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

Alice \rightarrow Bob

x \rightarrow y
Definitions : two-player model

\(d(x, y) < k\) ?
Definitions: simultaneous model

Referee

Alice

Bob

\[ x \leftrightarrow y \]
Definitions: simultaneous model
Definitions: simultaneous model

\[ ed(x, y) < k \]
Definitions: simultaneous model

\[ ed(x, y) < k \]
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

0(k logn log(n/k)) protocol
Previous Results: Hamming

Sparse Recovery Scheme

Alice
\[ Ax \]
\[ x \]

Referee
\[ A(x-y) \]

Bob
\[ Ay \]
\[ y \]

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

\[ O(k \log n \log(n/k)) \] protocol

\[ O(k \log n) \] bit protocol (also streaming algorithm)

it outputs \( x-y \)

Time complexity: \( O(s \log n) \) for \( s \)-sparse vectors
Previous Results

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

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

$O(k \log k \log(n/k))$ bit $O(\log n)$-round protocol

Time complexity: $\text{Poly}(n)$ Cormode, Paterson, Sahinalp, and Vishkin. SODA 2000

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

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

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

\[ f : \Sigma^n \rightarrow \{0,1\}^{\text{poly}(n)} \]

- \( f(x) \) is efficiently computable
- \( f(x) \) and \( f(y) \) are distinct with high probability
- \( H(f(x), f(y)) < C \cdot \text{ed}(x, y) \) for small \( C \), Small Expansion

There exists efficient decoding procedure \( R \) where given \( f(x) \), \( R \) returns \( x \) whp
General Framework II

The Hamming protocol with parameter $k' = C_k$

Communication Complexity
$\mathcal{O}(C_k \log n)$

Time Complexity
Encoding time $f(x)$
Decoding time $R(f(x))$
Mapping $f$ for Ulam

An injective mapping $f : \Sigma^n \rightarrow \{0,1\}^{\text{poly}(n)}$ such that if $\text{ed}(x,y) = 1$ then $H(f(x), f(y))$ is small.
Mapping $f$ for Ulam

An injective mapping $f : \Sigma^n \rightarrow \{0,1\}^{\text{poly}(n)}$ such that
if $\text{ed}(x,y) = 1$ then $H(f(x), f(y))$ is small

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

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

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

Decoding is trivial
Mapping $f$ for Edit distance

An injective mapping $f : \Sigma^n \rightarrow \{0,1\}^{\text{poly}(n)}$ such that if $\text{ed}(x,y) = 1$ then $H(f(x), f(y))$ is small

Locally Consistent Parsing    CPSV 2000, CM 2002

A hierarchical partitioning of the string $x$ into substrings

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

$\varepsilon_d(x, y) = 1$ then $|f(x) - f(y)|_1 = O(\log n \log^* n)$
Based on LCP, there is a mapping $f: \Sigma^n \rightarrow [m]^L$.

$e_d(x,y) = 1$ 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
[CM 2002] Based on LCP, there is a mapping $f : \Sigma^n \rightarrow [m]^{L}$

$\text{ed}(x, y) = 1$ 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 $\text{poly}(n)$
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

\[ f(x) - f(y) \]
A simultaneous Protocol for Ulam I

\[ T'(x) \]

\[ T'(y) \]

Alice

Bob

Referee

\[ f(x) - f(y) \]

\[ f(x) \]

\[ f(x) \]

\[ f(y) \]

\[ f(y) \]
The children are missing
Because they appear in $T'(x)$
The children are missing
Because they appear in $T(x)$

The referee finds a matching between the blocks of $x$ and $y$
A simultaneous Protocol for Ulam III

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

\[ ed(x, y) = ed(x', y') \]
Streaming Implications

Simultaneous Streams

streaming alg ed(x, y) < k ?

permutation x

permutation y
Streaming Implications

Simultaneous Streams

```
iti b e o r n o q f s h j m h w
```

streaming alg

```
ed(x,y) < k ?
```

```
t i b q o r n s h f b c e m z w
```

permutation y

Assymetric Streaming

```
0 1 1 0 1 1 1 1 1 0 1 1 1 1 1 0 0 0
```

streaming alg

```
ed(x,y) < k ?
```

```
1 1 0 0 0 1 1 1 1 0 1 1 1 1 0 0 0
```
Open Question

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

\[
\text{If } ed(x,y) = 1 \text{ then } H(f(x),f(y)) = O(1) \? 
\]

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