Constructions for Clumps Statistics

P. Nicodème

CNRS - LIX École polytechnique, Palaiseau

Joint work with F. Bassino, J. Clément and J. Fayolle

00100

00100
111100

00100
111100
10111111100

00100
111100
1011111100
01100

00100
111100
1011111100
01100
0100

00100
111100
1011111100
01100
0100
11100

00100
111100
10111111100
01100
0100
11100
11100

00100
111100
10111111100
01100
0100
11100
11100
010100

00100
111100
10111111100
01100
0100
11100
11100
010100
0110100

00100
111100
1011111111100
01100
0100
11100
11100
010100
0110100
00100

00100
111100
10111111100
01100
0100
11100
11100
010100
0110100
00100
10111111010111111100

00100
111100
10111111100
01100
0100
11100
11100
010100
0110100
00100
10111111010111111100
00010100

00100
111100
10111111100
01100
0100
11100
11100
010100
0110100
00100
101111110101111111100
00010100
0101100

00100
111100
10111111100
01100
0100
11100
11100
010100
0110100
00100
1011111101011111100
00010100
0101100

100010100111

00100
111100
10111111100
01100
0100
11100
11100
010100
0110100
00100
10111111010111111100
00010100
0101100

100010100111
1100010100111

00100
111100
10111111100
01100
0100
11100
11100
010100
0110100
00100
10111111010111111100
00010100
0101100

100010100111
1100010100111
1000010111

00100
111100
10111111100
01100
0100
11100
11100
010100
0110100
00100
10111111010111111100
00010100
0101100

100010100111
1100010100111
1000010111
0011000011000000111

00100
111100
1011111111100
01100
0100
11100
11100
010100
0110100
00100
1011111101011111100
00010100
0101100

100010100111
1100010100111
1000010111
001100001100000000111
0101010101010000111

00100
111100
10111111100
01100
0100
11100
11100
010100
0110100
00100
10111111010111111100
00010100
0101100

100010100111
1100010100111
1000010111
0011000011000000111
010101010101000111
110110110011011010100101100110000011010111

00100
111100
10111111100
01100
0100
11100
11100
010100
0110100
00100
10111111010111111100
00010100
0101100

100010100111
1100010100111
1000010111
0011000001100000111
010101010101000111
110110110011011010100101100110000011010111
0111

00100
111100
10111111100
01100
0100
11100
11100
010100
0110100
00100
10111111010111111100
00010100
0101100

100010100111
1100010100111
1000010111
0011000011000000111
010101010101000111
110110110011011010100101100110000011010111
0111
111

00**100**
111**100**
10111111**100**
01**100**
0**100**
11**100**
11**100**
010**100**
0110**100**
00**100**
10111111010111111**100**
00010**100**
0101**100**

100010100**111**
1100010100**111**
1000010**111**
001100000011000000**111**
01010101010101000**111**
1101101100110110101001011001100000011010**111**
0**111**
111
001100010**111**

00100
111100
10111111100
01100
0100
11100
11100
010100
0110100
00100
10111111010111111100
00010100
0101100

100010100111
1100010100111
1000010111
0011000011000000111
010101010101000111
110110110011011010100101100110000011010111
0111
111
001100010111
0100010010010100000100111

00100
111100
10111111100
01100
0100
11100
11100
010100
0110100
00100
10111111010111111100
00010100
0101100

100010100111
1100010100111
1000010111
0011000001100000111
010101010101000111
110110110011011010100101100110000011010111
0111
111
001100010111
0100010010010100000100111
00111

00100
111100
10111111100
01100
0100
11100
11100
010100
0110100
00100
10111111010111111100
00010100
0101100

100010100111
1100010100111
1000010111
0011000011000000111
010101010101000111
110110110011011010100101100110000011010111
0111
111
001100010111
0100010010010100000100111
00111
000101100101010000111

00100
111100
10111111100
01100
0100
11100
11100
010100
0110100
00100
10111111010111111100
00010100
0101100

100010100111
1100010100111
1000010111
0011000011000000111
0101010101010000111
1101101100110110101001011001100000011010111
0111
111
001100010111
0100010010010100000100111
00111
0001011001010100000111
1010101010001010110010101000000010000110111

00100
111100
10111111100
01100
0100
11100
11100
010100
0110100
00100
10111111010111111100
00010100
0101100

100010100111
1100010100111
1000010111
0011000011000000111
010101010101000111
110110110011011010100101100110000011010111
0111
111
001100010111
0100010010010100000100111
00111
000101100101010000111
101010101000101011001010100000010000110111

1	1	1000011100

1	1	1000011100
2	2	0001101000010110

1	1	1000011100
2	2	0001101000010110
3	3	1001110101

1	1	1000011100
2	2	0001101000010110
3	3	1001110101
4		111011110

1	1	1000011100
2	2	0001101000010110
3	3	1001110101
4		111011110
5		111011111

1	1	1000011100
2	2	0001101000010110
3	3	1001110101
4		111011110
5		111011111
6	4	001000101010

1	1	1000011100
2	2	0001101000010110
3	3	1001110101
4		111011110
5		111011111
6	4	001000101010
7	5	11001011011

1	1	1000011100
2	2	0001101000010110
3	3	1001110101
4		111011110
5		111011111
6	4	001000101010
7	5	11001011011
8	6	1001100010

1	1	1000011100
2	2	0001101000010110
3	3	1001110101
4		111011110
5		111011111
6	4	001000101010
7	5	11001011011
8	6	1001100010
9		10111010011

1	1	1000011100
2	2	0001101000010110
3	3	1001110101
4		111011110
5		111011111
6	4	001000101010
7	5	11001011011
8	6	1001100010
9		10111010011
10	7	0011000001010

1	1	1000011100
2	2	0001101000010110
3	3	1001110101
4		111011110
5		111011111
6	4	001000101010
7	5	11001011011
8	6	1001100010
9		10111010011
10	7	001100001010
11		0111110010

1	1	1000011100
2	2	0001101000010110
3	3	1001110101
4		111011110
5		111011111
6	4	001000101010
7	5	11001011011
8	6	1001100010
9		10111010011
10	7	001100001010
11		0111110010
12		11010111100010

1	1	1000011100
2	2	0001101000010110
3	3	1001110101
4		111011110
5		111011111
6	4	001000101010
7	5	11001011011
8	6	1001100010
9		10111010011
10	7	0011000001010
11		011110010
12		1101011100010
13		111010011

1	1	1000011100
2	2	0001101000010110
3	3	1001110101
4		111011110
5		111011111
6	4	001000101010
7	5	11001011011
8	6	1001100010
9		10111010011
10	7	0011000001010
11		0111110010
12		1101011100010
13		111010011
14	8	101001011000

1	1	1000011100
2	2	0001101000010110
3	3	1001110101
4		111011110
5		111011111
6	4	001000101010
7	5	11001011011
8	6	1001100010
9		10111010011
10	7	0011000001010
11		0111110010
12		1101011100010
13		111010011
14	8	101001011000
15	9	001000011011

1	1	1000011100
2	2	0001101000010110
3	3	1001110101
4		111011110
5		111011111
6	4	001000101010
7	5	11001011011
8	6	1001100010
9		10111010011
10	7	0011000001010
11		0111110010
12		11010111100010
13		111010011
14	8	101001011000
15	9	001000011011
16	10	001000011110

1	1	1000011100
2	2	0001101000010110
3	3	1001110101
4		111011110
5		111011111
6	4	001000101010
7	5	11001011011
8	6	1001100010
9		10111010011
10	7	0011000001010
11		0111110010
12		1101011100010
13		111010011
14	8	101001011000
15	9	001000011011
16	10	001000011110
17	11	000011000111111

1	1	1000011100
2	2	0001101000010110
3	3	1001110101
4		111011110
5		111011111
6	4	001000101010
7	5	11001011011
8	6	1001100010
9		10111010011
10	7	0011000001010
11		0111110010
12		11010111100010
13		111010011
14	8	101001011000
15	9	001000011011
16	10	001000011110
17	11	000011000111111
18	12	1001010011

1	1	1000011100
2	2	0001101000010110
3	3	1001110101
4		111011110
5		111011111
6	4	001000101010
7	5	11001011011
8	6	1001100010
9		10111010011
10	7	0011000001010
11		0111110010
12		1101011100010
13		111010011
14	8	101001011000
15	9	001000011011
16	10	001000011110
17	11	000011000111111
18	12	1001010011
19	13	00010101001110001

1	1	1000011100
2	2	0001101000010110
3	3	1001110101
4		111011110
5		111011111
6	4	001000101010
7	5	11001011011
8	6	1001100010
9		10111010011
10	7	0011000001010
11		0111110010
12		11010111100010
13		111010011
14	8	101001011000
15	9	001000011011
16	10	001000011110
17	11	000011000111111
18	12	1001010011
19	13	00010101001110001
20		111101011

1	1	1000011100
2	2	0001101000010110
3	3	1001110101
4		111011110
5		111011111
6	4	001000101010
7	5	11001011011
8	6	1001100010
9		10111010011
10	7	0011000001010
11		011110010
12		1101011100010
13		111010011
14	8	101001011000
15	9	001000011011
16	10	001000011110
17	11	000011000111111
18	12	1001010011
19	13	00010101001110001
20		111101011
21	14	10101001001000

1	1	1000011100
2	2	0001101000010110
3	3	1001110101
4		111011110
5		111011111
6	4	001000101010
7	5	11001011011
8	6	1001100010
9		10111010011
10	7	0011000001010
11		0111110010
12		11010111100010
13		111010011
14	8	101001011000
15	9	001000011011
16	10	001000011110
17	11	000011000111111
18	12	1001010011
19	13	00010101001110001
20		111101011
21	14	10101001001000
22	15	101001111100

1	1	1000011100
2	2	0001101000010110
3	3	1001110101
4		111011110
5		111011111
6	4	001000101010
7	5	11001011011
8	6	1001100010
9		10111010011
10	7	0011000001010
11		0111110010
12		1101011100010
13		111010011
14	8	101001011000
15	9	001000011011
16	10	001000011110
17	11	000011000111111
18	12	1001010011
19	13	00010101001110001
20		111101011
21	14	10101001001000
22	15	101001111100
23		0111010111

1	1	1000011100
2	2	0001101000010110
3	3	1001110101
4		111011110
5		111011111
6	4	001000101010
7	5	11001011011
8	6	1001100010
9		10111010011
10	7	0011000001010
11		0111110010
12		11010111100010
13		111010011
14	8	101001011000
15	9	001000011011
16	10	001000011110
17	11	000011000111111
18	12	1001010011
19	13	00010101001110001
20		111101011
21	14	10101001001000
22	15	101001111100
23		0111010111
24	16	0001001010110

1	1	1000011100
2	2	0001101000010110
3	3	1001110101
4		111011110
5		111011111
6	4	001000101010
7	5	11001011011
8	6	1001100010
9		10111010011
10	7	001100001010
11		0111110010
12		1101011100010
13		111010011
14	8	101001011000
15	9	00100011011
16	10	00100011110
17	11	000011000111111
18	12	1001010011
19	13	00010101001110001
20		111101011
21	14	10101001001000
22	15	101001111100
23		0111010111
24	16	0001001010110
25		111110111

1	1	1000011100
2	2	0001101000010110
3	3	1001110101
4		111011110
5		111011111
6	4	00100101010
7	5	11001011011
8	6	1001100010
9		10111010011
10	7	001100001010
11		0111110010
12		110101111000010
13		111010011
14	8	10100101100 0
15	9	00100011011
16	10	00100011110
17	11	0000110001111111
18	12	1001010011
19	13	00010101001110001
20		111101011
21	14	1010100100100 0
22	15	10100111110 0
23		0111010111
24	16	0001001010110
25		111110111
26		0110111000011

1	1	1000011100
2	2	0001101000010110
3	3	1001110101
4		111011110
5		111011111
6	4	001000101010
7	5	11001011011
8	6	1001100010
9		10111010011
10	7	0011000001010
11		0111110010
12		11010111100010
13		111010011
14	8	101001011000
15	9	001000011011
16	10	001000011110
17	11	000011000111111
18	12	1001010011
19	13	00010101001110001
20		111101011
21	14	10101001001000
22	15	101001111100
23		0111010111
24	16	0001001010110
25		111110111
26		0110111000011
27	17	1000000000

1	1	1000011100
2	2	0001101000010110
3	3	1001110101
4		111011110
5		111011111
6	4	001000101010
7	5	11001011011
8	6	1001100010
9		10111010011
10	7	0011000001010
11		0111110010
12		11010111100010
13		111010011
14	8	101001011000
15	9	001000011011
16	10	001000011110
17	11	000011000111111
18	12	1001010011
19	13	00010101001110001
20		111101011
21	14	10101001001000
22	15	101001111100
23		0111010111
24	16	0001001010110
25		111110111
26		0110111000011
27	17	1000000000
28		00111111011

1	1	1000011100
2	2	0001101000010110
3	3	1001110101
4		111011110
5		111011111
6	4	001000101010
7	5	11001011011
8	6	1001100010
9		10111010011
10	7	0011000001010
11		0111110010
12		11010111100010
13		111010011
14	8	101001011000
15	9	001000011011
16	10	001000011110
17	11	000011000111111
18	12	1001010011
19	13	00010101001110001
20		111101011
21	14	10101001001000
22	15	101001111100
23		0111010111
24	16	0001001010110
25		111110111
26		0110111000011
27	17	1000000000
28		00111111011
29	18	11000011001

1	1	1000011100
2	2	0001101000010110
3	3	1001110101
4		111011110
5		111011111
6	4	001000101010
7	5	11001011011
8	6	1001100010
9		10111010011
10	7	0011000001010
11		0111110010
12		11010111100010
13		111010011
14	8	101001011000
15	9	001000011011
16	10	001000011110
17	11	000011000111111
18	12	1001010011
19	13	00010101001110001
20		111101011
21	14	10101001001000
22	15	101001111100
23		0111010111
24	16	0001001010110
25		111110111
26		0110111000011
27	17	1000000000
28		00111111011
29	18	1100011001
30		0111001000

1	1	1000011100
2	2	0001101000010110
3	3	1001110101
4		111011110
5		111011111
6	4	001000101010
7	5	11001011011
8	6	1001100010
9		10111010011
10	7	0011000001010
11		0111110010
12		11010111100010
13		111010011
14	8	101001011000
15	9	001000011011
16	10	001000011110
17	11	000011000111111
18	12	1001010011
19	13	00010101001110001
20		111101011
21	14	10101001001000
22	15	101001111100
23		0111010111
24	16	0001001010110
25		111110111
26		0110111000011
27	17	1000000000
28		00111111011
29	18	11000011001
30		0111001000

What is going on?

- Probability of appearance at a given position

$$\mathbf{P}(100) = \frac{1}{8}$$

$$\mathbf{P}(111) = \frac{1}{8}$$

What is going on?

- Probability of appearance at a given position

$$\mathbf{P}(100) = \frac{1}{8}$$

$$\mathbf{P}(111) = \frac{1}{8}$$

- BUT the 111 occur often by CLUMPS

$$\begin{array}{ll} \cdots 0111110 & \qquad\qquad \cdots 011110 \\ \quad\ \ 111 & \qquad\qquad\quad\ \ 111\cdots \\ \quad\ \ \ \ 111\cdots & \end{array}$$

- while the 100 NEVER OVERLAP

What is going on?

- Probability of appearance at a given position

$$\mathbf{P}(100) = \frac{1}{8}$$

$$\mathbf{P}(111) = \frac{1}{8}$$

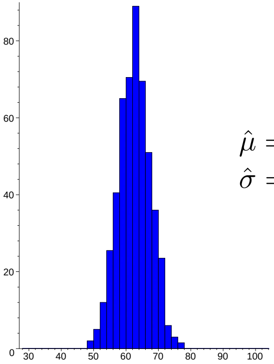
- BUT the 111 occur often by CLUMPS

$$\begin{array}{ll} \cdots 0111110 & \qquad \cdots 011110 \\ \quad 111 & \qquad \quad 111 \cdots \\ \quad\; 111 \cdots & \end{array}$$

- while the 100 NEVER OVERLAP

Expected waiting time: $111 - 14$ $\qquad 100 - 7$

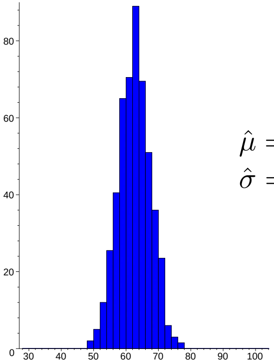
Number of occurrences 1000 texts of size 500



$$\hat{\mu} = 62.38$$
$$\hat{\sigma} = 4.90$$

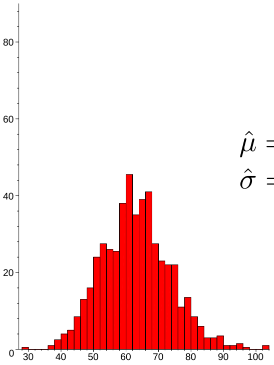
100

Number of occurrences 1000 texts of size 500



$$\hat{\mu} = 62.38$$
$$\hat{\sigma} = 4.90$$

100



$$\hat{\mu} = 62.07$$
$$\hat{\sigma} = 10.79$$

111

Aim of our work

- a **combinatorial** analysis for counting clumps of reduced sets of words
- an **algorithmic** contruction by **automata** that solves the counting problem in the general case and implies a normal limit law
- future extensions to **tandem repeats**?

$\ldots$GGGGATCGA|GATCGA|GATCGA|GATCGAGGG$\ldots$

Counting clumps

$w = aa \qquad T = bbbbaaaabbbbaaa$

- word counting with overlaps: 5 matches $(aa|a|a|, aa|a|)$
- clump counting: 2 matches $(aaaa|, aaa|)$

Definition of a clump of a word w

Set of **maximal contiguous** positions such that:

- each position is inside an occurrence of w
- if a position is at the beginning or at the end of an occurrence of w, it is covered by two occurrences of w, with exception of the first and last positions of the clump

Probability of start at a position i

(A) word $w = 1^k$ $\qquad$ $p = \mathbf{P}(1) = 1 - \mathbf{P}(0) = 1 - q$

$$\ldots \overbrace{\boxed{1}\boxed{1}\ldots\boxed{1}}^{k} \ldots$$

$\mathbf{P}(\mathsf{start}) = p^k$

Probability of start at a position i

(A) word $w = 1^k$ $\qquad$ $p = \mathbf{P}(1) = 1 - \mathbf{P}(0) = 1 - q$

$$\dots \overbrace{\boxed{1}\boxed{1}\dots\boxed{1}}^{k} \dots$$

$\mathbf{P}(\text{start}) = p^k$

(B) clump $\Gamma = 1^k.1^\star$

$$\begin{array}{ll}
\dots & \overbrace{\boxed{1}\boxed{1}\dots\boxed{1}}^{k}\boxed{0}\dots \\
\dots & \boxed{1}\boxed{1}\dots\boxed{1}\boxed{1}\boxed{0}\dots \\
\dots & \boxed{1}\boxed{1}\dots\boxed{1}\boxed{1}\boxed{1}\boxed{0}\dots \\
\dots & \dots
\end{array}$$

$$\mathbf{P}(\text{start}) = p^k \times q \times (1 + p + p^2 + \dots) = p^k \times \frac{q}{1-p} = p^k$$

Probability of start at a position i

(A) word $w = 1^k$ $\qquad$ $p = \mathbf{P}(1) = 1 - \mathbf{P}(0) = 1 - q$

$$\dots \overbrace{\boxed{1}\boxed{1}\dots\boxed{1}}^{k} \dots$$

$\mathbf{P}(\text{start}) = p^k$

(B) clump $\Gamma = 1^k.1^\star$

$$\begin{array}{ll}
\dots & \overbrace{\boxed{1}\boxed{1}\dots\boxed{1}}^{k}\boxed{0}\dots \\
\dots & \boxed{1}\boxed{1}\dots\boxed{1}\boxed{1}\boxed{0}\dots \\
\dots & \boxed{1}\boxed{1}\dots\boxed{1}\boxed{1}\boxed{1}\boxed{0}\dots \\
\dots & \dots
\end{array}$$

$$\mathbf{P}(\text{start}) = p^k \times q \times (1 + p + p^2 + \dots) = p^k \times \frac{q}{1-p} = p^k$$

FALSE

Probability of start at a position i

(A) word $w = 1^k$

$$\dots \overbrace{\boxed{1}\boxed{1}\dots\boxed{1}}^{k}\dots$$

$\mathbf{P}(\text{start}) = p^k$

(B) clump $\Gamma = 1^k.1^\star$

$\dots 1|\overbrace{1\dots1}^{k}\dots$ no clump **beginning** at position i

$\dots 0|\overbrace{1\dots1}^{k}\dots$ a clump **begins** at position i

$\mathbf{P}(\text{start}) = p^k \times q = p^k \times (1-p)$

Probabilistic approach

Prum, Reinert, Schbath, Pape, ...

$w = aaa$

$\mathbf{P}(a)$ small $\rightsquigarrow$ $\mathbf{P}$(start-of-a-clump-at-a-position) small

$\rightsquigarrow$

Poisson law for the number of clumps

Geometric law for the number of words in a clump

This extends to the general case.

Chen-Stein Poisson approximation provides bounds for the total variation distance to the composed law.

Combinatorial approach

Jacquet-Szpankowski

I- Combinatorial approach

1 word - Bernoulli model

$\mathcal{A}$ alphabet, w considered word

Polynomial of autocorrelation

$\mathcal{C}_w = \{\ h, \quad w.h = u.w \quad \text{and} \quad |u| < |w|\ \}$

$w = ababa$

$$\begin{array}{l} ababa \\ ababa| \\ \quad ababa \\ \qquad ababa \end{array}$$

$\mathcal{C}_{ababa} = \{\epsilon, ba, baba\}$

$$C_{ababa}(z) = \sum_{v \in \mathcal{C}_{ababa}} \mathbf{P}(v) z^{|v|} = 1 + \pi_a \pi_b z^2 + \pi_a^2 \pi_b^2 z^4$$

Régnier and Szpankowski decomposition

First $\mathcal{R} = \{\ t = u.w \quad \text{and} \quad \nexists r, s, \quad t = r.w.s\ \}$
$aaaaa\,ababa \in \mathcal{R}, \quad bbbbb\,abababa \notin \mathcal{R}$

Régnier and Szpankowski decomposition

First $\mathcal{R} = \{\ t = u.w$ and $\nexists r, s, \quad t = r.w.s\ \}$
$aaaaaababa \in \mathcal{R}, \quad bbbbbabababa \notin \mathcal{R}$
Minimal $\mathcal{M} \ = \{\ t, \quad w.t = u.w$ and $\nexists r, s, \quad w.t = r.w.s\ \}$

$ababa_{aaaaababa} \in \mathcal{M} \quad ababa_{babbbbbbbababa} \notin \mathcal{M} \quad ababa_{ba} \in \mathcal{M}$

Régnier and Szpankowski decomposition

First $\mathcal{R} = \{\ t = u.w$ and $\nexists r, s, \quad t = r.w.s\ \}$

$aaaaaababa \in \mathcal{R}, \quad bbbbbabababa \notin \mathcal{R}$

Minimal $\mathcal{M} = \{\ t, \quad w.t = u.w$ and $\nexists r, s, \quad w.t = r.w.s\ \}$

$ababa_{aaaaababa} \in \mathcal{M} \quad ababa_{babbbbbbbbababa} \notin \mathcal{M} \quad ababa_{ba} \in \mathcal{M}$

Ultimate $\mathcal{U} = \{t, \quad \nexists r, s, \quad w.t = r.w.s\}$

$ababa_{aabbbabbbbbbb} \in \mathcal{U} \qquad ababa_{babbbbbbbbbb} \notin \mathcal{U}$

Régnier and Szpankowski decomposition

First $\mathcal{R} = \{\ t = u.w \quad \text{and} \quad \nexists r, s, \quad t = r.w.s\ \}$
$aaaaaababa \in \mathcal{R}, \quad bbbbbabababa \notin \mathcal{R}$
Minimal $\mathcal{M} \ = \{\ t, \quad w.t = u.w \ \text{ and } \ \nexists r, s, \quad w.t = r.w.s\ \}$

$ababa_{aaaaababa} \in \mathcal{M} \quad ababa_{babbbbbbbbababa} \notin \mathcal{M} \quad ababa_{ba} \in \mathcal{M}$

Ultimate $\mathcal{U} = \{t, \quad \nexists r, s, \quad w.t = r.w.s\}$

$ababa_{aabbbabbbbbb} \in \mathcal{U} \qquad ababa_{babbbbbbbbb} \notin \mathcal{U}$

Not $\mathcal{N} = \overline{\mathcal{A}^{\star}.w.\mathcal{A}^{\star}} = \{t, \ \nexists r, s, \quad t = r.w.s\}$

Régnier and Szpankowski decomposition

First $\mathcal{R} = \{\ t = u.w \quad \text{and} \quad \nexists r, s, \quad t = r.w.s\ \}$

$aaaaababa \in \mathcal{R}, \quad bbbbbabababa \notin \mathcal{R}$

Minimal $\mathcal{M} \ = \{\ t, \quad w.t = u.w \ \text{ and } \ \nexists r, s, \quad w.t = r.w.s\ \}$

$ababa_{aaaaababa} \in \mathcal{M} \quad ababa_{babbbbbbbbababa} \notin \mathcal{M} \quad ababa_{ba} \in \mathcal{M}$

Ultimate $\mathcal{U} = \{t, \quad \nexists r, s, \quad w.t = r.w.s\}$

$ababa_{aabbbabbbbbbb} \in \mathcal{U} \qquad ababa_{babbbbbbbbbb} \notin \mathcal{U}$

Not $\mathcal{N} = \overline{\mathcal{A}^\star.w.\mathcal{A}^\star} = \{t, \ \nexists r, s, \quad t = r.w.s\}$

$$\mathcal{A}^\star = \mathcal{N} + \mathcal{R}.\left(\mathcal{M}\right)^\star.\mathcal{U} \quad \Rightarrow \quad \mathcal{L}_x = \mathcal{N} + \mathcal{R}x.\left(\mathcal{M}x\right)^\star.\mathcal{U}$$

Languages $\Rightarrow$ generating functions

$$
\begin{array}{llll}
\text{(I)} & \Sigma^\star = \mathcal{U} + \mathcal{M}\Sigma^\star & \Rightarrow & \dfrac{1}{1-z} = U(z) + \dfrac{M(z)}{1-z} \\
\text{(II)} & \Sigma^\star w = \mathcal{R}.\mathcal{C} + \mathcal{R}.\Sigma^\star.w & \Rightarrow & \dfrac{\pi_w z^m}{1-z} = R(z)\left(C(z) + \dfrac{\pi_w z^m}{1-z}\right) \\
\text{(III)} & \mathcal{M}^+ = \Sigma^\star.w + \mathcal{C} - \epsilon & \Rightarrow & \dfrac{M(z)}{1-M(z)} = \dfrac{\pi_w z^m}{1-z} + C(z) - 1 \\
\text{(IV)} & \mathcal{N}.\Sigma = \mathcal{R} + \mathcal{N} - \epsilon & \Rightarrow & zN(z) = R(z) + N(z) - 1
\end{array}
$$

Add the mark x after each match $\Rightarrow$ marked language $\mathcal{L}$

$$\mathcal{L} = \mathcal{N} + \mathcal{R}.x(\mathcal{M}.x)^\star\mathcal{U}$$

Translation into generating function

$$L(z,x) = N(z) + \frac{xR(z)U(z)}{1 - xM(z)}$$

Solving the system

$$R(z) = \frac{\pi_w z^{|w|}}{\pi_w z^{|w|} + (1-z)C(z)} \qquad U(z) = \frac{1}{\pi_w z^{|w|} + (1-z)C(z)}$$

$$N(z) = \frac{C(z)}{\pi_w z^{|w|} + (1-z)C(z)} \qquad M(z) = 1 + \frac{z-1}{\pi_w z^{|w|} + (1-z)C(z)}$$

$$L(z,x) = \frac{1}{1 - z + \pi_w z^{|w|} \dfrac{1-x}{x + (1-x)C(z)}}$$

Language equations for the clumps

Notation:

if $\mathcal{L} = \mathcal{W} \cdot w$ we write $\mathcal{L}^{-} = \mathcal{W}$

Combinatorial decomposition

$$\begin{aligned}
\mathcal{A}^{\star} &= \mathcal{N} + \mathcal{R}^{-} w\, \mathcal{C}^{\star} \Big((\mathcal{M} - \mathcal{K})^{-} w\, \mathcal{C}^{\star} \Big)^{\star} \mathcal{U} \\
&= \mathcal{N} + \mathcal{R}^{-} w\, \mathcal{K}^{\star} \Big((\mathcal{M} - \mathcal{K})^{-} w\, \mathcal{K}^{\star} \Big)^{\star} \mathcal{U} \\
&= \mathcal{N} + \mathcal{R}^{-} \mathbf{\Gamma} \Big((\mathcal{M} - \mathcal{K})^{-} \mathbf{\Gamma} \Big)^{\star} \mathcal{U}
\end{aligned}$$

Clump: $\mathbf{\Gamma} = w\, \mathcal{C}^{\star} = w\, \mathcal{K}^{\star}$

Some combinatorial properties

$w = aaaaa$
$\mathcal{C} - \{\epsilon\} = \{a, aa, aaa, aaaa\}$
$\mathcal{K} = \{a\}$
$\mathcal{M} = \{a, b(b + ab + aab + aaab + aaaab)^{\star}aaaaa\}$

Properties

- $\mathcal{K} \subset \mathcal{M}$
- $\mathcal{M} - \mathcal{K} = \mathcal{L}w$
- $\mathcal{K}^{\star} = \mathcal{C}^{\star}$ and $\mathcal{K}^{\star}$ is **unambiguous**

A Prefix Code generating the clumps

Lemma.
Let $\mathcal{C}_\circ = \mathcal{C} - \{\epsilon\}$ be the strict autocorrelation set of a word w

- the Prefix code $\mathcal{K} = \mathcal{C}_\circ - \mathcal{C}_\circ \mathcal{A}^+$ generates **unambiguously** $\mathcal{C}^+ - \{\epsilon\}$, which implies that $\mathcal{K}^\star = \mathcal{C}_\circ{}^\star$
- $\mathcal{K}^\star$ is **unambiguous**

Generating functions

$$F(z,\bullet,\bullet,\bullet) = \mathcal{N}(z) + \frac{\mathcal{R}(z)}{\pi_w z^{|w|}} \mathbf{\Gamma}(\bullet z,\bullet,\bullet) \frac{1}{1 - \frac{\mathcal{M}(z) - \mathcal{K}(z)}{\pi_w z^{|w|}} \times \mathbf{\Gamma}(\bullet z,\bullet,\bullet)} \mathcal{U}(z)$$

- number of w and number of clumps:

  $$\mathbf{\Gamma}(z,u,x) = ux\pi_w z^{|w|} \frac{1}{1 - x\mathcal{K}(z)}$$

- number of clumps and total number of positions inside clumps:

  $$\mathbf{\Gamma}(tz,u) = u\pi_w (tz)^{|w|} \frac{1}{1 - \mathcal{K}(tz)}$$

- number of w and total number of positions inside clumps:

  $$\mathbf{\Gamma}(tz,x) = x\pi_w (tz)^{|w|} \frac{1}{1 - x\mathcal{K}(tz)}$$

- number of "stuttering" w and total number of positions inside clumps:

  $$\mathbf{\Gamma}(tz,x) = \frac{x}{1-x}\pi_w (tz)^{|w|} \frac{1}{1 - \frac{x}{1-x}\mathcal{K}(tz)}$$

One word - Expectation - Variance

clumps

$$\mathbf{E}(O_n^{\mathfrak{K}}) = (\mathbf{n} - |w| + 1)\pi_w(1 - \mathcal{K}(1)) - \pi_w \mathcal{K}'(1)$$

$$\mathbf{Var}(O_n^{\mathfrak{K}}) = \mathbf{n} \times (1 - \mathcal{K}(1))^2 \mathbf{V}_w - \mathbf{n} \times \pi_w(1 - \mathcal{K}(1))(\mathcal{K}(1) - 2\pi_w \mathcal{K}'(1))$$

one word

$$\mathbf{E}(O_n^w) = (\mathbf{n} - |w| + 1)\pi_w, \qquad \mathbf{Var}(O_n^w) = \mathbf{n} \times \mathbf{V}_w + O(1).$$

$$\mathbf{V}_w = \pi_w\big(2\mathcal{C}(1) - 1 - (2|w| - 1)\pi_w\big)$$

Reduced compound patterns

$W = \{w_1, w_2\}$ and w_1 (resp. w_2) is **not factor** of w_2 (resp. w_1)

Correlation of words

$$\mathbb{C}(z) = (\mathcal{C}_{i,j}(z)) \qquad \mathcal{C}_{i,j} = \{h,\ w_i.h = u.w_j\}$$

Example $W = \{aab, abaa\}$

$$\mathbb{C}(z) = \begin{pmatrix} 1 & \pi_a^2 z^2 \\ \pi_b z & 1 + \pi_a^2 \pi_b z^3 \end{pmatrix}$$

Languages **Right**, **Minimal**, **Ultimate** $\mathcal{R}_i, \mathcal{M}_{i,j}, \mathcal{U}_i$

The set of language equations

$$\bigcup_{k\geq 1}\left(\mathbb{M}^k\right)_{i,i} = \mathcal{A}^\star \cdot w_i + \mathcal{C}_{ii} - \epsilon, \qquad 1 \leq i \leq 2,$$

$$\bigcup_{k\geq 1}\left(\mathbb{M}^k\right)_{i,j} = \mathcal{A}^\star \cdot w_j + \mathcal{C}_{ij}, \qquad 1 \leq i \neq j \leq 2,$$

$$\mathcal{U}_i \cdot \mathcal{A} = \bigcup_j \mathcal{M}_{ij} + \mathcal{U}_i - \epsilon, \qquad 1 \leq i \leq 2,$$

$$\mathcal{A} \cdot \mathcal{R}_j - (\mathcal{R}_j - w_j) = \bigcup_i w_i \mathcal{M}_{ij}, \qquad 1 \leq j \leq 2,$$

$$\mathcal{N} \cdot w_j = \mathcal{R}_j + \bigcup_i \mathcal{R}_i \left(\mathcal{C}_{ij} - \delta_{ij}\epsilon\right), \qquad 1 \leq i, j \leq 2$$

$$F(z,x,y) = \mathbb{R}(z,x,y)\left(\mathbb{I} - \mathbb{M}(z,x,y)\right)^{-1}\mathbb{U}(z)$$

Putting up equations for clumps of two words

Minimal Correlation Language: $\mathcal{K}_{ij} = \mathcal{C}_{ij} - \mathcal{C}_{ij}\mathcal{A}^{+}$
Lemma: $\mathcal{M}_{ij} - \mathcal{K}_{ij} = \mathcal{L}w_j$

$$\mathbb{K} = \begin{pmatrix} \mathcal{K}_{11} & \mathcal{K}_{12} \\ \mathcal{K}_{21} & \mathcal{K}_{22} \end{pmatrix}$$

$$\boldsymbol{\Xi} = \mathbb{K}^{\star} \qquad \mathbb{G} = \begin{pmatrix} w_1 \boldsymbol{\Xi}_{11} & w_1 \boldsymbol{\Xi}_{12} \\ w_2 \boldsymbol{\Xi}_{21} & w_2 \boldsymbol{\Xi}_{22} \end{pmatrix}$$

$$\mathcal{A}^{\star} = \mathcal{N} + (\mathcal{R}_1^{-}, \mathcal{R}_2^{-})\, \mathbb{G} \left((\mathbb{M} - \mathbb{K})^{-} \mathbb{G} \right)^{\star} \begin{pmatrix} \mathcal{U}_1 \\ \mathcal{U}_2 \end{pmatrix}$$

II- Automaton approach

Clumps of the set of words $\mathcal{U} = \{aabaa, baab\}$

$\mathcal{E}_{w_1,w_2}$ extension set from w_1 to w_2

$$\begin{aligned} \mathcal{E}_{aabaa,aabaa} &= \{baa, abaa\} & \mathcal{E}_{baab,baab} &= \{aab\} \\ \mathcal{E}_{aabaa,baab} &= \{b\} & \mathcal{E}_{baab,aabaa} &= \{aa\} \end{aligned}$$

Clumps of the set of words $\mathcal{U} = \{aabaa, baab\}$

$\mathcal{E}_{w_1,w_2}$ extension set from w_1 to w_2

$$\mathcal{E}_{aabaa,aabaa} = \{baa, abaa\} \qquad \mathcal{E}_{baab,baab} = \{aab\}$$
$$\mathcal{E}_{aabaa,baab} = \{b\} \qquad \mathcal{E}_{baab,aabaa} = \{aa\}$$

Algorithm

1. Build the set of strings

Clumps of the set of words $\mathcal{U} = \{aabaa, baab\}$

$\mathcal{E}_{w_1,w_2}$ extension set from w_1 to w_2

$$\begin{array}{llll} \mathcal{E}_{aabaa,aabaa} & = \{baa, abaa\} & \mathcal{E}_{baab,baab} & = \{aab\} \\ \mathcal{E}_{aabaa,baab} & = \{b\} & \mathcal{E}_{baab,aabaa} & = \{aa\} \end{array}$$

Algorithm

1. Build the set of strings

$$\begin{array}{rllll} X = & & \{aabaa.(\epsilon + \mathcal{E}_{aabaa,aabaa})\} & \cup & \{aabaa.\mathcal{E}_{aabaa,baab}\} \\ & \cup & \{baab.(\epsilon + \mathcal{E}_{baab,baab})\} & \cup & \{baab.\mathcal{E}_{baab,aabaa}\} \end{array}$$

Clumps of the set of words $\mathcal{U} = \{aabaa, baab\}$

$\mathcal{E}_{w_1,w_2}$ extension set from w_1 to w_2

$$\begin{aligned} \mathcal{E}_{aabaa,aabaa} &= \{baa, abaa\} & \mathcal{E}_{baab,baab} &= \{aab\} \\ \mathcal{E}_{aabaa,baab} &= \{b\} & \mathcal{E}_{baab,aabaa} &= \{aa\} \end{aligned}$$

Algorithm

1. Build the set of strings
$$\begin{aligned} X = \quad & \{aabaa.(\epsilon + \mathcal{E}_{aabaa,aabaa})\} \quad \cup \quad \{aabaa.\mathcal{E}_{aabaa,baab}\} \\ \cup \quad & \{baab.(\epsilon + \mathcal{E}_{baab,baab})\} \quad \cup \quad \{baab.\mathcal{E}_{baab,aabaa}\} \end{aligned}$$
2. Build a trie $\mathcal{T}$ on X

Clumps of the set of words $\mathcal{U} = \{aabaa, baab\}$

$\mathcal{E}_{w_1,w_2}$ extension set from w_1 to w_2

$$\begin{aligned} \mathcal{E}_{aabaa,aabaa} &= \{baa, abaa\} & \mathcal{E}_{baab,baab} &= \{aab\} \\ \mathcal{E}_{aabaa,baab} &= \{b\} & \mathcal{E}_{baab,aabaa} &= \{aa\} \end{aligned}$$

Algorithm

1. Build the set of strings

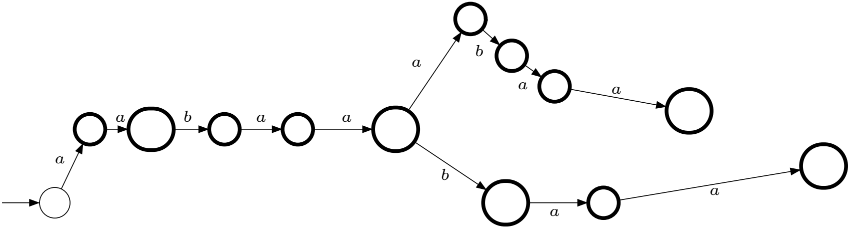
$$\begin{aligned} X = \quad & \{aabaa.(\epsilon + \mathcal{E}_{aabaa,aabaa})\} \cup \{aabaa.\mathcal{E}_{aabaa,baab}\} \\ \cup \quad & \{baab.(\epsilon + \mathcal{E}_{baab,baab})\} \cup \{baab.\mathcal{E}_{baab,aabaa}\} \end{aligned}$$

2. Build a trie $\mathcal{T}$ on X

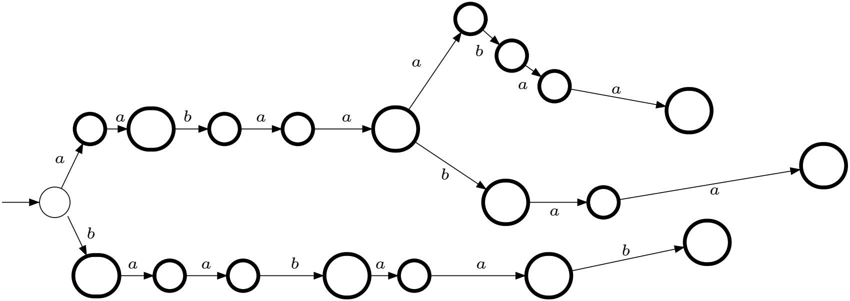
3. Build a Aho-Corasick like automaton upon $\mathcal{T}$. For each node ν of $\mathcal{T}$ with "access word" v, use the transition function δ

   $\delta(\nu, \ell) =$ node accessed by the **longuest prefix** in X that is **suffix** of $v.\ell$

$X = \{aabaa, aabaabaa, aabaaabaa, aabaab, \qquad\qquad\qquad \}$

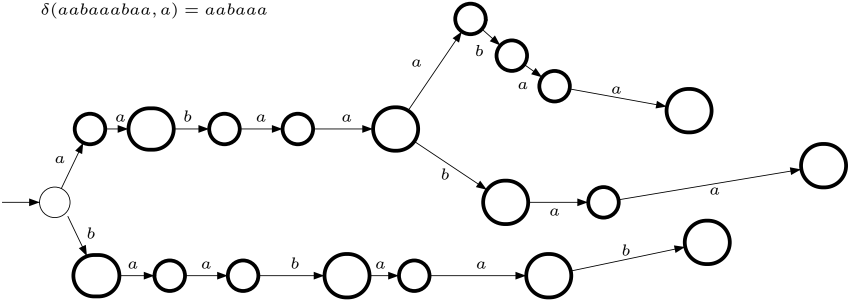


$X = \{aabaa, aabaabaa, aabaaabaa, aabaab, baab, baabaab\}$



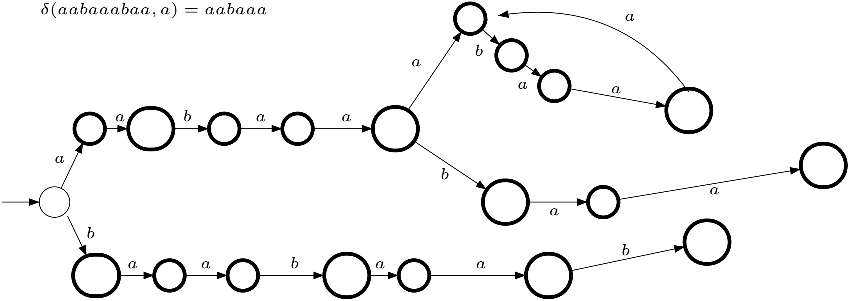
$$X = \{aabaa, aabaabaa, aabaaabaa, aabaab, baab, baabaab\}$$

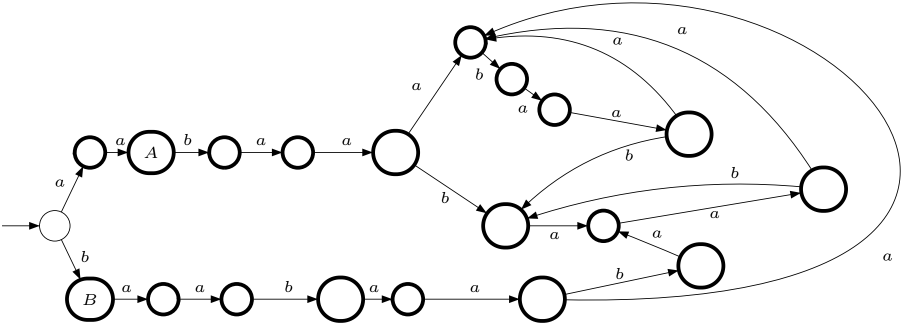
$$\delta(aabaaabaa, a) = aabaaa$$



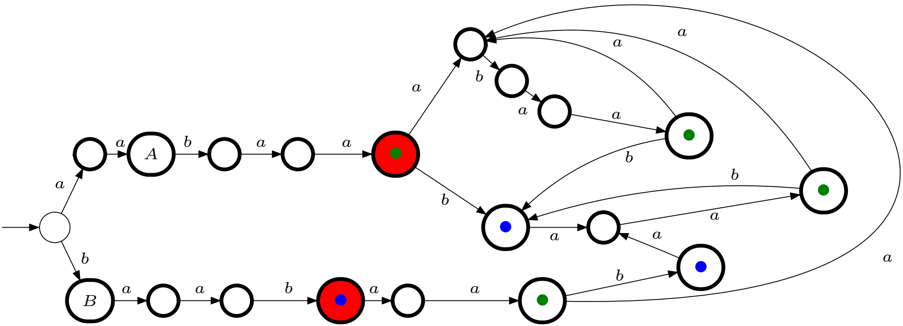
$$X = \{aabaa, aabaabaa, aabaaabaa, aabaab, baab, baabaab\}$$

$$\delta(aabaaabaa, a) = aabaaa$$



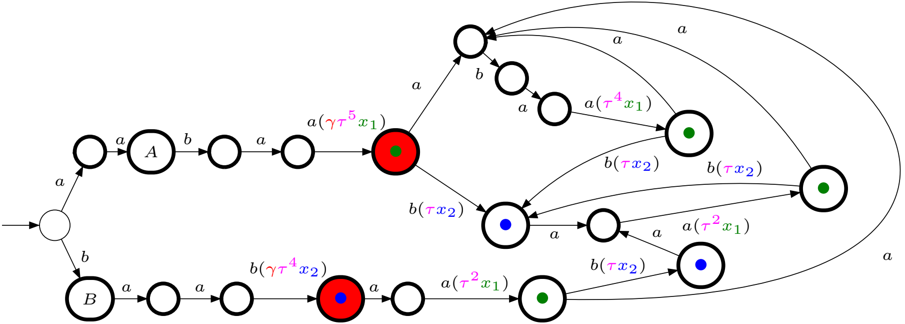


An automaton for $\mathcal{V} = \{v_1 = aabaa, v_2 = baab\}$. All transitions labeled by a and b ending respectively on state A and B are omitted.



An automaton for $\mathcal{V} = \{v_1 = aabaa, v_2 = baab\}$. All transitions labeled by a and b ending respectively on state A and B are omitted.

- ●, ● → the corresponding prefix (or state) ends with some occurrence of $aabaa$, $baab$.
- **red states** → states where we have entered a **new clump**



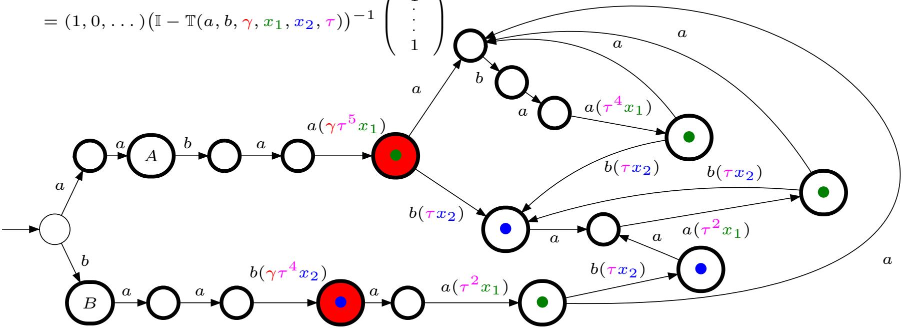
An automaton for $\mathcal{V} = \{v_1 = aabaa, v_2 = baab\}$. All transitions labeled by a and b ending respectively on state A and B are omitted.

- ●, ● $\longrightarrow$ the corresponding prefix (or state) ends with some occurrence of $aabaa, baab$.
- **red states** $\longrightarrow$ states where we have entered a **new clump**

Formal weights on transitions

- $\gamma \longrightarrow$ the number of clumps
- $\tau \longrightarrow$ total length of clumps
- $x_1, x_2 \longrightarrow$ occurrences of $aabaa, baab$

$$F(a,b,\gamma,x_1,x_2,\tau) = (1,0,\dots)(\mathbb{I} - \mathbb{T}(a,b,\gamma,x_1,x_2,\tau))^{-1}\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix}$$



An automaton for $\mathcal{V} = \{v_1 = aabaa, v_2 = baab\}$. All transitions labeled by a and b ending respectively on state A and B are omitted.

- ●, ● $\rightarrow$ the corresponding prefix (or state) ends with some occurrence of $aabaa, baab$.
- **red states** $\rightarrow$ states where we have entered a **new clump**

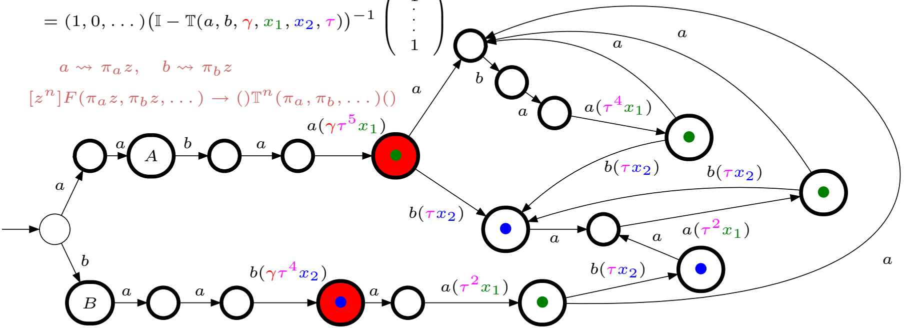
Formal weights on transitions

- $\gamma \rightarrow$ the number of clumps
- $\tau \rightarrow$ total length of clumps
- $x_1, x_2 \rightarrow$ occurrences of $aabaa, baab$

$$F(a,b,\gamma,x_1,x_2,\tau) = (1,0,\dots)(\mathbb{I}-\mathbb{T}(a,b,\gamma,x_1,x_2,\tau))^{-1}\begin{pmatrix}1\\ \vdots\\ \vdots\\ 1\end{pmatrix}$$

$$a \rightsquigarrow \pi_a z, \quad b \rightsquigarrow \pi_b z$$

$$[z^n]F(\pi_a z, \pi_b z, \dots) \to ()\mathbb{T}^n(\pi_a, \pi_b, \dots)()$$



An automaton for $\mathcal{V} = \{v_1 = aabaa, v_2 = baab\}$. All transitions labeled by a and b ending respectively on state A and B are omitted.

- ●, ● $\to$ the corresponding prefix (or state) ends with some occurrence of $aabaa$, $baab$.
- **red states** $\to$ states where we have entered a **new clump**

Formal weights on transitions

- $\gamma \to$ the number of clumps
- $\tau \to$ total length of clumps
- $x_1, x_2 \to$ occurrences of $aabaa$, $baab$

Asymptotic Limit Laws

- **One word**
  - $O_n^{\mathfrak{K}} = O(1)$ Poisson law for the number of clumps
- **General non-reduced sets**
  - $O_n^{\mathfrak{K}} = \Theta(n)$ Normal limit law (number of clumps, size covered)

Proofs

- Poisson law: Rouché theorem, singularity analysis
- Normal law: automaton, Perron-Frobenius for $\mathbb{T}(\dots)$, singularity analysis, large powers theorem

Complexity of computing the prefix code(s)

- **one word**

$$|\mathcal{K}| \log(|\mathcal{K}|)$$

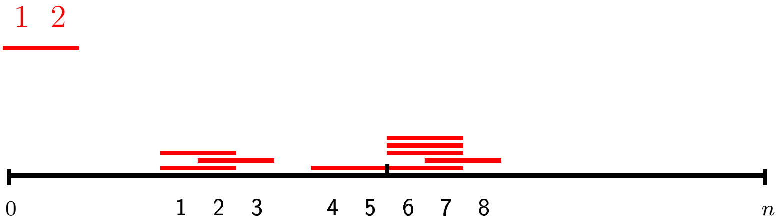
- **several words** (reduced case)

$$\left(\sum_{i,j} |\mathcal{K}_{i,j}|\right) \log \left(\sum_{i,j} |\mathcal{K}_{i,j}|\right)$$

Complexity of **insertion** of random keys in **tries**

Throwing flat dimers on an “integral” segment

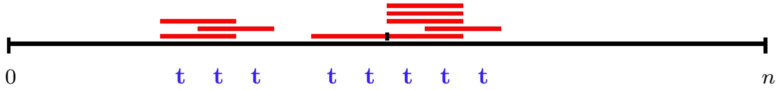
Throwing flat dimers on an “integral” segment



Size covered by the dimers?

Throwing flat dimers on an "integral" segment

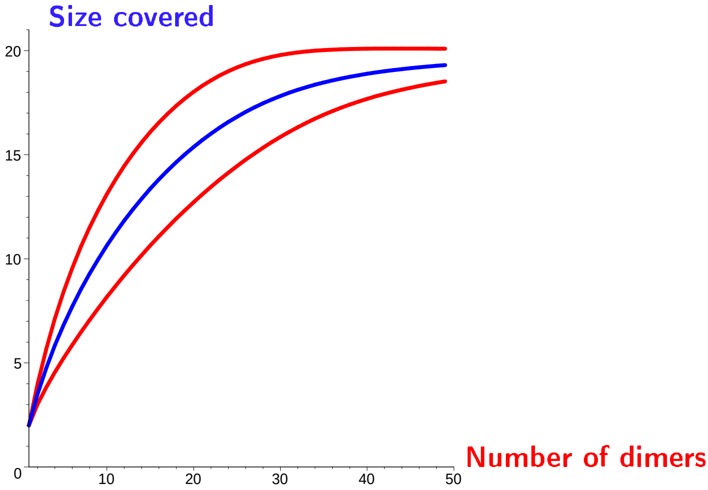
$a \quad a$ 　 **stuttering** occurrences of **aa** on $(\mathbf{a}+\mathbf{b})^{\mathbf{n}}$



$$\mathbf{P}(a) = \mathbf{P}(b) = \frac{1}{2} \qquad \mathbf{P}(aa) = \frac{1}{4}$$

stuttering by **substituting** $\quad \mathbf{x} \rightsquigarrow \dfrac{\mathbf{x}}{1-\frac{1}{4}\mathbf{x}} \quad$ **in** $\quad \Gamma(\mathbf{z}\mathbf{t}, \mathbf{x})$

Throwing flat dimers on an “integral” segment



$$\mathbf{n} = |\text{segment}| = \mathbf{20}$$

The Prefix Code $\mathcal{K} = \mathcal{C}_\circ - \mathcal{C}_\circ\mathcal{A}^+$ generates $\mathcal{C}_\circ{}^+$

w word, $\quad \mathcal{C}_\circ = \mathcal{C} - \epsilon, \qquad v \in \mathcal{C}_\circ \implies v = \kappa_1 \dots \kappa_r, \quad \kappa_i \in \mathcal{K}$

- $v \notin \mathcal{K} \implies v = \kappa_1 v'$ with $\kappa_1 \in \mathcal{K}$ and $v' \in \mathcal{A}^+$.
- $v, \kappa_1 \in \mathcal{C}_\circ \implies \exists p, p', s'$ such that $p = p's'$, $wv = pw$ and $w\kappa_1 = p'w$. Therefore $wv' = s'w$ and $v' \in \mathcal{C}_\circ$.
- iterate a finite number of time