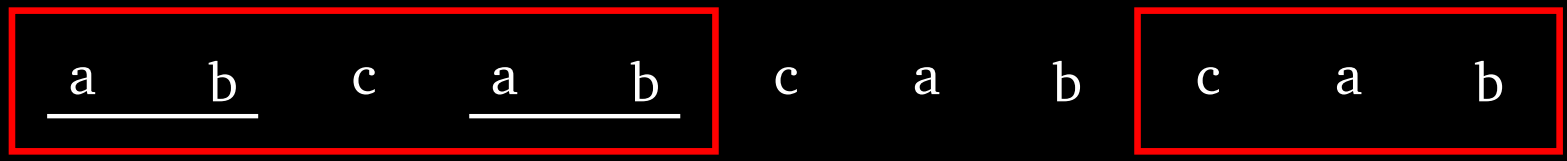
On Maximal Unbordered Factors

Alexander Loptev
Higher School of Economics, Russia

Gregory Kucherov
LIGM, Université Paris-Est Marne-la-Vallée & CNRS, France

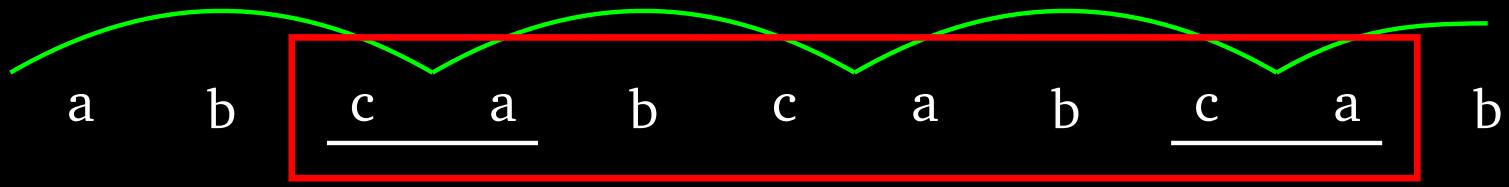
Tatiana Starikovskaya
University of Bristol, United Kingdom



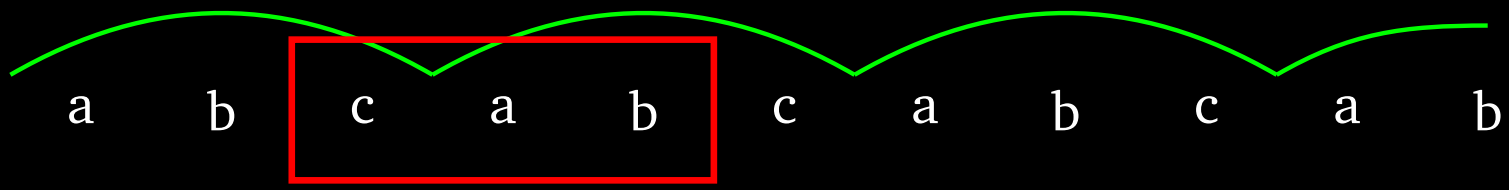
bordered factor:
prefix = suffix

border

unbordered factor:
$\nexists$ prefix = suffix



min period $\geq$ |max unbordered factor|

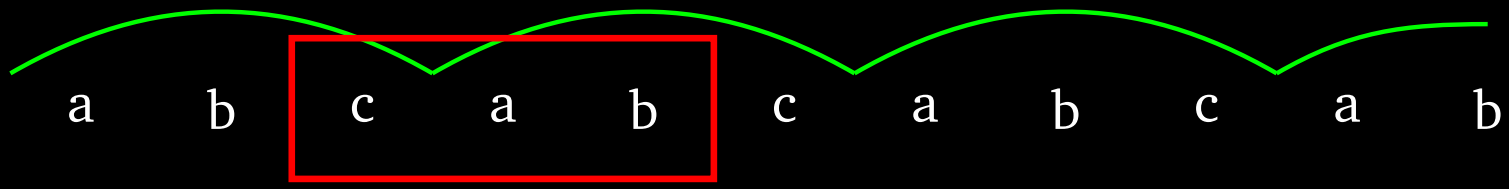


When min period $=$ |max unbordered factor| ($\clubsuit$) ?

Theorem: min period $\leq 0.5n \Rightarrow (\clubsuit)$

Conjecture: |max unbordered factor| $\leq 0.5n \Rightarrow (\clubsuit)$

[Ehrenfeucht, Silberger 1979]



When min period $=$ |max unbordered factor| ($\clubsuit$) ?

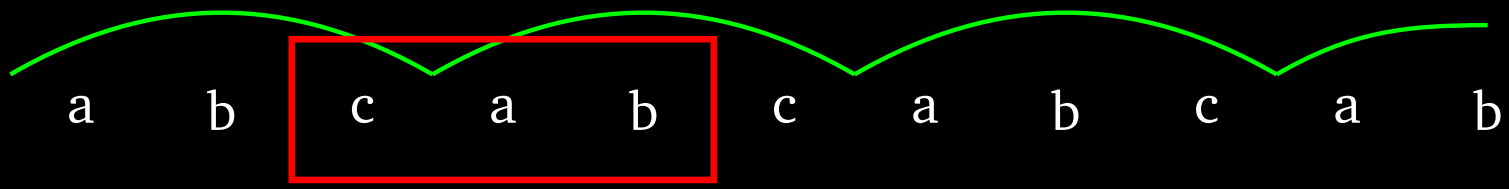
$a^m b\ a^{m+1} b\ a^m b\ a^{m+2} b\ a^m b\ a^{m+1} b\ a^m$

$n = 7m + 10$

min period $= 4m + 7$

|max unbordered factor| $= 3m + 6 < \frac{1}{2} n$ $(a^{m+2} b\ a^m b\ a^{m+1} b)$

[Assous, Pouzet 1979]



When min period $=$ $|$max unbordered factor$|$ ($\clubsuit$) ?

$|\text{max unbordered factor}| \leq \frac{3}{7}n \Rightarrow (\clubsuit)$ [Holub, Nowotka 2012]

a	b	c	a	b	c	a	b	c	a	b

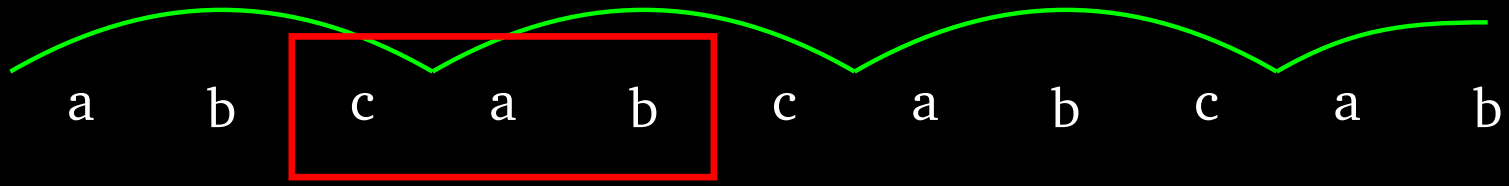
How to compute max length of unbordered factor?

Border arrays: $\mathcal{O}(n^2)$-time, $\mathcal{O}(n)$-space [Knuth, Morris, Pratt 1977]

- i^{th} entry $=$ max border of i^{th} prefix

0	0	0	1	2	3	4	5	6	7	8
1	2	3	4	5	6	7	8	9	10	11

- i^{th} entry is $0 \Rightarrow i^{th}$ prefix is unbordered
- Rightmost $0 \Rightarrow$ max length of unbordered prefix
- Border array can be built in $\mathcal{O}(n)$ time

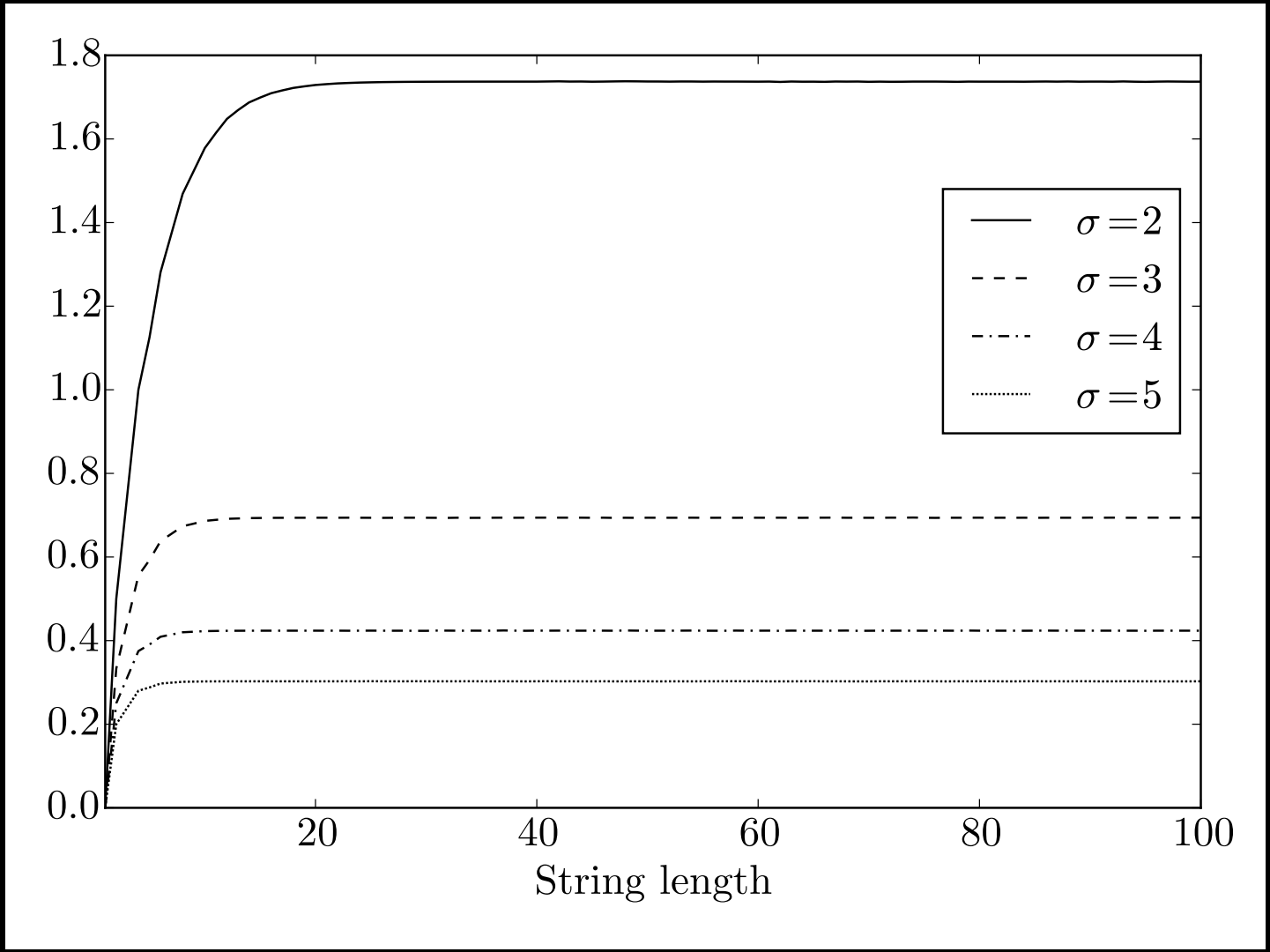


How to compute max length of unbordered factor?

($\clubsuit$) $\Rightarrow$ $\mathcal{O}(n)$-time, $\mathcal{O}(n)$-space [Duval et al. 2014]

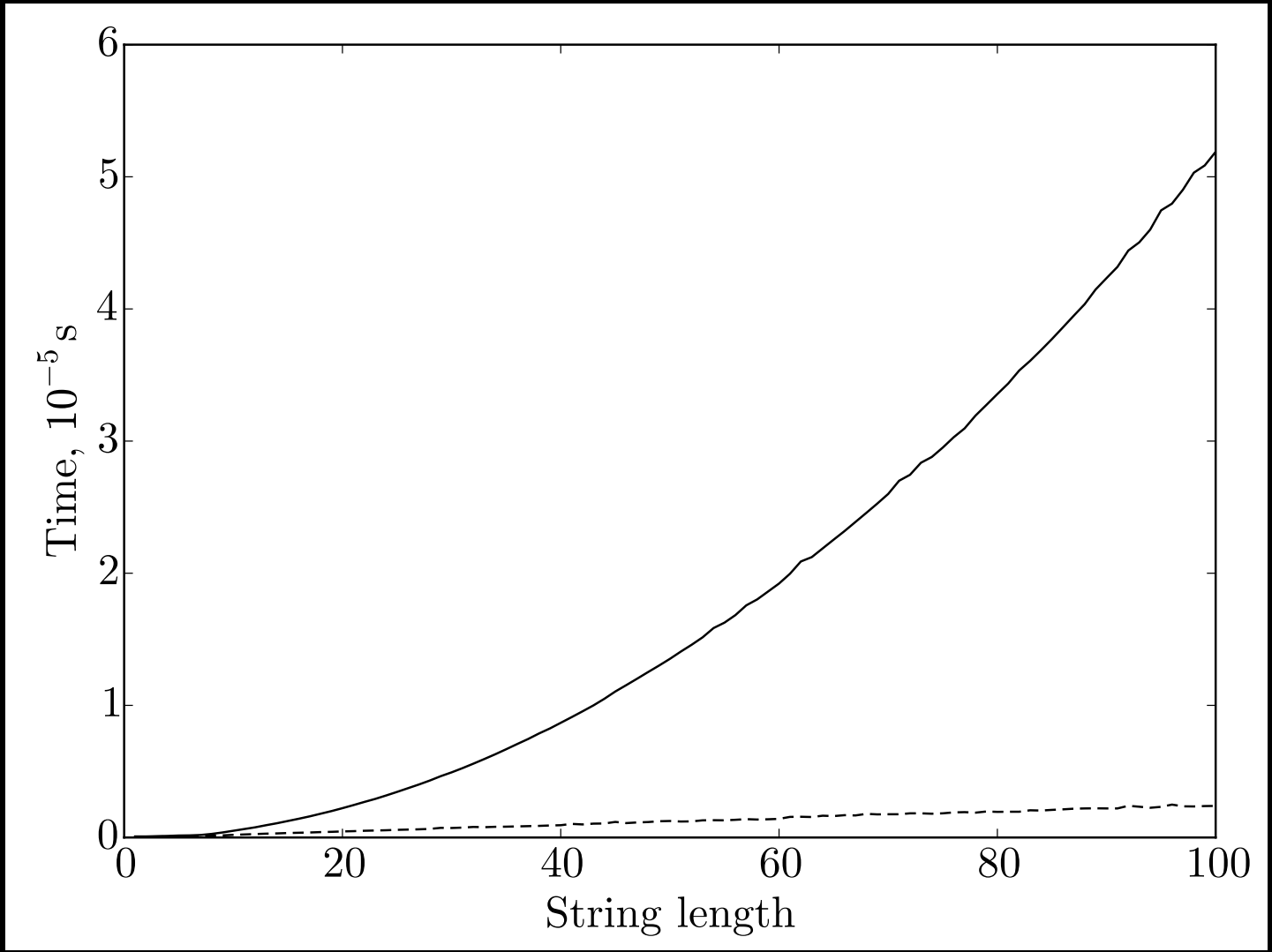
- Longest unbordered factor $=$ cyclic shift of min period
- Min period can be computed in $\mathcal{O}(n)$ time and space
- Unbordered cyclic shift can be computed in $\mathcal{O}(n)$ time and space

Exp. difference between n and $|$max unbordered factor$|$



Conjecture: $\mathbb{E}[|\text{max unbordered factor}|] = n - \mathcal{O}(1)$

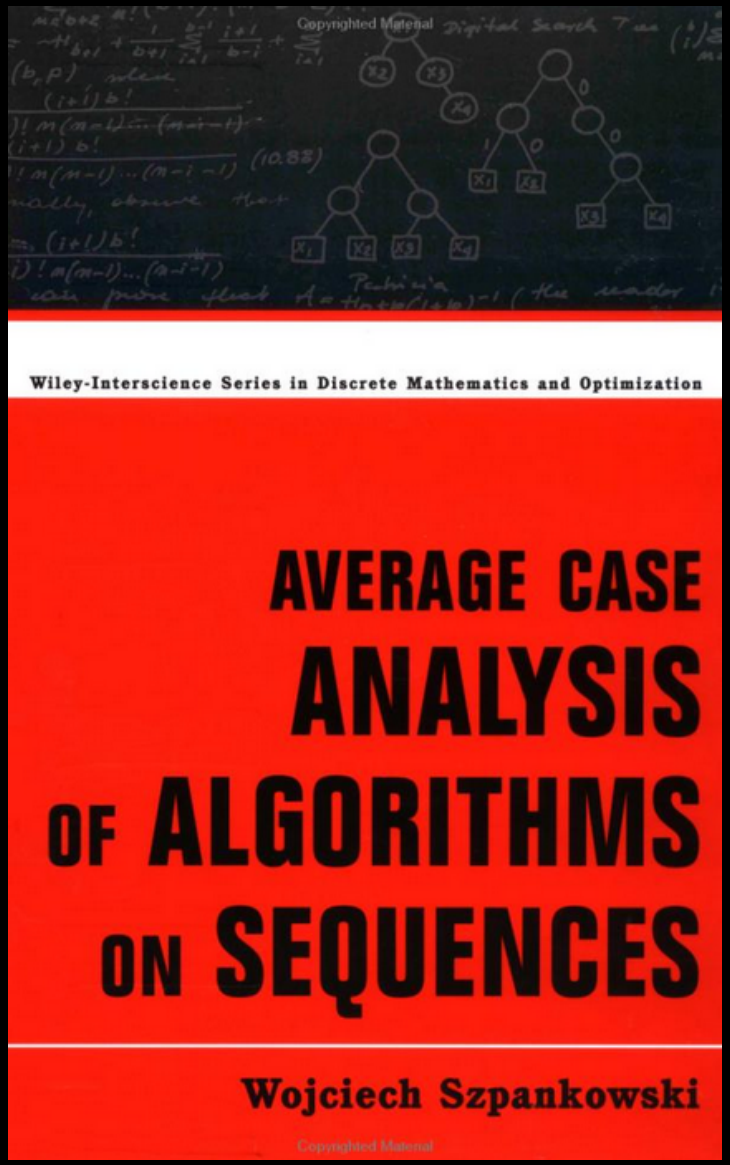
Border-array algorithm vs. ”improved” border-array algorithm



Improvement: Stop when $|\text{suffix}| \leq |\text{max unbordered factor}|$

$$\mathbb{E}[|\text{max unbordered factor}|] \geq (1 - \tfrac{4}{\sigma^4}) \cdot n \text{ for } \sigma \geq 5$$

$$\mathbb{E}[\text{time}] = \mathcal{O}(\tfrac{n^2}{\sigma^4} + n) \text{ for } \sigma \geq 5$$



- Pattern matching
- Edit distance
- LZ factorizations
- Shortest common superstring
- . . .

To prove that $\mathbb{E}[|\text{max unbordered factor}|]$ is big, we need many strings with long unbordered factors.

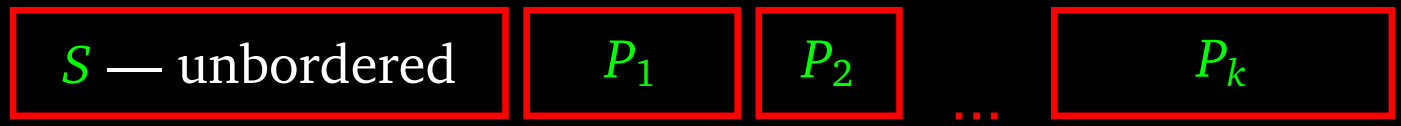
of i-length unbordered strings $\geq \sigma^i - \sigma^{i-1} - \sigma^{i-2}$ [Nielsen 1973]

To prove that $\mathbb{E}[|\text{max unbordered factor}|]$ is big, we need many strings with long unbordered factors.

of i-length unbordered strings $\geq \sigma^i - \sigma^{i-1} - \sigma^{i-2}$ [Nielsen 1973]

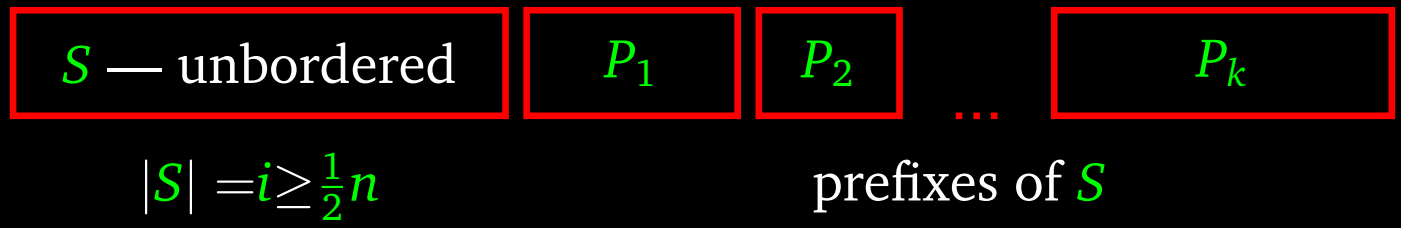
of strings with unbordered factors of length $n \geq \sigma^n - \sigma^{n-1} - \sigma^{n-2}$

$$\mathbb{E}[|\text{max unbordered factor}|] \geq (1 - \tfrac{1}{\sigma} - \tfrac{1}{\sigma^2}) \cdot n$$

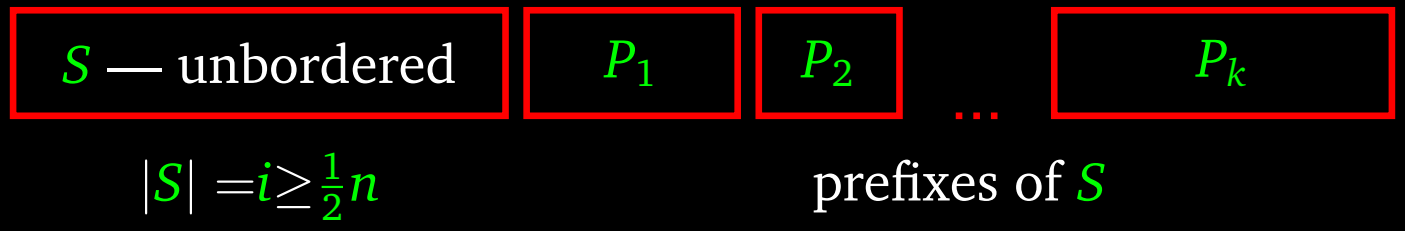


$|S| = i \geq \frac{1}{2}n$

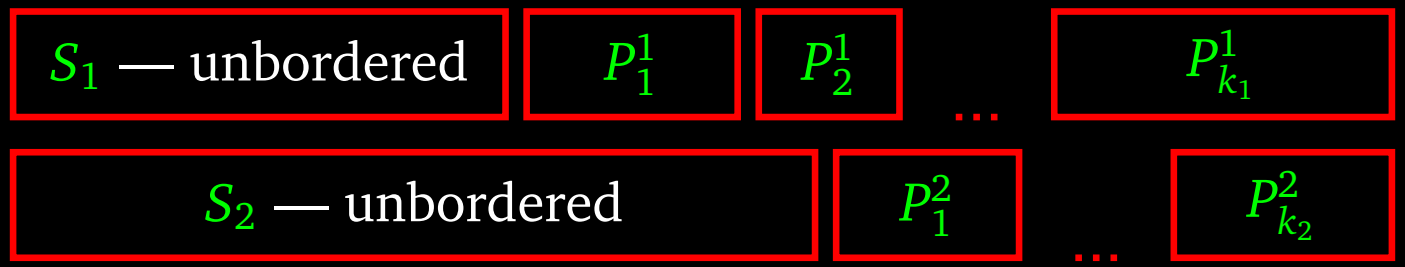
prefixes of S

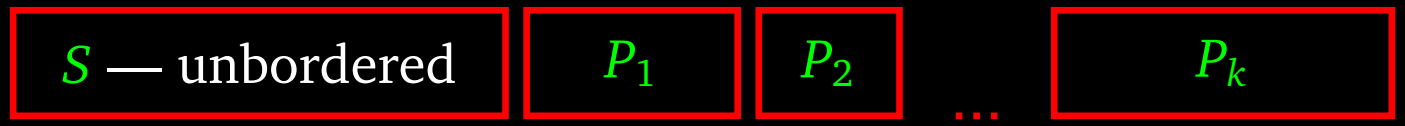


- $S \to 2^{n-1-i}$ strings with max unbordered factor of length $\geq i$



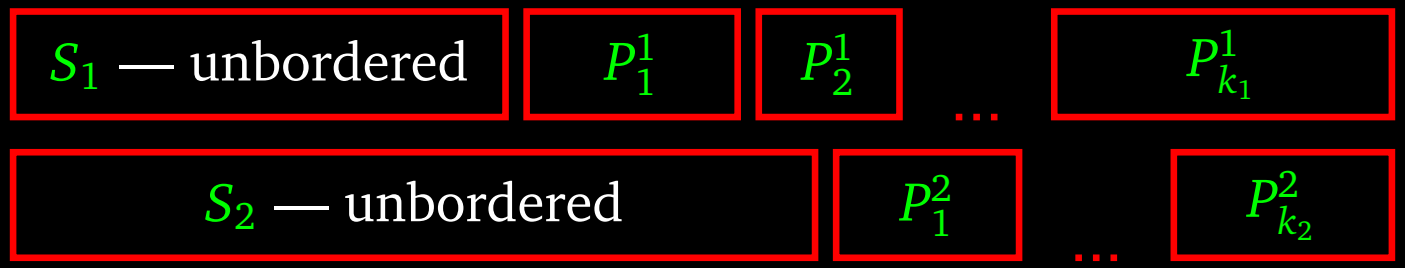
- $S \to 2^{n-1-i}$ strings with max unbordered factor of length $\geq i$
- $S_1 \neq S_2$ — unbordered $\Rightarrow$ any two gen. strings are distinct



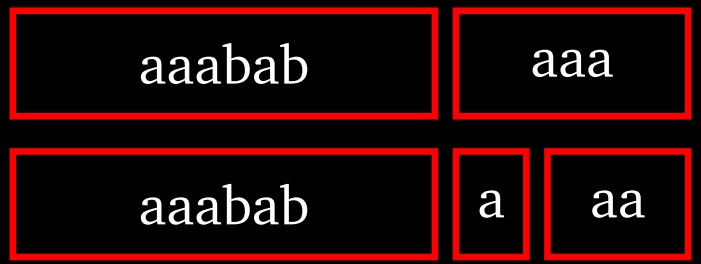


$|S| = i \geq \frac{1}{2}n$ prefixes of S

- $S \to 2^{n-1-i}$ strings with max unbordered factor of length $\geq i$
- $S_1 \neq S_2$ — unbordered $\Rightarrow$ any two gen. strings are distinct



- We can produce two equal strings from S



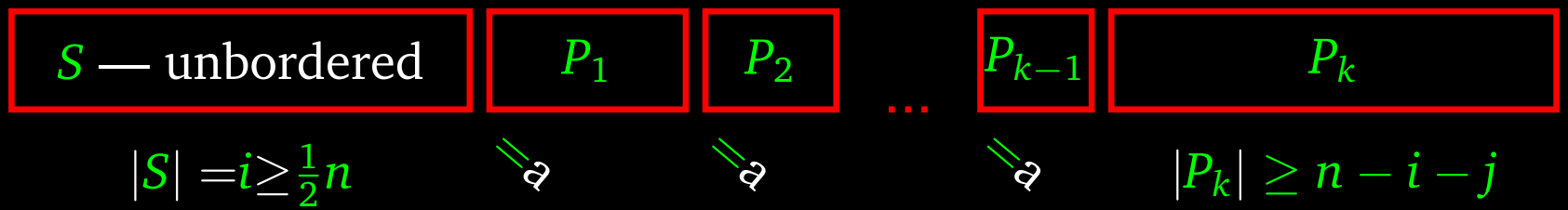
S — unbordered

- $S[2], S[3], \ldots, S[j+1] \neq S[1]$

S — unbordered

- $S[2], S[3], \ldots, S[j+1] \neq S[1]$
- # of such unbordered strings is big for all j

- $S[2], S[3], \ldots, S[j+1] \neq S[1]$
- # of such unbordered strings is big for all j



- $|P_1|, |P_2|, \ldots, |P_{k-1}| \leq j + 1$
- Each S gives $\geq 2^j$ distinct strings

Summary

- Method of generating strings with long unbordered factors
- Lower bound on expected length of max unbordered factor
- Lower bound on min period
- Algorithm for computing max unbordered factor
- Library for unbordered factors:
  `http://github.com/avlonger/unbordered`
- **Conjecture:** $\mathbb{E}[|\text{max unbordered factor}|] = n - \mathcal{O}(1)$

Stay tuned!

- P. Gawryhowski, G. Kucherov, B. Sach, T. Starikovskaya.
  Computing the longest unbordered substring.
  Accepted to SPIRE 2015.

- $\mathcal{O}(n \log n)$-time algorithm (av.-case complexity)

- $\mathcal{O}(n^{1.5})$-time algorithm (worst-case complexity)