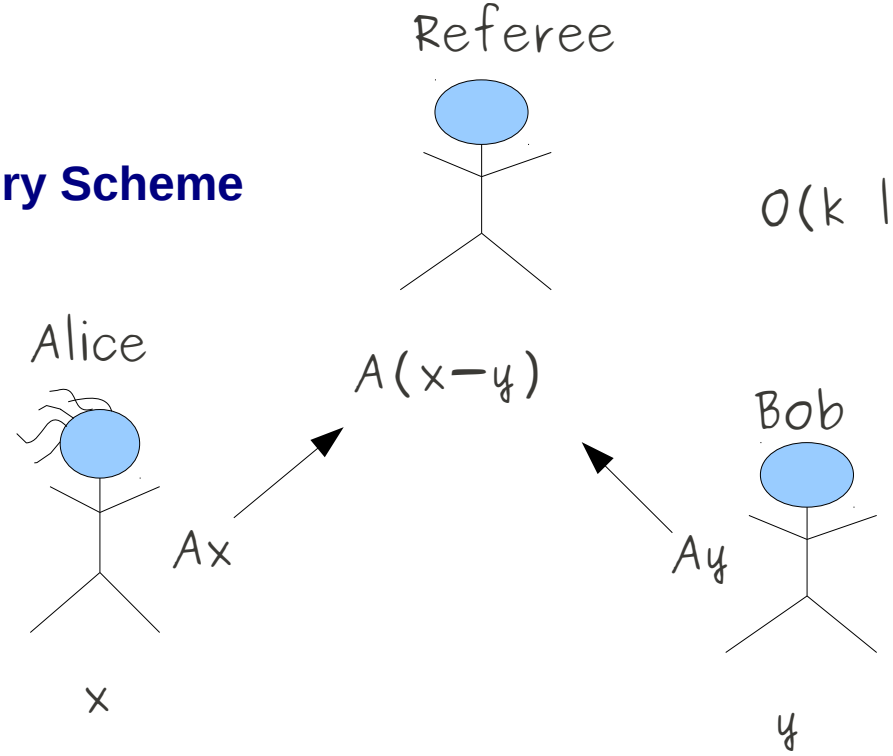
Sparse Recovery Scheme



O(k logn log(n/k)) protocol

Previous Results: Hamming

Sparse Recovery Scheme



O(k logn log(n/k)) protocol

There are more efficient (non-linear) methods: Lipsky-Porat CPM 2007

O(klog n) bit protocol (also streaming algorithm)

it outputs x−y

Time complexity: O(slog n) for s−sparse vectors

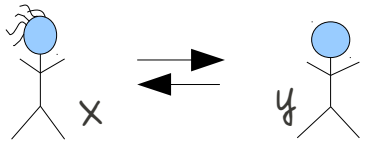
Previous Results

O(klog n) bit 1-way protocol

Time complexity: $n^{O(k)}$ Orlitsky FOCS 91

Edit distance

k vs k+1



O(klog k log(n/k)) bit O(log n)-round protocol

Time complexity: Poly(n)

Cormode, Paterson, Sahinalp, and Vishkin. SODA 2000

O(klog k log(n/k)) bit 1-way protocol

Time complexity: Poly(n)

Irmak, Mihaylov, and TSuel. INFOCOM 2005

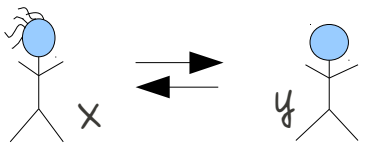
New Results

Edit distance

k vs k+1

$\tilde{O}(k\log^2 n)$ bit 1-way protocol

Time complexity: $\tilde{O}(n\log n + k^2 \log^2 n)$

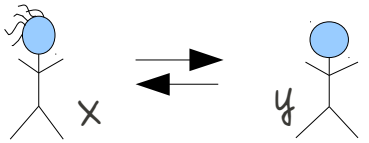


Ulam distance

k vs k+1

$O(k\log n)$ bit 1-way protocol

Time complexity: $O(n\log n)$



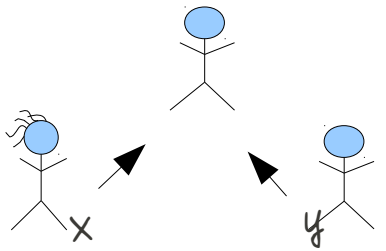
Ulam distance

k vs k+1

$\tilde{O}(k\log^2 n)$ bit protocol

Alice and Bob's time complexity: $O(n\log n)$

Referee's time complexity: $\tilde{O}(k^2 \log^2 n)$



General Framework : Reduction to Hamming

$$f : \Sigma^n \rightarrow \{0,1\}^{poly(n)}$$

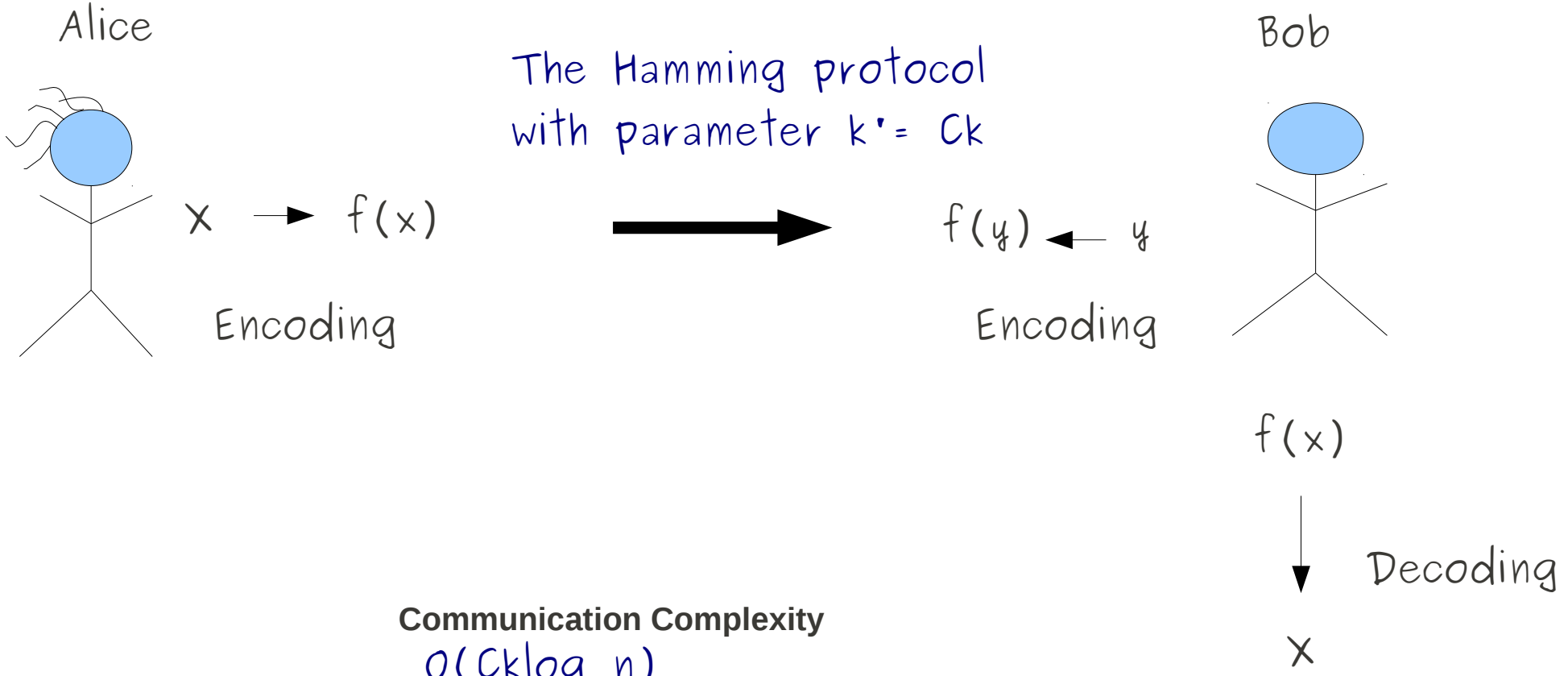
$f(x)$ is efficiently computable

$f(x)$ and $f(y)$ are **distinct** with high probability

$H(f(x), f(y)) < C \cdot ed(x,y)$ for small C, **Small Expansion**

There exists efficient **decoding procedure** R
where given $f(x)$, R returns x whp

General Framework II



Communication Complexity

$O(Ck \log n)$

Time Complexity

Encoding time $f(x)$

Decoding time $R(f(x))$

Mapping f for Ulam

An injective mapping $f : \Sigma^n \rightarrow \{0,1\}^{poly(n)}$ such that

if $ed(x,y) = 1$ then $H(f(x),f(y))$ is small

Mapping f for Ulam

An injective mapping $f : \Sigma^n \rightarrow \{0,1\}^{poly(n)}$ such that

if $ed(x,y) = 1$ then $H(f(x),f(y))$ is small

$$f : S^n \rightarrow \{0,1\}^{n^2}$$

$f(x)_{a,b} = 1$ iff a and b are adjacent in x

If $ed(x,y) = 1$ then $H(f(x),f(y)) \leq 6$

Decoding is trivial

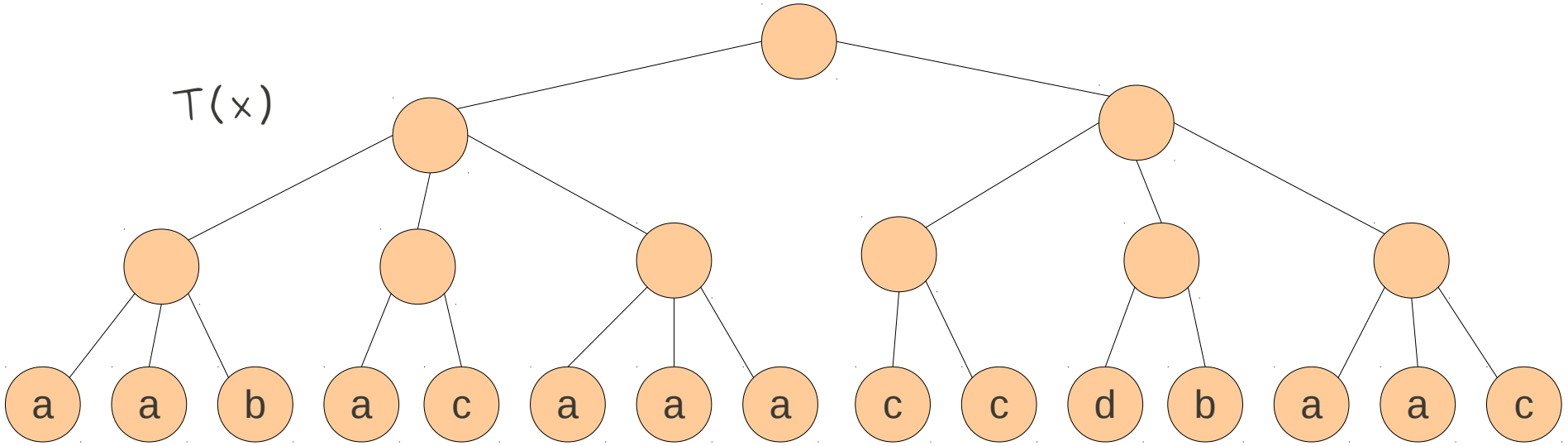
Mapping f for Edit distance

An injective mapping $f : \Sigma^n \rightarrow \{0,1\}^{poly(n)}$ such that

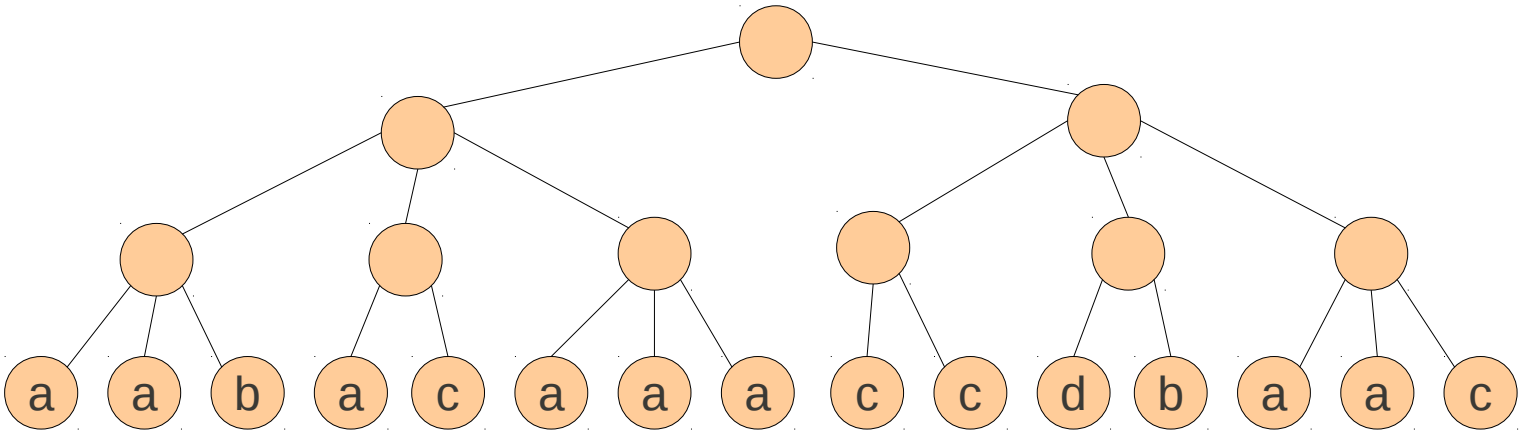
if $ed(x,y) = 1$ then $H(f(x),f(y))$ is small

Locally Consistent Parsing CPSV 2000, CM 2002

A hierarichal partitioning of the string x into substrings



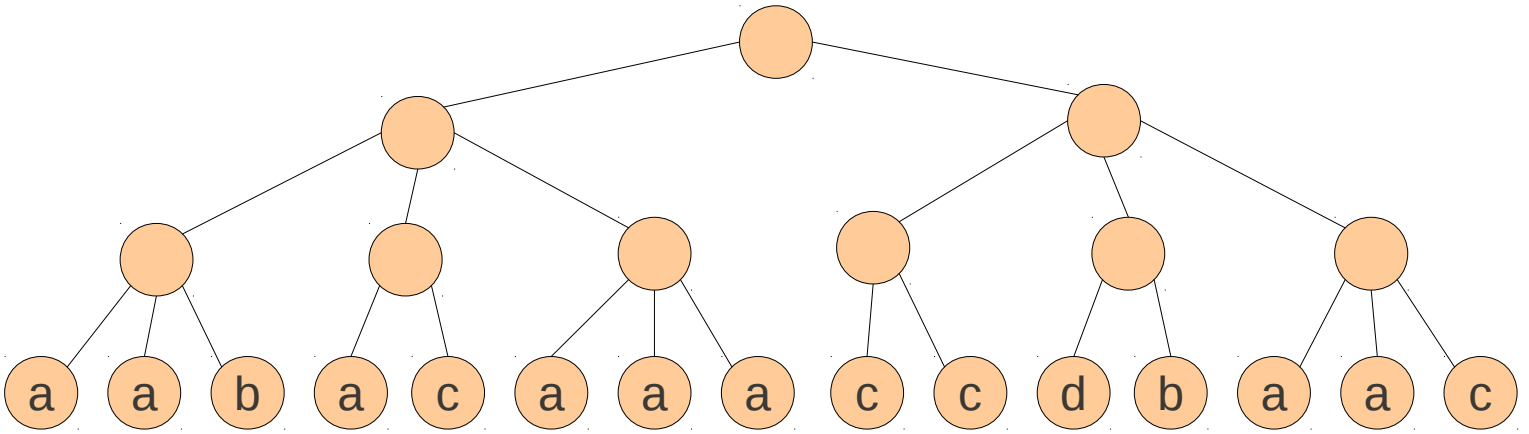
Encoding: Locally Consistent Parsing



[CM 2002] Based on LCP, there is a mapping $f:\Sigma^n \to [m]^L$

$ed(x,y)=1$ then $|f(x)-f(y)|_1 = O(\log n \log^* n)$

Encoding: Locally Consistent Parsing



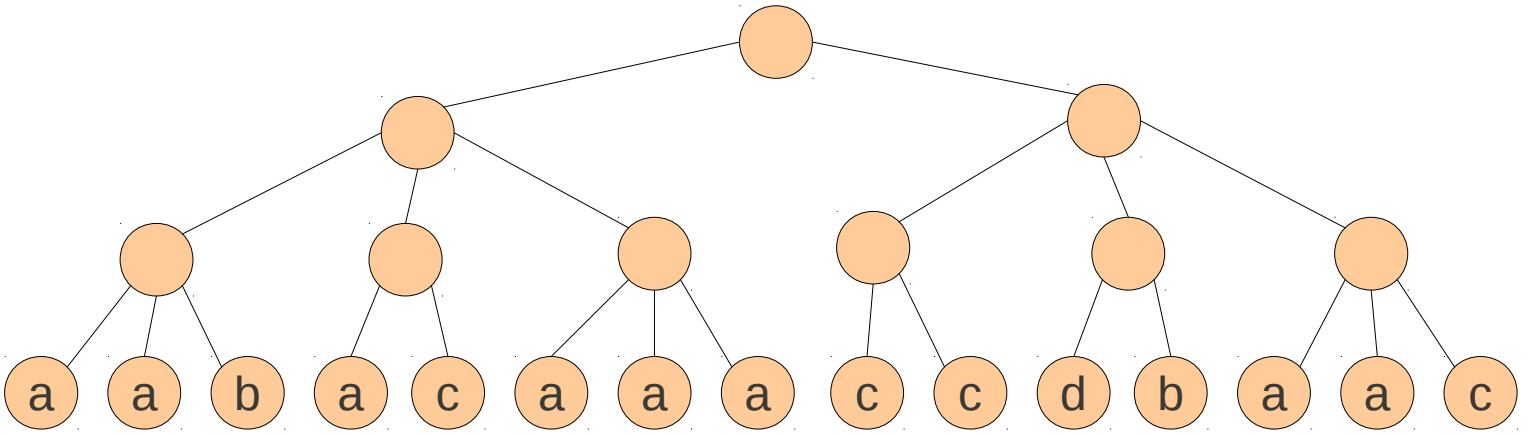
[CM 2002] Based on LCP, there is a mapping $f:\Sigma^n \to [m]^L$

$ed(x,y)=1$ then $|f(x)-f(y)|_1 = O(\log n \log^* n)$

L is exponential in n

For every possible substring there is an associated coordinate

Encoding: Locally Consistent Parsing



[CM 2002] Based on LCP, there is a mapping $f:\Sigma^n \to [m]^L$

$ed(x,y)=1$ then $|f(x)-f(y)|_1 = O(\log n \log^* n)$

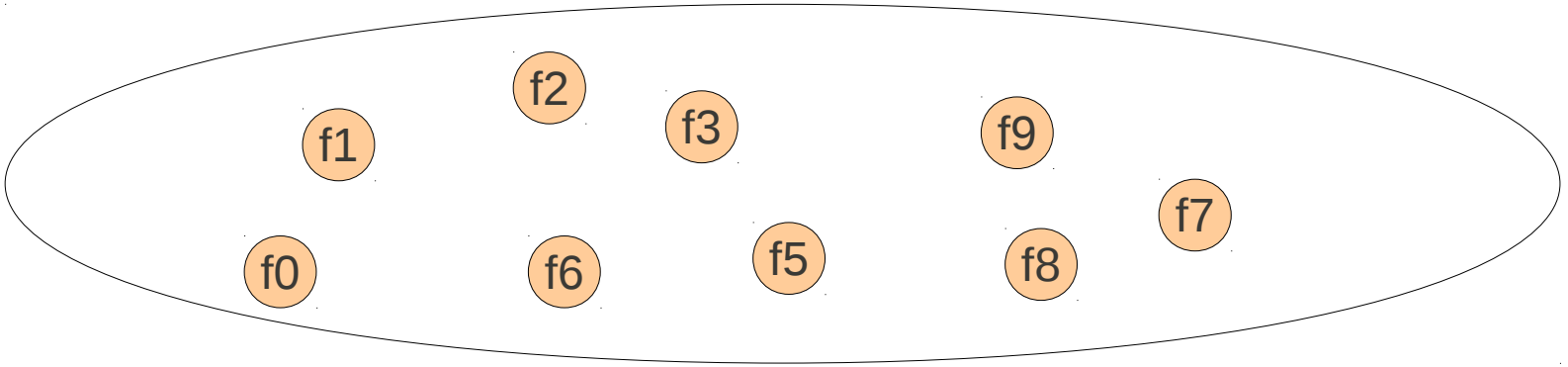
L is exponential in n

For every possible substring there is an associated coordinate

Since f is 2n-sparse we can use Rabin-Karp Fingerprinting to reduce number of dimensions to poly(n)

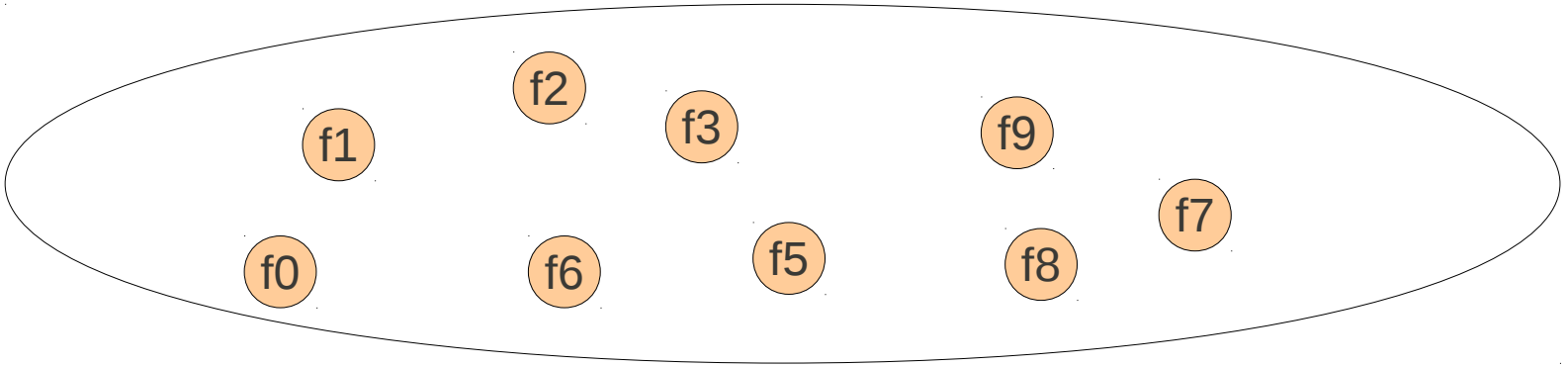
Decoding Procedure

After Bob computes f(x), he has a collection of fingerprints

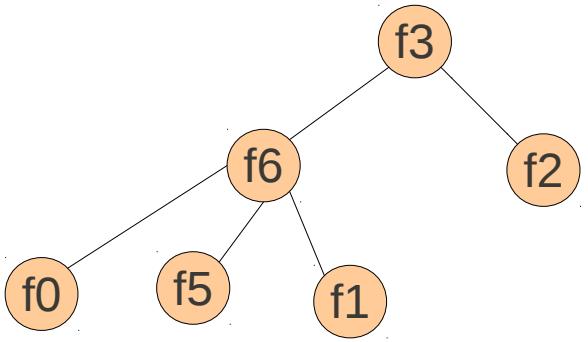


Decoding Procedure

After Bob computes f(x), he has a collection of fingerprints

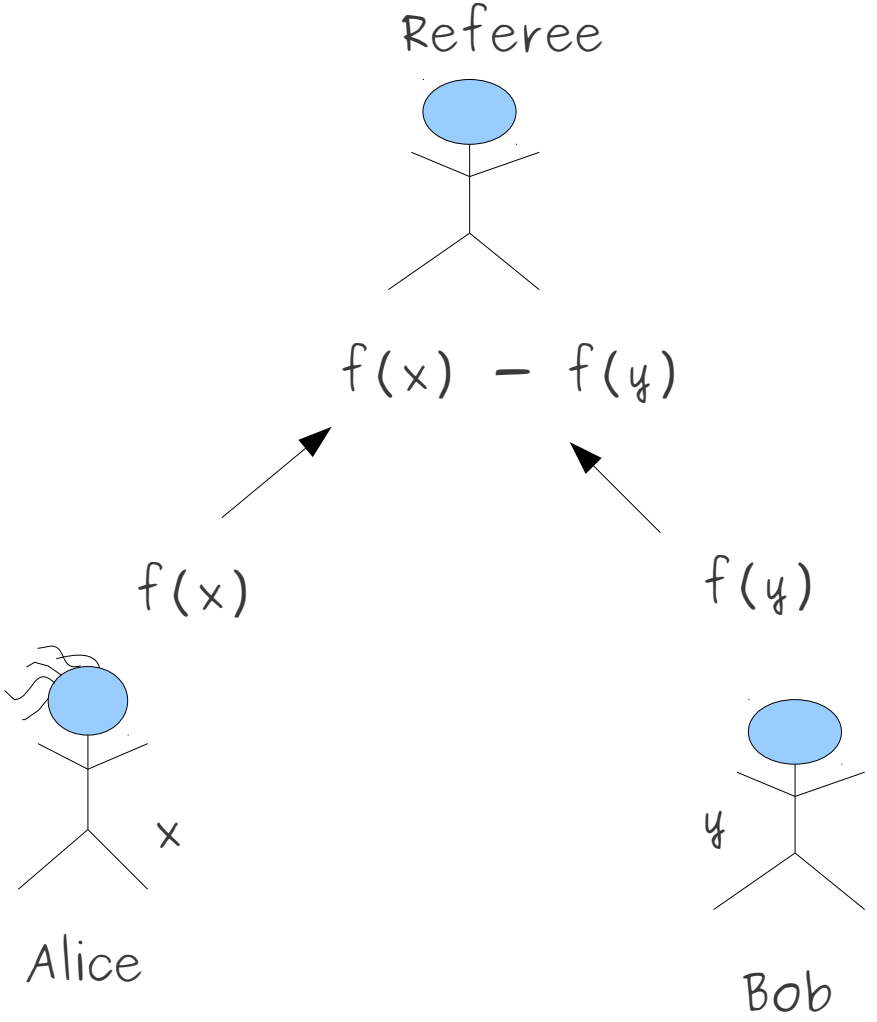


Bob reconstructs the tree T(x) in a top-down manner using the information from fingerprints

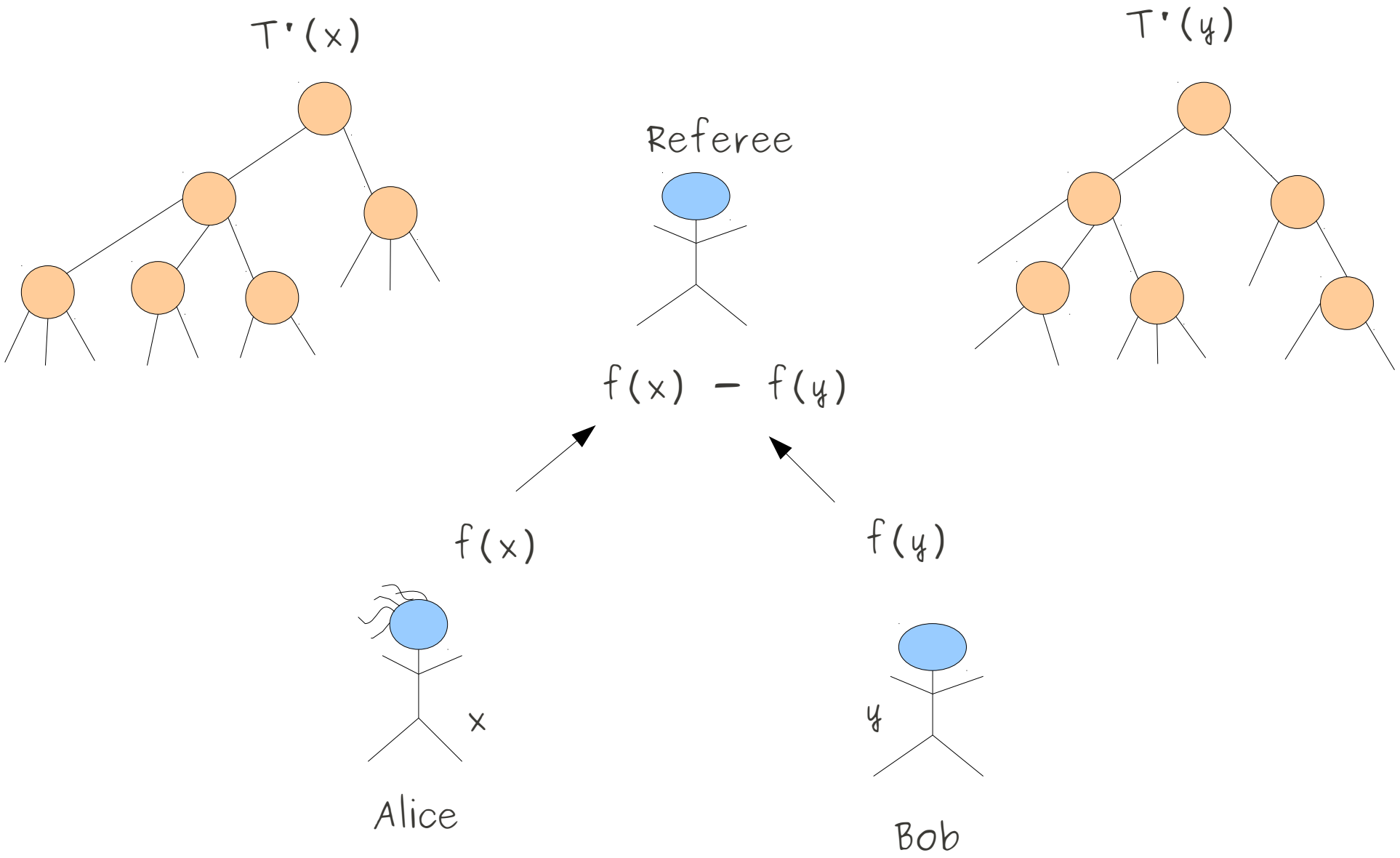


Extracting characters from the fingerprints of the leaves is straightforward.

A simultaneous Protocol for Ulam I

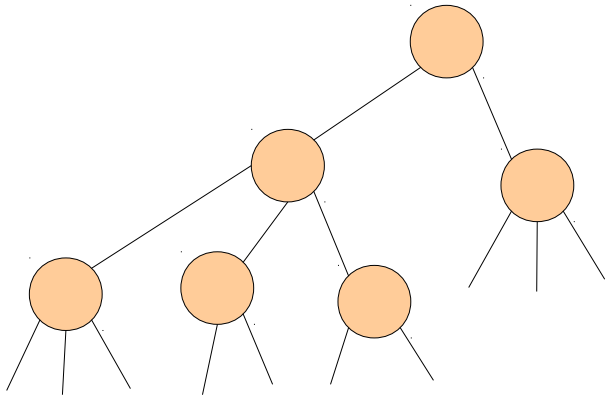


A simultaneous Protocol for Ulam I

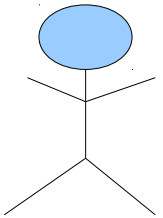


A simultaneous Protocol for Ulam II

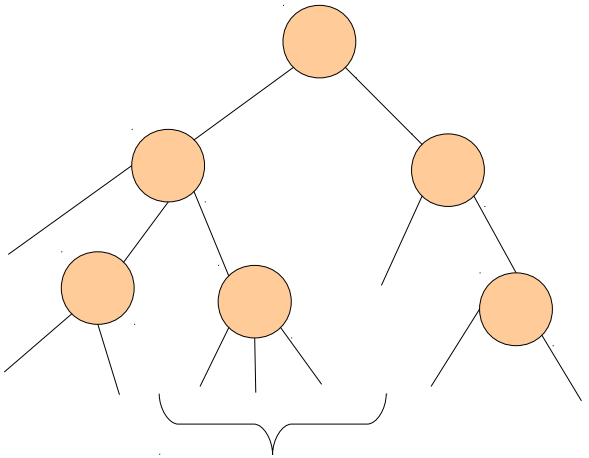
T'(x)



Referee

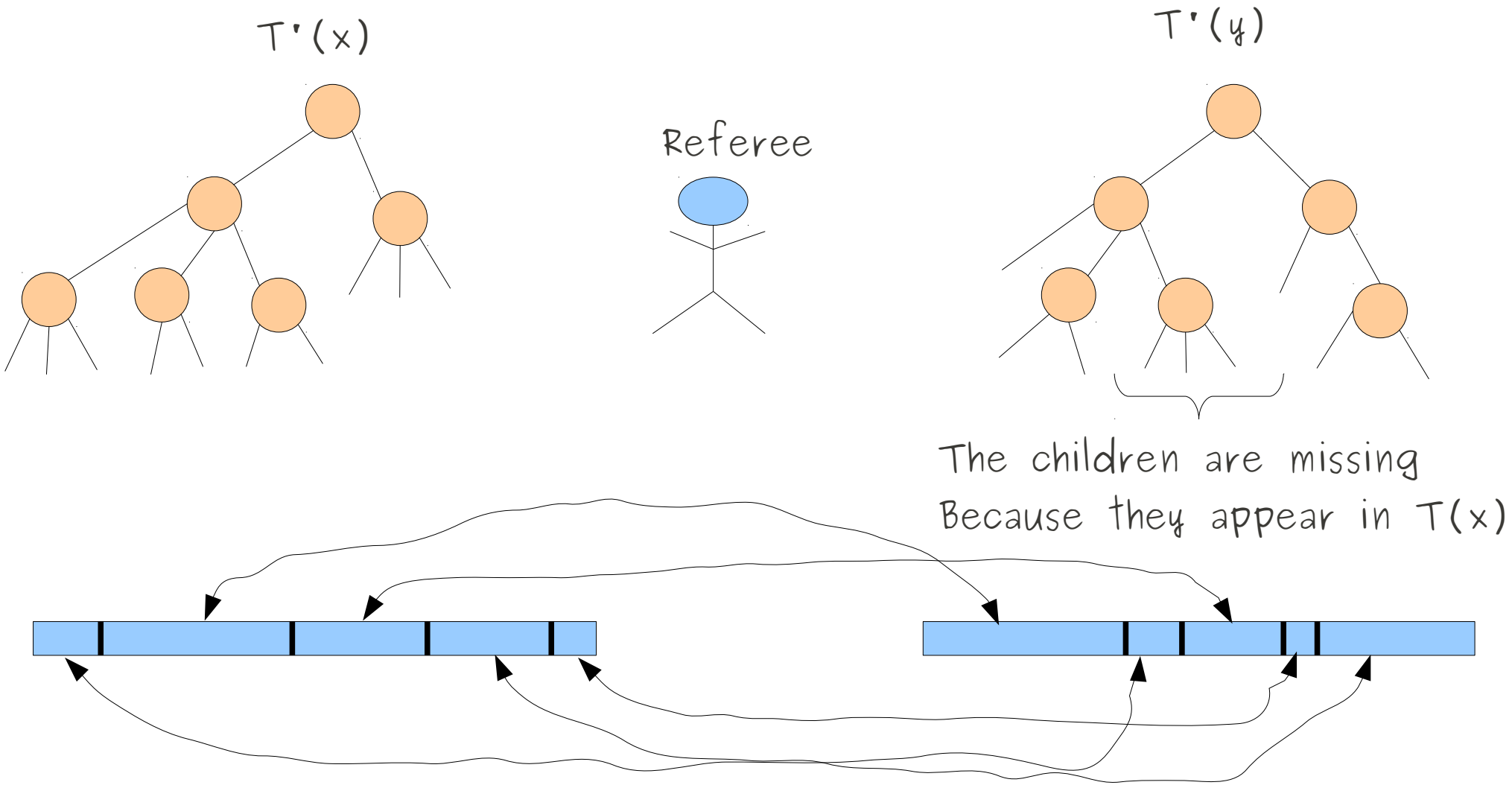


T'(y)

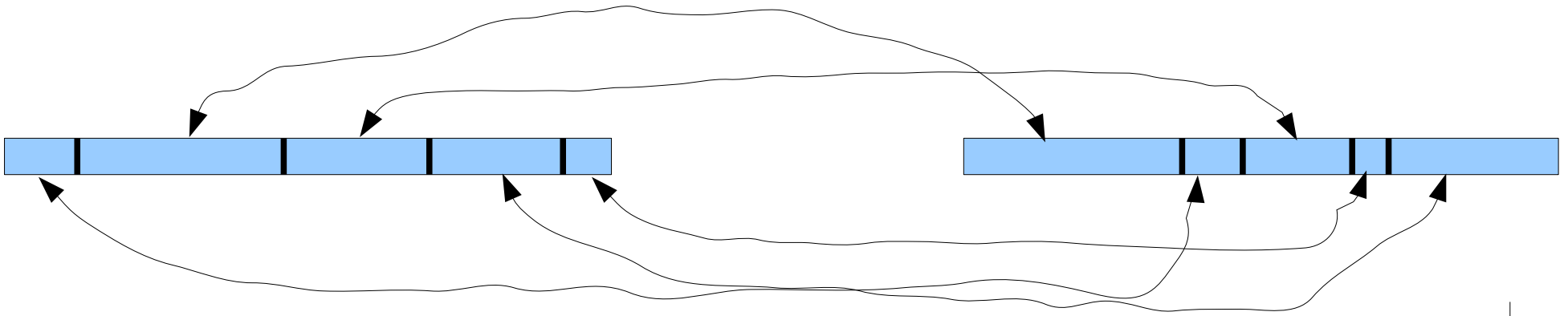


The children are missing
Because they appear in T(x)

A simultaneous Protocol for Ulam II



A simultaneous Protocol for Ulam III



Since the strings are non-repetitive,
The corresponding blocks can be relabeled accordingly
With arbitrary non-repeating characters

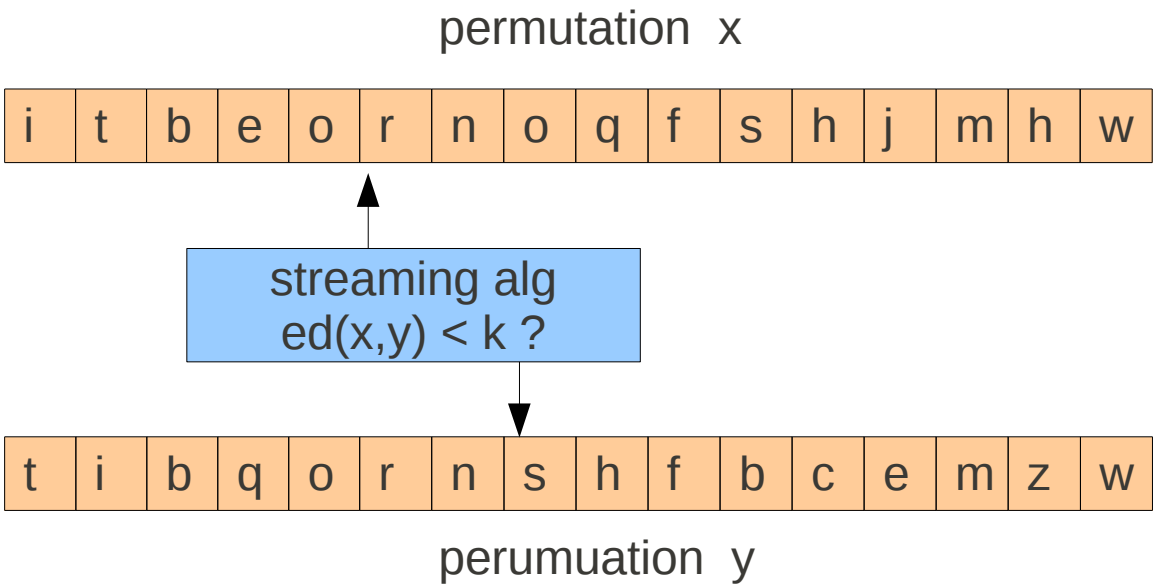
x'

y'

$$ed(x,y) = ed(x',y')$$

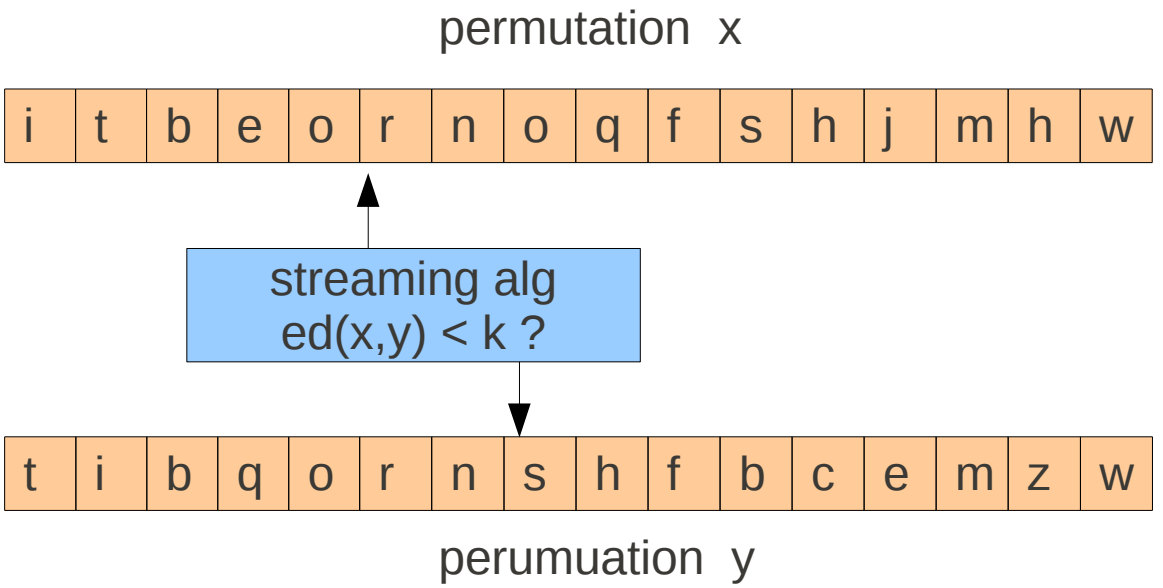
Streaming Implications

Simultaneous Streams

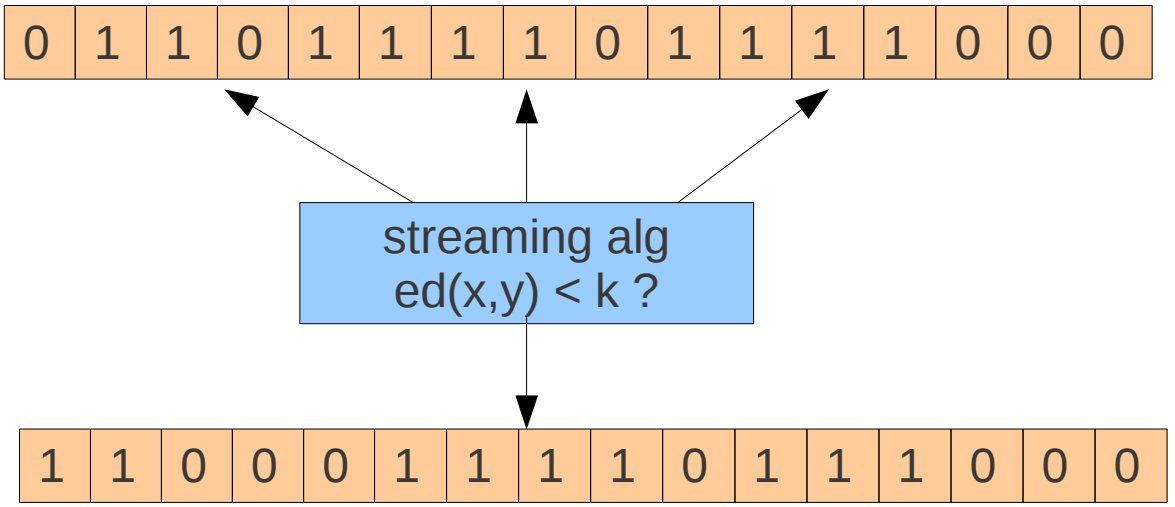


Streaming Implications

Simultaneous Streams



Assymetric Streaming



Open Question

Is there an injective mapping $f : \Sigma^n \rightarrow \{0,1\}^m$ such that $m = poly(n)$ and

If $ed(x,y) = 1$ then $H(f(x),f(y)) = O(1)$?

The best upper bound is $O(\log n \log^* n)$

Thanks