International Journal of Foundations of Computer Science
Vol. 23, No. 5 (2012) 969–984
© World Scientific Publishing Company
DOI: 10.1142/S0129054112400400

Int. J. Found. Comput. Sci. 2012.23:969-984. Downloaded from www.worldscientific.com
by UNIVERSITY OF WATERLOO on 02/11/15. For personal use only.

ON THE AVERAGE SIZE OF GLUSHKOV AND PARTIAL DERIVATIVE AUTOMATA*

SABINE BRODA†, ANTÓNIO MACHIAVELO‡,
NELMA MOREIRA§ and ROGÉRIO REIS¶

CMUP, Faculdade de Ciências, Universidade do Porto
Rua do Campo Alegre 687, 4169-007 Porto, Portugal
†*sbb@dcc.fc.up.pt*
‡*ajmachia@fc.up.pt*
§*nam@dcc.fc.up.pt*
¶*rvr@dcc.fc.up.pt*

Received 29 October 2011
Accepted 16 February 2012
Communicated by Giancarlo Mauri

In this paper, the relation between the Glushkov automaton ($\mathcal{A}_{\mathrm{pos}}$) and the partial derivative automaton ($\mathcal{A}_{\mathrm{pd}}$) of a given regular expression, in terms of transition complexity, is studied. The average transition complexity of $\mathcal{A}_{\mathrm{pos}}$ was proved by Nicaud to be linear in the size of the corresponding expression. This result was obtained using an upper bound of the number of transitions of $\mathcal{A}_{\mathrm{pos}}$. Here we present a new quadratic construction of $\mathcal{A}_{\mathrm{pos}}$ that leads to a more elegant and straightforward implementation, and that allows the exact counting of the number of transitions. Based on that, a better estimation of the average size is presented. Asymptotically, and as the alphabet size grows, the number of transitions per state is on average 2. Broda *et al.* computed an upper bound for the ratio of the number of states of $\mathcal{A}_{\mathrm{pd}}$ to the number of states of $\mathcal{A}_{\mathrm{pos}}$, which is about 1/2 for large alphabet sizes. Here we show how to obtain an upper bound for the number of transitions in $\mathcal{A}_{\mathrm{pd}}$, which we then use to get an average case approximation. In conclusion, assymptotically, and for large alphabets, the size of $\mathcal{A}_{\mathrm{pd}}$ is half the size of the $\mathcal{A}_{\mathrm{pos}}$. This is corroborated by some experiments, even for small alphabets and small regular expressions.

Keywords: Regular languages; regular expressions; partial derivatives; conversion between regular expressions and nondeterministic finite automata; analytic combinatorics; average case analysis.

1. Introduction

The conversion methods of regular expressions into equivalent nondeterministic finite automata (NFA) are normally divided in two classes depending on whether

*This work was partially funded by the European Regional Development Fund through the programme COMPETE and by the Portuguese Government through the FCT – Fundação para a Ciência e a Tecnologia under the project PEst-C/MAT/UI0144/2011 and by project CANTE-PTDC/EIA-CCO/101904/2008.

Int. J. Found. Comput. Sci. 2012.23:969-984. Downloaded from www.worldscientific.com
by UNIVERSITY OF WATERLOO on 02/11/15. For personal use only.

ε-transitions are allowed or not in the resulting NFA. Paradigmatic methods of each class are the Thompson's and Glushkov's constructions, respectively. Several optimizations and worst-case descriptional and computational complexity results were obtained for both methods (see Holzer and Kutrib [12], and the works cited therein). Given a regular expression with n letters the size of an ε-NFA can be, in the worst-case $\Theta(n)$. While the size of a Glushkov automaton can be $\Theta(n^2)$, $\Omega(n \log n^2)$ was proved to be a lower bound for the size of an ε-free NFA. In this context, and for practical purposes, it is useful to carry out average-case analysis, both for descriptional and computational complexities, of these methods.

The framework of analytic combinatorics, by relating the enumeration of combinatorial objects to the algebraic and complex analytic properties of generating functions, provides a powerful tool for asymptotic average-case analysis. Using this framework, Nicaud [17] proved that the average transition complexity of the Glushkov automaton ($\mathcal{A}_{\text{pos}}$) of a regular expression α of size n is $\Theta(n)$. This result was obtained using an upper bound of the number of transitions of $\mathcal{A}_{\text{pos}}$. Here we present a new quadratic construction of the $\mathcal{A}_{\text{pos}}$ that leads to a more elegant and straightforward implementation, and that allows the exact counting of the number of transitions. Based on that, a better estimation of the average size is presented. Asymptotically, and as the alphabet size grows, the number of transitions per state is on average 2.

The partial derivative automaton ($\mathcal{A}_{\text{pd}}$) is a quotient of the $\mathcal{A}_{\text{pos}}$, and thus the states of the former can be seen as mergings of states of the latter. In a previous paper [3], we presented a technique for estimating some of those state mergings. This enabled us, in the framework of analytic combinatorics, to compute an upper bound for the ratio of the number of states of $\mathcal{A}_{\text{pd}}$ to the number of states of $\mathcal{A}_{\text{pos}}$, which is about 1/2 for large alphabet sizes. This upper bound was obtained by estimating the number of regular expressions that have ε as a partial derivative. In this paper, we use an analogous approach to compute an upper bound for the number of transitions in $\mathcal{A}_{\text{pd}}$, and study its asymptotic behaviour. As the alphabet size grows, this upper bound tends to the number of letters of the regular expression, thus it is half the number of transitions in $\mathcal{A}_{\text{pos}}$.

2. Regular Expressions and Automata

In this section we briefly review some basic definitions about regular expressions and finite automata. For more details, we refer the reader to Kozen [13] or Sakarovitch [19].

Given an alphabet (set of *letters*) $\Sigma = \{\sigma_1, \ldots, \sigma_k\}$ of size k, the set $\mathcal{R}$ of *regular expressions*, α, over Σ is defined by the following grammar:

$$\alpha := \emptyset \mid \varepsilon \mid \sigma_1 \mid \cdots \mid \sigma_k \mid (\alpha + \alpha) \mid (\alpha \cdot \alpha) \mid \alpha^\star \tag{1}$$

where the operator $\cdot$ (concatenation) is often omitted. The language associated to α is denoted by $\mathcal{L}(\alpha)$ and defined as usual. The *size* $|\alpha|$ of $\alpha \in \mathcal{R}$ is the number

Int. J. Found. Comput. Sci. 2012.23:969-984. Downloaded from www.worldscientific.com
by UNIVERSITY OF WATERLOO on 02/11/15. For personal use only.

of symbols in α (parentheses not counted); the *alphabetic size* $|\alpha|_\Sigma$ is its number of letters. For example, for $\tau = a(bc + a^\star)^\star$ one has $|\tau| = 9$ (do not forget the ommitted concatenation symbols) and $|\tau|_\Sigma = 4$. We define $\varepsilon(\alpha)$ by $\varepsilon(\alpha) = \varepsilon$ if $\varepsilon \in \mathcal{L}(\alpha)$, and $\varepsilon(\alpha) = \emptyset$ otherwise. Also, we denote by α_ε and $\alpha_{\overline{\varepsilon}}$, respectively, the regular expressions such that $\varepsilon(\alpha_\varepsilon) = \varepsilon$ and $\varepsilon(\alpha_{\overline{\varepsilon}}) = \emptyset$.

A *non-deterministic automaton* (NFA) $\mathcal{A}$ is a quintuple $(Q, \Sigma, \delta, q_0, F)$, where Q is a finite set of states, Σ is the alphabet, $\delta \subseteq Q \times \Sigma \times Q$ the transition relation, q_0 the initial state, and $F \subseteq Q$ the set of final states. The *size* of a NFA is $|Q| + |\delta|$. For $q \in Q$ and $\sigma \in \Sigma$, we denote the set $\{\, p \mid (q, \sigma, p) \in \delta \,\}$ by $\delta(q, \sigma)$, and we can extend this notation to $w \in \Sigma^\star$, and to $R \subseteq Q$. The *language* accepted by $\mathcal{A}$ is $\mathcal{L}(\mathcal{A}) = \{\, w \in \Sigma^\star \mid \delta(q_0, w) \cap F \neq \emptyset \,\}$.

2.1. *The Glushkov automaton*

The Glushkov, or position, automaton was independently introduced by Glushkov [11] and McNaughton and Yamada [15]. The states in the Glushkov automaton, representing a regular expression α, correspond to the positions of letters in α plus an additional initial state. Let $\tilde{\alpha}$ denote the regular expression obtained by marking each letter with its position in α. The marked version of the regular expression τ, used above, is $\tilde{\tau} = a_1(b_2c_3 + a_4^\star)^\star$. Now, let $\mathsf{Pos}(\alpha) = \{1, 2, \ldots, |\alpha|_\Sigma\}$ be the set of positions for $\alpha \in \mathcal{R}$, and let $\mathsf{Pos}_0(\alpha) = \mathsf{Pos}(\alpha) \cup \{0\}$. Then, the construction of the Glushkov automaton is based on the position sets $\mathsf{First}(\tilde{\alpha}) = \{\, i \mid (\exists w)\ \sigma_i w \in \mathcal{L}(\tilde{\alpha}) \,\}$, $\mathsf{Last}(\tilde{\alpha}) = \{\, i \mid (\exists w)\ w\sigma_i \in \mathcal{L}(\tilde{\alpha}) \,\}$, and $\mathsf{Follow}(\tilde{\alpha}) = \{\, (i, j) \mid (\exists u, v)\ u\sigma_i\sigma_j v \in \mathcal{L}(\tilde{\alpha}) \,\}$. These sets can be inductively defined as follows:

$$
\begin{array}{ll}
\mathsf{First}(\emptyset) \;= \mathsf{First}(\varepsilon) = \emptyset & \mathsf{First}(\alpha + \beta) = \mathsf{First}(\alpha) \cup \mathsf{First}(\beta) \\
\mathsf{First}(\sigma_i) \;= \{i\} & \mathsf{First}(\alpha \cdot \beta) \;\; = \begin{cases} \mathsf{First}(\alpha) \cup \mathsf{First}(\beta) & \text{if } \varepsilon(\alpha) = \varepsilon \\ \mathsf{First}(\alpha) & \text{otherwise.} \end{cases} \\
\mathsf{First}(\alpha^\star) = \mathsf{First}(\alpha) &
\end{array}
$$

The definition of Last is almost identical and differs only for the case of concatenation, which is

$$
\mathsf{Last}(\alpha \cdot \beta) = \begin{cases} \mathsf{Last}(\alpha) \cup \mathsf{Last}(\beta) & \text{if } \varepsilon(\beta) = \varepsilon \\ \mathsf{Last}(\beta) & \text{otherwise.} \end{cases}
$$

The set Follow can be computed by

$$
\begin{aligned}
\mathsf{Follow}(\emptyset) &= \mathsf{Follow}(\varepsilon) = \mathsf{Follow}(\sigma_j) = \emptyset \\
\mathsf{Follow}(\alpha + \beta) &= \mathsf{Follow}(\alpha) \cup \mathsf{Follow}(\beta) \\
\mathsf{Follow}(\alpha \cdot \beta) &= \mathsf{Follow}(\alpha) \cup \mathsf{Follow}(\beta) \cup \mathsf{Last}(\alpha) \times \mathsf{First}(\beta) \\
\mathsf{Follow}(\alpha^\star) &= \mathsf{Follow}(\alpha) \cup \mathsf{Last}(\alpha) \times \mathsf{First}(\alpha).
\end{aligned}
$$

For $\tilde{\tau}$ as before we have $\mathsf{First}(\tilde{\tau}) = \{1\}$, $\mathsf{Last}(\tilde{\tau}) = \{1, 3, 4\}$, and $\mathsf{Follow}(\tilde{\tau}) = \{(1, 2), (1, 4), (2, 3), (3, 2), (3, 4), (4, 2), (4, 4)\}$.

The *Glushkov automaton* for α is $\mathcal{A}_{\text{pos}}(\alpha) = (\mathsf{Pos}_0(\alpha), \Sigma, \delta_{\text{pos}}, 0, F)$, with $\delta_{\text{pos}} = \{\, (0, \tilde{\sigma_j}, j) \mid j \in \mathsf{First}(\tilde{\alpha}) \,\} \cup \{\, (i, \tilde{\sigma_j}, j) \mid \; (i, j) \in \mathsf{Follow}(\tilde{\alpha}) \,\}$ and $F = \mathsf{Last}(\tilde{\alpha}) \cup \{0\}$

Int. J. Found. Comput. Sci. 2012.23:969-984. Downloaded from www.worldscientific.com
by UNIVERSITY OF WATERLOO on 02/11/15. For personal use only.

if $\varepsilon(\alpha) = \varepsilon$, and $F = \mathsf{Last}(\tilde{\alpha})$, otherwise. Note that the number of states of $\mathcal{A}_{\text{pos}}(\alpha)$ is exactly $n+1$, where $n = |\alpha|_\Sigma$. On the other hand, the number of transitions in $\mathcal{A}_{\text{pos}}(\alpha)$ is in the worst case $n^2 + n$. Consequently, the time-complexity of any construction algorithm for $\mathcal{A}_{\text{pos}}(\alpha)$ must be at least $O(n^2)$. Considering the simplicity of the recursive definitions of the position sets used for the construction of $\mathcal{A}_{\text{pos}}(\alpha)$, an algorithm of this complexity should not be hard to find. Nevertheless, a naive implementation leads to a $O(n^3)$ algorithm, such as the one proposed by Berry and Sethi [2]. This is due to possibly non-disjoint unions of sets in the rule for $\alpha^\star$ in the recursive definition of $\mathsf{Follow}(\alpha)$. To overcome this problem, several techniques for the construction of $\mathcal{A}_{\text{pos}}(\alpha)$ were proposed over the years. The first one, of order $O(m + n^2)$, where $m = |\alpha|$, was proposed by Brüggemann-Klein in 1993 [5] and it is primarily based on the prior transformation of α into *star-normal form*. Other quadratic, however sophisticated, algorithms have been introduced in 1996 and 1997, respectively by Ponty *et al.* [18] and Chang and Paige [9].

Our goal in the next section, is to present an alternative recursive definition of $\mathsf{Follow}(\alpha)$, that only involves disjoint unions of sets, allowing for simple implementations of that construction in time $O(n^2)$. This definition also allows us to define a cost generating function of the exact number of transitions in the Glushkov automaton in Section 4.

3. A New Algorithm for Computing $\mathsf{Follow}(\alpha)$

In this section we define a new function E, such that for every marked regular expression α we have $\mathsf{Follow}(\alpha) = \mathsf{E}(\alpha)$. This function has the advantage that all unions in its definition are clearly disjoint. Our definition of E was inspired by the construction of $\mathcal{A}_{\text{pos}}(\alpha)$ by Leiss [14] and shows some similarities to the transformation algorithm of α into star-normal-form by Brüggemann-Klein. Let E and $\mathsf{E}^\star$ be given by

$$
\begin{aligned}
\mathsf{E}(\emptyset) = \mathsf{E}(\varepsilon) = \mathsf{E}(\sigma_i) &= \emptyset \\
\mathsf{E}(\alpha+\beta) &= \mathsf{E}(\alpha) \cup \mathsf{E}(\beta) \\
\mathsf{E}(\alpha\cdot\beta) &= \mathsf{E}(\alpha) \cup \mathsf{E}(\beta) \cup \mathsf{Last}(\alpha) \times \mathsf{First}(\beta) \\
\mathsf{E}(\alpha^\star) &= \mathsf{E}^\star(\alpha)
\end{aligned} \tag{2}
$$

$$
\begin{aligned}
\mathsf{E}^\star(\emptyset) = \mathsf{E}^\star(\varepsilon) &= \emptyset \\
\mathsf{E}^\star(\sigma_i) &= \{(i,i)\} \\
\mathsf{E}^\star(\alpha+\beta) &= \mathsf{E}^\star(\alpha) \cup \mathsf{E}^\star(\beta) \cup \mathsf{Cross}(\alpha,\beta) \\
\mathsf{E}^\star(\alpha\cdot\beta) &= \begin{cases}
\mathsf{E}^\star(\alpha) \ \cup \ \mathsf{E}^\star(\beta) \ \cup \ \mathsf{Cross}(\alpha,\beta) & \text{if } \varepsilon(\alpha) = \varepsilon(\beta) = \varepsilon \\
\mathsf{E}^\star(\alpha) \ \cup \ \mathsf{E}(\beta) \ \cup \ \mathsf{Cross}(\alpha,\beta) & \text{if } \varepsilon(\beta) = \varepsilon \\
\mathsf{E}(\alpha) \ \cup \ \mathsf{E}^\star(\beta) \ \cup \ \mathsf{Cross}(\alpha,\beta) & \text{if } \varepsilon(\alpha) = \varepsilon \\
\mathsf{E}(\alpha) \ \cup \ \mathsf{E}(\beta) \ \cup \ \mathsf{Cross}(\alpha,\beta) & \text{otherwise}
\end{cases} \\
\mathsf{E}^\star(\alpha^\star) &= \mathsf{E}^\star(\alpha),
\end{aligned} \tag{3}
$$

with $\mathsf{Cross}(\alpha,\beta) = \mathsf{Last}(\alpha) \times \mathsf{First}(\beta) \cup \mathsf{Last}(\beta) \times \mathsf{First}(\alpha)$.

Int. J. Found. Comput. Sci. 2012.23:969-984. Downloaded from www.worldscientific.com
by UNIVERSITY OF WATERLOO on 02/11/15. For personal use only.

In the following proposition we prove the correctness of the function E.

Proposition 1. *For every regular expression* γ *we have* $\mathsf{Follow}(\tilde{\gamma}) = \mathsf{E}(\tilde{\gamma})$.

Proof. The proof follows by induction on the structure of γ.The result is trivially true for $\tilde{\gamma} = \emptyset, \varepsilon, \sigma_i, \alpha + \beta, \alpha \cdot \beta$. For $\tilde{\gamma} = \delta^\star$, it is sufficient to show that one has $\mathsf{Follow}(\delta) \cup \mathsf{Last}(\delta) \times \mathsf{First}(\delta) = \mathsf{E}^\star(\delta)$.

For $\delta = \emptyset$, $\delta = \varepsilon$ and $\delta = \sigma_i$ this equation evaluates to $\emptyset = \emptyset$, $\emptyset = \emptyset$ and $\{(i,i)\} = \{(i,i)\}$, respectively. For $\delta = \alpha + \beta$ we have

$$\begin{aligned}
&\mathsf{Follow}(\alpha+\beta) \cup \mathsf{Last}(\alpha+\beta) \times \mathsf{First}(\alpha+\beta)\\
&\quad = (\mathsf{Follow}(\alpha) \cup \mathsf{Last}(\alpha) \times \mathsf{First}(\alpha)) \cup (\mathsf{Follow}(\beta) \cup \mathsf{Last}(\beta) \times \mathsf{First}(\beta))\\
&\qquad \cup \mathsf{Last}(\alpha) \times \mathsf{First}(\beta) \cup \mathsf{Last}(\beta) \times \mathsf{First}(\alpha)\\
&\quad = \mathsf{E}^\star(\alpha) \cup \mathsf{E}^\star(\beta) \cup \mathsf{Last}(\alpha) \times \mathsf{First}(\beta) \cup \mathsf{Last}(\beta) \times \mathsf{First}(\alpha) = \mathsf{E}^\star(\alpha+\beta).
\end{aligned}$$

We illustrate the proof for $\delta = \alpha \cdot \beta$ with the case where $\varepsilon(\alpha) \neq \varepsilon$ and $\varepsilon(\beta) = \varepsilon$:

$$\begin{aligned}
&\mathsf{Follow}(\alpha \cdot \beta) \cup \mathsf{Last}(\alpha \cdot \beta) \times \mathsf{First}(\alpha \cdot \beta)\\
&\quad = \mathsf{Follow}(\alpha) \cup \mathsf{Follow}(\beta) \cup \mathsf{Last}(\alpha) \times \mathsf{First}(\beta)\\
&\qquad \cup \mathsf{Last}(\alpha) \times \mathsf{First}(\alpha) \cup \mathsf{Last}(\beta) \times \mathsf{First}(\alpha)\\
&\quad = \mathsf{E}^\star(\alpha) \cup \mathsf{E}(\beta) \cup \mathsf{Last}(\alpha) \times \mathsf{First}(\beta) \cup \mathsf{Last}(\beta) \times \mathsf{First}(\alpha) = \mathsf{E}^\star(\alpha \cdot \beta).
\end{aligned}$$

Finally, for $\delta = \alpha^\star$ we have

$$\begin{aligned}
&\mathsf{Follow}(\alpha^\star) \cup \mathsf{Last}(\alpha^\star) \times \mathsf{First}(\alpha^\star)\\
&\quad = \mathsf{Follow}(\alpha) \cup \mathsf{Last}(\alpha) \times \mathsf{First}(\alpha) \cup \mathsf{Last}(\alpha) \times \mathsf{First}(\alpha)\\
&\quad = \mathsf{Follow}(\alpha) \cup \mathsf{Last}(\alpha) \times \mathsf{First}(\alpha) = \mathsf{E}^\star(\alpha) = \mathsf{E}^\star(\alpha^\star).
\end{aligned}$$

□

Note that the fact that no pair in $\mathsf{Last}(\alpha) \times \mathsf{First}(\beta)$ or in $\mathsf{Last}(\beta) \times \mathsf{First}(\alpha)$ can occur in $\mathsf{E}(\alpha)$, $\mathsf{E}(\beta)$, $\mathsf{E}^\star(\alpha)$ or in $\mathsf{E}^\star(\beta)$, guarantees that all unions in the definition of E and $\mathsf{E}^\star$ are disjoint.

4. The Average Number of Transitions in $\mathcal{A}_{\mathbf{pos}}$

Nicaud [17] showed that the average number of transitions in the Glushkov automaton $\mathcal{A}_{pos}(\alpha)$ is $O(|\alpha|)$. However, his computation of the number of transitions was not exact because the definition used for the Follow function did not take into account the possible non-disjoint unions of its results. In this section, based on the algorithm E we compute the exact number of transitions in $\mathcal{A}_{pos}(\alpha)$, $E_k(z)$, as well as its average cardinality, $T_k(z)$. This is done by the use of the standard methods of analytic combinatorics as expounded by Flajolet and Sedgewick [10]. These apply to generating functions $A(z) = \sum_n a_n z^n$ for a combinatorial class $\mathcal{A}$ with a_n objects of size n, or cost generating functions $C(z) = \sum_\alpha c(\alpha) z^{|\alpha|}$, where $c(\alpha)$ is some measure of the object $\alpha \in \mathcal{A}$. In this section we compute and study the cost generating functions $E_k(z)$ and $T_k(z)$, and their asymptotic behaviours. The other functions used herein, as well as some details on how to obtain them, can be found in the above cited article and in Broda *et al.* [3].

For counting purposes, we will consider regular expressions as defined in (1), but without $\emptyset$. Note that this limitation only excludes the empty language.

The functions that count the cardinalities of $\mathsf{First}(\tilde{\alpha})$, $\mathsf{Last}(\tilde{\alpha})$, and $\mathsf{E}(\tilde{\alpha})$, are respectively denoted by $\mathsf{f}(\alpha)$, $\mathsf{s}(\alpha)$, and $\mathsf{e}(\alpha)$. Given the definitions of $\mathsf{f}(\alpha)$ and $\mathsf{s}(\alpha)$, $\mathsf{e}(\alpha)$ satisfies the following:

$$\begin{aligned} \mathsf{e}(\sigma) &= \mathsf{e}(\varepsilon) = 0, \\ \mathsf{e}(\alpha+\beta) &= \mathsf{e}(\alpha) + \mathsf{e}(\beta), \\ \mathsf{e}(\alpha\cdot\beta) &= \mathsf{e}(\alpha) + \mathsf{e}(\beta) + \mathsf{s}(\alpha)\,\mathsf{f}(\beta), \\ \mathsf{e}(\alpha^\star) &= \mathsf{e}^\star(\alpha), \end{aligned} \tag{4}$$

where $\mathsf{e}^\star(\alpha)$ is given by,

$$\begin{aligned} \mathsf{e}^\star(\varepsilon) &= 0, \qquad \mathsf{e}^\star(\sigma) = 1, \\ \mathsf{e}^\star(\alpha+\beta) &= \mathsf{e}^\star(\alpha) + \mathsf{e}^\star(\beta) + \mathsf{c}(\alpha,\beta), \\ \mathsf{e}^\star(\alpha_\varepsilon\cdot\beta_\varepsilon) &= \mathsf{e}^\star(\alpha_\varepsilon) + \mathsf{e}^\star(\beta_\varepsilon) + \mathsf{c}(\alpha_\varepsilon,\beta_\varepsilon), \\ \mathsf{e}^\star(\alpha_{\overline{\varepsilon}}\cdot\beta_\varepsilon) &= \mathsf{e}^\star(\alpha_{\overline{\varepsilon}}) + \mathsf{e}(\beta_\varepsilon) + \mathsf{c}(\alpha_{\overline{\varepsilon}},\beta_\varepsilon), \\ \mathsf{e}^\star(\alpha_\varepsilon\cdot\beta_{\overline{\varepsilon}}) &= \mathsf{e}(\alpha_\varepsilon) + \mathsf{e}^\star(\beta_{\overline{\varepsilon}}) + \mathsf{c}(\alpha_\varepsilon,\beta_{\overline{\varepsilon}}), \\ \mathsf{e}^\star(\alpha_{\overline{\varepsilon}}\cdot\beta_{\overline{\varepsilon}}) &= \mathsf{e}(\alpha_{\overline{\varepsilon}}) + \mathsf{e}(\beta_{\overline{\varepsilon}}) + \mathsf{c}(\alpha_{\overline{\varepsilon}},\beta_{\overline{\varepsilon}}), \\ \mathsf{e}^\star(\alpha^\star) &= \mathsf{e}^\star(\alpha)\,, \end{aligned} \tag{5}$$

with $\mathsf{c}(\alpha,\beta) = \mathsf{s}(\alpha)\,\mathsf{f}(\beta) + \mathsf{s}(\beta)\,\mathsf{f}(\alpha)$. Then, the function

$$\mathsf{t}(\alpha) = \mathsf{f}(\alpha) + \mathsf{e}(\alpha)$$

computes the number of transitions in the Glushkov automaton of α. The cost generating function associated to t is given by

$$T_k(z) = F_k(z) + E_k(z),$$

where $F_k(z)$ and $E_k(z)$ are the cost generating functions associated to f and e, respectively. By symmetry, the cost generating functions $F_k(z)$ and $S_k(z)$ associated to f and s, respectively, are both equal to

$$S_k(z) = F_k(z) = \frac{kz}{1 - z - 3zR_k(z) - zR_{k,\varepsilon}(z)}. \tag{6}$$

In this last expression $R_k(z)$ and $R_{k,\varepsilon}(z)$ denote respectively the generating functions for regular expressions, and for regular expressions whose languages contain ε and are given by

$$R_k(z) = \frac{1 - z - \sqrt{\Delta_k(z)}}{4z} \quad \text{and} \quad R_{k,\varepsilon}(z) = \frac{z + zR_k(z)}{1 - 2zR_k(z)}, \tag{7}$$

Int. J. Found. Comput. Sci. 2012.23:969-984. Downloaded from www.worldscientific.com
by UNIVERSITY OF WATERLOO on 02/11/15. For personal use only.

Int. J. Found. Comput. Sci. 2012.23:969-984. Downloaded from www.worldscientific.com
by UNIVERSITY OF WATERLOO on 02/11/15. For personal use only.

where

$$\Delta_k(z) = 1 - 2z - (7+8k)z^2. \tag{8}$$

Hence, for the number of regular expressions of size n, one has

$$[z^n]R_k(z) \sim \frac{\sqrt{2(1-\rho_k)}}{8\rho_k\sqrt{\pi}}\rho_k^{-n}n^{-3/2}, \text{ where } \rho_k = \frac{1}{1+\sqrt{8k+8}}. \tag{9}$$

From the equations in (4) one can compute the associated cost generating functions $E_k(z)$ and $E_k^\star(z)$. For instance, the equation for concatenation contributes with the term

$$\begin{aligned}
\sum_{\alpha,\beta\in\mathcal{R}} \mathsf{e}(\alpha\cdot\beta)z^{|\alpha\cdot\beta|} &= z\sum_{\alpha\in\mathcal{R}}\sum_{\beta\in\mathcal{R}}(\mathsf{e}(\alpha)+\mathsf{e}(\beta)+\mathsf{s}(\alpha)\,\mathsf{f}(\beta))z^{|\alpha|}z^{|\beta|}\\
&= z\sum_{\alpha\in\mathcal{R}}\sum_{\beta\in\mathcal{R}}\mathsf{e}(\alpha)z^{|\alpha|}z^{|\beta|} + z\sum_{\alpha\in\mathcal{R}}\sum_{\beta\in\mathcal{R}}\mathsf{e}(\beta)z^{|\alpha|}z^{|\beta|}\\
&\quad + z\sum_{\alpha\in\mathcal{R}}\sum_{\beta\in\mathcal{R}}\mathsf{s}(\alpha)\,\mathsf{f}(\beta)z^{|\alpha|}z^{|\beta|}\\
&= z\sum_{\alpha\in\mathcal{R}}\mathsf{e}(\alpha)z^{|\alpha|}\sum_{\beta\in\mathcal{R}}z^{|\beta|} + z\sum_{\beta\in\mathcal{R}}\mathsf{e}(\beta)z^{|\beta|}\sum_{\alpha\in\mathcal{R}}z^{|\alpha|}\\
&\quad + z\sum_{\alpha\in\mathcal{R}}\mathsf{s}(\alpha)z^{|\alpha|}\sum_{\beta\in\mathcal{R}}\mathsf{f}(\beta)z^{|\beta|}\\
&= zE_k(z)R_k(z) + zE_k(z)R_k(z) + zS_k(z)F_k(z)\\
&= 2zE_k(z)R_k(z) + zF_k(z)^2
\end{aligned}$$

in the equation for $E_k(z)$. Collecting all terms the following equations must be satisfied

$$E_k(z) = 4zE_k(z)R_k(z) + zF_k(z)^2 + zE_k^\star(z)$$

$$E_k^\star(z) = kz+2zE_k^\star(z)R_k(z)+2zE_k^\star(z)R_{k,\varepsilon}(z)+4zF_k(z)^2+2zE_k(z)R_{k,\overline{\varepsilon}}(z)+zE_k^\star(z).$$

After simplification one gets

$$E_k(z) = \frac{kz^2 + zF_k(z)^2\Lambda_k(z) + 4z^2F_k(z)^2}{(1-4zR_k(z))\Lambda_k(z) - 2z^2R_{k,\overline{\varepsilon}}(z)}, \tag{10}$$

where $\Lambda_k(z) = 1 - z - 2zR_{k,\varepsilon}(z) - 2zR_k(z)$. After substituting the functions in the previous equation by their expressions (6) and (7) in terms of z and k, one obtains

$$T_k(z) = \frac{P_k(z)}{Q_k(z)\sqrt{\Delta_k(z)}}, \tag{11}$$

where

$$P_k(z) = 2kz\left(1+z+\sqrt{\Delta_k(z)}\right)(6\sqrt{\Delta_k(z)}^5 + (20+41\,z)\sqrt{\Delta_k(z)}^4$$
$$+ \left(24+12\,z^2+68\,z+8\,kz^2\right)\sqrt{\Delta_k(z)}^3$$
$$+ \left(88\,z^3k+24\,kz^2-64\,z^2+12+30\,z-122\,z^3\right)\sqrt{\Delta_k(z)}^2$$
$$+ \left(88\,z^4k-52\,z^3+2-32\,z^2+24\,kz^2+14\,z^4+4\,z+144\,z^3k\right)\sqrt{\Delta_k(z)}$$
$$+28\,z^4+z+8\,z^5k+49\,z^5+56\,z^3k+56\,z^4k-10\,z^3+8\,kz^2-4\,z^2) \tag{12}$$

and

$$Q_k(z) = \left(1-2\,z-7\,z^2+4\,(1+z)\sqrt{\Delta_k(z)}+3\,\Delta_k(z)\right)^2$$
$$\left(1-5\,z^2+2\,(1+2z)\sqrt{\Delta_k(z)}+\Delta_k(z)\right). \tag{13}$$

This function $Q_k(z)$ is positive for all values of z in the real segment $[0, \rho_k]$, because $1-2z-7z^2 = 8kz^2+\Delta_k(z)$ and $1-5z^2 = 2z+2z^2+8kz^2+\Delta_k(z)$, and $\Delta_k(z)$ is non-negative in that segment. By Pringsheim's Theorem (Theorem IV.6 of [10], p. 240) one can conclude that $T_k(z)$ has radius of convergence equal to ρ_k. Moreover, it can be shown that $T_k(z)$ has no singularities on the boundary of its disc of convergence, $\|z\| = \rho_k$, besides the one at $z = \rho_k$. In order to do that we will use the following straightforward observation.

Lemma 2. *For all $(a_n)_{n\geq 0} \in \mathbb{R}$ one has*

$$\Re\left(\sum_{n\geq 0} a_n z^n\right) \geq a_0 - \sum_{n\geq 1} |a_n|\,\|z^n\|.$$

In particular, if $f(z) = c - \sum_{n\geq 1} c_n z^n$ with $c \in \mathbb{R}$ and $(c_i)_{i\geq 1} \in \mathbb{R}_0^+$, in a neighbourhood of the origin containing the circle $\|z\| \leq \rho$ then $\Re(f(z)) \geq f(\rho)$, for all $\|z\| = \rho$.

Proof. This results directly from the fact that $\Re(z) \geq -\|z\|$. □

Using this one can now show the following result.

Lemma 3. *Both factors of $Q_k(z)$ are non-zero on the circle $\|z\| = \rho_k$.*

Proof. Let us deal first with the first factor in (13), disregarding its multiplicity. Note that it can be written in the form

$$4((1+z)\sqrt{\Delta_k(z)}+\Delta_k(z)+2kz^2). \tag{14}$$

Now $\Delta_k(z)+2kz^2 = 1-2z-7z^2-6kz^2$ and for $||z|| = \rho_k$ one has

$$\Re(1-2z-7z^2-6kz^2) \geq 1-2\rho_k-7\rho_k^2-6k\rho_k^2 = 2k\rho_k^2 > 0, \tag{15}$$

using the fact that $1-2\rho_k-7\rho_k^2-8k\rho_k^2 = 0$. Therefore $\Re(\Delta_k(z)+2kz^2) > 0$.

Int. J. Found. Comput. Sci. 2012.23:969-984. Downloaded from www.worldscientific.com
by UNIVERSITY OF WATERLOO on 02/11/15. For personal use only.

Int. J. Found. Comput. Sci. 2012.23:969-984. Downloaded from www.worldscientific.com
by UNIVERSITY OF WATERLOO on 02/11/15. For personal use only.

On the other hand, by Eq. (7) it follows that $(1+z)\sqrt{\Delta_k(z)} = 1 - z^2 - 4z(1+z)R_k(z)$, and since $R_k(z)$ is represented by a series with non negative coefficients, the Lemma 2 is applicable to $(1+z)\sqrt{\Delta_k(z)}$ yielding $\Re((1+z)\sqrt{\Delta_k(z)}) \geq 0$. From this, and the result of the previous paragraph one concludes that the first factor has no zero in the circle $\|z\| = \rho_k$.

As for the second factor, it can be written as

$$2\left(1 - z - 6z^2 - 4kz^2 + (2z+1)\sqrt{\Delta_k(z)}\right).$$

As done in (15), it is easy to see that

$$\Re(1 - z - 6z^2 - 4kz^2) \geq \rho_k + \rho_k^2 + 4k\rho_k^2 > 0. \tag{16}$$

And thus one gets the desired result in the same way as was just done for the other factor. □

Similarly to what was done in [4], using Proposition 1 in [17] which follows easily for Theorem VI.1 in [10], one obtains

$$T_k(z) = \frac{P_k(\rho_k)}{\sqrt{2-2\rho_k}\, Q_k(\rho_k)} \frac{1}{\sqrt{1 - z/\rho_k}} + o\left(\frac{1}{\sqrt{1 - z/\rho_k}}\right) \tag{17}$$

from which it follows that

$$[z^n]T_k(z) \sim \frac{P_k(\rho_k)}{\sqrt{\pi}\,\sqrt{2-2\rho_k}\, Q_k(\rho_k)} \rho_k^{-n} n^{-1/2}. \tag{18}$$

Using the actual expression of P_k one has

$$[z^n]T_k(z) \sim \frac{(1+\rho_k)(2 + 16\rho_k + 10\rho_k^2 - 12\rho_k^3)}{8\rho_k\sqrt{\pi}(1 - 5\rho_k^2)\sqrt{2-2\rho_k}} \rho_k^{-n} n^{-1/2}. \tag{19}$$

Considering the cost generating function for the number of letters in an element $\alpha \in \mathcal{R}$, computed by Nicaud to be equal to

$$Let_k(z) = \frac{kz}{\sqrt{\Delta_k(z)}},$$

and for which

$$[z^n]Let_k(z) \sim \frac{k\rho_k}{\sqrt{\pi(2-2\rho_k)}} \rho_k^{-n} n^{-1/2},$$

one gets the asymptotic expression for the average number of transitions per state, and for the average number of transitions per regular expression, stated in the following result.

Theorem 4. *With the same notation as above,*

$$\frac{[z^n]T_k(z)}{[z^n]Let_k(z)} \sim \frac{(1+\rho_k)(2 + 16\rho_k + 10\rho_k^2 - 12\rho_k^3)}{(1 - 2\rho_k - 7\rho_k^2)(1 - 5\rho_k^2)}, \tag{20}$$

$$\frac{[z^n]T_k(z)}{[z^n]R_k(z)} \sim \frac{(1+\rho_k)(1 + 8\rho_k + 5\rho_k^2 - 6\rho_k^3)}{(1 - \rho_k)(1 - 5\rho_k^2)}\, n. \tag{21}$$

Int. J. Found. Comput. Sci. 2012.23:969-984. Downloaded from www.worldscientific.com
by UNIVERSITY OF WATERLOO on 02/11/15. For personal use only.

Since ρ_k tends to 0 as k goes to ∞, it follows that for large k the average number of transitions per state is approximately 2, while the average number of transitions per automaton is approximately the size of the original regular expression.

5. The Average Number of Transitions in $\mathcal{A}_{\text{pd}}$

The partial derivative automaton $\mathcal{A}_{\text{pd}}(\alpha)$ of a regular expression α was defined independently by Mirkin's [16] and Antimirov [1]. Champarnaud and Ziadi stated the equivalence of the two formulations [7], and proved that $\mathcal{A}_{\text{pd}}$ is a quotient of the Glushkov automaton $\mathcal{A}_{\text{pos}}$ [8]. This means that $\mathcal{A}_{\text{pd}}(\alpha)$ can be obtained from $\mathcal{A}_{\text{pos}}(\alpha)$ by the merging of states belonging to the same equivalence class. That, on the other hand, may lead to the merging of transitions. In this section, we estimate the average number of transitions of $\mathcal{A}_{\text{pd}}(\alpha)$ when compared with the ones of $\mathcal{A}_{\text{pos}}(\alpha)$. For this, it is essential to have the exact counting of the number of transitions of $\mathcal{A}_{\text{pos}}(\alpha)$ obtained in Section 4.

The $\mathcal{A}_{\text{pd}}(\alpha)$ can be defined using the notion of partial derivative, introduced by Antimirov as a non-deterministic version of Brzozowski's derivative [6].

For a regular expression α and a letter $\sigma \in \Sigma$, the set $\partial_\sigma(\alpha)$ of *partial derivatives* of α w.r.t. σ is defined inductively as follows:

$$\begin{aligned}
&\partial_\sigma(\emptyset) = \partial_\sigma(\varepsilon) = \emptyset & \partial_\sigma(\alpha+\beta) &= \partial_\sigma(\alpha) \cup \partial_\sigma(\beta)\\
&\partial_\sigma(\sigma') = \begin{cases} \{\varepsilon\}, & \text{if } \sigma' = \sigma \\ \emptyset, & \text{otherwise} \end{cases} & \partial_\sigma(\alpha_\varepsilon \cdot \beta) &= \partial_\sigma(\alpha_\varepsilon)\beta \cup \partial_\sigma(\beta)\\
&\partial_\sigma(\alpha^\star) = \partial_\sigma(\alpha)\alpha^\star & \partial_\sigma(\alpha_{\overline{\varepsilon}} \cdot \beta) &= \partial_\sigma(\alpha_{\overline{\varepsilon}})\beta
\end{aligned}$$

where for any $S \subseteq \mathcal{R}$, $S\emptyset = \emptyset S = \emptyset$, and $S\varepsilon = \varepsilon S = S$. This definition can be extended to sets of regular expressions, to words, and to languages in the obvious way. The *set of partial derivatives* of α, $\{\, \partial_w(\alpha) \mid w \in \Sigma^\star \,\}$, is denoted by $\mathsf{P}(\alpha)$. The *partial derivative automaton* $\mathcal{A}_{\text{pd}}(\alpha)$ is defined by $\mathcal{A}_{\text{pd}}(\alpha) = (\mathsf{P}(\alpha), \Sigma, \delta_{\text{pd}}, \alpha, \{\, q \in \mathsf{P}(\alpha) \mid \varepsilon(q) = \varepsilon \,\})$, where $\delta_{\text{pd}}(q, \sigma) = \partial_\sigma(q)$, for all $q \in \mathsf{P}(\alpha)$ and $\sigma \in \Sigma$. Antimirov proved that $\mathcal{L}(\mathcal{A}_{\text{pd}}(\alpha)) = \mathcal{L}(\alpha)$.

5.1. *An alternative recursive definition of $\mathcal{A}_{\text{pd}}(\alpha)$*

Using Mirkin's formulation one has $\mathsf{P}(\alpha) = \pi(\alpha) \cup \{\alpha\}$, where the set $\pi(\alpha)$ is inductively defined as follows:

$$\begin{aligned}
\pi(\emptyset) &= \emptyset & \pi(\alpha+\beta) &= \pi(\alpha) \cup \pi(\beta)\\
\pi(\varepsilon) &= \emptyset & \pi(\alpha \cdot \beta) &= \pi(\alpha)\beta \cup \pi(\beta)\\
\pi(\sigma) &= \{\varepsilon\} & \pi(\alpha^\star) &= \pi(\alpha)\alpha^\star.
\end{aligned} \tag{22}$$

The algorithm presented by Mirkin to compute the $\pi(\alpha)$, and thus the set of partial derivatives of α, provides also an inductive definition of the set of transitions of

Int. J. Found. Comput. Sci. 2012.23:969-984. Downloaded from www.worldscientific.com
by UNIVERSITY OF WATERLOO on 02/11/15. For personal use only.

$\mathcal{A}_{\mathrm{pd}}$. By the definition of $\mathcal{A}_{\mathrm{pd}}$, and because $\mathsf{P}(\alpha) = \pi(\alpha) \cup \{\alpha\}$, one has

$$\delta_{\mathrm{pd}} = \{\, (\alpha, \sigma, \gamma) \mid \gamma \in \partial_\sigma(\alpha),\ \sigma \in \Sigma \,\} \ \cup$$
$$\bigcup_{\alpha' \in \pi(\alpha)} \{\, (\alpha', \sigma, \gamma') \mid \gamma' \in \partial_\sigma(\alpha'),\ \sigma \in \Sigma \,\} = \{\alpha\} \times \varphi(\alpha) \cup \mathsf{F}(\alpha)$$

where $\varphi(\alpha) = \{\, (\sigma, \gamma) \mid \gamma \in \partial_\sigma(\alpha),\ \sigma \in \Sigma \,\}$, with the result of the $\times$ operation seen as a set of triples. The set F can be inductively defined by

$$\begin{aligned}
\mathsf{F}(\emptyset) = \mathsf{F}(\varepsilon) = \mathsf{F}(\sigma) &= \emptyset,\ \sigma \in \Sigma \\
\mathsf{F}(\alpha + \beta) &= \mathsf{F}(\alpha) \cup \mathsf{F}(\beta) \\
\mathsf{F}(\alpha \cdot \beta) &= \mathsf{F}(\alpha)\beta \cup \mathsf{F}(\beta) \cup \lambda(\alpha)\beta \times \varphi(\beta) \\
\mathsf{F}(\alpha^\star) &= \mathsf{F}(\alpha)\alpha^\star \cup (\lambda(\alpha) \times \varphi(\alpha))\alpha^\star,
\end{aligned} \tag{23}$$

where $\lambda(\alpha) = \{\, \alpha' \mid \alpha' \in \pi(\alpha),\ \varepsilon(\alpha') = \varepsilon \,\}$, and the concatenation of a transition (α, σ, β) with a regular expression is defined by $(\alpha, \sigma, \beta)\gamma = (\alpha\gamma, \sigma, \beta\gamma)$, if $\gamma \notin \{\emptyset, \varepsilon\}$, $(\alpha, \sigma, \beta)\emptyset = \emptyset$, and $(\alpha, \sigma, \beta)\varepsilon = (\alpha, \sigma, \beta)$.

It is important to note that the sets F, λ, and φ correspond, respectively, to the sets Follow, Last, and First, modulo the equivalence relation that defines $\mathcal{A}_{\mathrm{pd}}$ as a quotient of $\mathcal{A}_{\mathrm{pos}}$ [8]. From this, it follows that $|\lambda(\alpha)| \leq \mathsf{s}(\alpha)$ and $|\varphi(\alpha)| \leq \mathsf{f}(\alpha)$. Moreover the functions E_{pd} and $\mathsf{E}^\star_{\mathrm{pd}}$ corresponding to E and $\mathsf{E}^\star$ for this construction of the partial derivative automaton, are defined by

$$\begin{aligned}
\mathsf{E}_{\mathrm{pd}}(\emptyset) = \mathsf{E}_{\mathrm{pd}}(\varepsilon) = \mathsf{E}_{\mathrm{pd}}(\sigma) &= \emptyset \\
\mathsf{E}_{\mathrm{pd}}(\alpha + \beta) &= \mathsf{E}_{\mathrm{pd}}(\alpha) \cup \mathsf{E}_{\mathrm{pd}}(\beta) \\
\mathsf{E}_{\mathrm{pd}}(\alpha \cdot \beta) &= \mathsf{E}_{\mathrm{pd}}(\alpha)\beta \cup \mathsf{E}_{\mathrm{pd}}(\beta) \cup (\lambda(\alpha)\beta \times \varphi(\beta)) \\
\mathsf{E}_{\mathrm{pd}}(\alpha^\star) &= \mathsf{E}^\star_{\mathrm{pd}}(\alpha)\alpha^\star
\end{aligned} \tag{24}$$

$$\begin{aligned}
\mathsf{E}^\star_{\mathrm{pd}}(\emptyset) = \mathsf{E}^\star_{\mathrm{pd}}(\varepsilon) &= \emptyset \\
\mathsf{E}^\star_{\mathrm{pd}}(\sigma) &= \{(\varepsilon, \sigma, \varepsilon)\} \\
\mathsf{E}^\star_{\mathrm{pd}}(\alpha + \beta) &= \mathsf{E}^\star_{\mathrm{pd}}(\alpha) \cup \mathsf{E}^\star_{\mathrm{pd}}(\beta) \cup (\lambda(\alpha) \times \varphi(\beta)) \cup (\lambda(\beta) \times \varphi(\alpha)) \\
\mathsf{E}^\star_{\mathrm{pd}}(\alpha_\varepsilon \cdot \beta_\varepsilon) &= \mathsf{E}^\star_{\mathrm{pd}}(\alpha_\varepsilon)\beta_\varepsilon \ \cup\ \mathsf{E}^\star_{\mathrm{pd}}(\beta_\varepsilon) \ \cup\ \mathsf{Cross}_1(\alpha_\varepsilon, \beta_\varepsilon) \\
\mathsf{E}^\star_{\mathrm{pd}}(\alpha_{\overline{\varepsilon}} \cdot \beta_\varepsilon) &= \mathsf{E}^\star_{\mathrm{pd}}(\alpha_{\overline{\varepsilon}})\beta_\varepsilon \ \cup\ \mathsf{E}_{\mathrm{pd}}(\beta_\varepsilon) \ \cup\ \mathsf{Cross}_1(\alpha_{\overline{\varepsilon}}, \beta_\varepsilon) \\
\mathsf{E}^\star_{\mathrm{pd}}(\alpha_\varepsilon \cdot \beta_{\overline{\varepsilon}}) &= \mathsf{E}_{\mathrm{pd}}(\alpha_\varepsilon)\beta_{\overline{\varepsilon}} \ \cup\ \mathsf{E}^\star_{\mathrm{pd}}(\beta_{\overline{\varepsilon}}) \ \cup\ \mathsf{Cross}_1(\alpha_\varepsilon, \beta_{\overline{\varepsilon}}) \\
\mathsf{E}^\star_{\mathrm{pd}}(\alpha_{\overline{\varepsilon}} \cdot \beta_{\overline{\varepsilon}}) &= \mathsf{E}_{\mathrm{pd}}(\alpha_{\overline{\varepsilon}})\beta_{\overline{\varepsilon}} \ \cup\ \mathsf{E}_{\mathrm{pd}}(\beta_{\overline{\varepsilon}}) \ \cup\ \mathsf{Cross}_1(\alpha_{\overline{\varepsilon}}, \beta_{\overline{\varepsilon}}) \\
\mathsf{E}^\star_{\mathrm{pd}}(\alpha^\star) &= \mathsf{E}^\star_{\mathrm{pd}}(\alpha)\alpha^\star,
\end{aligned} \tag{25}$$

with $\mathsf{Cross}_1(\alpha, \beta) = (\lambda(\alpha)\beta \times \varphi(\beta)) \cup (\lambda(\beta) \times \varphi(\alpha)\beta)$.

The unions in these equalities are not necessarily disjoint. A lower bound for the cardinality of these sets will be computed in the next section.

5.2. *A lower bound for the number of transitions*

Broda *et al.* [3] gave a lower bound of the number of mergings of states from $\mathsf{Pos}(\alpha)$ to $\mathcal{A}_{\mathrm{pd}}(\alpha)$ computed with $\pi(\alpha)$, which allowed to obtain an upper bound for the

Int. J. Found. Comput. Sci. 2012.23:969-984. Downloaded from www.worldscientific.com
by UNIVERSITY OF WATERLOO on 02/11/15. For personal use only.

average state complexity of $\mathcal{A}_{\text{pd}}(\alpha)$. There, it was observed that the merging of states is primarily caused by sub-expressions γ of α such that $\varepsilon \in \pi(\gamma)$. In fact, the presence of sub-expressions with this property, denoted by α_{π_ε}, make some unions not disjoint, during the computation of $\pi(\alpha)$.

In this section, and using an analogous technique, we determine a lower bound for the number of transitions.

Considering the definition of E_{pd} (24), and in particular the concatenation case, it is easy to see that a merging of two states in the set $\lambda(\alpha)$, relative to $\mathsf{Last}(\alpha)$, gives origin to $|\varphi(\beta)| \leq \mathsf{f}(\beta)$ mergings of transitions. Although there can be merging of states corresponding to $\varphi(\beta)$, they will not be considered in the computation of that lower bound. We first compute a lower bound for the number of mergings $i_\ell(\alpha)$ of states in $\lambda(\alpha)$ with respect to $\mathsf{Last}(\alpha)$.

The definition of $\lambda(\alpha)$ coincides with that of $\pi(\alpha)$ except for the case of concatenation (i.e corresponds to the definition of Last), which is:

$$\begin{aligned} \lambda(\alpha \cdot \beta_\varepsilon) &= \lambda(\alpha)\beta_\varepsilon \cup \lambda(\beta_\varepsilon) \\ \lambda(\alpha \cdot \beta_{\overline{\varepsilon}}) &= \lambda(\beta_{\overline{\varepsilon}}). \end{aligned} \tag{26}$$

Following an observation by Broda *et al.*, whenever $\epsilon \in \lambda(\alpha) \cap \lambda(\beta)$,

$$|\lambda(\alpha + \beta)| \leq |\lambda(\alpha)| + |\lambda(\beta)| - 1 \tag{27}$$

and, if β belongs to the subclass of regular expressions $\alpha_{r,\varepsilon}$ such that $\alpha_{r,\varepsilon} \in \lambda(\alpha_{r,\varepsilon})$, we have

$$|\lambda(\alpha \cdot \beta)| = |\lambda(\alpha)\beta \cup \lambda(\beta)| \leq |\lambda(\alpha)| + |\lambda(\beta)| - 1. \tag{28}$$

The set of regular expressions for which $\epsilon \in \lambda(\alpha)$ coincides trivially with the set of regular expressions for which $\epsilon \in \pi(\alpha)$, which was defined in Broda *et al.* and denoted by α_{π_ε}. The associated generating function is

$$R_{k,\pi_\varepsilon}(z) = \frac{z^2 + 3zR_k(z) - 1 + \sqrt{\left(z^2 + 3zR_k(z) - 1\right)^2 + 4kz^2}}{2z}.$$

For $\alpha_{r,\varepsilon}$ one has

$$\alpha_{r,\varepsilon} := \alpha^\star_{\pi_\varepsilon} \mid \alpha_{r,\varepsilon} \cdot \alpha_\varepsilon, \quad \text{and} \quad R_{k,r,\varepsilon}(z) = \frac{zR_{k,\pi_\varepsilon}(z)}{1 - zR_{k,\varepsilon}(z)}.$$

Thus, the number of mergings of states in $\lambda(\alpha)$ due to (27) and (28), denoted by $i_\ell(\alpha)$, can be defined by the following:

$$\begin{aligned} i_\ell(\emptyset) = i_\ell(\varepsilon) &= i_\ell(\sigma) = 0, \\ i_\ell(\alpha_{\pi_\varepsilon} + \alpha_{\pi_\varepsilon}) &= i_\ell(\alpha_{\pi_\varepsilon}) + i_\ell(\alpha_{\pi_\varepsilon}) + 1, \\ i_\ell(\alpha_{\pi_\varepsilon} + \alpha_{\overline{\pi_\varepsilon}}) &= i_\ell(\alpha_{\pi_\varepsilon}) + i_\ell(\alpha_{\overline{\pi_\varepsilon}}), \\ i_\ell(\alpha_{\overline{\pi_\varepsilon}} + \alpha) &= i_\ell(\alpha_{\overline{\pi_\varepsilon}}) + i_\ell(\alpha), \\ i_\ell(\alpha^\star) &= i_\ell(\alpha), \end{aligned}$$

$$\begin{aligned} i_\ell(\alpha_{\pi_\varepsilon} \cdot \alpha_{r,\varepsilon}) &= i_\ell(\alpha_{\pi_\varepsilon}) + i_\ell(\alpha_{r,\varepsilon}) + 1, \\ i_\ell(\alpha_{\pi_\varepsilon} \cdot \alpha_{r,\overline{\varepsilon}}) &= i_\ell(\alpha_{r,\overline{\varepsilon}}), \\ i_\ell(\alpha_{\pi_\varepsilon} \cdot \alpha_{\overline{r},\varepsilon}) &= i_\ell(\alpha_{\pi_\varepsilon}) + i_\ell(\alpha_{\overline{r},\varepsilon}), \\ i_\ell(\alpha_{\pi_\varepsilon} \cdot \alpha_{\overline{r},\overline{\varepsilon}}) &= i_\ell(\alpha_{\overline{r},\overline{\varepsilon}}), \\ i_\ell(\alpha_{\overline{\pi_\varepsilon}} \cdot \alpha_\varepsilon) &= i_\ell(\alpha_{\overline{\pi_\varepsilon}}) + i_\ell(\alpha_\varepsilon), \\ i_\ell(\alpha_{\overline{\pi_\varepsilon}} \cdot \alpha_{\overline{\varepsilon}}) &= i_\ell(\alpha_{\overline{\varepsilon}}), \end{aligned}$$

where $\gamma_{\overline{x}}$ denotes the complement of γ_x.

Int. J. Found. Comput. Sci. 2012.23:969-984. Downloaded from www.worldscientific.com
by UNIVERSITY OF WATERLOO on 02/11/15. For personal use only.

To illustrate the previous rules, consider the case of $i_\ell(\alpha_{\pi_\varepsilon}\alpha_{r,\varepsilon})$. Here, one has $\lambda(\alpha_{\pi_\varepsilon}\alpha_{r,\varepsilon}) = \lambda(\alpha_{\pi_\varepsilon})\alpha_{r,\varepsilon} \cup \lambda(\alpha_{r,\varepsilon})$. By rule (28) there is a merging of two states, i.e the two occurrences of $\alpha_{r,\varepsilon}$, which is accounted for by the 1 in the definition of $i_\ell(\alpha_{\pi_\varepsilon}\alpha_{r,\varepsilon})$.

The generating function, $Il_k(z)$, of i_ℓ satisfies the following:

$$Il_k(z) = \frac{zR_{k,\pi_\varepsilon}(z)^2 + zR_{k,\pi_\varepsilon}(z)R_{k,r,\varepsilon}(z)}{1 - z - 3zR_k(z) - zR_{k,\varepsilon}(z)}.$$

Using (24) and (25), one can easily define a function that computes a lower bound for the number of transition mergings:

$$\begin{aligned}
i_t(\varepsilon) = i_t(\sigma) &= 0 \\
i_t(\alpha+\beta) &= i_t(\alpha) + i_t(\beta) \\
i_t(\alpha\cdot\beta) &= i_t(\alpha) + i_t(\beta) + i_\ell(\alpha)\,\mathsf{f}(\beta) \\
i_t(\alpha^\star) &= i_t^\star(\alpha) \\
i_t^\star(\varepsilon) &= 0 \\
i_t^\star(\sigma) &= 1
\end{aligned}$$

$$\begin{aligned}
i_t^\star(\alpha+\beta) &= i_t^\star(\alpha) + i_t^\star(\beta) + \mathsf{c}_t(\alpha,\beta) \\
i_t^\star(\alpha_\varepsilon\cdot\beta_\varepsilon) &= i_t^\star(\alpha_\varepsilon) + i_t^\star(\beta_\varepsilon) + \mathsf{c}_t(\alpha_\varepsilon,\beta_\varepsilon) \\
i_t^\star(\alpha_{\overline{\varepsilon}}\cdot\beta_\varepsilon) &= i_t^\star(\alpha_{\overline{\varepsilon}}) + i_t(\beta_\varepsilon) + \mathsf{c}_t(\alpha_{\overline{\varepsilon}},\beta_\varepsilon) \\
i_t^\star(\alpha_\varepsilon\cdot\beta_{\overline{\varepsilon}}) &= i_t(\alpha_\varepsilon) + i_t^\star(\beta_{\overline{\varepsilon}}) + \mathsf{c}_t(\alpha_\varepsilon,\beta_{\overline{\varepsilon}}) \\
i_t^\star(\alpha_{\overline{\varepsilon}}\cdot\beta_{\overline{\varepsilon}}) &= i_t(\alpha_{\overline{\varepsilon}}) + i_t(\beta_{\overline{\varepsilon}}) + \mathsf{c}_t(\alpha_{\overline{\varepsilon}},\beta_{\overline{\varepsilon}}) \\
i_t^\star(\alpha^\star) &= i_t^\star(\alpha)
\end{aligned}$$

where $\mathsf{c}_t(\alpha,\beta) = i_\ell(\alpha)\,\mathsf{f}(\beta) + i_\ell(\beta)\,\mathsf{f}(\alpha)$. The corresponding generating function satisfies the following:

$$It_k(z) = \frac{z\Lambda_k(z)Il_k(z)F_k(z) + kz^2 + 4z^2Il_k(z)F_k(z)}{(1-4zR_k(z))\Lambda_k(z) - 2z^2R_{k,\overline{\varepsilon}}(z)},$$

where $\Lambda_k(z) = 1 - z - 2zR_k(z) - 2zR_{k,\varepsilon}(z)$.

Lemma 5. *The function $It_k(z)$ does not have any other singularity on the circle $\|z\| \geq \rho_k$, besides ρ_k.*

Proof. It turns out that $It_k(z)$ can be written as a quotient of two polynomials[a] in $\sqrt{\Delta_k(z)}$ and $\sqrt{(z^2+3zR_k(z)-1)^2+4kz^2}$ where its denominator is

$$Q_k(z)\left(2+z-3z^2+(z+2)\sqrt{\Delta_k(z)}\right). \tag{29}$$

It is clear that $\sqrt{(z^2+3zR_k(z)-1)^2+4kz^2}$ does not introduce any singularity in the real axis since it is built from a sum of two squares. Therefore, by Pringsheim Theorem, there can be no singularity in the interior of the circle.

For the circle $\|z\| = \rho_k$ we need to show that $\left(z^2+3zR_k(z)-1\right)^2+4kz^2$ is not null and thus does not introduce any new singularity. For this is enough to prove that its real part is always positive. Observe that

$$\begin{aligned}
&\left(z^2+3zR_k(z)-1\right)^2+4kz^2 \\
&\quad= \frac{5}{8} - \frac{3}{4}z - \frac{31}{8}z^2 - \frac{1}{2}kz^2 - \frac{3}{2}z^3 + z^4 + \frac{3}{8}(1-z)(1+4z)\sqrt{\Delta_k(z)}.
\end{aligned} \tag{30}$$

[a]For more details see [4] and the companion Maple file there referred.

Int. J. Found. Comput. Sci. 2012.23:969-984. Downloaded from www.worldscientific.com
by UNIVERSITY OF WATERLOO on 02/11/15. For personal use only.

Using the fact that

$$(1-z)\sqrt{\Delta_k(z)} = (1-z)^2 - (1-z)4zR_k(z),$$

and that the coefficients of $R_k(z)$, the generating function for the number of regular expressions, form an increasing sequence of non-negative integers, one can see that the power series of the expression in (30) has all but the first coefficient non-negative. Thus the Lemma 2 is applicable and, recalling that $\Delta_k(\rho_k) = 0$, one gets

$$\Re\left(\left(z^2 + 3zR_k(z) - 1\right)^2 + 4kz^2\right) \geq \frac{5}{8} - \frac{3}{4}\,\rho_k - \frac{31}{8}{\rho_k}^2 - \frac{1}{2}k{\rho_k}^2 - \frac{3}{2}{\rho_k}^3 + {\rho_k}^4.$$

This is always positive, since $\rho_k \leq 1/5$ and $k\rho_k^2 \leq 1/8$, for all k.

By Lemma 3 one already knows that no new singularities are introduced by $Q_k(z)$. To deal with the remaining factor of the denominator of $It_k(z)$, using exactly what was done in that same Lemma, one only needs to remark that, because $-1 + 2\rho_k + (7+8k)\rho_k^2) = 0$,

$$\Re(2 + z - 3z^2) \geq 2 - \rho_k - 3\rho_k^2 = 1 + \rho_k + 4\rho_k^2 + 8k\rho_k^2 > 0 \tag{31}$$

ensuring that this new factor introduces no new singularity. □

Using the same results mentioned regarding Equation 17 (pag. 977) one has

$$[z^n]It_k(z) \sim \frac{(1+\rho_k)\,(a(\rho_k)b(\rho_k) + c(\rho_k))}{16\,\sqrt{\pi}\,\rho_k\,\sqrt{2-2\rho_k}\,(1-5\rho_k^2)d(\rho_k)}\rho_k^{-n}n^{-1/2} \tag{32}$$

where

$$\begin{aligned}
a(z) &= -2 - 23z - 77z^2 - 50z^3 + 92z^4 + 77z^5 - 13z^6 - 4z^7\\
b(z) &= \sqrt{9 - 10z - 55z^2 - 24z^3 + 16z^4}\\
c(z) &= 10 + 89z + 54z^2 - 603z^3 - 1114z^4 - 349z^5\\
&\quad + 130z^6 - 209z^7 - 40z^8 - 16z^9\\
d(z) &= (1 - 2z - 7z^2)(2 + z - 3z^2)\,.
\end{aligned}$$

From all this, one obtains the lower bound for the average number of mergings per transition of the Glushkov automaton expressed in the following result.

Theorem 6. *With the same notation used above,*

$$[z^n]\frac{It_k(z)}{T_k(z)} \sim \frac{a(\rho_k)b(\rho_k) + c(\rho_k)}{4(1 + 8\rho_k + 5\rho_k^2 - 6\rho_k^3)d(\rho_k)}. \tag{33}$$

This means that, asymptotically with respect to k, the number of transitions in $\mathcal{A}_{\mathrm{pd}}$ is at most half the number of transitions in $\mathcal{A}_{pos}$.

Int. J. Found. Comput. Sci. 2012.23:969-984. Downloaded from www.worldscientific.com
by UNIVERSITY OF WATERLOO on 02/11/15. For personal use only.

6. Comparison with Experimental Results

We compared the estimates obtained in the previous sections with some experimental results. For each $k \in \{2, 3, 10, 30, 50\}$, the experiment consisted of the comparison of the sizes of $\mathcal{A}_{\text{pos}}$ and $\mathcal{A}_{\text{pd}}$, that were computed for each regular expression in the samples of 1000 uniform random generated regular expressions of size 1000. Table 1 presents the average values obtained, and columns seven and nine give the asymptotic ratios obtained in (21) and in (33), respectively. The quality of the approximation of the asymptotic average number of transitions per state for $\mathcal{A}_{\text{pos}}$, and that the actual values are close to the limit even for relatively small alphabets is evident from the table. The experimental values also suggest that the number of transitions of $\mathcal{A}_{\text{pd}}$ is on average, and as k grows, the alphabetic size of the original regular expression.

Table 1. Experimental results for uniform random generated regular expressions.

k	$\lvert\alpha\rvert_\Sigma$	$\lvert\delta_{\text{pos}}\rvert$	$\lvert\mathsf{P}(\alpha)\rvert$	$\lvert\delta_{\text{pd}}\rvert$	$\frac{\lvert\delta_{\text{pos}}\rvert}{\lvert\alpha\rvert_\Sigma+1}$	$\frac{[z^n]T_k(z)}{[z^n]Let_k(z)+1}$	$\frac{\lvert\delta_{\text{pd}}\rvert}{\lvert\delta_{\text{pos}}\rvert}$	$1-\frac{[z^n]It_k(z)}{[z^n]T_k(z)}$
2	276	3345	187	1806	**12.1**	**12.2**	**0.54**	**0.68**
3	318	2997	206	1564	**9.4**	**9.6**	**0.52**	**0.64**
10	405	2203	236	1079	**5.4**	**5.3**	**0.49**	**0.58**
30	453	1676	247	796	**3.7**	**3.6**	**0.47**	**0.54**
50	466	1516	250	718	**3.3**	**3.2**	**0.47**	**0.53**
100	–	–	–	–	–	**2.8**	–	**0.53**
1000	–	–	–	–	–	**2.2**	–	**0.49**

7. Conclusions

In this paper we presented a new algorithm for computing Follow, which is quadratic in the alphabetic size of the original regular expression, and that leads to a straightforward and direct implementation. This algorithm allowed us to exactly count the number of transitions of the Glushkov automaton. Using this, we computed the average number of transitions of that automaton, concluding that, for large alphabets, it is approximately the double of the original regular expression alphabetic size.

We present a recursive definition of the set of transitions of the partial derivative automaton, already implicit in Mirkin's construction, and that allowed the extension to $\mathcal{A}_{\text{pd}}$ of the new algorithm for computing $\mathcal{A}_{\text{pos}}$. Considering special sub-classes of regular expressions that are primarily responsible for state mergings, we computed an upper bound for the number of transitions in the partial derivative automaton. We, then, used analytic combinatorial methods to obtain average values and asymptotic limits for that number, concluding that, on average and asymptotically, the partial derivative automaton has at most half the number of transitions of the Glushkov's. Experimental figures corroborate these results.

Int. J. Found. Comput. Sci. 2012.23:969-984. Downloaded from www.worldscientific.com
by UNIVERSITY OF WATERLOO on 02/11/15. For personal use only.

Acknowledgments

The authors would like to thank the valuable remarks and corrections suggested by the anonymous referees.

References

[1] V. M. Antimirov. Partial derivatives of regular expressions and finite automaton constructions. *Theoret. Comput. Sci.*, 155(2):291–319, 1996.

[2] G. Berry and R. Sethi. From regular expressions to deterministic automata. *Theoret. Comput. Sci.*, 48(3):117–126, 1986.

[3] S. Broda, A. Machiavelo, N. Moreira, and R. Reis. On the average state complexity of partial derivative automata: An analytic combinatorics approach. *International Journal of Foundations of Computer Science*, 22(7):1593–1606, 2011.

[4] S. Broda, A. Machiavelo, N. Moreira, and R. Reis. A study on the average size of Glushkov and partial derivative automata. Technical Report DCC-2011-03, FCUP and CMUP, Universidade do Porto, April 2011.

[5] A. Brüggemann-Klein. Regular expressions into finite automata. *Theoret. Comput. Sci.*, 120(2):197–213, 1993.

[6] J. A. Brzozowski. Derivatives of regular expressions. *JACM*, 11(4):481–494, October 1964.

[7] J. M. Champarnaud and D. Ziadi. From Mirkin's prebases to Antimirov's word partial derivatives. *Fundam. Inform.*, 45(3):195–205, 2001.

[8] J. M. Champarnaud and D. Ziadi. Canonical derivatives, partial derivatives and finite automaton constructions. *Theoret. Comput. Sci.*, 289(1):137–163, 2002.

[9] C.-H. Chang and R. Paige. From regular expressions to DFA's using compressed NFA's. *Theor. Comput. Sci.*, 178(1-2):1–36, 1997.

[10] P. Flajolet and R. Sedgewick. *Analytic Combinatorics*. Cambridge University Press, 2008.

[11] V. M. Glushkov. The abstract theory of automata. *Russian Math. Surveys*, 16(5): 1–53, 1961.

[12] M. Holzer and M. Kutrib. The complexity of regular(-like) expressions. In Y. Gao, H. Lu, S.Seki, and S. Yu, editors, *Proc. 14th DLT 2010*, volume 6224 of *LNCS*, pages 16–30. Springer, 2010.

[13] D. C. Kozen. *Automata and Computability*. Springer, 1997.

[14] E. Leiss. Constructing a finite automaton for a given regular expression. *SIGACT News*, 12(3):81–87, Sep 1980.

[15] R. McNaughton and H. Yamada. Regular expressions and state graphs for automata. *IEEE Transactions on Electronic Computers*, 9(1):39–47, 1960.

[16] B. G. Mirkin. An algorithm for constructing a base in a language of regular expressions. *Engineering Cybernetics*, 5:110—116, 1966.

[17] C. Nicaud. On the average size of Glushkov's automata. In A. Dediu, A.-M. Ionescu, and C. M. Vide, editors, *3rd LATA 2009*, volume 5457 of *LNCS*, pages 626–637. Springer, 2009.

[18] J.-L. Ponty, D. Ziadi, and J.-M. Champarnaud. A new quadratic algorithm to convert a regular expression into an automaton. In D. R. Raymond, D. Wood, and S. Yu, editors, *Proc. 1st WIA'96*, volume 1260 of *LNCS*, pages 109–119. Springer, 1997.

[19] J. Sakarovitch. *Elements of Automata Theory*. Cambridge University Press, 2009.