Usefulness of Directed Acyclic Subword Graphs in Problems Related to Standard Sturmian Words

Paweł Baturo Marcin Piątkowski Wojciech Rytter

Faculty of Mathematics and Computer Science
Nicolaus Copernicus University

13th Prague Stringology Conference
Prague 2008

1. Sturmian words
2. Subword graphs
3. Number of subwords
4. Structure of occurreces of subwords
5. Critical factorization and maximal suffixes
6. Relation to dual Ostrovski numeration system

Sturmian word

- Alphabet: $\Sigma = \{a, b\}$
- Directive sequence: $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_n)$, where $\gamma_0 \geqslant 0$ and $\gamma_1, \ldots, \gamma_n > 0$
- Recurrence:
  - $x_{-1} = b$
  - $x_0 = a$
  - $x_k = (x_{k-1})^{\gamma_{k-1}} \cdot x_{k-2}$
- $\mathrm{Word}(\gamma_0, \gamma_1, \ldots, \gamma_n) = x_{n+1}$

Sturmian word example

- $\gamma = (1, 2, 1, 3, 1)$

- $x_{-1} = b$
- $x_0 = a$
- $x_1 = (x_0)^1 \cdot x_{-1} = a \cdot b$
- $x_2 = (x_1)^2 \cdot x_0 = ab \cdot ab \cdot a$
- $x_3 = (x_2)^1 \cdot x_1 = ababa \cdot ab$
- $x_4 = (x_3)^3 \cdot x_2 = ababaab \cdot ababaab \cdot ababaab \cdot ababa$
- $x_5 = (x_4)^1 \cdot x_3 = ababaababaababaabababa \cdot ababaab$

- $\mathrm{Word}(1, 2, 1, 3, 1) = ababaababaababaabababaababaab$

Notation

- $\hat{w}$ – prefix of w of size 2
- y_k – (basic subword) reverse of some x_k

Example:

$$x_0 = a \qquad y_0 = a$$

$$x_1 = ab \qquad y_1 = ba$$

$$x_2 = ababa \qquad y_2 = ababa$$

$$x_3 = ababaab \qquad y_3 = baababa$$

DAWG

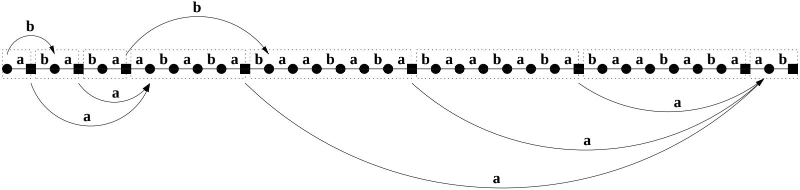
The **Direct Acyclic Subword Graph** of the word w is the minimal deterministic automaton (not necessarily complete) that accepts all suffixes of w.

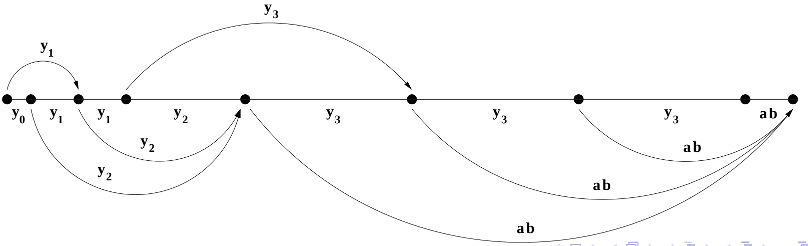
CDAWG

The **Compacted Subword Graph** results from the subword graph by removing all nodes of out-degree one and replacing each chain by a single edge with the label representing the path label of this chain (except the node creating last edge labelled "*ab*" or "*ba*").

Example:

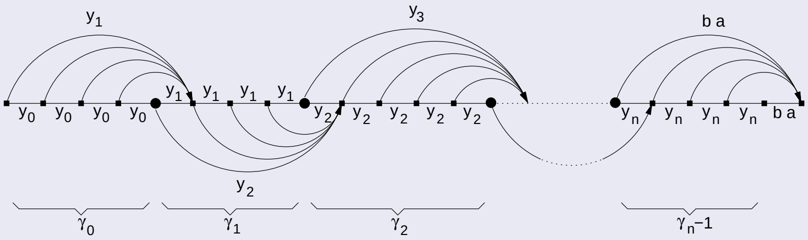
Subword graph and its compacted version for $\mathrm{Word}(1,2,1,3,1)$.





Theorem

Compacted subword graph of $\mathrm{Word}(\gamma_0, \gamma_1, \ldots, \gamma_n)$ has the following structure:



Definition

Let G be a compacted subword graph and v a vertex in G.
Define:

- $mult(v)$ – multiplicity of vertex v – number of paths leading from source node to v
- $edges(v)$ – sum of weight of all edges outgoing from v

Lemma

Let $w = \mathrm{Word}(\gamma_0, \gamma_1, \ldots, \gamma_n)$ and G be CDAWG of w. Then

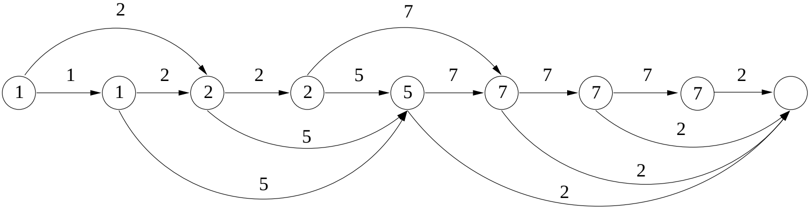
$$\left|Subwords(w)\right| = \sum_{v \in G} mult(v) \cdot edges(v)$$

Theorem

Let $x_{n+1} = \mathrm{Word}(\gamma_0, \gamma_1, ..., \gamma_n)$. Then

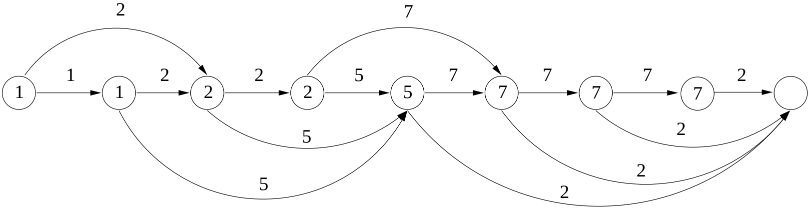
$$\left| Subwords(x_{n+1}) \right| = |x_n| \cdot |x_{n-1}| + 2 \cdot |x_n| - 1$$

Example:
$x_5 = \mathrm{Word}(1, 2, 1, 3, 1)$



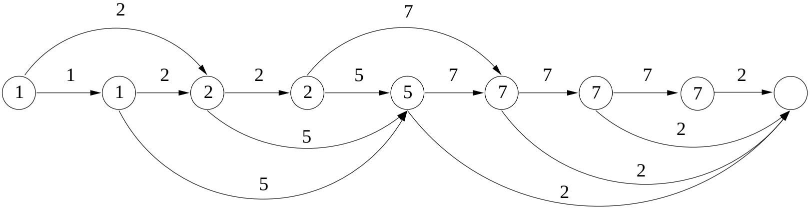
Example:
$x_5 = \mathrm{Word}(1, 2, 1, 3, 1)$
Number of subwords:

1. $\sum_{v \in G} mult(v) \cdot edges(v) = 3+7+14+24+45+63+63+14 = 233$



Example:
$x_5 = \mathrm{Word}(1,2,1,3,1)$
Number of subwords:

1. $\sum_{v \in G} mult(v) \cdot edges(v) = 3+7+14+24+45+63+63+14 = 233$
2. $|Subwords(x_5)| = |x_4| \cdot |x_3| + 2 \cdot |x_4| - 1 = 26 \cdot 7 + 2 \cdot 26 - 1 = 233$



Definition

A **right special factor** of the word w is any word z such that za and zb are subwords of w.

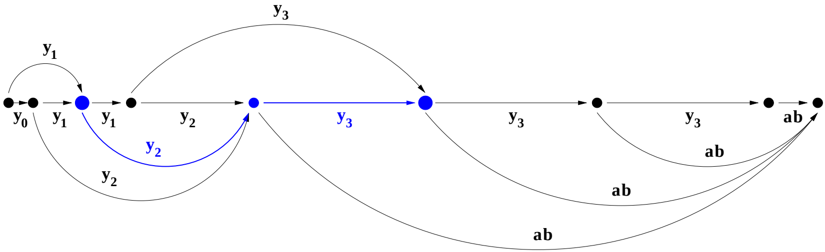
Example:

$$w = \mathrm{a\,b\,a\,b\,a\,a\,b\,a}\,\underbrace{\mathbf{b\,a\,b\,a\,a}}_{za}\,\mathrm{b\,a\,b\,a\,b\,a\,a}\,\underbrace{\mathbf{b\,a\,b\,a\,b}}_{zb}\,\mathrm{a\,a\,b\,a\,b\,a\,a\,b}$$

$z = baba$ is a right special factor of w.

Example:

Path in DAWG of $\mathrm{Word}(1,2,1,3,1)$ corresponding to right special factor y_2y_3.



Theorem

Let $w = \mathrm{Word}(\gamma_0, \gamma_1, \ldots, \gamma_n)$ be a standard Sturmian word.

1. For each $k > 0$ there is at most one right special factor of length k of w.
2. Every right special factor of w has the form:

$$z = y_i^{\alpha_i} \cdot y_{i+1}^{\alpha_{i+1}} \cdots y_{i+k}^{\alpha_{i+k}}$$

   where $0 \leqslant \alpha_j \leqslant \gamma_j$ for $j < n-1$ and $0 \leqslant \alpha_j \leqslant \gamma_j - 1$ for $j = n-1$.
3. For a given $k > 0$ the right special factor of w of length k has grammar-representation of size $O(n)$ that can be computed in $O(n)$ time.

Example:

Right special factors of $\mathrm{Word}(1,2,1,3,1)$ with their lengths.
Special prefixes are **marked**.

1	$\mathbf{y_0}$			11	$y_1^2 y_3$	18	$y_1^2 y_3^2$
				12	$y_2 y_3$	19	$y_2 y_3^2$
2	y_1	6	$y_0 y_2$	13	$y_0 y_2 y_3$	20	$y_0 y_2 y_3^2$
3	$\mathbf{y_0 y_1}$	7	$y_1 y_2$	14	$y_1 y_2 y_3$	21	$y_1 y_2 y_3^2$
		8	$y_0 y_1 y_2$	15	$y_0 y_1 y_2 y_3$	22	$y_0 y_1 y_2 y_3^2$
4	y_1^2	9	$y_1^2 y_2$	16	$y_1^2 y_2 y_3$	23	$y_1^2 y_2 y_3^2$
5	$\mathbf{y_0 y_1^2}$	10	$\mathbf{y_0 y_1^2 y_2}$	17	$\mathbf{y_0 y_1^2 y_2 y_3}$	24	$\mathbf{y_0 y_1^2 y_2 y_3^2}$

Definition

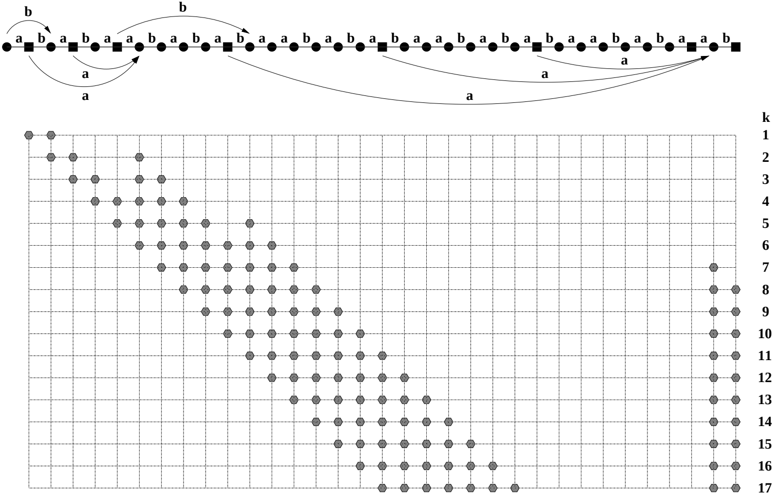
FIN(k, w) – set of the last positions of the first occurrences of all subwords of length k in word w.

Example:

$$FIN\Big(3, \mathrm{Word}(1,2,1,3,1)\Big) = \{3,4,6,7\}$$

```
    ▼ ▼   ▼ ▼
a b a b a a b a b a b a a b a b a b a a b a b a b a a b a b a a b
a b a
  b a b
      b a a
        a a b
```

Example: Structure of sets $FIN(k, w)$ for $w = \mathrm{Word}(1, 2, 1, 3, 1)$.



Theorem

Let $w = \mathrm{Word}(\gamma_0, \gamma_1, \ldots, \gamma_n)$ be a standard Sturmian word.

1. The set $FIN(k, w)$ consists of a single interval or of two disjoint intervals.
2. For a given k we can compute the intervals representing $FIN(k, w)$ in $O(n)$ time.

Definition

Minimal local period in a word w at position k is a positive integer p such that $w[i-p] = w[i]$ for every $k < i \leqslant k+p$, where $w[i]$ and $w[i-p]$ are defined.

Example:

Minimal local period in $w = \mathrm{Word}(1,2,1,3,1)$ at position $k = 9$ equals 2 and at position $k = 27$ equals 5.

$$w = \mathrm{a\,b\,a\,b\,a\,a\,b}\,\underbrace{\mathbf{a\,b}}_{2}\,\underbrace{\mathbf{a\,b}}_{2}\,\mathrm{a\,a\,b\,a\,b\,a\,b\,a\,a\,b\,a}\,\underbrace{\mathbf{b\,a\,b\,a\,a}}_{5}\,\underbrace{\mathbf{b\,a\,b\,a\,a}}_{5}\,\mathrm{b}$$

Definition

Critical factorization point in a word w is a position k in w for which minimal local period at k equals (global) minimal period of w.

Example:

Minimal local periods of $\mathrm{Word}(1, 2, 1, 3, 1)$.

Critical factorization point is **marked**.

i		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17 ......
		a	b	a	b	a	a	b	a	b	a	b	a	a	b	a	b	a ......
p(i)	1	2	2	2	5	1	7	2	2	2	2	7	1	7	2	2	2	2 ......

i	......	18	19	20	21	22	23	24	25	|	26	27	28	29	30	31	32	33
	......	b	a	a	b	a	b	a	b	|	a	a	b	a	b	a	a	b
p(i)	......	2	7	1	7	2	2	2	4	**33**	1	5	2	2	5	1	3	1

Fact (M. Crochemore, D. Perrin 1991)

The critical factorization point of word w is given as the starting position of a lexicographically maximal suffix, maximized over all possible orders of alphabet.

Definition

For a word w define two paths in DAWG of w:

- $\pi_a(w)$ – starts in root, ends in sink and uses letter a whenever it is possible.
- $\pi_b(w)$ – starts in root, ends in sink and uses letter b whenever it is possible.

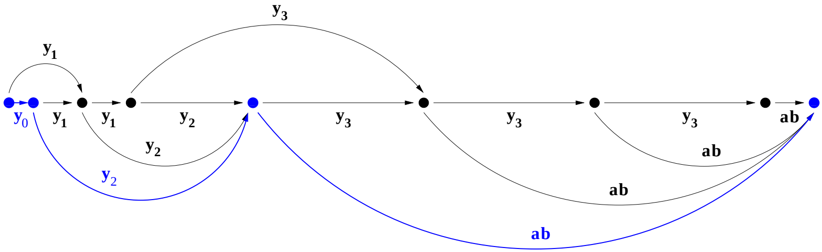
Similarly $\pi_a(w)$ and $\pi_b(w)$ are defined in CDAWG of w.

Observation

1. Label of the path $\pi_a(w)$ is lexicographically maximal suffix of w with respect to the letter ordering "$a > b$".
2. Label of the path $\pi_b(w)$ is lexicographically maximal suffix of w with respect to the letter ordering "$a < b$".

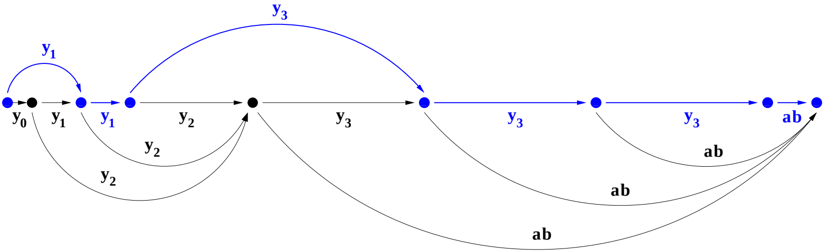
Example:

$$\pi_a\Big(\mathrm{Word}(1,2,1,3,1)\Big) = y_0\, y_2\, ab.$$



Example:

$$\pi_b\Big(\mathrm{Word}(1,2,1,3,1)\Big) = y_1^2\, y_3^3\, ab.$$



Lemma

Let $w = \mathrm{Word}(\gamma_0, \gamma_1, \ldots, \gamma_n)$ be a standard Sturmian word.
Then:

$$
\begin{aligned}
\pi_a(w) &= y_0^{\gamma_0} \cdot y_2^{\gamma_2} \cdots y_{2k}^{\gamma_{2k}} \cdot \hat{y}_{n-1} \\
\pi_b(w) &= y_1^{\gamma_1} \cdot y_3^{\gamma_3} \ldots y_{2l+1}^{\gamma_{2l+1}} \cdot \hat{y}_{n-1}
\end{aligned}
$$

where $k = \lfloor \frac{n-1}{2} \rfloor$ and $l = \lfloor \frac{n-2}{2} \rfloor$.

Theorem

Let $w = \mathrm{Word}(\gamma_0, \gamma_1, \ldots, \gamma_n)$ be a standard Sturmian word.

1. The critical factorization point of w is at position

$$k = |w| - \min\left\{\ |\pi_a(w)|,\ |\pi_b(w)|\ \right\}$$

2. The lexicographically maximal suffix of w has grammar-based representation of size $O(n)$.
3. The compressed representation of the lexicographically maximal suffix of w and its critical factorization point can be computed in time $O(n)$.

Example:

- $w = \mathrm{Word}(1, 2, 1, 3, 1)$
- $\pi_a(w) = y_0\, y_2\, ab$
- $\pi_b(w) = y_1^2\, y_3^3\, ab$
- critical factorization point at position:

$$k = |w| - |y_0\, y_2\, ab| = 33 - 8 = 25$$

$$w = \mathrm{a\,b\,a\,b\,a\,a}\,\overbrace{\mathrm{b\,a\,b\,a\,b\,a\,a\,b\,a\,b\,a\,b\,a\,a\,b\,a\,b\,a}\,\mathbf{b}\,\underbrace{\mathrm{a\,a\,b\,a\,b\,a\,a\,b}}_{\pi_a}}^{\pi_b}$$

Definition

For infinite directive sequence $\gamma = (\gamma_0, \gamma_1, \ldots)$ define **base sequence** q as:

$$q = (q_0, q_1, \ldots) = \big(|x_0|, |x_1|, \ldots\big)$$

and

$$\mathrm{val}_\gamma(\alpha_0, \alpha_1, \ldots, \alpha_n) = \alpha_0 \cdot q_0 + \alpha_1 \cdot q_1 + \ldots + \alpha_n \cdot q_n$$

Note:

Base sequence can be defined directly:

$$q_{-1} = q_0 = 1, \qquad q_{i+1} = q_i \cdot \gamma_i + q_{i-1} \quad \text{for } i \geqslant 0$$

Definition

For $0 \leqslant i < |x_n|$ the representation of i in the **dual Ostrovski numeration system** is defined as: $[\hat{i}]_\gamma = (\alpha_0, \alpha_1, \ldots, \alpha_n)$, where:

1. $\mathrm{val}_\gamma(\alpha_0, \alpha_1, \ldots, \alpha_n) = i$
2. $\forall_{0 \leqslant j < n}\ \alpha_j \leqslant \gamma_j$
3. $(\ \alpha_j < \gamma_j \text{ and } \exists_{i > j}\ \alpha_i > 0\ ) \Rightarrow \alpha_{j+1} > 0$

Example:

- directive sequence: $\gamma = (1, 2, 1, 3, 1, \ldots)$.
- base sequence: $q = (|x_0|, |x_1|, ...) = (1, 2, 5, 7, 26, 33, \ldots)$

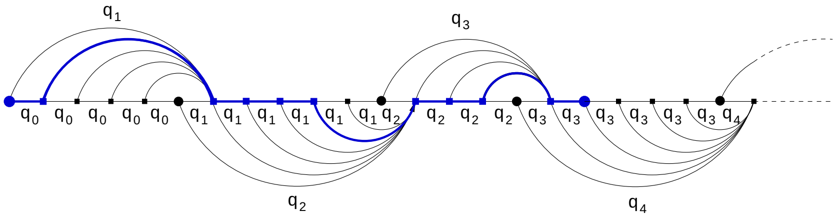
- $[\hat{29}]_\gamma = (1, 1, 1, 3)$, because

$$29 = 1 \cdot 1 + 1 \cdot 2 + 1 \cdot 5 + 3 \cdot 7$$

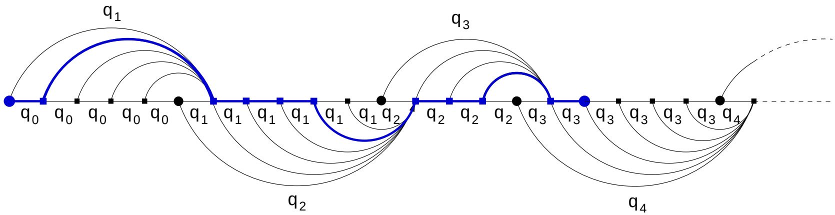
- $[\hat{58}]_\gamma = (0, 2, 0, 3, 0, 1)$, because

$$58 \;=\; 0 \cdot 1 + 2 \cdot 2 + 0 \cdot 5 + 3 \cdot 7 + 0 \cdot 26 + 1 \cdot 33$$

$\mathcal{G}_\infty$ – infinite compacted subword graph of $\mathrm{Word}(\gamma_0, \gamma_1, \ldots)$

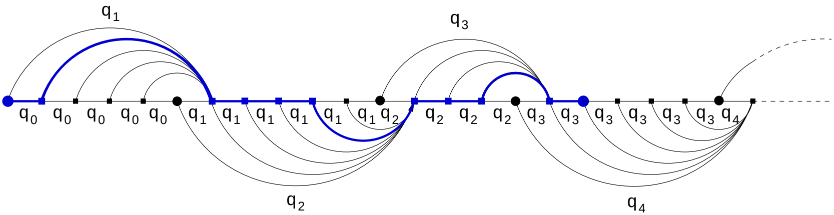


$\mathcal{G}_\infty$ – infinite compacted subword graph of $\mathrm{Word}(\gamma_0, \gamma_1, \ldots)$



- $|\pi| = 1 \cdot q_0 + 4 \cdot q_1 + 3 \cdot q_2 + 2 \cdot q_3$

$\mathcal{G}_\infty$ – infinite compacted subword graph of $\mathrm{Word}(\gamma_0, \gamma_1, \ldots)$



- $|\pi| = 1 \cdot q_0 + 4 \cdot q_1 + 3 \cdot q_2 + 2 \cdot q_3$
- $\left[\widehat{|\pi|}\right] = (1, 4, 3, 2)$

Theorem

Let $\mathcal{G}_\infty$ be the infinite compacted subword graph corresponding to directive sequence $\gamma = (\gamma_0, \gamma_1, \ldots)$.

1. Let π be a path from the root to another node of $\mathcal{G}_\infty$. Let $\mathrm{rep}(\pi) = (h_0, h_1, \ldots)$, where h_i is the number of edges of weight q_i on the path π. Then $\mathrm{rep}(\pi)$ is the representation of the length $|\pi|$ of this path in the dual Ostrovski numeration system corresponding to the directive sequence of $\mathcal{G}_\infty$.
2. For each $k > 1$ there is exactly one path of length k in $\mathcal{G}_\infty$.

Definition

For directive sequence $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_n)$ define $SD(\gamma)$, the **set of representations** in dual Ostrovski numeration system of all numbers not exceeding $|x_{n+1}| + |x_n| - 2$.

Definition

The minimal deterministic finite automaton accepting language

$$L(\gamma) = \left\{ a_0^{i_0}\, a_1^{i_1} \cdots a_n^{i_n} \;:\; (i_0, i_1, \ldots, i_n) \in SD(\gamma) \right\}$$

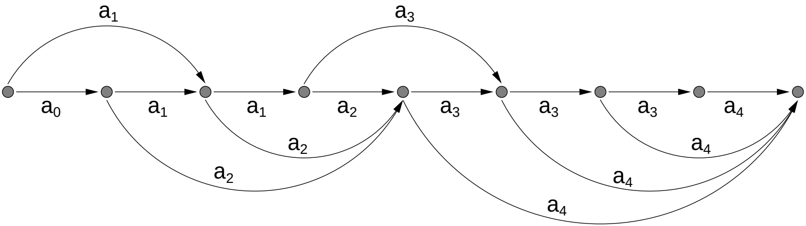
for alphabet $\Sigma = \{a_0, a_1, \ldots, a_n\}$ is called **Ostrovski automaton** and denoted $OA(\gamma)$.

Remark: $a^0 = \varepsilon$ for any letter a.

Example:

Minimal deterministic automaton $OA(1,2,1,3,1)$ accepting

$$L(1,2,1,3,1) = \left\{ a_0^{i_0}\ a_1^{i_1}\ a_2^{i_2}\ a_3^{i_3}\ a_4^{i_4}\ :\ (i_0, i_1, i_2, i_3, i_4) \in SD(1,2,1,3,1) \right\}$$



Theorem

The minimal Ostrovski automaton, without the dead state, for directive sequence $(\gamma_0, \gamma_1, \ldots, \gamma_n)$ is isomorphic as a graph to the compact directed acyclic subword graph of $\mathrm{Word}(\gamma_0, \gamma_1, \ldots, \gamma_n)$.

Thank You
For Your Attention