On the compressibility of sketches

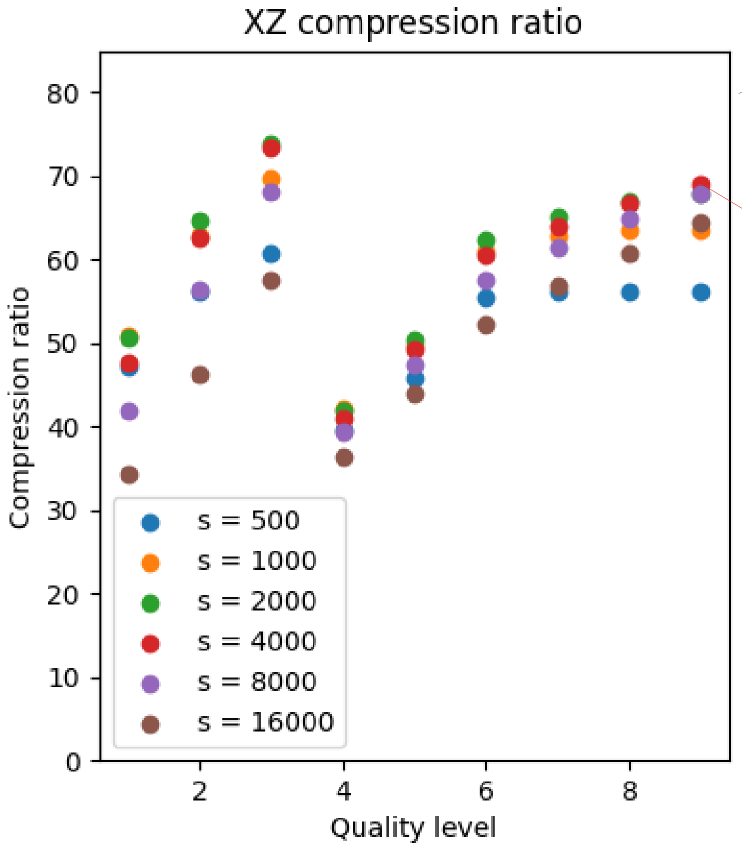
RLE-compression & Hamming distance

Marie Picard

Supervisors : Karel Břinda,
Leo Ackermann

November 20, 2025

Quick recap of previous results



Compression rate up to 80

Best compression for s = 4000

H (informal)

There is a value of sketch size s, around 2000 to 4000, such that the compressibility of sketches is optimal.

H (formal)

$\forall \mathcal{G}, \exists s \in \mathbb{N}, \frac{1}{s} K(\mathit{Sketches}(\mathcal{G}, s))$ is minimal where

- K is Kolmogorov's complexity
- $\mathcal{G}$ is the set of genomes to sketch
- $\mathit{Sketches}(\mathcal{G}, s)$ is the archive of all sketches of $g \in \mathcal{G}$ of size s

H (informal)

There is a value of sketch size s, around 2000 to 4000, such that the compressibility of sketches is optimal.

H (formal)

$\forall \mathcal{G}, \exists s \in \mathbb{N}, \frac{1}{s} K(Sketches(\mathcal{G}, s))$ is minimal where

Not computable

- K is Kolmogorov's complexity
- $\mathcal{G}$ is the set of genomes to sketch
- $Sketches(\mathcal{G}, s)$ is the archive of all sketches of $g \in \mathcal{G}$ of size s

H

$$\forall \mathcal{G}, \exists s \in \mathbb{N}, \frac{1}{s} RLE(Sketches(\mathcal{G}, s)) \text{ is minimal}$$

where

- RLE is run-length encoding
- $\mathcal{G}$ is the set of genomes to sketch
- $Sketches(\mathcal{G}, s)$ is the archive of all mash sketches of $g \in \mathcal{G}$ of size s

Input

List of mash sketches :

$$\mathcal{S} = (S_1, ..., S_{|\mathcal{S}|})$$

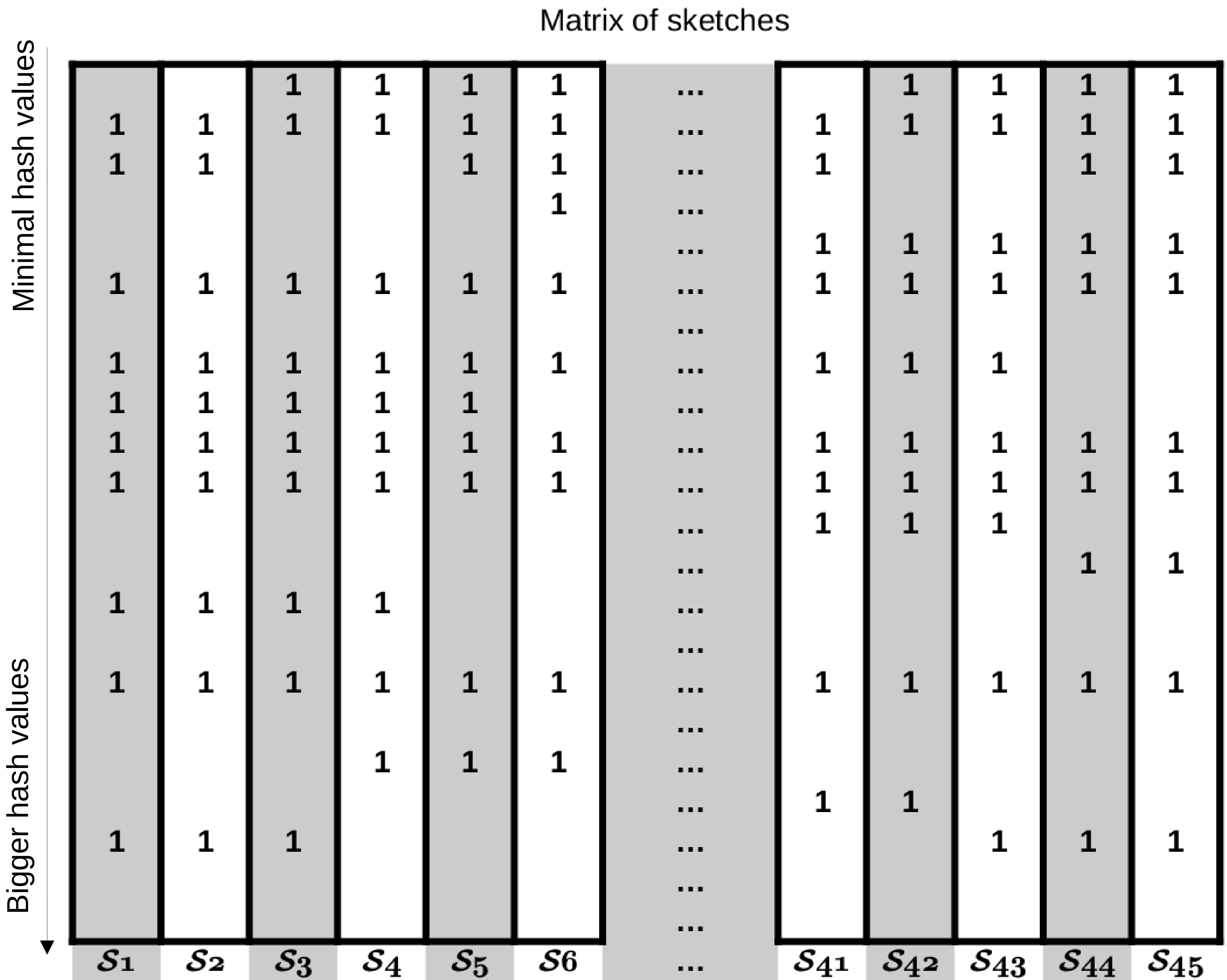
such that $s = |S_1| = ... = |S_{|\mathcal{S}|}|$

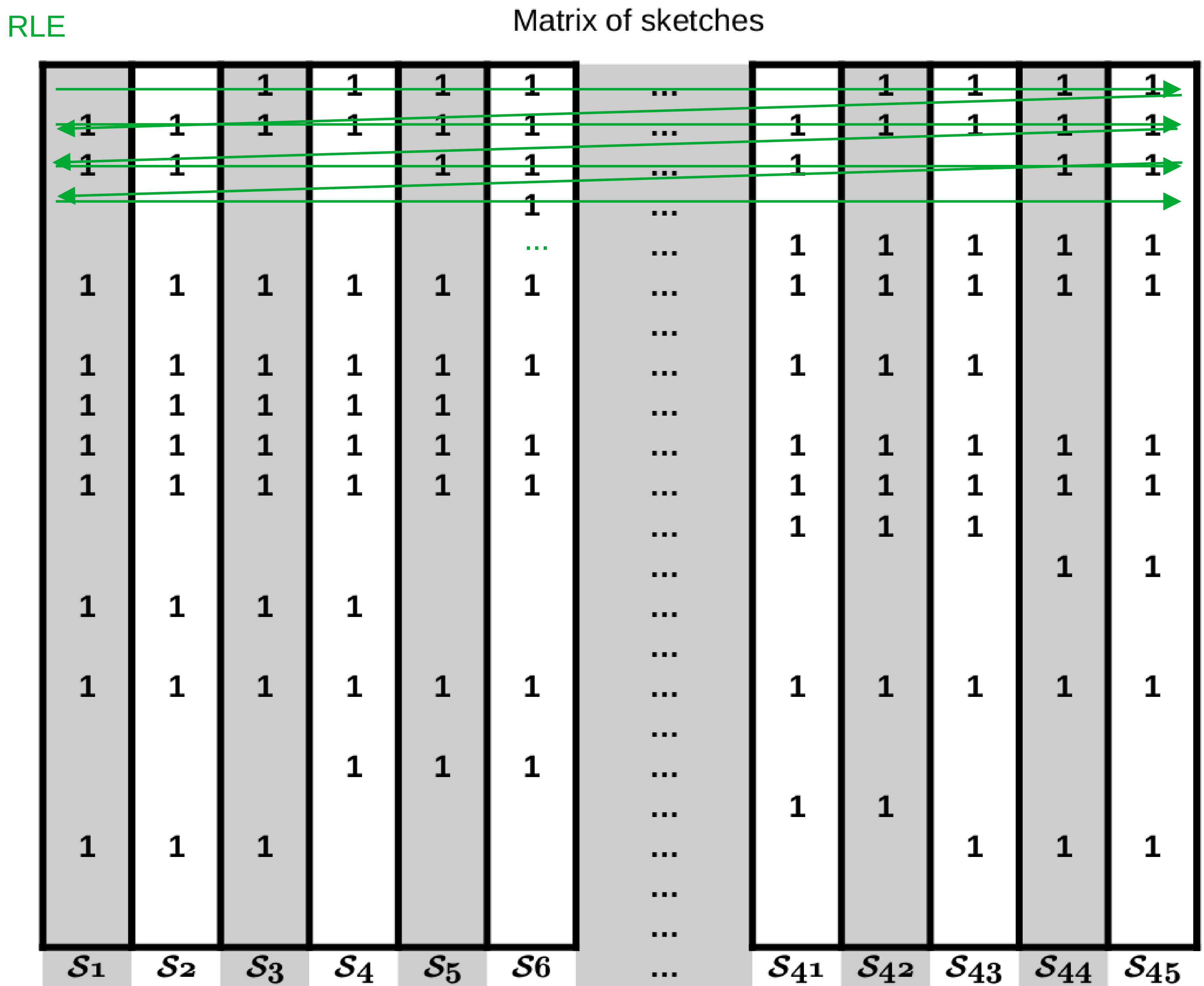
Compressed representation

- Union of all sketches (sorted)

$$\mathcal{S}_U := \bigcup_{S \in \mathcal{S}} S$$

- presence matrix $\mathcal{M}$
  $\mathcal{M}_{i,j} = 1$ iff the i$^{\text{th}}$ element of $\mathcal{S}_U \in S_j$


Matrix of sketches
Minimal hash values
Bigger hash values
$\mathcal{S}_1$ $\mathcal{S}_2$ $\mathcal{S}_3$ $\mathcal{S}_4$ $\mathcal{S}_5$ $\mathcal{S}_6$ ... $\mathcal{S}_{41}$ $\mathcal{S}_{42}$ $\mathcal{S}_{43}$ $\mathcal{S}_{44}$ $\mathcal{S}_{45}$


RLE
Matrix of sketches
S1
S2
S3
S4
S5
S6
...
S41
S42
S43
S44
S45

distance δ

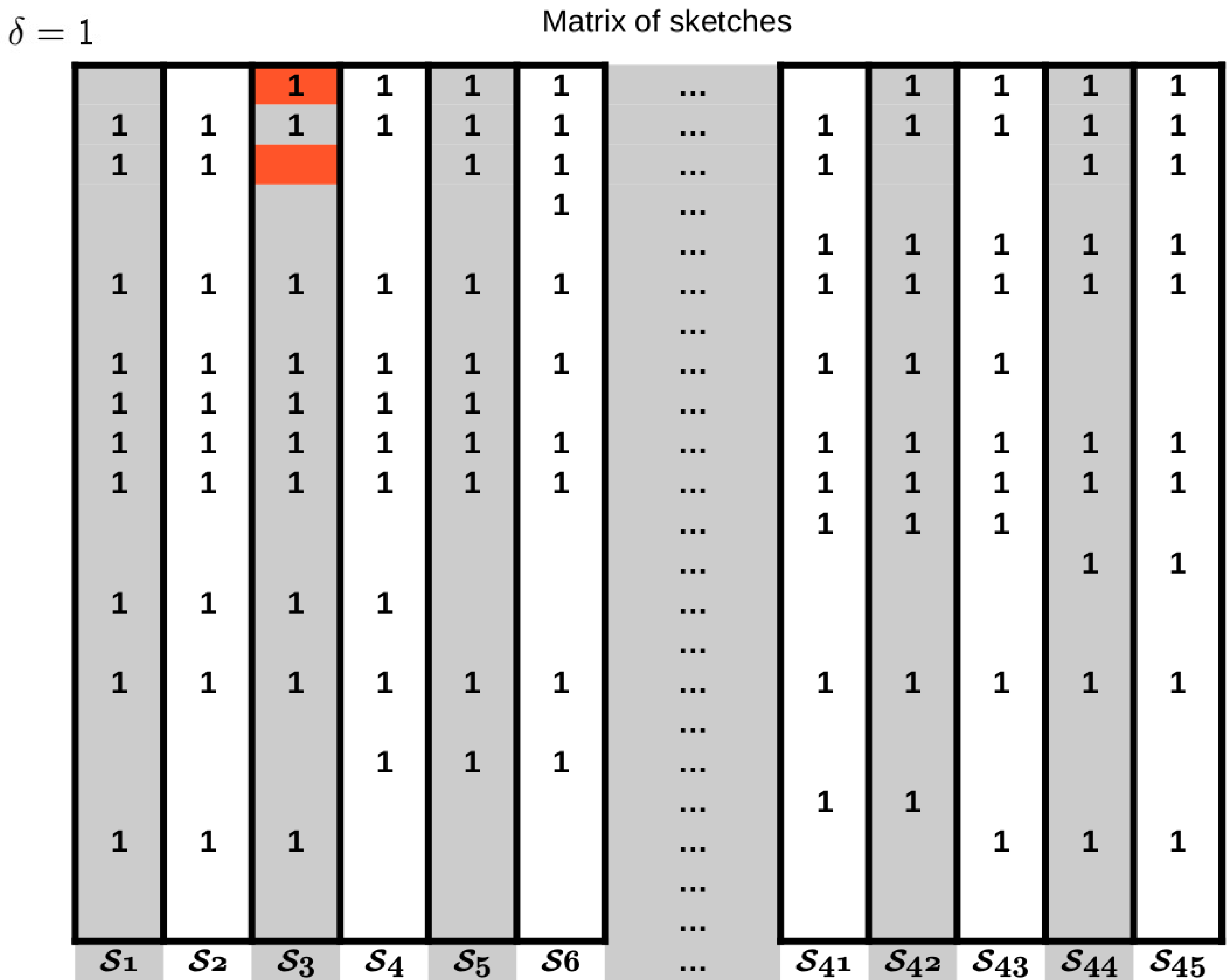
$$\delta = \frac{1}{2} \sum_{i=1}^{|\mathcal{S}|-1} d_{Hamming}(S_i, S_{i+1})$$

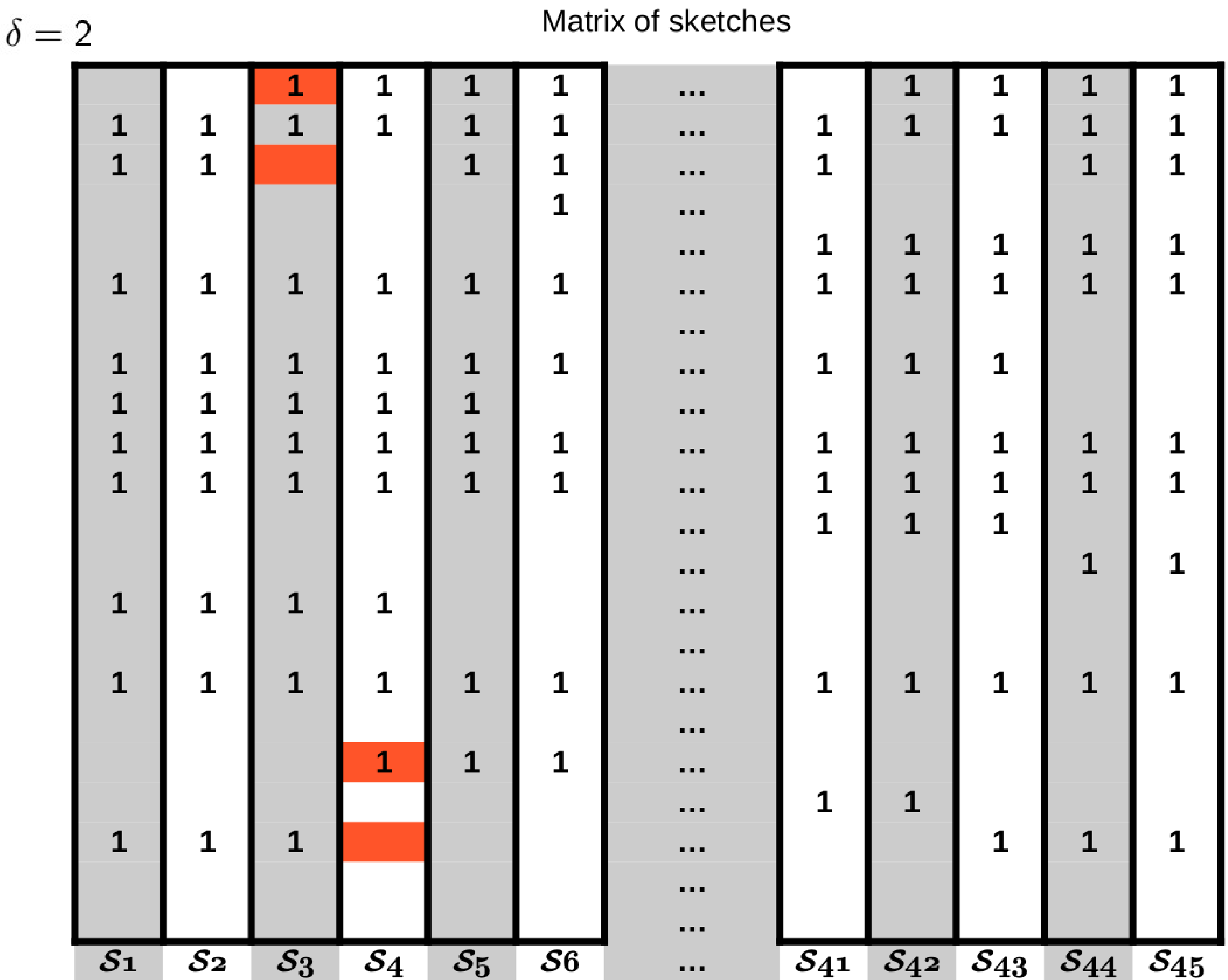
- Easy to compute
- Determines the RLE compression rate

$\delta = 0$

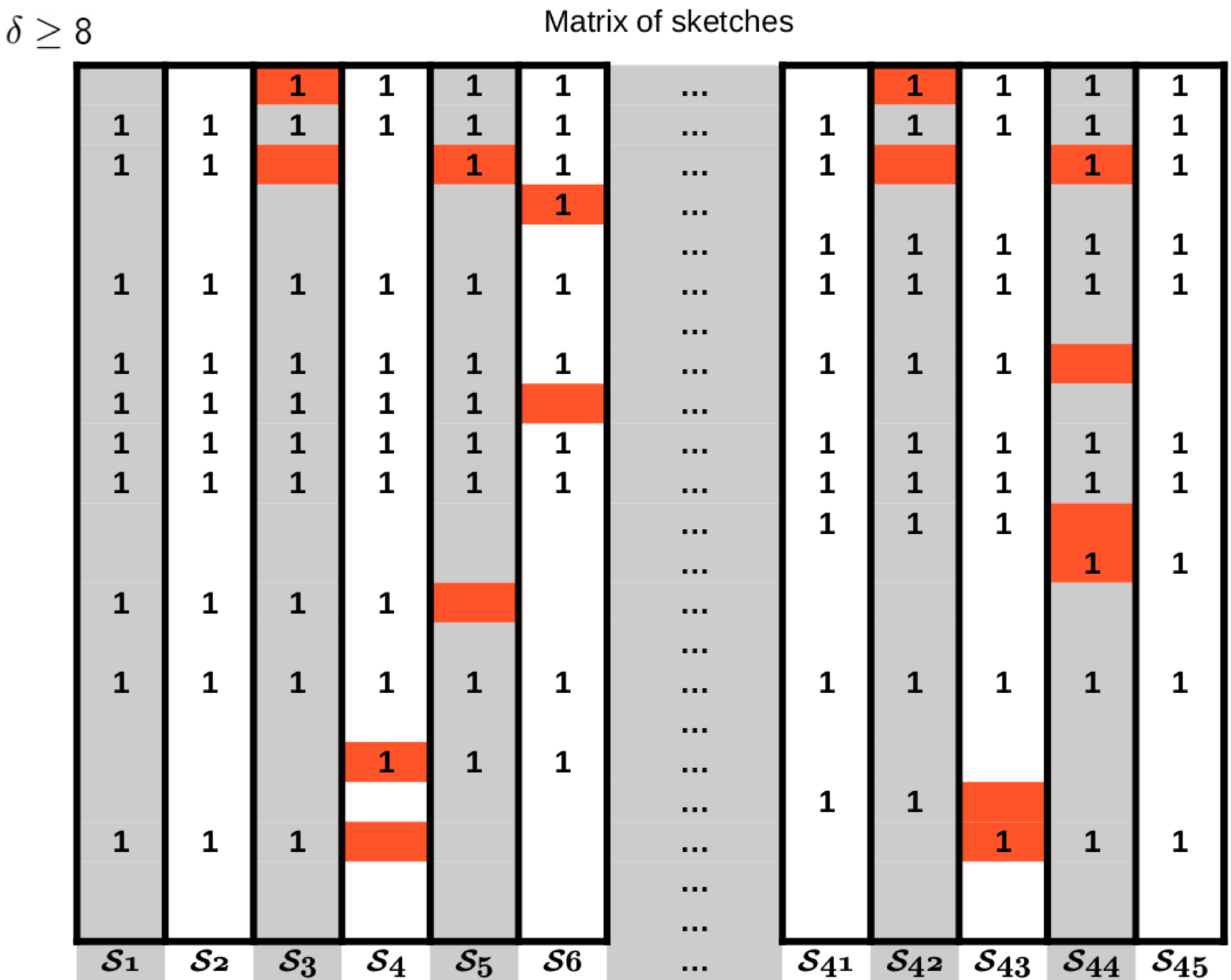
Matrix of sketches

$\mathcal{S}_1$	$\mathcal{S}_2$	$\mathcal{S}_3$	$\mathcal{S}_4$	$\mathcal{S}_5$	$\mathcal{S}_6$	...	$\mathcal{S}_{41}$	$\mathcal{S}_{42}$	$\mathcal{S}_{43}$	$\mathcal{S}_{44}$	$\mathcal{S}_{45}$
		1	1	1	1	...		1	1	1	1
1	1	1	1	1	1	...	1	1	1	1	1
1	1			1	1	...	1			1	1
					1	...					
						...	1	1	1	1	1
1	1	1	1	1	1	...	1	1	1	1	1
						...					
1	1	1	1	1	1	...	1	1	1		
1	1	1	1	1		...					
1	1	1	1	1	1	...	1	1	1	1	1
1	1	1	1	1	1	...	1	1	1	1	1
						...	1	1	1		
						...				1	1
1	1	1	1			...					
						...					
1	1	1	1	1	1	...	1	1	1	1	1
						...					
			1	1	1	...					
						...	1	1			
1	1	1				...			1	1	1
						...					
						...					


δ = 1
Matrix of sketches
S1
S2
S3
S4
S5
S6
...
S41
S42
S43
S44
S45



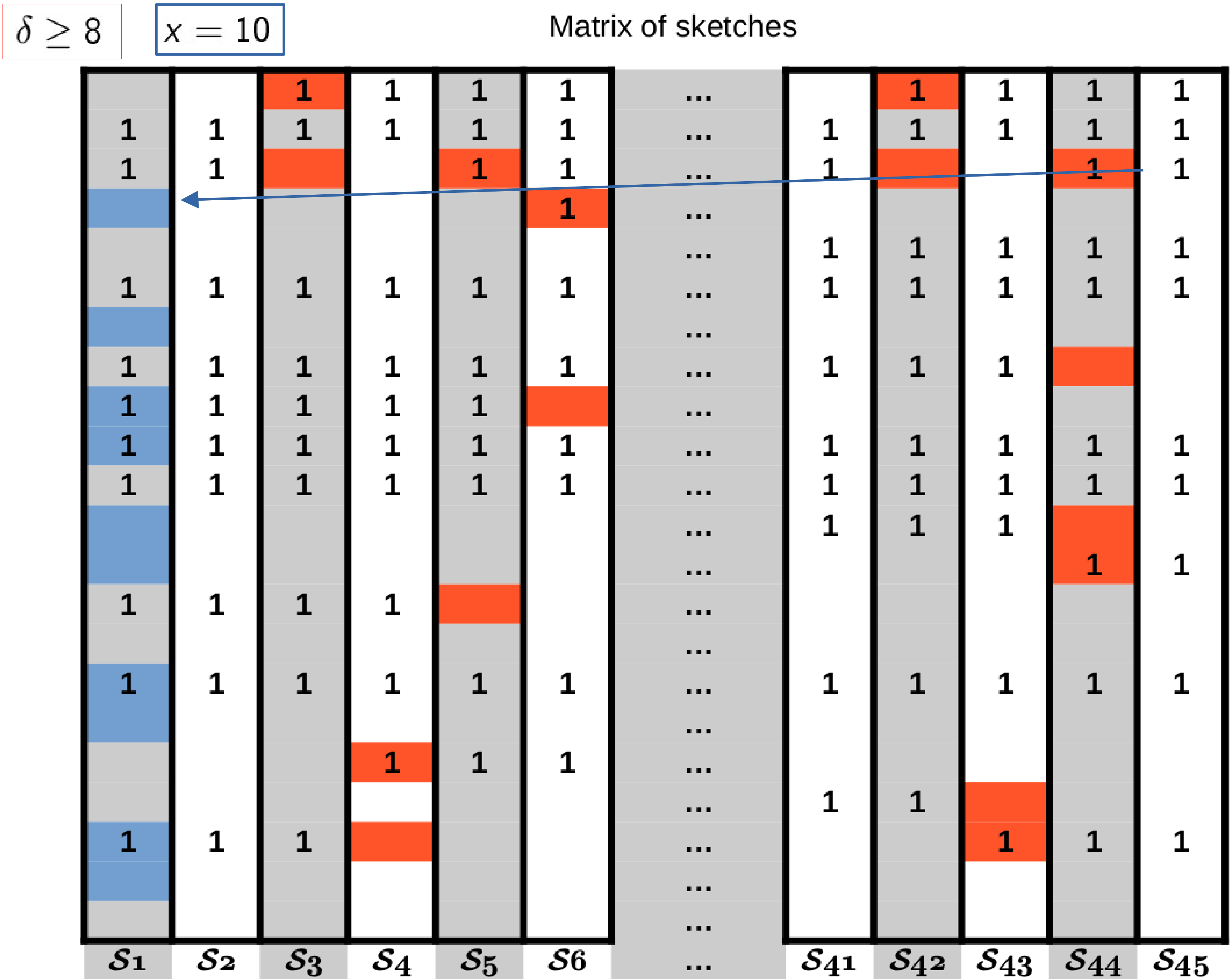
δ = 2
Matrix of sketches
S1
S2
S3
S4
S5
S6
...
S41
S42
S43
S44
S45


$\delta \geq 8$
Matrix of sketches
$\mathcal{S}_1$
$\mathcal{S}_2$
$\mathcal{S}_3$
$\mathcal{S}_4$
$\mathcal{S}_5$
$\mathcal{S}_6$
...
$\mathcal{S}_{41}$
$\mathcal{S}_{42}$
$\mathcal{S}_{43}$
$\mathcal{S}_{44}$
$\mathcal{S}_{45}$

Line changes

Number of bit changes upon line skips :

$$x = \sum_{i=1}^{|\mathcal{S}_U|-1} |\mathcal{M}_{i,|\mathcal{S}|} - \mathcal{M}_{i+1,1}|$$


$\delta \geq 8$
$x = 10$
Matrix of sketches
$\mathcal{S}_1$
$\mathcal{S}_2$
$\mathcal{S}_3$
$\mathcal{S}_4$
$\mathcal{S}_5$
$\mathcal{S}_6$
...
$\mathcal{S}_{41}$
$\mathcal{S}_{42}$
$\mathcal{S}_{43}$
$\mathcal{S}_{44}$
$\mathcal{S}_{45}$

- $2\delta + x$ bit changes in RLE
- $x \leq 2s - 1$

So at most $2\delta + 2s - 1$ bit changes.

Size of output :

$$|\mathcal{S}_U| + |RLE(\mathcal{M})|$$

- $2\delta + x$ bit changes in RLE
- $x \leq 2s - 1$

So at most $2\delta + 2s - 1$ bit changes.

Size of output :

$$\boxed{|\mathcal{S}_U|} + |RLE(\mathcal{M})|$$

$$\boxed{64s \leq |\mathcal{S}_U| \leq 64(s + \delta)}$$

- 64 bits for each hash
- between s and $s + \delta$ hashes

- $2\delta + x$ bit changes in RLE
- $x \leq 2s - 1$

So at most $2\delta + 2s - 1$ bit changes.

Size of output :

$$\boxed{|\mathcal{S}_U|} + \boxed{|RLE(\mathcal{M})|}$$

$$\boxed{64s \leq |\mathcal{S}_U| \leq 64(s + \delta)} \qquad \boxed{(1 + 2\delta + x)(|bit| + |int|)}$$

- $2\delta + x$ bit changes in RLE
- $x \leq 2s - 1$

So at most $2\delta + 2s - 1$ bit changes.

Size of output :

$$\boxed{|\mathcal{S}_U|} + \boxed{|RLE(\mathcal{M})|}$$

$$\boxed{64s \leq |\mathcal{S}_U| \leq 64(s + \delta)}$$

$$\boxed{(1 + 2\delta + x)(|bit| + |int|)}$$

$$\log_2(s \cdot |\mathcal{S}|)$$

For $s \leq 16000$, $|\mathcal{S}| \leq 131000$, $\log_2(s \cdot |\mathcal{S}|) \leq 31$

- $2\delta + x$ bit changes in RLE
- $x \leq 2s - 1$

So at most $2\delta + 2s - 1$ bit changes.

Size of output :

$$\boxed{|\mathcal{S}_U|} + \boxed{|RLE(\mathcal{M})|}$$

$$\boxed{64s \leq |\mathcal{S}_U| \leq 64(s + \delta)}$$

$$\boxed{(1 + 2\delta + x)(|bit| + |int|)}$$

$$\leq 64(\delta + s)$$

Size of input :

$$\geq 64 \cdot |\mathcal{S}| \cdot s$$

- 64 bits per hash
- s hashes per sketch
- $|\mathcal{S}|$ sketches in $\mathcal{S}$

Compression ratio r :

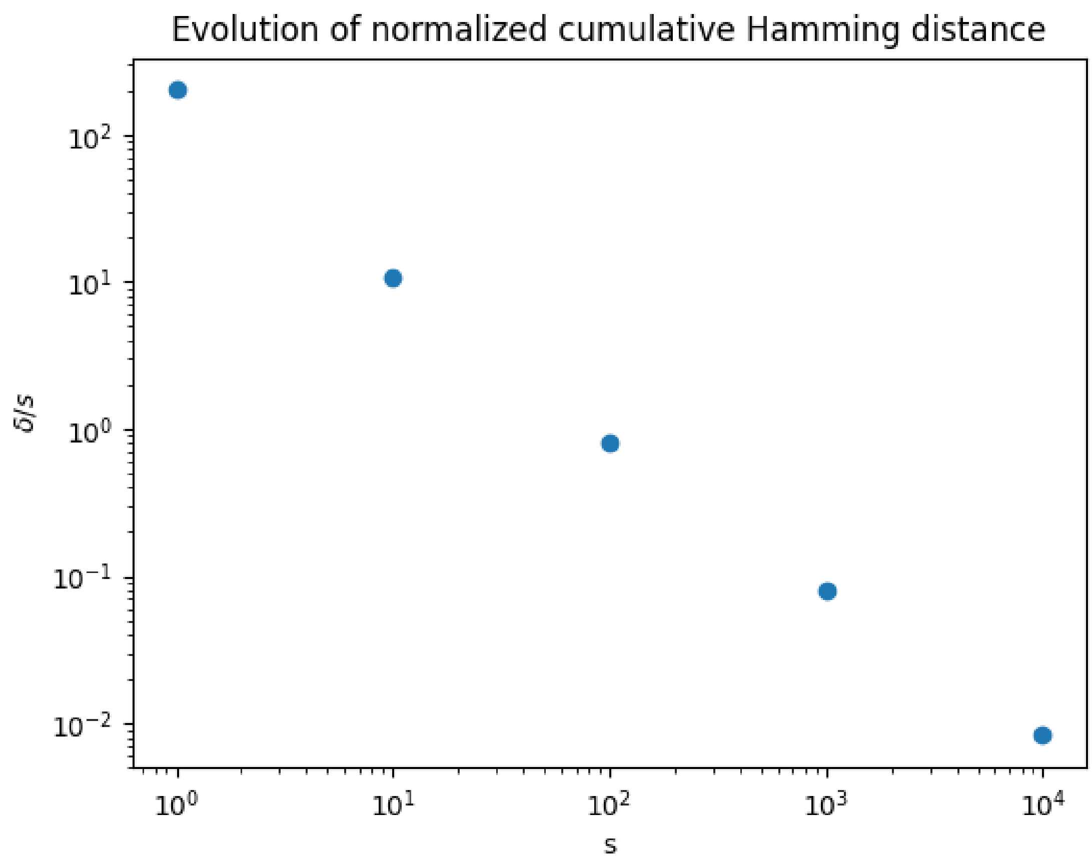
$$r = \frac{|Input|}{|Output|}$$

$$\geq \frac{64 \cdot |\mathcal{S}| \cdot s}{2 \cdot 64(\delta + s)}$$

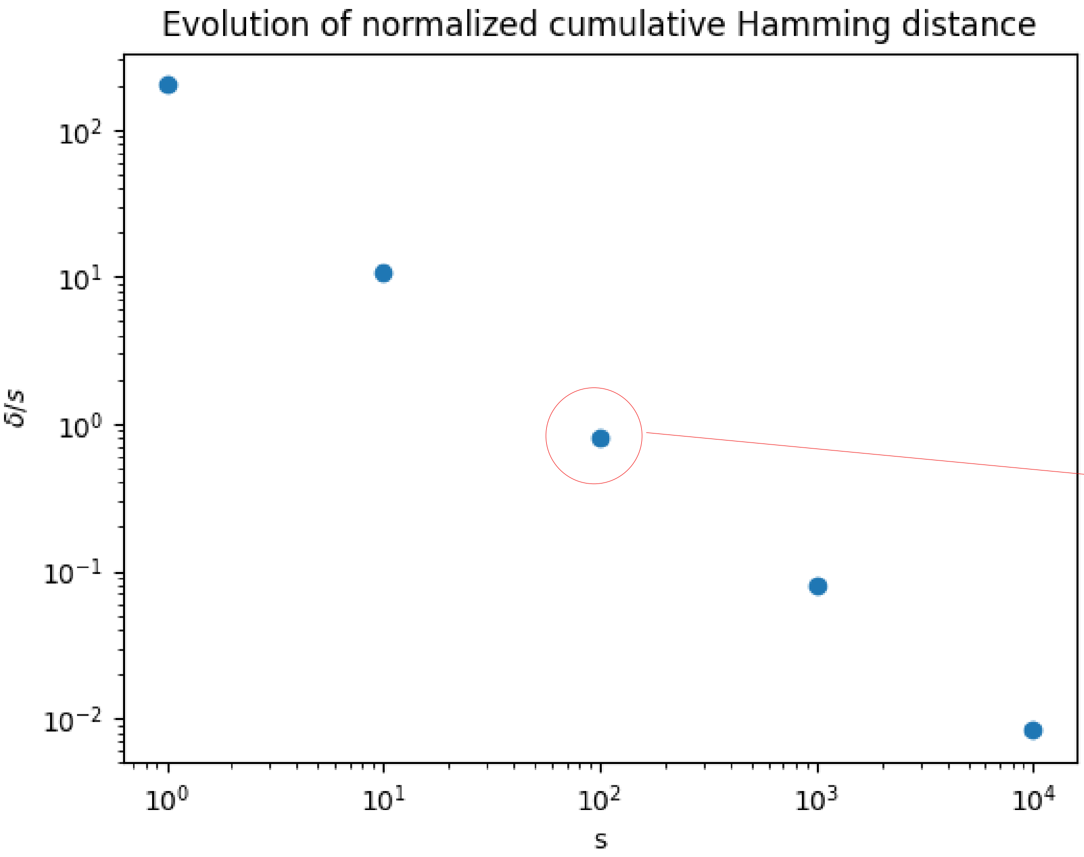
$$\geq \frac{|\mathcal{S}|}{2(1 + \frac{\delta}{s})}$$

Experimental evaluation

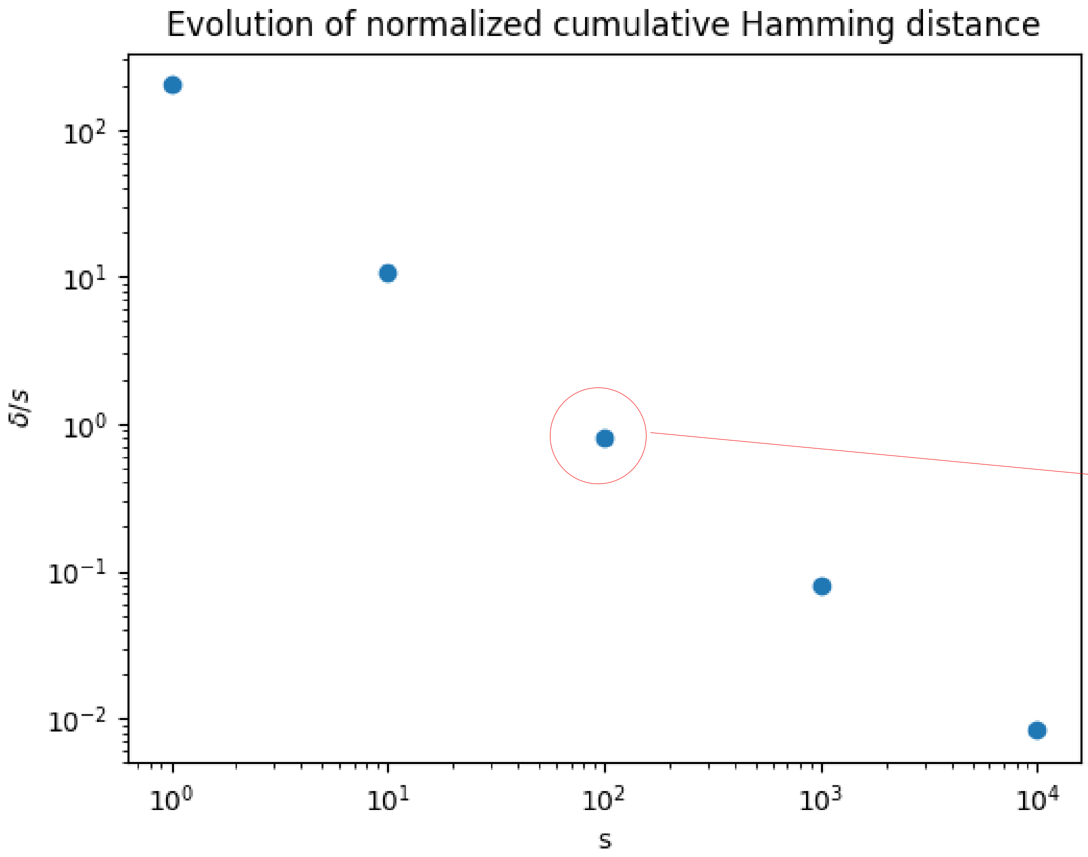
- Genomes to sketch : neisseria_gonorrhoeae__01.tar.xz in part 54 - https://zenodo.org/records/15367750
- Values of s : 1, 10, 100, 1000



Evolution of normalized cumulative Hamming distance
δ/s
10^2
10^1
10^0
10^{-1}
10^{-2}
10^0
10^1
10^2
10^3
10^4
s



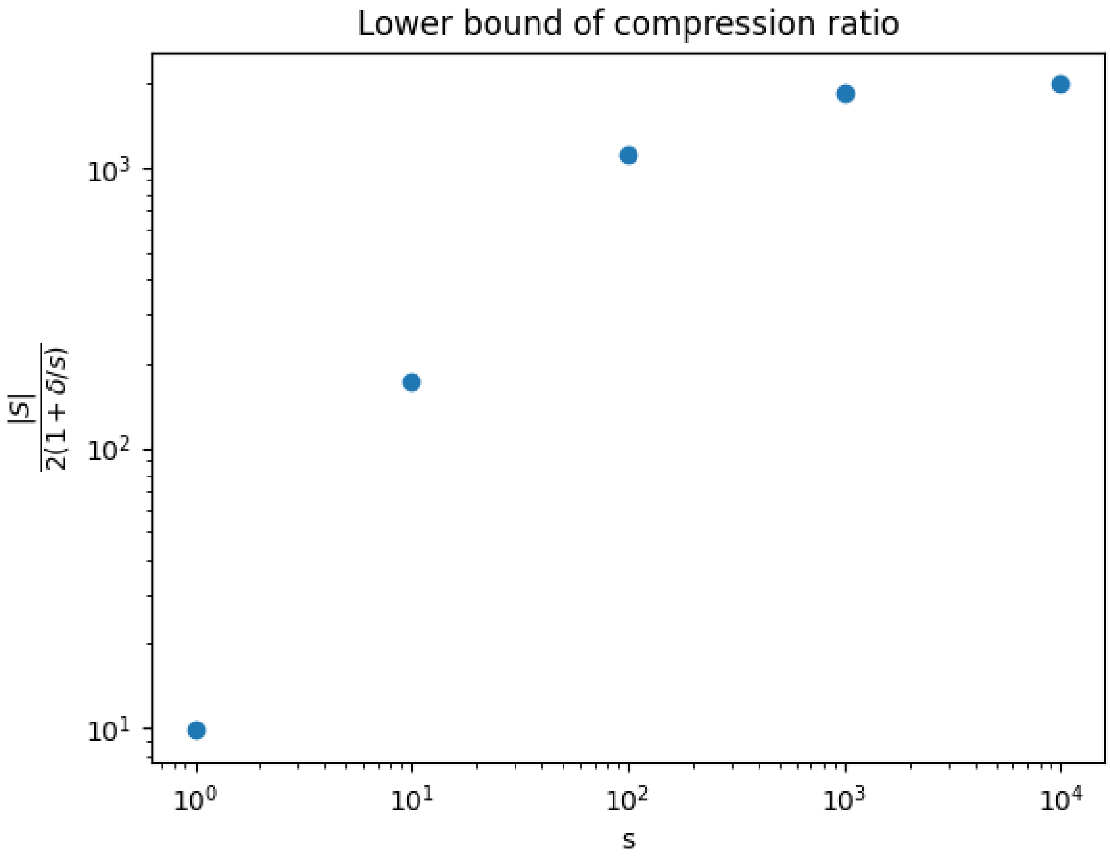
$$\frac{\delta}{s} \leq 1$$



$$\frac{\delta}{s} \leq 1$$

$$\boxed{r \geq \frac{4000}{4} \geq 1000}$$

!!!



Lower bound of compression ratio
$\frac{|S|}{2(1+\delta/s)}$
10^1
10^2
10^3
10^0
10^1
10^2
10^3
10^4
s

Conclusions :

- Compression rate of RLE increases with s for ngono
- Much better compression rates than with XZ/GZIP !

To do :

- Look into biological model of $\frac{\delta}{s}$
- Finer lower bound of r
- Compress $\mathcal{S}_U$ (Elias Fano ?)
- Implement compression