# Average Analysis of Glushkov Automata under a BST-Like Model

C. Nicaud, C. Pivoteau, B. Razet

FSTTCS, December 2010

# Introduction

## What is the...

# average number of transitions in large Glushkov automata ?

1. What is a Glushkov automata ?
2. What does mean **average** number of transitions ?
3. What is the **shape** of a large random regular expression ?
4. What is the appropriate probabilistic distribution on regular expressions ?
5. Why is this question interesting ?

# Introduction

## What is the...

# average number of transitions in large Glushkov automata ?

1. What is a Glushkov automata ?
2. What does mean **average** number of transitions ?
3. What is the **shape** of a large random regular expression ?
4. What is the appropriate probabilistic distribution on regular expressions ?
5. Why is this question interesting ?

# Introduction

## What is the...

# average number of transitions in large Glushkov automata ?

1. What is a Glushkov automata ?
2. What does mean **average** number of transitions ?
3. What is the **shape** of a large random regular expression ?
4. What is the appropriate probabilistic distribution on regular expressions ?

5. Why is this question interesting ?

# Introduction

## What is the...

# average number of transitions in large Glushkov automata ?

1. What is a Glushkov automata ?
2. What does mean **average** number of transitions ?
3. What is the **shape** of a large random regular expression ?
4. What is the appropriate probabilistic distribution on regular expressions ?
5. Why is this question interesting ?

# Introduction

## What is the...

# average number of transitions in large Glushkov automata ?

1. What is a Glushkov automata ?
2. What does mean **average** number of transitions ?
3. What is the **shape** of a large random regular expression ?
4. What is the appropriate probabilistic distribution on regular expressions ?

5. Why is this question interesting ?

# Introduction

## What is the...

# average number of transitions in large Glushkov automata ?

1. What is a Glushkov automata ?
2. What does mean **average** number of transitions ?
3. What is the **shape** of a large random regular expression ?
4. What is the appropriate probabilistic distribution on regular expressions ?

▷ Average Analysis of Glushkov Automata under a BST-Like Model

5. Why is this question interesting ?

# Introduction

## What is the...

# average number of transitions in large Glushkov automata ?

1. What is a Glushkov automata ?
2. What does mean **average** number of transitions ?
3. What is the **shape** of a large random regular expression ?
4. What is the appropriate probabilistic distribution on regular expressions ?

▷ Average Analysis of Glushkov Automata under a BST-Like Model

5. Why is this question interesting ?

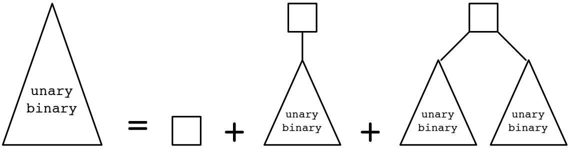
# Motivations

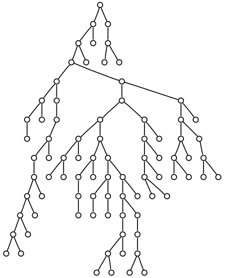
**Why are we interested in ...**

- **... the number of transitions in Glushkov automata ?**
  - bounds on time and space complexity of the algorithm compiling the Glushkov automaton
  - to compare different algorithms compiling regular expressions into automata
- **... average analysis ?**
  - average analysis of algorithms
  - to give more relevant information on practical running times of algorithms (in comparison with worst case analysis)
- **... the BST-like model ?**
  - easy random sampling
  - often used in practice
  - better modeling of regular expressions
    ▷ e.g. the number of nested stars in expressions

# Random regular expressions

# Random unary-binary trees : the BST-like model



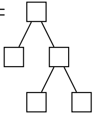
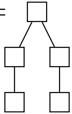



**Size : number of □-nodes.**

BST-like distribution of probabilities over unary-binary trees :

$$
\begin{cases}
\mathbb{P}(\square) = \mathbb{P}\left(\begin{array}{c}\square \\ | \\ \square\end{array}\right) = 1 \\
\mathbb{P}\left(\begin{array}{c}\square \\ | \\ T\end{array}\right) = q \cdot \mathbb{P}(T) \\
\mathbb{P}\left(\begin{array}{c}\square \\ / \backslash \\ T_1 \; T_2\end{array}\right) = (1-q) \cdot \frac{1}{n-2} \cdot \mathbb{P}(T_1) \cdot \mathbb{P}(T_2) \quad \text{if } |T_1| + |T_2| + 1 = n
\end{cases}
$$

# The BST-like distribution is not uniform

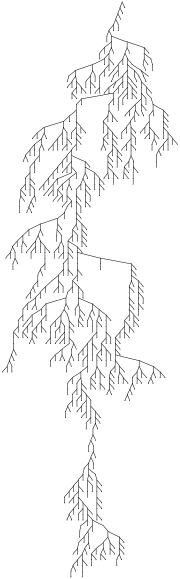
$T_1 =$  $T_2 =$ 

Uniform : $\mathbb{P}(T_1) = \mathbb{P}(T_2)$

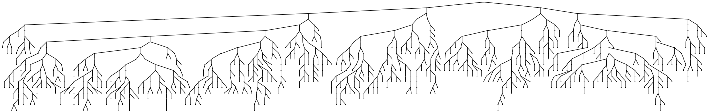
BST-like : $(1-q)^2/3 \overset{?}{=} (1-q)/3 \quad \rightarrow$ No solution !

---

Uniform random unary-binary tree (1021 nodes) $\triangleright$
$\sim$ height : $\Theta(\sqrt{n})$ [Flajolet, Odlyzko 82]



$\triangledown$ BST-like random unary-binary tree (1000 nodes)
$\sim$ height : $\Theta(\log n)$ [Robson79, Devroye86, Drmota01]



# Random regular expressions

Proba. of a random size $n$ reg. exp. in the BST-like model :

$$\begin{array}{lcll} \mathbb{P}(T^*) & = & \mathbb{P}(T) & \text{if } n = 2 \\ \mathbb{P}(T^*) & = & q \cdot \mathbb{P}(T) & \text{if } n > 2 \\ \mathbb{P}\left(T_1 \cup T_2\right) & = & \mathbb{P}\left(T_1 \bullet T_2\right) \;=\; \frac{1}{2}\frac{1-q}{(n-2)}\mathbb{P}(T_1)\mathbb{P}(T_2) & \text{if } |T_1| + |T_2| + 1 = n \end{array}$$

When n=1 (for the leaves) : $\mathbb{P}(\varepsilon) = p_\varepsilon$ and $\sum_{a \in A} \mathbb{P}(a) = 1 - p_\varepsilon$.

```
RE(n) ---------------------------------------- Random Sampler ----
    if n=1 then return ε with proba p_ε or a letter ℓ with proba ℙ(ℓ)
    if n=2 then return (RE(1))*
    else, choose "unary" with proba q or "binary" with proba 1 − q
           if "unary" then return (RE(n − 1))*
           else choose k uniformly at random between 1 and n − 2
                 return RE(k) ∪ RE(n − k − 1) with proba 1/2
                 or return RE(k) • RE(n − k − 1) with proba 1/2
-------------------------------------------------------------------
```

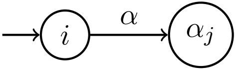
# Glushkov Automaton

# Glushkov Automaton

Glushkov (1961) ; McNaughton and Yamada (1960) ;
Berry and Sethi (1986).

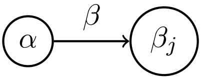
$$T = b^* \bullet (a \cup b \bullet b)^* \quad \xrightarrow{\textit{Relabeling}} \quad \widetilde{T} = b_1^* \bullet (a_2 \cup b_3 \bullet b_4)^*$$

`First` $(T) = \{\ \alpha_j \mid$ a word of $L(\widetilde{T})$ begins with $\alpha_j\ \}$

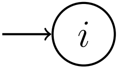



`Last` $(T) = \{\ \alpha_j \mid$ a word of $L(\widetilde{T})$ ends with $\alpha_j\ \}$


`Follow` $(T, \alpha) = \{\ \beta_j \mid \beta_j$ can follow $\alpha$ in a word of $L(\widetilde{T})\ \}$

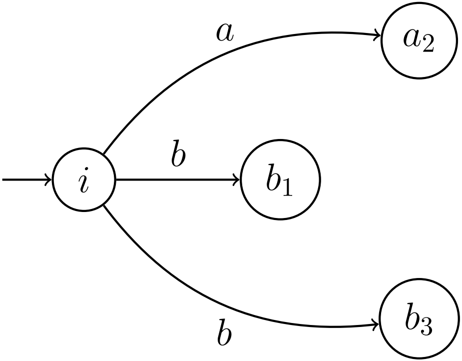


# Glushkov Automaton for $\widetilde{T} = b_1^* \bullet (a_2 \cup b_3 \bullet b_4)^*$



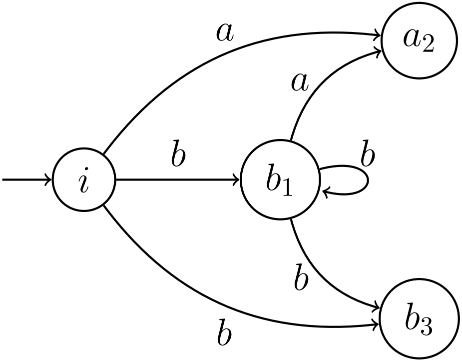
# Glushkov Automaton for $\widetilde{T} = b_1^* \bullet (a_2 \cup b_3 \bullet b_4)^*$

$\texttt{First}\,(T) = \{b_1, a_2, b_3\}$



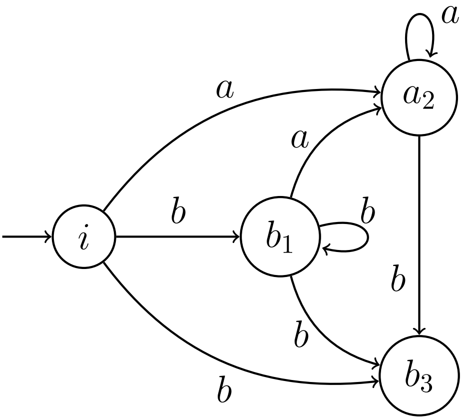
# Glushkov Automaton for $\widetilde{T} = b_1^* \bullet (a_2 \cup b_3 \bullet b_4)^*$

$\texttt{Follow}\,(T, b_1) = \{b_1, a_2, b_3\}$



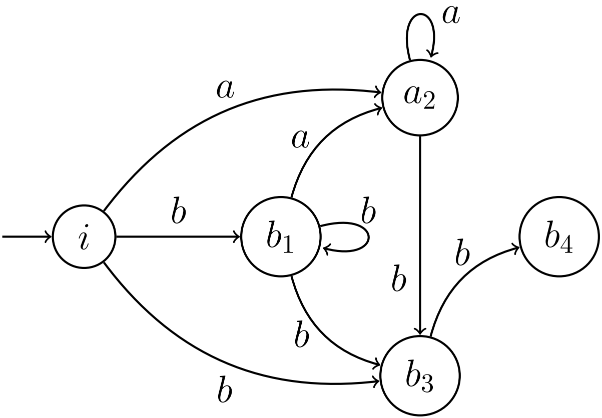
# Glushkov Automaton for $\widetilde{T} = b_1^* \bullet (a_2 \cup b_3 \bullet b_4)^*$

$\texttt{Follow}\,(T, a_2) = \{a_2, b_3\}$



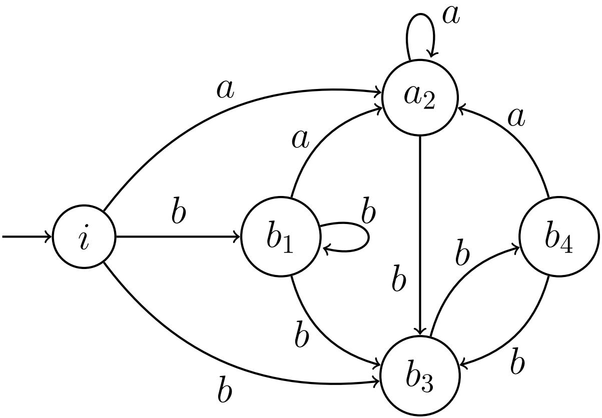
# Glushkov Automaton for $\widetilde{T} = b_1^* \bullet (a_2 \cup b_3 \bullet b_4)^*$

$\texttt{Follow}\,(T, b_3) = \{b_4\}$



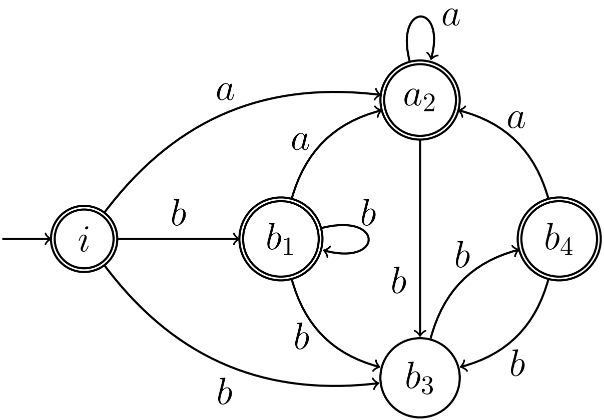
# Glushkov Automaton for $\widetilde{T} = b_1^* \bullet (a_2 \cup b_3 \bullet b_4)^*$

$\texttt{Follow}\,(T, b_4) = \{a_2, b_3\}$



# Glushkov Automaton for $\widetilde{T} = b_1^* \bullet (a_2 \cup b_3 \bullet b_4)^*$

$\texttt{Last}\,(T) = \{b_1, a_2, b_4\}$



# Average analysis

# Average number of transitions

**Theorem**

**In the BST-like model, the average number of transitions in the Glushkov automaton of a size n regular expression is quadratic, i.e., in $\boldsymbol{\Theta(n^2)}$.**

*Rmk : in the worst case, the number of transitions is also quadratic.*

Recall that :

**Theorem (Nicaud 09)**

*The average number of transitions of the Glushkov automaton associated to a regular expression of size $n$, for the uniform distribution, is in $\Theta(n)$.*

# Sketch of proof

The (non initial) transitions in the Glushkov Automaton of $T$ :

$$\begin{cases} \texttt{Edges}\,(\varepsilon) = \texttt{Edges}\,(a) = 0 \\ \texttt{Edges}\,(T^*) = \texttt{Edges}\,(T) \cup \texttt{Last}\,(T) \times \texttt{First}\,(T) \\ \texttt{Edges}\,(T_1 \cup T_2) = \texttt{Edges}\,(T_1) \cup \texttt{Edges}\,(T_2) \\ \texttt{Edges}\,(T_1 \bullet T_2) = \texttt{Edges}\,(T_1) \cup \texttt{Edges}\,(T_2) \cup \texttt{Last}\,(\mathbf{T_1}) \times \texttt{First}\,(\mathbf{T_2}) \end{cases}$$

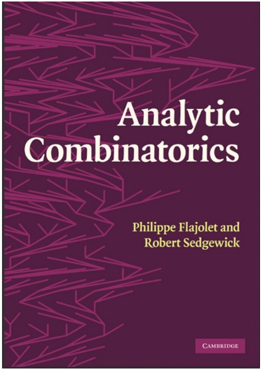
- The number of *new* transitions produced by $T_1 \bullet T_2$ is $|\texttt{Last}\,(T_1)| \cdot |\texttt{First}\,(T_2)|$
- The average size of **First** (or **Last**) is linear.

  ▷ There is a *non zero* probability that a size $n$ expression leads to an automaton with at least $\beta n^2$ transitions, $\beta > 0$.

  ▷ By Markov inequality : $\mathbb{E}[X] \geq a \cdot \mathbb{P}\,(X \geq a)$,

  **the average number of transitions is in $\mathbf{\Omega(n^2)}$.**

# Analytic Combinatorics




Ph. Flajolet,
R. Sedgewick.

- Study of the asymptotic behavior of counting sequences of the form : $(a_n)_{n\in\mathbb{N}}$
- Use its **generating function** $A(z)$, the formal power series defined by

$$A(z) = \sum_{n\in\mathbb{N}} a_n z^n.$$

- Recursive descriptions of sequences can **automatically** be translated into (differential) equations on generating functions.
- Many powerful results of Analytic Combinatorics to compute asymptotic estimates for the coefficients (the $a_n$'s).

# The average size of First is linear

**Theorem**

*The average size of* $\texttt{First}$ *for a size $n$ regular expression, according to the BST-like model, is asymptotically equivalent to $K\,n$, for some real constant $K \in ]0,1[$.*

$$\begin{cases} \texttt{First}\left(\begin{smallmatrix} & \bullet & \\ & / \backslash & \\ T_1 & & T_2 \end{smallmatrix}\right) = \texttt{First}\,(T_1) \cup \texttt{First}\,(T_2) & \forall T_1, T_2 \in \mathcal{T}, \varepsilon \in L(T_1) \\ \texttt{First}\left(\begin{smallmatrix} & \bullet & \\ & / \backslash & \\ T_1 & & T_2 \end{smallmatrix}\right) = \texttt{First}\,(T_1) & \forall T_1, T_2 \in \mathcal{T}, \varepsilon \notin L(T_1). \end{cases}$$

$f_n$ : average size of $\texttt{First}\,(T)$ when $|T| = n$. $\quad f_1 = f_2 = 1 - p_\varepsilon$

$$f_{n+2} = q f_{n+1} + \frac{2(1-q)}{n} \sum_{\ell=1}^{n} f_\ell - \frac{1-q}{2n} \sum_{\ell=1}^{n} \boldsymbol{r_\ell} f_{n+1-\ell}, \quad n \geq 1.$$

$\triangleright$ differential equation for $F(z)$ $\quad \triangleright$ asymptotic estimate of $f_n$.

# Recognizing the empty word

The size of `First` (and `Last`) is highly related to the probability of recognizing the empty word.

## Theorem

*A large random regular expression recognizes the empty word with high probability. More precisely, in the BST-like model, the probability that a size $n$ regular expression does not recognize $\varepsilon$ is asymptotically equivalent to*

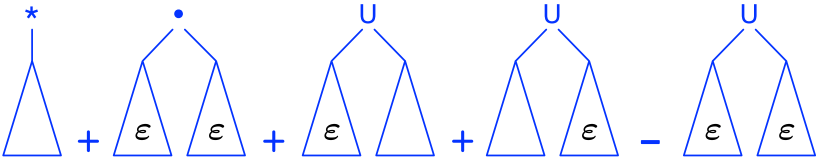
$$r_n \sim \frac{C}{n^q}$$

*with* $C = \frac{(1-p_\varepsilon)}{e^{1-q}\Gamma(1-q)}\left(1 - \int_0^1 \frac{e^{(1-q)t}(1-t)^{1-q}-1}{t^2}dt\right)$.

$r_n$ : the probability that a size $n$ regular expression does not recognize $\varepsilon$ ($r_0 = 0$)

# Recognizing the empty word (sketch of proof)

When does a regular expression recognize the empty word ?

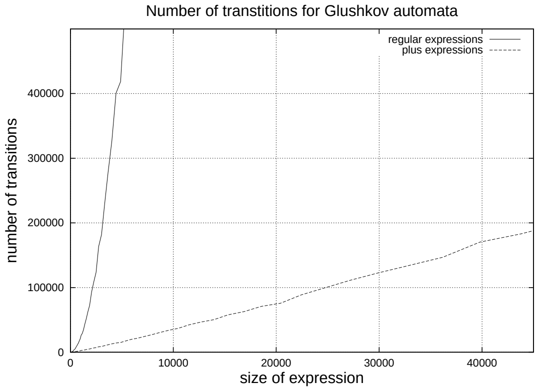



The sequence $(r_n)_{n\in\mathbb{N}}$ satisfies $r_1 = 1 - p_\varepsilon$, $r_2 = 0$ and

$$r_{n+2} = \frac{1-q}{n} \sum_{\ell=1}^{n} r_\ell, \quad n \geq 1.$$

- ▷ differential equation for $R(z) = \sum_{n\in\mathbb{N}} r_n z^n$ ;
  - ▷ asymptotic equivalent for $r_n$.

# Experiments




- $x$-axis : size of expressions defined on the alphabet $\{a, b\}$
- $y$-axis : number of transitions of Glushkov automata
- parameters : $q = \frac{1}{3}$, $p_\varepsilon = \frac{1}{100}$ and $\mathbb{P}(a) = \mathbb{P}(b)$

# Perspectives

- Study of regular expressions where the Kleene Star operator $*$ has been replaced by a $+$ operator :
  $\triangleright$ prove the linear behavior empirically observed (work in progress).

- Consider average analysis of other constructions related to Glushkov automata, such as :
  - the Follow automaton by Ilie and Yu,
  - Antimirov automaton.