Sequence Reconstruction for Non-Identical Channels

Michal Horovitz and Eitan Yaakobi

Computer Science Department
Technion, Haifa, Israel

Coding Seminar, May, 2017
Sequence Reconstruction

Motivation:
- Chemical and biological processes where the information is replicated and can be read from different noisy sources.
- Storage technologies, where the stored information has multiple copies or a single copy is read by several different read heads, e.g., DNA storage.

Sequence Reconstruction

A channel system of size $N$

A $t$-error channel - a channel causes at most $t$ errors.

Goal: $\forall x \in U: x' = x$ - Exact reconstruction.

Given: $U$, the channels are $t$-error.
Q: What is the minimal $N$ for exact reconstruction?

$L = \max \{|B_t(x) \cap B_t(z)|: x, z \in U\}$

Solution: $L + 1$ (Levenshtein'01)
Sequence Reconstruction

Goal: \( \forall x \in U: x' = x \) - Exact reconstruction.

Given: \( U \), data about the errors

Q: What is the minimal \( N \) for exact reconstruction?

Examples for data about the errors:
1. \( 1/2 \) of the channels are \( t_1 \)-error, and \( 1/2 \) are \( t_2 \)-error.
2. The first \( 1/2 \) of the channels are \( t_1 \)-error, and the rest are \( t_2 \)-error.
3. The average number of errors is given
Sequence Reconstruction

References

- Problem presentation. substitution errors, Johnson graphs, deletions, insertions, and more general metric distances. Le’01.

- Some general error graphs
  - LeKM’08, LeSi’09

- Permutations – K’07, K’08, KLeSi’07,

- Kendall’s $\tau$: YSwLaB’13

- Insertions – SGSD’15, GY’16

- Deletions – YG’16

D=Dolecek
G=Gabrys
K=Konstantinova,
La=Langberg,
Le=Levenshtein
M=Molodtsov
Sa=Sala
Si=Siemons,
So=Schoeny
Sw=Schwartz
Y=Yaakobi
Sequence Reconstruction for Non-Identical Channels

Outline

- Problem setup
- Two types of channels
  - General case
  - Substitutions errors
    - Explicit solution
    - Examples
  - Special systems
- $\ell$ types of channels
- Open problems
Sequence Reconstruction - Problem

A channel system of size $N$

$\mathcal{C}_1 \rightarrow y_1$
$\mathcal{C}_2 \rightarrow y_2$
$\mathcal{C}_3 \rightarrow y_3$
$\vdots$
$\mathcal{C}_{N-1} \rightarrow y_{N-1}$
$\mathcal{C}_N \rightarrow y_N$

$\mathbf{X} \in \mathbb{U}$

$V$ - a finite set, $\mathbb{U} \subseteq V$.
$\rho : V \times V \rightarrow \mathbb{N}$: a distance function
$T = (t_1, t_2, \ldots, t_\ell), t_1 < t_2 < \ldots < t_\ell \in \mathbb{N}$
$P = (p_1, p_2, \ldots, p_\ell), 0 < p_1 < p_2 < \ldots < p_\ell = 1$

Models:

A $(T, P)$-channel system:
$[p_iN]$ channels are $t_i$-error.

$N^u(T, P, U)$

A $(T, P)$-sequenced channel system:
The first $[p_iN]$ channels are $t_i$-error.

$N^k(T, P, U)$

A $t$-channel system:

$N(t, U)$

Problem:
The average number of errors is $t$

For these three models - find the minimum size of a channel system required for exact reconstruction

- Find $N^u(T, P, U), N^k(T, P, U), N(t, U)$
- Clearly: $N^u(T, P, U) \geq N^k(T, P, U)$
Sequence Reconstruction - Problem

A channel system of size $N$

A finite set $U \subseteq V$. $\rho : V \times V \to \mathbb{N}$: a distance function

$T = (t_1, t_2, ..., t_\ell), t_1 < t_2 < \cdots < t_\ell \in \mathbb{N}$

$P = (p_1, p_2, ..., p_\ell), 0 < p_1 < p_2 < \cdots < p_\ell = 1$

Models:

A $(T, P)$-channel system:

$[p_iN]$ channels are $t_i$-error.

$x'$ A $(T, P)$-sequenced channel system:

The first $[p_iN]$ channels are $t_i$-error.

A $t$-channel system:

The average number of errors is $t$

$$N^g(T, P, U) = \max\{N^g(T, P, \{x, z\}) : x, z \in U\}, \ g \in \{u, k\}$$

$$N(t, U) = \max\{N(t, \{x, z\}) : x, z \in U\}$$

Problem: For $U = \{x, z\}$

Find $N^u(T, P, U), N^k(T, P, U), N(t, U)$
Sequence Reconstruction - Notations

A channel system of size $N$

- $V$: a finite set
- $\rho: V \times V \to \mathbb{N}$: a distance function
- $U \subseteq V$
- $B_t(x) = \{y : \rho(x, y) \leq t\}$
- $I(x, z, t_1, t_2) = B_{t_1}(x) \cap B_{t_2}(z)$
- $I(x, z, t) = B_t(x) \cap B_t(z)$
- $N(x, z, t_1, t_2) = |I(x, z, t_1, t_2)|$
- $N(x, z, t) = |I(x, z, t)|$
Sequence Reconstruction for Non-Identical Channels

Outline
- Problem setup
- Two types of channels
  - General case
  - Substitutions errors
    - Explicit solution
    - Examples
  - Special systems
- $\ell$ types of channels
- Open problems
Sequence Reconstruction - $\ell = 2$

A channel system of size $N$

\[ U = \{x, z\} \subseteq V, \rho, \]
\[ T = (t_1, t_2), \, t_1 < t_2 \in \mathbb{N} \]
\[ P = (p, 1), \, 0 < p < 1 \]

Models:

A $(T, P)$-channel system:
- $[pN]$ channels are $t_1$-error
- All channels are $t_2$-error

A $(T, P)$-sequenced channel system:
- The first $[pN]$ channels are $t_1$-error
- All channels are $t_2$-error
Sequence Reconstruction - $\ell = 2$

Sequenced Model

$x, z \in V, \rho, T = (t_1, t_2), t_1 < t_2 \in \mathbb{N}, P = (p, 1), 0 < p < 1$

A $(T, P)$-sequenced channel system:

- The first $[pN]$ channels are $t_1$-error
- All channels are $t_2$-error

Problem: $N^k(T, P, \{x, z\}) = ?$

Solution:

$$N^k(T, P, \{x, z\}) = L + 1$$

$$L = \min \left\{ \left\lfloor \frac{N(x, z, t_1)}{p} \right\rfloor, N(x, z, t_2) \right\}$$

Proof: first part: $N \geq L + 1$ is sufficient:

- $N > N(x, z, t_2)$

OR

- $N > \left\lfloor \frac{N(x, z, t_1)}{p} \right\rfloor \Rightarrow [pN] > N(x, z, t_1)$
Sequence Reconstruction - $\ell = 2$

**Sequenced Model**

$x, z \in V, \rho, T = (t_1, t_2), t_1 < t_2 \in \mathbb{N}, P = (p, 1), 0 < p < 1$

A $(T, P)$-sequenced channel system:

- The first $[pN]$ channels are $t_1$-error
- All channels are $t_2$-error

**Problem:** $N^k(T, P, \{x, z\}) = ?$

**Solution:**

$$N^k(T, P, \{x, z\}) = L + 1$$

$$L = \min \left\{ \left\lfloor \frac{N(x, z, t_1)}{p} \right\rfloor, N(x, z, t_2) \right\}$$

**Proof:** second part: $N \leq L$ is not-sufficient:

- $N \leq N(x, z, t_2)$
- AND
- $N \leq \left\lfloor \frac{N(x, z, t_1)}{p} \right\rfloor \Rightarrow [pN] \leq N(x, z, t_1)$
Sequence Reconstruction - $\ell = 2$

Non-Sequenced Model

$x, z \in V, \rho, T = (t_1, t_2), t_1 < t_2 \in \mathbb{N}, P = (p, 1), 0 < p < 1$

A $(T, P)$-channel system:
- $[pN]$ channels are $t_1$-error
- All channels are $t_2$-error

**Problem:** $N^u(T, P, \{x, z\}) = ?$

**Solution:**

$$N^u(T, P, \{x, z\}) = L + 1$$

$$L = \min \left\{ \left[ \frac{N(x, z, t_1, t_2)}{p} \right], \left[ \frac{N(z, x, t_1, t_2)}{p} \right], N(x, z, t_2), N'(x, z, t_1, p) \right\}$$

$$N'(x, z, t_1, p) = \min \{ J : 2[pJ] - J > N(x, z, t_1), J \geq 1 \} - 1$$

**Proof:** first part: $N \geq L + 1$ is sufficient:

1) $N > N(x, z, t_2)$

2) $N > \left[ \frac{N(x, z, t_1, t_2)}{p} \right] \implies [pN] > N(x, z, t_1, t_2)$

3) $N > \left[ \frac{N(z, x, t_1, t_2)}{p} \right] \implies [pN] > N(z, x, t_1, t_2)$

4) $N = L + 1 \implies 2[pN] - N > N(x, z, t_1) \implies 2[pN] - N(x, z, t_1) > N$
Sequence Reconstruction - \( \ell = 2 \)

Non-Sequenced Model

\( x, z \in V, \rho, T = (t_1, t_2), t_1 < t_2 \in \mathbb{N}, P = (p, 1), 0 < p < 1 \)

A \((T, P)\)-channel system:

\[ [pN] \text{ channels are } t_1\text{-error} \]

All channels are \( t_2\)-error

**Problem:** \( N^u(T, P, \{x, z\}) = ? \)

**Solution:**

\[ N^u(T, P, \{x, z\}) = L + 1 \]

\[ L = \min \left\{ \left\lfloor \frac{N(x, z, t_1, t_2)}{p} \right\rfloor, \left\lfloor \frac{N(z, x, t_1, t_2)}{p} \right\rfloor, N(x, z, t_2), N'(x, z, t_1, p) \right\} \]

\[ N'(x, z, t_1, p) = \min\{J : 2[pJ] - J > N(x, z, t_1), J \geq 1\} - 1 \]

Proof: second part: \( N \leq L \) is not sufficient:

1) \( N \leq N(x, z, t_2) \)
2) \( N \leq \left\lfloor \frac{N(x, z, t_1, t_2)}{p} \right\rfloor \Rightarrow [pN] \leq N(x, z, t_1, t_2) \)
3) \( N \leq \left\lfloor \frac{N(z, x, t_1, t_2)}{p} \right\rfloor \Rightarrow [pN] \leq N(z, x, t_1, t_2) \)
4) \( N \leq N'(x, z, t_1, p) \Rightarrow 2[pN] - N(x, z, t_1) \leq N \)
Sequence Reconstruction - $\ell = 2$

Summarize - a $(T, P)$-Channel System

$x, z \in V, \rho, T = (t_1, t_2), t_1 < t_2 \in \mathbb{N}, P = (p, 1), 0 < p < 1$

A $(T, P)$-channel system:
- $[pN]$ channels are $t_1$-error
- All channels are $t_2$-error

$N^u(T, P, \{x, z\}) = L + 1$

$L = \min \left\{ \left[ \frac{N(x, z, t_1, t_2)}{p} \right], \left[ \frac{N(z, x, t_1, t_2)}{p} \right], N(x, z, t_2), N'(x, z, t_1, p) \right\}$

$N'(x, z, t_1, p) = \min\{J : 2[pJ] - J > N(x, z, t_1), J \geq 1\} - 1$

A $(T, P)$-sequenced channel system:
- The first $[pN]$ channels are $t_1$-error
- All channels are $t_2$-error

$N^k(T, P, \{x, z\}) = L + 1$

$L = \min \left\{ \left[ \frac{N(x, z, t_1)}{p} \right], N(x, z, t_2) \right\}$
Sequence Reconstruction for Non-Identical Channels

Outline

- Problem setup
- Two types of channels
  - General case
  - Substitutions errors
    - Explicit solution
    - Examples
  - Special systems
- $\ell$ types of channels
- Open problems
Sequence Reconstruction - $\ell = 2$
Substitution (Hamming)

$x, z \in V, \rho, d = \rho(x, z), T = (t_1, t_2), t_1 < t_2 \in \mathbb{N}, P = (p, 1), 0 < p < 1$

$[pN]$ channels are $t_1$-error, all channels are $t_2$-error.

- $N(d, t_1, t_2) = N(x, z, t_1, t_2)$
- $N^g(T, P, d) = N^g(T, P, \{x, z\})$
- $N^g(T, P, d) \geq N^g(T, P, d + 1)$
- $N^k(T, P, d) = \min\left\{\left\lfloor \frac{N(d, t_1)}{p} \right\rfloor, N(d, t_2)\right\}$
- $N^u(T, P, d) = \min\left\{\left\lfloor \frac{N(d, t_1, t_2)}{p} \right\rfloor, N(d, t_2), N'(d, t_1, p)\right\}$

$N'(d, t_1, p) = \min\{J : 2[pJ] - J > N(d, t_1), J \geq 1\}$
Sequence Reconstruction - $\ell = 2$

Substitution (Hamming) - Non-Sequenced Model

$x, z \in V, \rho, d = \rho(x, z), T = (t_1, t_2), t_1 < t_2 \in \mathbb{N}, P = (p, 1), 0 < p < 1$

$[pN]$ channels are $t_1$-error, all channels are $t_2$-error.

$N^u(T, P, d) = \min \left\{ \left\lfloor \frac{N(d, t_1, t_2)}{p} \right\rfloor, N(d, t_2), N'(d, t_1, p) \right\}$

$N'(d, t_1, p) = \min \{J : 2[pJ] - J > N(d, t_1), J \geq 1\}$

- **For $0 < p \leq 1/2$:**

  $N^u(T, P, d) = \begin{cases} 
  1 & \text{if } d > 2t_1 \\
  N(d, t_2) + 1 & \text{otherwise.}
  \end{cases}$

- **For $1/2 < p < 1$:**

  $N^u(T, P, d) = \begin{cases} 
  \left\lfloor \frac{N(d, t_1, t_2)}{p} \right\rfloor + 1 & \text{if } d \text{ is even, } t_2 = t_1 + 1, \\
  \Theta(n^{\frac{t_1 + t_2 - d}{2}}) & \text{if } d > 2t_1, \text{ otherwise.}
  \end{cases}$

$N'(d, t_1, p)$:

- **For $0 < p \leq 1/2$:**

  $N'(d, t_1, p) = \Theta(n^{\frac{2t_1 - d}{2}})$

- **For $1/2 < p < 1$:**

  $N'(d, t_1, p) = \Theta(n^{\frac{2t_1 - d}{2}})$
Sequence Reconstruction - $\ell = 2$

Substitution (Hamming) - Non-Sequenced Model

$x, z \in V, \rho, d = \rho(x, z), T = (t_1, t_2), t_1 < t_2 \in \mathbb{N}, P = (p, 1), 0 < p < 1$

$[pN]$ channels are $t_1$-error, all channels are $t_2$-error.

$N_u(T, P, d) = \min \left\{ \left\lfloor \frac{N(d, t_1, t_2)}{p} \right\rfloor, N(d, t_2), N'(d, t_1, p) \right\}$

$N'(d, t_1, p) = \min \{J : 2[pJ] - J > N(d, t_1), J \geq 1\}$

For $0 < p \leq 1/2$:

$N_u(T, P, d) = \Theta\left( n^{\frac{2t_1-d}{2}} \right)$

For $1/2 < p < 1$:

$N_u(T, P, d) = \Theta\left( n^{\left\lfloor \frac{t_1+t_2-d}{2} \right\rfloor} \right)$

All channels are identical | A $(T, P)$-channel system

| $0 < p \leq \frac{1}{2}, d = 1, T = (2,4)$ | $\Theta(n^3)$ | $\Theta(n^2)$ |
| $0 < p \leq \frac{1}{2}, d = 1, T = (2,8)$ | $\Theta(n^7)$ | $\Theta(n^4)$ |
| $\frac{1}{2} < p \leq \frac{2}{3}, d = 2, T = (4,5)$ | $\Theta(n^4)$ | $\Theta(n^3)$ |
Sequence Reconstruction for Non-Identical Channels

Outline

- Problem setup
- Two types of channels
  - General case
  - Substitutions errors
    - Explicit solution
    - Examples
  - Special systems
- \( \ell \) types of channels
- Open problems
Sequence Reconstruction - A $(T, i, b)$-Channel System

$x, z \in V, \rho, T = (t_1, t_2), t_1 < t_2 \in \mathbb{N}, b \in \mathbb{N}, i \in \{1, 2\}$

A $(T, i, b)$-sequenced channel system:
- The first $b$ channels are $t_i$-error
- All channels are $t_{3-i}$-error
$N^k(T, i, b, \{x, z\})$ - minimum size of a $(T, i, b)$-sequenced channel system required for exact reconstruction

A $(T, i, b)$-channel system:
- $b$ channels are $t_i$-error
- All channels are $t_{3-i}$-error
$N^u(T, i, b, \{x, z\})$ - minimum size of a $(T, i, b)$-channel system required for exact reconstruction
Sequence Reconstruction

A \((T, 1, b)\)-Channel System

\(x, z \in V, \rho, T = (t_1, t_2), t_1 < t_2 \in \mathbb{N}, b \in \mathbb{N}\)

**A \((T, 1, b)\)-sequenced channel system:**
- The first \(b\) channels are \(t_1\)-error
- All channels are \(t_2\)-error

**A \((T, 1, b)\)-channel system:**
- \(b\) channels are \(t_1\)-error
- All channels are \(t_2\)-error

**Problem:** \(N^k(T, 1, b, \{x, z\}) = \) ? , \(N^u(T, 1, b, \{x, z\}) = \) ?

**Solution:** \(N^k(T, 1, b, \{x, z\}) = N^u(kT, 1, b, \{x, z\}) = L + 1\)

\[
L = \begin{cases} 
N(x, z, t_2), & \text{if } N(x, z, t_1) \geq b \\
N(x, z, t_1), & \text{else}
\end{cases}
\]
Sequence Reconstruction

A \((T, 2, b)\)-Sequenced Channel System

\[ x, z \in V, \rho, T = (t_1, t_2), t_1 < t_2 \in \mathbb{N}, b \in \mathbb{N} \]

A \((T, 2, b)\)-sequenced channel system:

- The first \(b\) channels are \(t_2\)-error
- All channels are \(t_1\)-error

**Problem:** \(N^k(T, 2, b, \{x, z\}) = ?\)

**Solution:** \(N^k(T, 2, b, \{x, z\}) = L + 1\)

\[ L = \min\{N(x, z, t_1) + b, N(x, z, t_2)\} \]
Sequence Reconstruction

A \((T, 2, b)\)-Non-Sequenced Channel System

\[ x, z \in V, \rho, T = (t_1, t_2), t_1 < t_2 \in \mathbb{N}, b \in \mathbb{N} \]

A \((T, 2, b)\)-channel system:
- \(b\) channels are \(t_2\)-error
- All channels are \(t_1\)-error

Problem: \(N^u(T, 2, b, \{x, z\}) = ?\)

Solution:
\[ N^u(T, 2, b, \{x, z\}) = L + 1 \]

\[ L = \min \left\{ \begin{array}{l}
N(x, z, t_1, t_2) + b, \\
N(z, x, t_1, t_2) + b, \\
N(x, z, t_2), \\
N(x, z, t_1) + 2b
\end{array} \right\} \]
Sequence Reconstruction - $\ell = 2$

$x, z \in V, \rho, T = (t_1, t_2), t_1 < t_2 \in \mathbb{N}, P = (p, 1), 0 < p < 1, b \in \mathbb{N}$.

Solution: $N = L + 1$, where $L =$

<table>
<thead>
<tr>
<th></th>
<th>Sequenced</th>
<th>Non-Sequenced</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A (T,P)$-channel system</td>
<td>$\min \left{ \left[ \frac{N(x, z, t_1)}{p} \right], N(x, z, t_2) \right}$</td>
<td>$\min \left{ \left[ \frac{N(x, z, t_1, t_2)}{p} \right], N(x, z, t_2), N'(x, z, t_1, p) \right}$</td>
</tr>
<tr>
<td>$A (T,1,b)$-channel system</td>
<td>$\begin{cases} N(x, z, t_2), &amp; \text{if } N(x, z, t_1) \geq b \ N(x, z, t_1), &amp; \text{else} \end{cases}$</td>
<td>$\begin{cases} N(x, z, t_2), &amp; \text{if } N(x, z, t_1) \geq b \ N(x, z, t_1), &amp; \text{else} \end{cases}$</td>
</tr>
<tr>
<td>$A (T,2,b)$-channel system</td>
<td>$\min { N(x, z, t_1) + b, N(x, z, t_2) }$</td>
<td>$\min \left{ N(x, z, t_1, t_2) + b, N(z, x, t_1, t_2) + b, N(x, z, t_2), N(x, z, t_1) + 2b \right}$</td>
</tr>
</tbody>
</table>
Sequence Reconstruction for Non-Identical Channels

Outline

- Problem setup
- Two types of channels
  - General case
  - Substitutions errors
    - Explicit solution
    - Examples
  - Special systems
- $\ell$ types of channels
- Open problems
Sequence Reconstruction - $\ell$ Types of Channels

Sequenced Model

$x, z \in V, \rho,$

$T = (t_1, t_2, ..., t_\ell), t_1 < t_2 < ... < t_\ell \in \mathbb{N},$

$P = (p_1, p_2, ..., p_\ell), 0 < p_1 < p_2 < ... < p_\ell = 1$

A $(T, P)$-sequenced channel system:

The first $[p_iN]$ channels are $t_i$-error

$N^k(T, P, U) = L + 1$

$L = \min \left\{ \left\lfloor \frac{N(x, z, t_i)}{p_i} \right\rfloor : 1 \leq i \leq \ell \right\}$
Sequence Reconstruction - $\ell$ Types of Channels
Non-Sequenced Model

$x, z \in V, \rho,$
$T = (t_1, t_2, \ldots, t_\ell), t_1 < t_2 < \ldots < t_\ell \in \mathbb{N},$
$P = (p_1, p_2, \ldots, p_\ell), 0 < p_1 < p_2 < \ldots < p_\ell = 1$

A $(T, P)$-channel system:

$[p_i N]$ channels are $t_i$-error

$N^u(T, P, U) = \min\{L_1, L_2, L_3\} + 1$

$L_1 = \min \left\{ \left[ \frac{N(x, z, t_i, t_\ell)}{p_i} \right] : 1 \leq i < \ell \right\}$

$L_2 = \min \left\{ \left[ \frac{N(z, x, t_i, t_\ell)}{p_i} \right] : 1 \leq i < \ell \right\}$

$L_2 = N(x, z, t_\ell)$

$L_4 = \min\{N'(x, z, t_i, t_j, p_i, p_j) : 1 \leq i, j < \ell\}$

$N'(x, z, t_i, t_j, p_i, p_j) = \min\{J : [p_i J] + [p_j J] - J > N(x, z, t_i, t_j), J \geq 1\} - 1$
Sequence Reconstruction - $\ell$ Types of Channels

A $t$-channel system

$x, z \in V, \rho, t \in \mathbb{N}, d = \rho(x, z)$

A $t$-channel system:

The average number of errors is $t$

Solution: If $d > 2t$ then $N(t, U) = 1$.

Otherwise, exact reconstruction is not supported (for any $N$).
Sequence Reconstruction - $\ell$ Types of Channels

A $t$-channel system

$x, z \in V, \rho, t \in \mathbb{N}, d = \rho(x, z)$

A $t$-channel system:

The average number of errors is $t$

Solution: If $d > 2t$ then $N(t, U) = 1$.

Otherwise, exact reconstruction is not supported (for any $N$).

$$N(t, \{x, z\}) = \min \left\{ N : tN \leq \left\lfloor \frac{dN - 1}{2} \right\rfloor \right\}$$
Sequence Reconstruction for Non-Identical Channels

Outline

- Problem setup
- Two types of channels
  - General case
  - Substitutions errors
    - Explicit solution
    - Examples
  - Special systems
- \( \ell \) types of channels
- Open problems
Sequence Reconstruction for non-identical Channels

Open Problems

Given: $U$, channels types.

Q: What is the minimal size of a channel system?
   - Combination of types of errors.
   - Other data about the number of errors.

Given: A channel system of size $N$.

Q: What is the minimal distance of the input?
   - Combination of types of errors.
   - Some data about the number of errors.

Thank You!