On the structure of the presence/absence matrix

Elias-Fano and RLE

Marie Picard

–

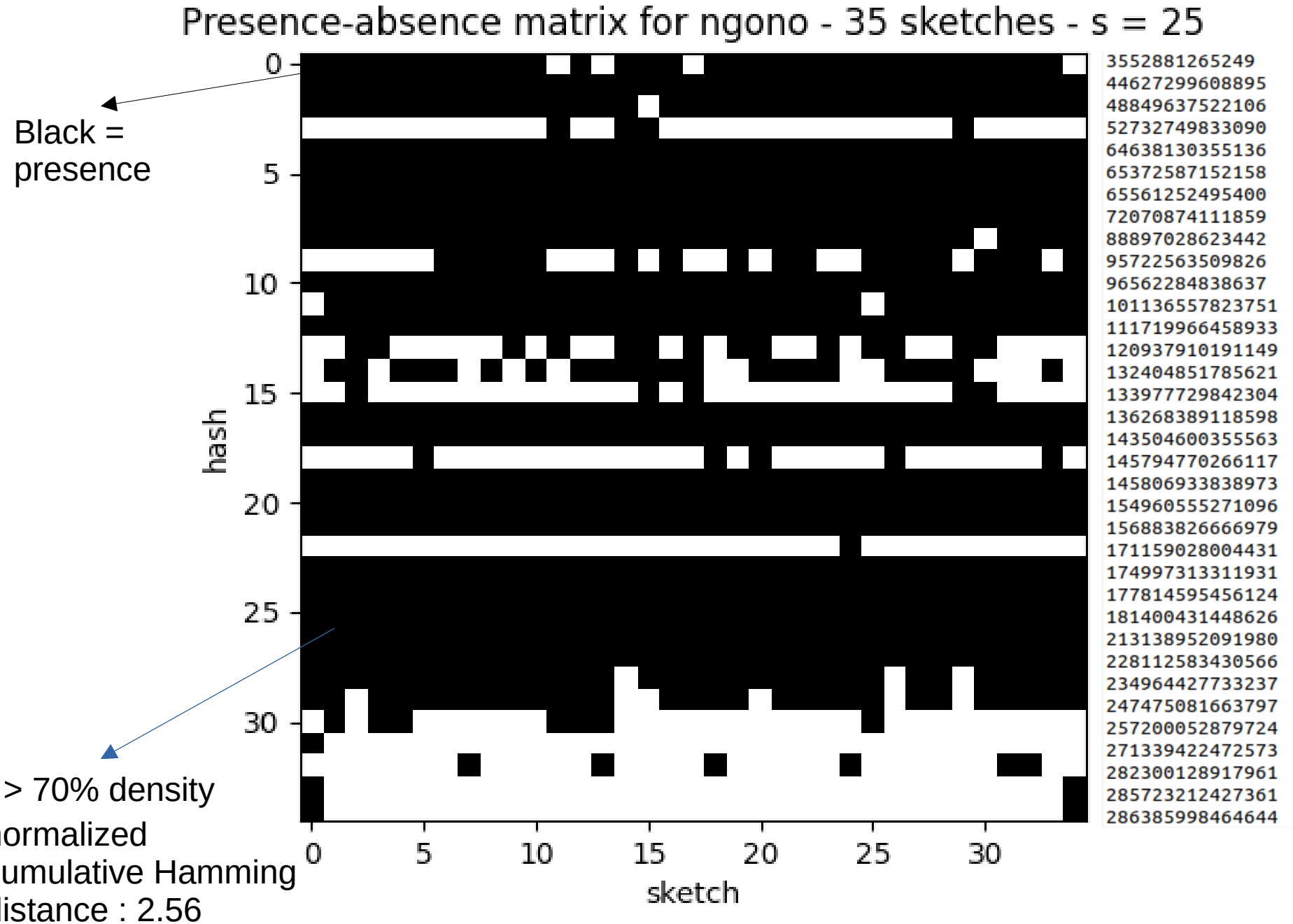
Supervisors : Karel Břinda, Leo Ackermann

December 11, 2025

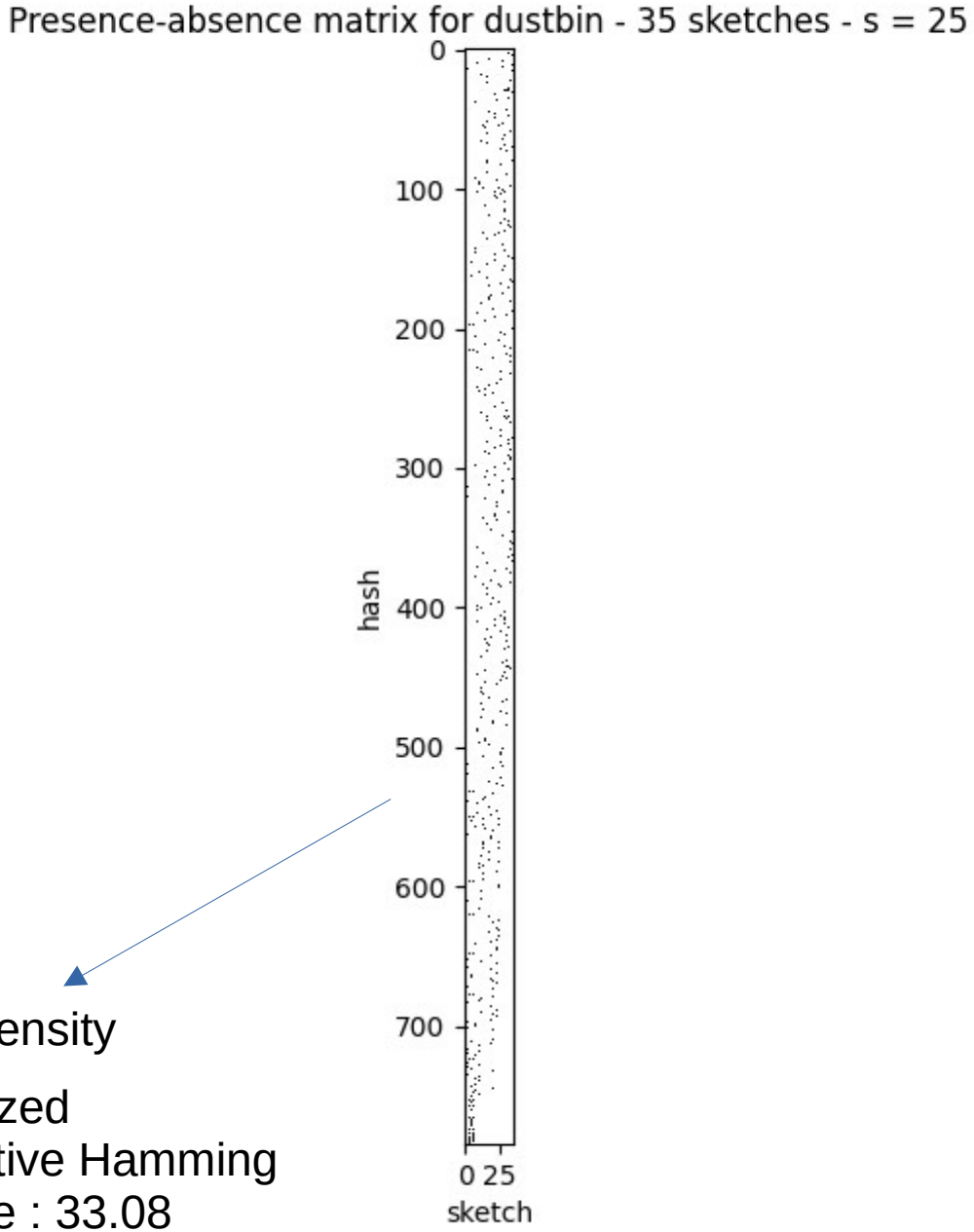
Last time’s results

- |S| = 35 (number of genomes sketched)
- s = 25 (number of hashes per sketch)
- Types of genomes sketched :
  - Neisseria gonorrhoeae (part 54, n°01)
  - Dustbin (part 24, n°23)
- Phylogenetic order

Last time's results



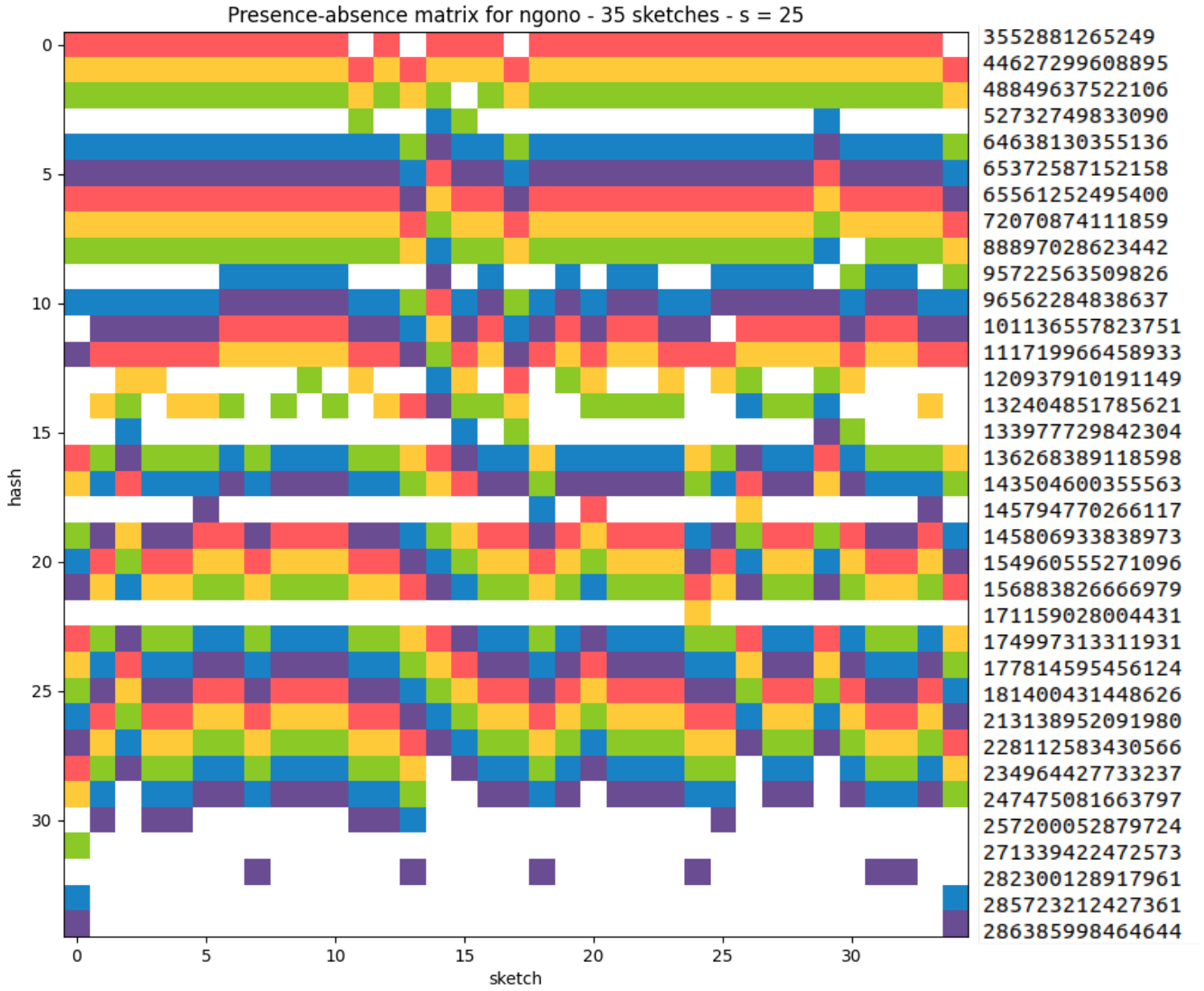
Last time’s results

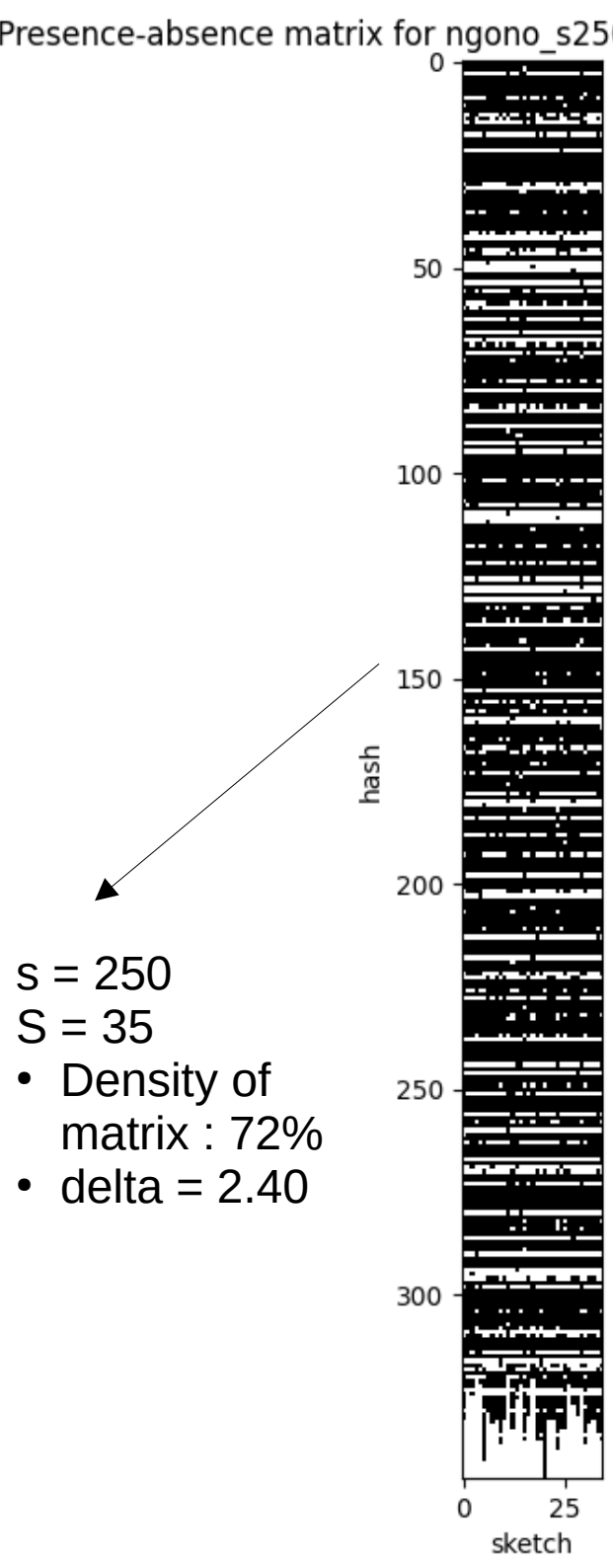


~ 3% density

normalized cumulative Hamming distance : 33.08

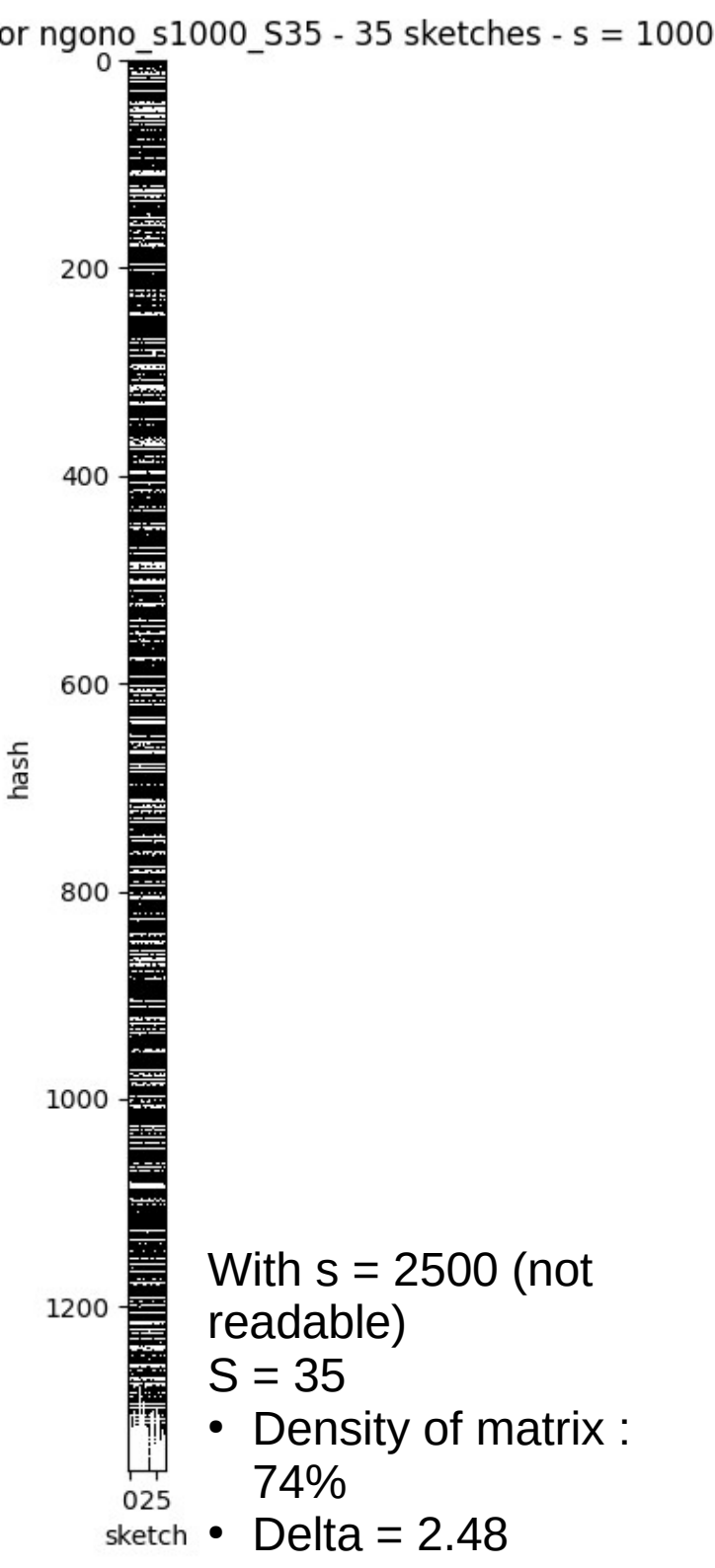
Structure of matrices





s = 250
S = 35

- Density of matrix : 72%
- delta = 2.40

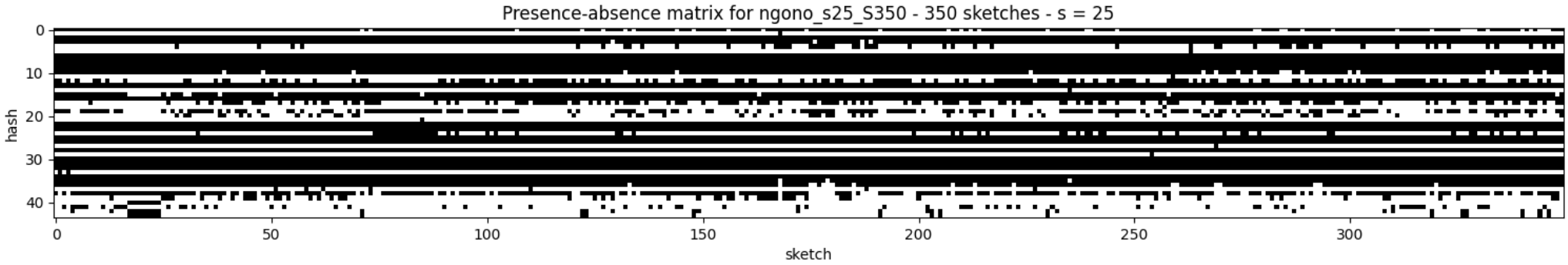


s = 1000
S = 35

- Density of matrix : 74%
- delta = 2.32

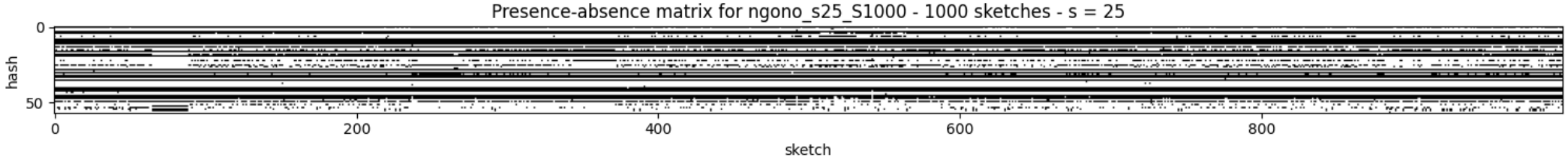
With s = 2500 (not readable)
S = 35

- Density of matrix : 74%
- Delta = 2.48



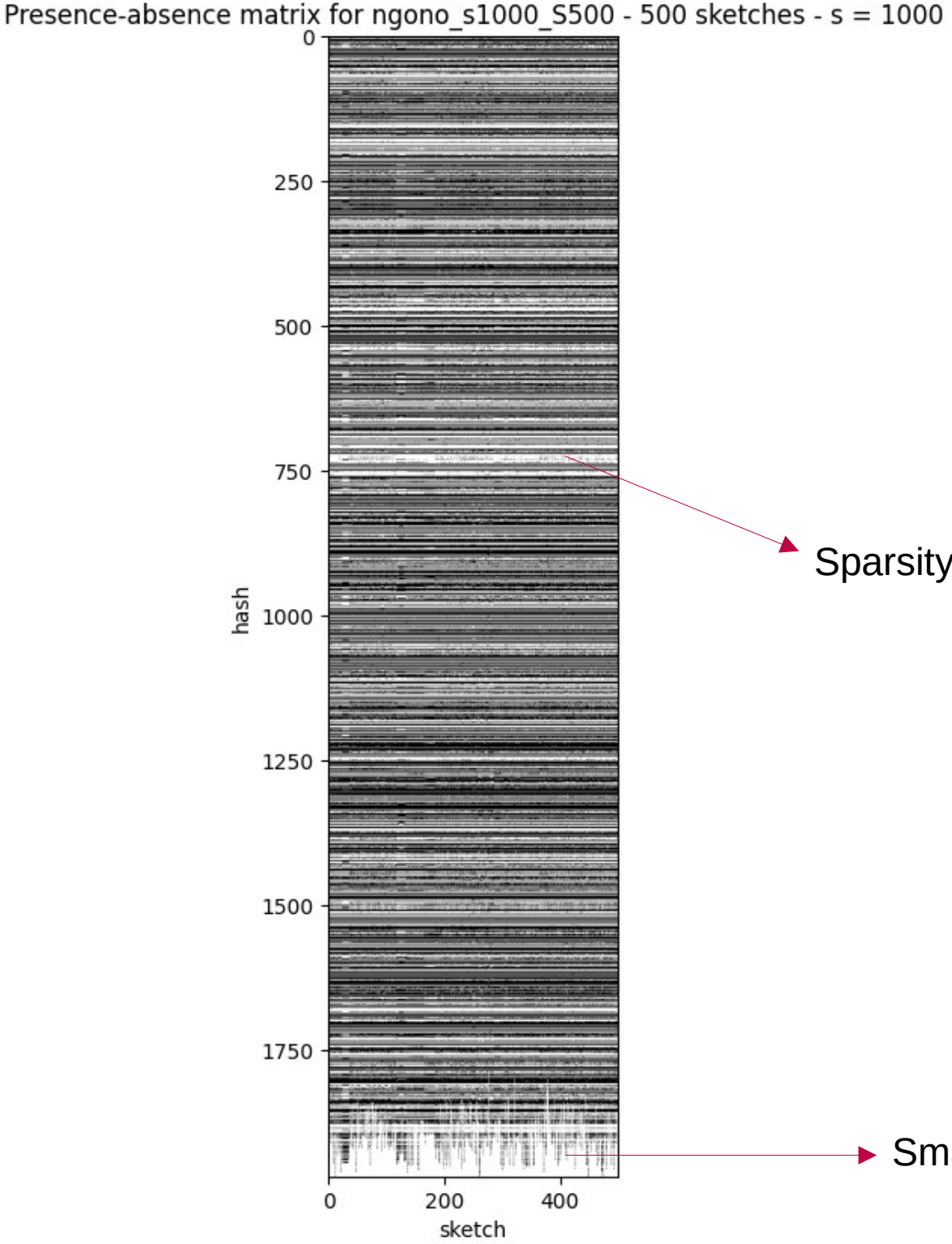
S = 350 - s = 25

- Density of matrix : 57%
- Delta = 24.0



S = 1000 - s = 25

- Density of matrix : 44%
- Delta = 65.36



S = 500 - s = 1000

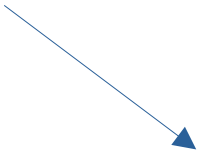
- Density of matrix : 51%
- Delta = 29.754

Sparsity comes from sparse rows

Small sparse area at the end

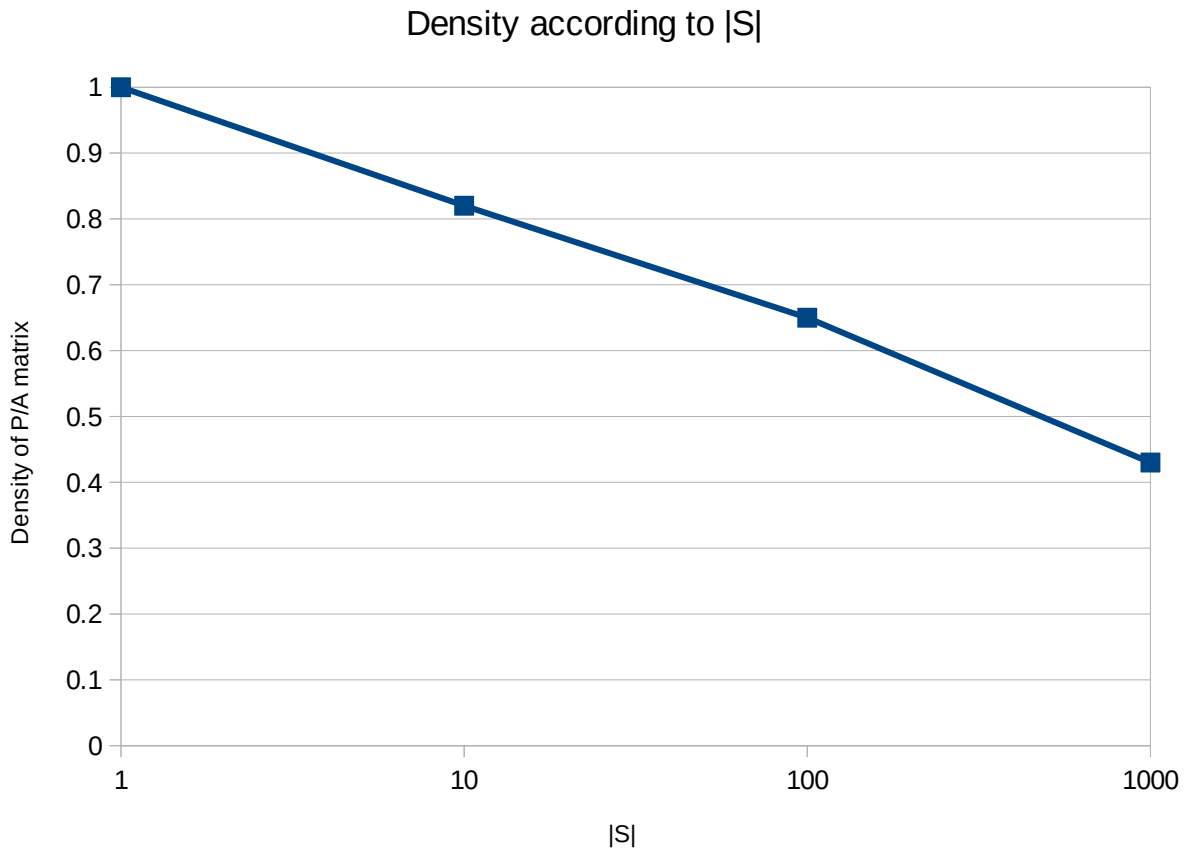
Conjectures :

- δ and density constant in s
- δ and density decreasing proportionately with $|\mathcal{S}|$
- constant compression ratio



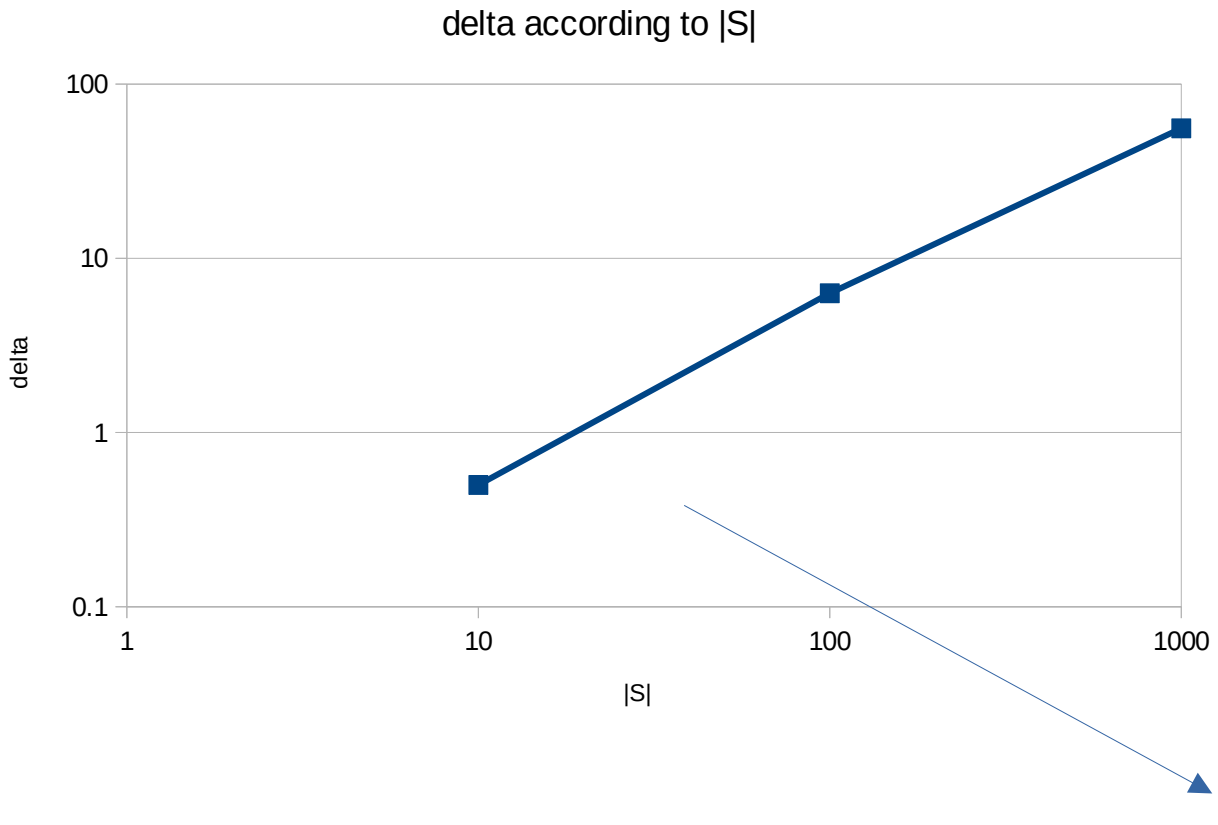
$$\frac{|\mathcal{S}|}{\delta}$$

Density according to |S|



- s = 1000
- Neisseria gonorrhoeae

Delta according to |S|



- s = 1000
- Neisseria gonorrhoeae

Conclusion

Conclusions :

- ~ constant compression rate
- Sparse rows

Todo :

- Increase |S|
- Look into Elias-Fano more closely (XZ 2 to 3 times more efficient)