On varieties of automata

Ondřej Klíma and Libor Polák

Department of Mathematics and Statistics
Masaryk University, Brno
Czech Republic

SSAOS 2016

I. Algebraic Theory of Regular Languages

Examples

- Goal of the study: effective characterizations of certain natural classes of regular languages.
- Typical result: a language belongs to a given class iff its syntactic monoid belongs to a certain class of monoids.

Examples

- Goal of the study: effective characterizations of certain natural classes of regular languages.
- Typical result: a language belongs to a given class iff its syntactic monoid belongs to a certain class of monoids.

Theorem (Schützenberger – 1966)

A regular language L is star-free if and only if its syntactic monoid is aperiodic.

Theorem (Simon — 1972)

A regular language L is piecewise testable if and only if the syntactic monoid of L is $\mathcal{J}$-trivial.

- General framework – Eilenberg correspondence.

Varieties of Languages

Definition

A variety of languages $\mathcal{V}$ associates to every finite alphabet A a class $\mathcal{V}(A)$ of regular languages over A in such a way that

- $\mathcal{V}(A)$ is closed under finite unions, finite intersections and complements (in particular $\emptyset, A^* \in \mathcal{V}(A)$),
- $\mathcal{V}(A)$ is closed under quotients, i.e.
  $L \in \mathcal{V}(A),\ u, v \in A^*$ implies
  $u^{-1}Lv^{-1} = \{\, w \in A^* \mid uwv \in L \,\} \in \mathcal{V}(A)$,
- $\mathcal{V}$ is closed under preimages in morphisms, i.e.
  $f : B^* \to A^*,\ L \in \mathcal{V}(A)$ implies
  $f^{-1}(L) = \{\, v \in B^* \mid f(v) \in L \,\} \in \mathcal{V}(B)$.

A Formal Definition of a DFA

Definition

A deterministic finite automaton over the alphabet A is a five-tuple $\mathcal{A} = (Q, A, \cdot, i, F)$, where

- Q is a nonempty set of states,
- $\cdot : Q \times A \to Q$ is a complete transition function, which can be extended to a mapping $\cdot : Q \times A^* \to Q$ by $q \cdot \lambda = q,\ \ q \cdot (ua) = (q \cdot u) \cdot a$,
- $i \in Q$ is the initial state,
- $F \subseteq Q$ is the set of final states.

The automaton $\mathcal{A}$ accepts a word $u \in A^*$ iff $i \cdot u \in F$. The automaton $\mathcal{A}$ recognizes the language $L_{\mathcal{A}} = \{u \in A^* \mid i \cdot u \in F\}$.

Motivations for a Notion of a Variety of Automata

- Why monoids instead of automata?
  - An equational description of pseudovarieties of monoids by pseudoidentities.
  - Other algebraic constructions, e.g. products (semidirect, wreath, Mal'cev).

Motivations for a Notion of a Variety of Automata

- Why monoids instead of automata?
  - An equational description of pseudovarieties of monoids by pseudoidentities.
  - Other algebraic constructions, e.g. products (semidirect, wreath, Mal'cev).
- Why are we still interested in automata characterizations?
  - Usually, a regular language is given by an automaton. And computation of the syntactic monoid need not to be effective (can be exponentially larger).
  - Sometimes a “graph condition” on automata can be easier to test than an equational condition on monoids.

Motivations for a Notion of a Variety of Automata

- Why monoids instead of automata?
  - An equational description of pseudovarieties of monoids by pseudoidentities.
  - Other algebraic constructions, e.g. products (semidirect, wreath, Mal'cev).
- Why are we still interested in automata characterizations?
  - Usually, a regular language is given by an automaton. And computation of the syntactic monoid need not to be effective (can be exponentially larger).
  - Sometimes a "graph condition" on automata can be easier to test than an equational condition on monoids.

So, basically there are three worlds: classes of languages, classes of (enriched) semiautomata (no initial and no final states) and those of appropriate algebraic structures.

Generalizations of the Eilenberg Correspondence

Since not all natural classes of regular languages are varieties, one of the recent directions of the research in algebraic theory of regular languages is devoted to generalizations of the Eilenberg correspondence.

Generalizations of the Eilenberg Correspondence

Since not all natural classes of regular languages are varieties, one of the recent directions of the research in algebraic theory of regular languages is devoted to generalizations of the Eilenberg correspondence.

- Pin (1995): Positive varieties of regular languages — closure under complementation is not required.
  Algebraic counterparts are pseudovarieties of finite ordered monoids.
  (Syntactic monoid is implicitly ordered.)

Generalizations of the Eilenberg Correspondence

Since not all natural classes of regular languages are varieties, one of the recent directions of the research in algebraic theory of regular languages is devoted to generalizations of the Eilenberg correspondence.

- Pin (1995): Positive varieties of regular languages — closure under complementation is not required. Algebraic counterparts are pseudovarieties of finite ordered monoids. (Syntactic monoid is implicitly ordered.)
- Polák (1999): Conjunctive (and disjunctive) varieties.
- Straubing (2002): $\mathcal{C}$-varieties of languages.
- Ésik, Larsen (2003): literal varieties of languages.
- Gehrke, Grigorieff, Pin (2008): Lattices of regular languages.

II. Varieties of Automata

The Construction of a Minimal DFA by Brzozowski

- For a language $L \subseteq A^*$ and $u \in A^*$, we define a left quotient $u^{-1}L = \{\, w \in A^* \mid uw \in L \,\}$.

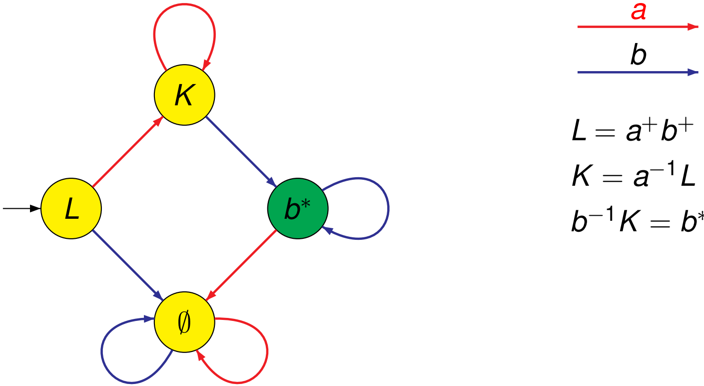
Definition

The canonical deterministic automaton of L is $\mathcal{D}_L = (D_L, A, \cdot, L, F)$, where

- $D_L = \{\, u^{-1}L \mid u \in A^* \,\}$,
- $q \cdot a = a^{-1}q$, for each $q \in D_L$, $a \in A$,
- $q \in F$ iff $\lambda \in q$.

- Each state $q = u^{-1}L$ is formed by all words transforming the state q into a final state.

An Example of a Canonical Automaton



$L = a^+b^+$

$K = a^{-1}L = a^*b^+$

$b^{-1}K = b^*$

Preimages in Morphisms, Varieties of Automata

- Let $f : B^* \to A^*$ be a morphism, We say that $(P, B, \circ)$ is an f-subautomaton of $(Q, A, \cdot)$ if $P \subseteq Q$ and $q \circ b = q \cdot f(b)$ for every $q \in P$, $b \in B$.

Definition

A variety of semiautomata $\mathbb{V}$ associates to every finite alphabet A a class $\mathbb{V}(A)$ of semiautomata (no initial nor final states) over alphabet A in such a way that

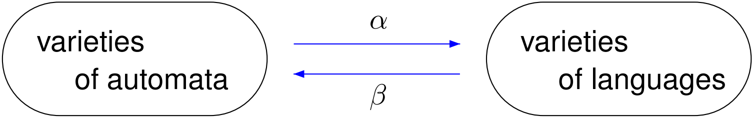
- $\mathbb{V}(A) \neq \emptyset$ is closed under disjoint unions, finite direct products and morphic images,
- $\mathbb{V}$ is closed under f-subautomata.

An Eilenberg Type Correspondence

- For each variety of automata $\mathbb{V}$ we denote by $\alpha(\mathbb{V})$ the variety of regular languages given by

$$(\alpha(\mathbb{V}))(A) = \{L \subseteq A^* \mid \exists \mathcal{A} = (Q, A, \cdot, i, F) : L = L_{\mathcal{A}} \wedge (Q, A, \cdot) \in \mathbb{V}(A)\}.$$

- For each variety of regular languages $\mathcal{L}$ we denote by $\beta(\mathcal{L})$ the variety of automata generated by all DFAs $\mathcal{D}_L$, where $L \in \mathcal{L}(A)$ for some alphabet A.



Theorem (Ésik and Ito, Chaubard, Pin and Straubing)

The mappings α and β are mutually inverse isomorphisms between the lattice of all varieties of automata and the lattice of all varieties of regular languages.

- A version for $\mathcal{C}$-varieties is obvious: we consider f-subautomata (etc.) just for $f \in \mathcal{C}$.
- Ésik and Ito were working with literal varieties (morphisms map letters to letters, i.e $f(B) \subseteq A$) and used disjoint union.
- Chaubard, Pin and Straubing called the automata $\mathcal{C}$-actions and used trivial automata.

An Examples – Acyclic Automata

- One of the conditions in Simon's characterization of piecewise testable languages is that a minimal DFA is acyclic.
- A content $\mathrm{c}(u)$ of a word $u \in A^*$ is the set of all letters occurring in u.
- We say that $(Q, A, \cdot)$ is a acyclic if for each $u \in A^*$ and $q \in Q$ we have

$$q \cdot u = q \implies (\forall a \in \mathrm{c}(u) : q \cdot a = q) .$$

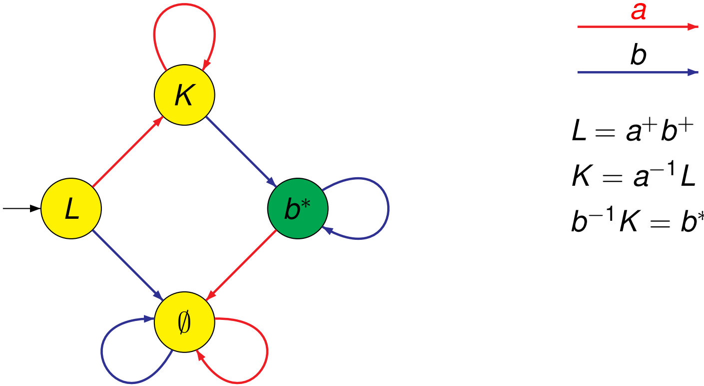
- The class of all acyclic automata is a variety.
- The corresponding variety of languages (well-known): (disjoint) unions of the languages of the form

$$A_0^* a_1 A_1^* a_2 A_2^* \dots A_{n-1}^* a_n A_n^*, \quad \text{where} \quad a_i \notin A_{i-1} \subseteq A .$$

An Example – Piecewise Testable Languages

- In DLT'13 we gave an alternative condition for automata recognizing piecewise testable languages.
- We call an acyclic automaton $(Q, A, \cdot)$ locally confluent, if for each state $q \in Q$ and every pair of letters $a, b \in A$, there is a word $w \in \{a, b\}^*$ such that $(q \cdot a) \cdot w = (q \cdot b) \cdot w$.
- A stronger condition: an acyclic automaton $(Q, A, \cdot)$ is confluent, if for each state $q \in Q$ and every pair of words $u, v, \in \{a, b\}^*$, there is a word $w \in \{a, b\}^*$ such that $(q \cdot u) \cdot w = (q \cdot v) \cdot w$.
- Each acyclic automaton is confluent iff it is locally confluent.
- The class of all acyclic confluent automata is a variety which corresponds to the variety of piecewise testable languages.

An Example of a Piecewise Testable Language

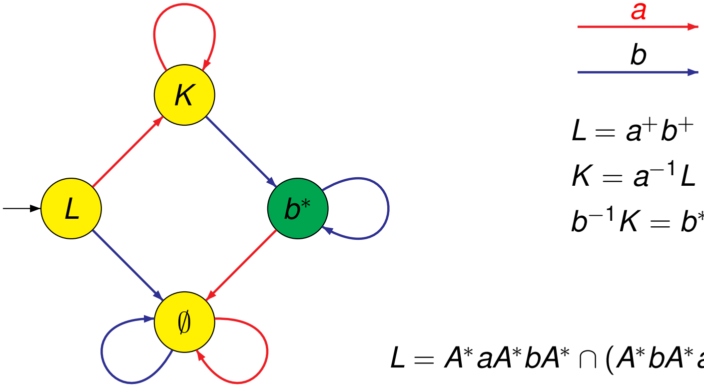


$L = a^+ b^+$

$K = a^{-1}L = a^* b^+$

$b^{-1}K = b^*$

An Example of a Piecewise Testable Language



$L = a^+ b^+$

$K = a^{-1}L = a^* b^+$

$b^{-1}K = b^*$

$L = A^* a A^* b A^* \cap (A^* b A^* a A^*)^c$

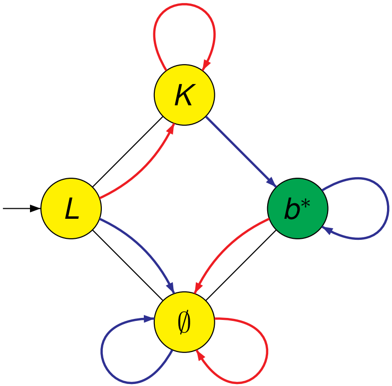
III. Automata Enriched with an Algebraic Structure

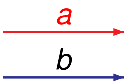
III.1 Ordered Automata

A Natural Ordering of the Canonical Automaton

- For a language $L \subseteq A^*$, we have defined a the canonical deterministic automaton: $\mathcal{D}_L = (D_L, A, \cdot, L, F)$, where
  - $D_L = \{\, u^{-1}L \mid u \in A^* \,\}$,
  - $q \cdot a = a^{-1}q$, for each $q \in D_L$, $a \in A$,
  - $q \in F$ iff $\lambda \in q$.
- Therefore states are ordered by inclusion, which means that each minimal automaton is implicitly equipped with a partial order.
- The action by each letter a is an isotone mapping: for all states p, q such that $p \subseteq q$ we have $p \cdot a = a^{-1}p \subseteq a^{-1}q = q \cdot a$.
- The final states form an upward closed subset w.r.t. $\subseteq$.

An Example of an Ordered Automaton



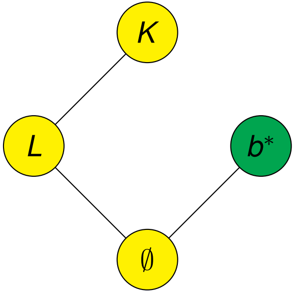


$L = a^+ b^+$

$K = a^{-1}L = a^* b^+$

$L \subseteq K$

An Example of an Ordered Automaton



$L = a^+ b^+$

$K = a^{-1} L = a^* b^+$

$L \subseteq K$

An Ordered Automaton

Definition

An ordered automaton over the alphabet A is a six-tuple $\mathcal{A} = (Q, A, \cdot, \leq, i, F)$, where

- $\mathcal{A} = (Q, A, \cdot, i, F)$ is a usual DFA;
- $\leq$ is a partial order;
- an action by every letter is an isotone mapping from the partial ordered set $(Q, \leq)$ to itself;
- F is an upward closed set, i.e. $p \leq q, p \in F \implies q \in F$.

Algebraic Constructions on Ordered Automata

Definition

A variety of ordered semiautomata $\mathbb{V}$ associates to every finite alphabet A a class $\mathbb{V}(A)$ of ordered semiautomata over alphabet A in such a way that

- $\mathbb{V}(A) \neq \emptyset$ is closed under disjoint union, finite direct products and morphic images,
- $\mathbb{V}$ is closed under f-subautomata.

An Eilenberg Type Correspondence

Theorem (Pin)

There are mutually inverse isomorphisms between the lattice of all varieties of ordered automata and the lattice of all positive varieties of regular languages.

The Level $1/2$

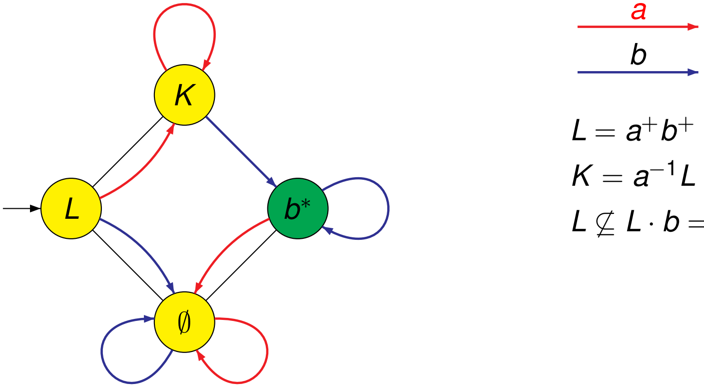
- Piecewise testable languages are Boolean combinations of languages of the form

$$A^* a_1 A^* a_2 A^* \dots A^* a_\ell A^*, \text{ where } a_1, \dots, a_\ell \in A, \ \ell \geq 0 \, .$$

- Piecewise testable languages form level 1 in Straubing-Thérien hierarchy.
- Level $1/2$ is formed just by finite unions of intersections of languages above.
- The corresponding variety of ordered automata is the class of all ordered automata where actions by letters are increasing mappings. I.e. ordered automata satisfying:

$$\forall q \in Q, a \in A : q \cdot a \geq q \, .$$

An Example of an Ordered Automaton outside 1/2



$L = a^+b^+$

$K = a^{-1}L = a^*b^+$

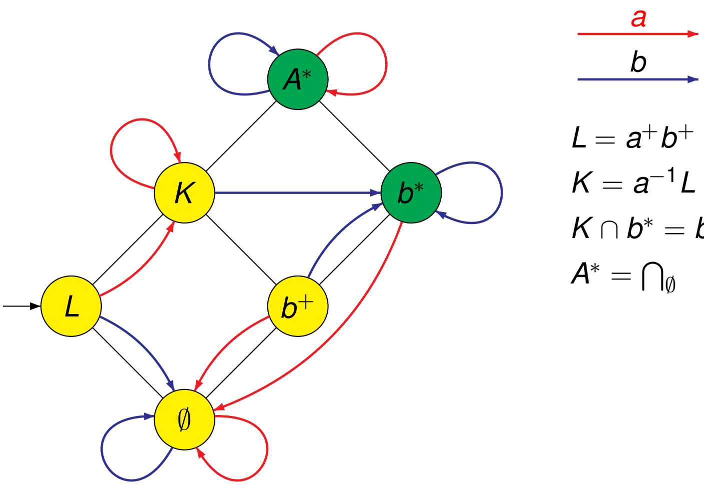
$L \not\subseteq L \cdot b = \emptyset$

III.2 Meet Automata

Intersections of Left Quotients

- For a language $L \subseteq A^*$, we extend the canonical semiautomaton $(D_L, A, \cdot)$, where states are subsets of A^*.
- We can consider intersections of states: $U_L = \{\bigcap_{j \in I} K_j \mid I \text{ finite set }, K_j \in D_L\}$. If $I = \emptyset$ then we put $\bigcap_{j \in I} K_j = A^*$.
- The finite set U_L is equipped with the operation intersection $\cap$ and we can define $(\bigcap_{j \in I} K_j) \cdot a = \bigcap_{j \in I}(K_j \cdot a)$.
- We have the semiautomaton $(U_L, A, \cdot)$ with semilattice operation $\cap$. Moreover, A^* is the largest element in the semilattice $(U_L, \cap)$ and it is an absorbing state in $(U_L, A, \cdot)$.
- Naturally, $F = \{K \mid \lambda \in K\}$ is a main filter.

An Example of a Meet Automaton



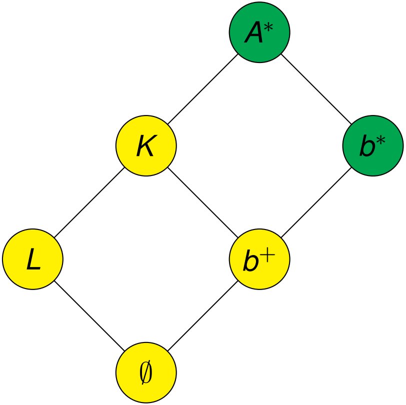
$L = a^+ b^+$

$K = a^{-1} L = a^* b^+$

$K \cap b^* = b^+$

$A^* = \bigcap_\emptyset$

An Example of a Meet Automaton



$$L = a^+b^+$$
$$K = a^{-1}L = a^*b^+$$
$$K \cap b^* = b^+$$
$$A^* = \bigcap_{\emptyset}$$

Meet Automata

Definition

A structure $(Q, A, \cdot, \wedge, \top)$ is a meet semiautomaton if

- $(Q, A, \cdot)$ is a DFA,
- $(Q, \wedge)$ is a semilattice with the largest element $\top$,
- actions by letters are endomorphisms of the semilattice $(Q, \wedge)$, i.e. $\forall p, q \in Q, a \in A : (p \wedge q) \cdot a = p \cdot a \wedge q \cdot a$
- $\top$ is an absorbing state.

This meet semiautomaton recognizes a language L if there are $i, f \in Q$ such that $L = \{u \in A^* \mid i \cdot u \wedge f = f\}$.

Varieties of Meet Automata

Definition

A variety of meet semiautomata $\mathbb{V}$ associates to every finite alphabet A a class $\mathbb{V}(A)$ of meet semiautomata over alphabet A in such a way that

- $\mathbb{V}(A) \neq \emptyset$ is closed under finite direct products and morphic images,
- $\mathbb{V}$ is closed under f-subautomata.

An Eilenberg Type Correspondence

Theorem (Klíma, Polák)

There are mutually inverse isomorphisms between the lattice of all varieties of meet semiautomata and the lattice of all conjunctive varieties of regular languages.

Varieties of Meet Automata – An Example

Example

For each alphabet A, a meet automata $(Q, A, \cdot, \wedge, \top)$ belongs to $\mathbb{S}(A)$ if $\forall q \in Q, a \in A \;:\; q \cdot a = q \cdot a \wedge q$ and

$$\forall q \in Q, a, b \in A \;:\; q \cdot ab = q \cdot a \wedge q \cdot b\,. \quad (*)$$

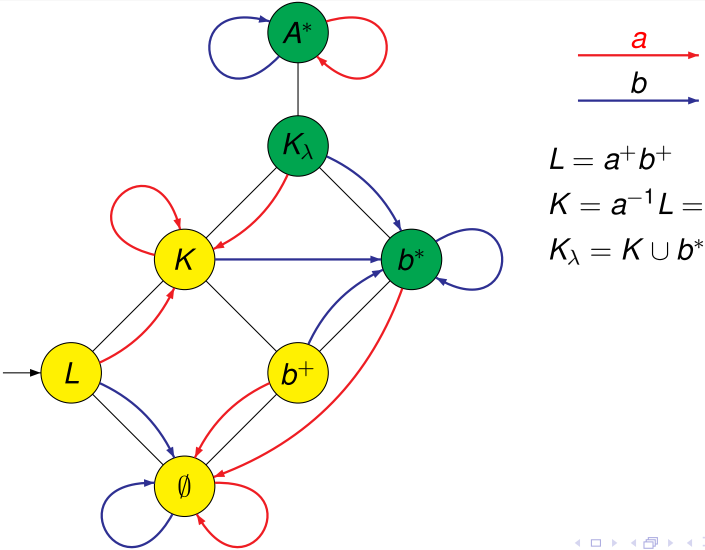
Then $\mathbb{S}$ is a variety of meet automata and the corresponding conjunctive variety of languages $\mathcal{S}$ is given by $\mathcal{S}(A) = \{B^* \mid B \subseteq A\} \cup \{\emptyset\}$.

III.3 Lattice automata

The Canonical Lattice Automaton of a Language

- For a language $L \subseteq A^*$ we extend the canonical meet semiautomaton $(U_L, A, \cdot, \wedge, A^*)$ by unions of states: $W_L = \{\bigcup_{j \in I} M_j \mid I \text{ finite set }, M_j \in U_L\}$. If $I = \emptyset$ then we put $\bigcup_{j \in I} M_j = \emptyset$.
- The finite set W_L is equipped with the operations intersection $\cap$ and union $\cup$ (due to distributive laws). We can define $(\bigcup_{j \in I} M_j) \cdot a = \bigcup_{j \in I}(M_j \cdot a)$.
- We have the semiautomaton $(W_L, A, \cdot)$ and a distributive lattice $(W_L, \cap, \cup)$. Moreover, A^* is the largest element, $\emptyset$ is the smallest element – both are absorbing states in $(W_L, A, \cdot)$.
- Naturally $F = \{M \mid \lambda \in M\}$ is closed w.r.t. $\cap$, upward closed, and $M_1 \cup M_2 \in F$ implies $M_1 \in F$ or $M_2 \in F$. I.e. F is an ultrafilter. In other words the intersection of all elements in F (the minimum in F) is join-irreducible.

An Example of a Canonical Lattice Automaton

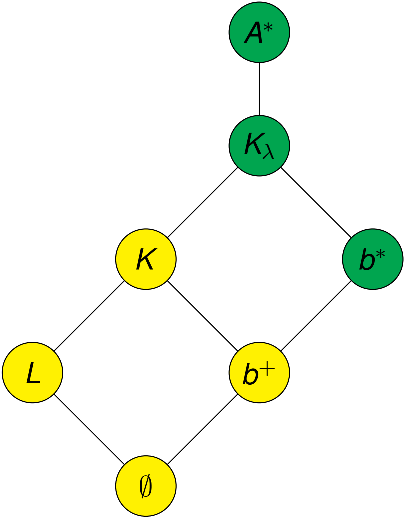


$$L = a^+ b^+$$

$$K = a^{-1}L = a^* b^+ = L \cup b^+$$

$$K_\lambda = K \cup b^* = K + \lambda$$

An Example of a Canonical Lattice Automaton



$$L = a^+b^+$$

$$K = a^{-1}L = a^*b^+ = L \cup b^+$$

$$K_\lambda = K \cup b^* = K + \lambda$$

A Lattice Automata – a Formal Definition

Definition (new)

A structure $(i, P, Q, A, \cdot, \wedge, \vee, \bot, \top)$ is a lattice semiautomaton if

- $i \in P \subseteq Q$,
- $(Q, A, \cdot)$ is a DFA,
- $(Q, \wedge, \vee)$ is a distributive lattice with the minimum element $\bot$ and the largest element $\top$,
- actions by letters are endomorphisms of the lattice $(Q, \wedge, \vee)$,
- $\top$ and $\bot$ are absorbing states,
- P is the set of all states reachable from i,
- the lattice Q is generated by the set P.

Languages Recognized by a Lattice Semiautomaton

- A DL-semiautomaton $(i, P, Q, A, \cdot, \wedge, \vee, \perp, \top)$ recognizes a language L if there are $j \in P, f \in Q$ such that f is a join-irreducible and $L = \{u \in A^* \mid j \cdot u \geq f\}$.

An Eilenberg Type Correspondence

Definition (new)

Let $\mathcal{C}$ be a "Straubing" class of morphisms. A weak $\mathcal{C}$-variety of languages $\mathcal{V}$ associates to every finite alphabet A a class $\mathcal{V}(A)$ of regular languages over A in such a way that

- $\mathcal{V}(A)$ is closed under quotients,
- $\mathcal{V}$ is closed under preimages in morphisms from $\mathcal{C}$.

Theorem (new)

There are mutually inverse isomorphisms between the lattice of all $\mathcal{C}$-varieties of lattice semiautomata and the lattice of all weak $\mathcal{C}$-varieties of regular languages.

An Eilenberg Type Correspondence – An Example

Example

Let $\mathcal{V}$ be a class of languages such that
$\mathcal{V}(A) = \{A^*aA^* \mid a \in A\} \cup \{A^*\}$.

- $\mathcal{V}(A)$ is not closed under intersections nor unions, i.e. $\mathcal{V}$ is not a conjunctive (nor disjunctive) variety of languages.
- Let $f : B^* \to A^*$, $a \in A$, $L = A^*aA^*$, then $f^{-1}(L) = B^*DB^*$ where $D = \{d \in B \mid f(d)$ contains $a\}$.
  Therefore we should consider only f's such that

$$\forall b, c \in B : b \neq c \implies \mathrm{c}(f(b)) \cap \mathrm{c}(f(c)) = \emptyset\,.$$

- $\mathcal{V}$ is a weak $\mathcal{C}$-variety for such morphisms.

An Example

One can show that the corresponding variety of DL-semiautomata is given by

$$\forall q \in Q, a, b \in A \cup \{\lambda\} \ : \ q \cdot ab = q \cdot a \vee q \cdot b \,. \quad (*)$$

and

$$\forall q \in Q, a, b \in A : a \neq b \implies q \cdot a \wedge q \cdot b = q \,. \quad (**)$$