Automata, semigroups and duality

Mai Gehrke[1] Serge Grigorieff[2] Jean-Éric Pin[2]

[1]Radboud Universiteit

[2]LIAFA, CNRS and University Paris Diderot

TANCL'07, August 2007, Oxford

Outline

(1) Four ways of defining languages
(2) The profinite world
(3) Duality
(4) Back to the future

Part I

Four ways of defining languages

Words and languages

Words over the alphabet $A = \{a, b, c\}$: a, $babb$, cac, the empty word 1, etc.

The set of all words A^* is the free monoid on A. A language is a set of words.

Recognizable (or regular) languages can be defined in various ways:

- by (extended) regular expressions
- by finite automata
- in terms of logic
- by finite monoids

Basic operations on languages

- Boolean operations: union, intersection, complement.
- Product: $L_1L_2 = \{u_1u_2 \mid u_1 \in L_1, u_2 \in L_2\}$
  Example: $\{ab, a\}\{a, ba\} = \{aa, aba, abba\}$.
- Star: L^* is the submonoid generated by L

$$L^* = \{u_1u_2 \cdots u_n \mid n \geqslant 0 \text{ and } u_1, \ldots, u_n \in L\}$$

$$\{a, ba\}^* = \{1, a, aa, ba, aaa, aba, \ldots\}.$$

Various types of expressions

- Regular expressions: union, product, star:

$$(ab)^* \cup (ab)^* a$$

- Extended regular expressions (union, intersection, complement, product and star):

$$A^* \setminus (bA^* \cup A^*aaA^* \cup A^*bbA^*)$$

- Star-free expressions (union, intersection, complement, product but no star):

$$\emptyset^c \setminus (b\emptyset^c \cup \emptyset^c aa\emptyset^c \cup \emptyset^c bb\emptyset^c)$$

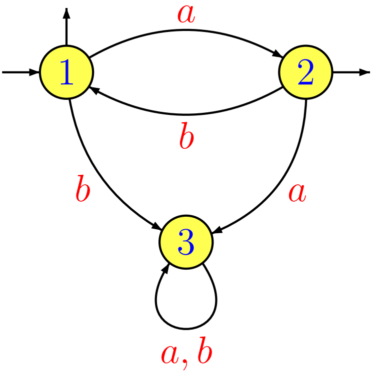
Finite automata

The set of states is $\{1, 2, 3\}$.

The initial state is 1.

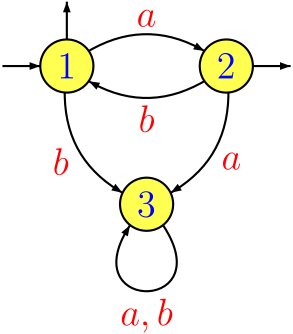
The final states are 1 and 2.

The transitions are



$$1 \cdot a = 2 \qquad 2 \cdot a = 3 \qquad 3 \cdot a = 3$$
$$1 \cdot b = 3 \qquad 2 \cdot b = 1 \qquad 3 \cdot b = 3$$

Language recognized by $\mathcal{A}$



Transitions extend to words: $1 \cdot aba = 2$, $1 \cdot abb = 3$.
The language accepted by $\mathcal{A}$ is the set of words u such that $1 \cdot u$ is a final state. Here:

$$L(\mathcal{A}) = (ab)^* \cup (ab)^* a$$

Büchi's logic to describe properties of words

- The formula $\exists x\ \mathbf{a}x$ is interpreted as:

  There exists an integer x such that, in the word, the letter in position x is an a.

It defines the language A^*aA^*.

- The first letter (of a word) is an a

$$\exists x\ \forall y\ ((x < y) \vee (x = y)) \wedge \mathbf{a}x$$

defines the language aA^*.

- The formula $\exists x\ \exists y\ (x < y) \wedge \mathbf{a}x \wedge \mathbf{b}y$ defines the language $A^*aA^*bA^*$.

Recognition by monoids

A language L of A^* is recognized by a monoid M if there exists a surjective monoid morphism $\varphi : A^* \to M$ and a subset P of M such that $L = \varphi^{-1}(P)$.

Fact 1. There is a way of associating with each finite automaton a finite monoid which recognizes the same language.

Fact 2. There is a natural notion of minimal automaton and a corresponding notion of syntactic monoid.

Recognizable languages

Definition

A language is recognizable if it is recognized by some finite automaton, or, equivalently, by a finite monoid.

A language is recognizable if and only if its syntactic monoid is finite.

The syntactic monoid is an important algebraic invariant. Its usage to classify recognizable languages is reminiscent to the use of homotopy groups in algebraic topology.

Kleene's theorem

Theorem (Kleene 1954)

Let L be a language. The following conditions are equivalent:

(1) L *is recognizable,*

(2) L *can be represented by a regular expression,*

(3) L *can be represented by an extended regular expression.*

Back to logic

Monadic second order: set variables (unary relations) are allowed.

Theorem (Büchi 1960, Elgot 1961)

Monadic second order of Büchi's logic captures recognizable languages.

Two fundamental results

Theorem (McNaughton-Papert 1971)

First order captures star-free languages (defined by star-free expressions).

Theorem (Schützenberger 1965)

A language is star-free iff its syntactic monoid is aperiodic (for all $x \in M$, there exists $n > 0$ such that $x^n = x^{n+1}$).

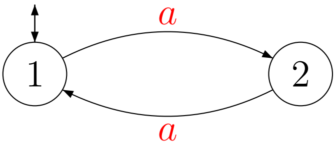
Examples of star-free languages

(1) $A^* = \emptyset^c$ is star-free.

(2) $b^* = (A^*aA^*)^c$ is star-free.

(3) $(ab)^* = \big(b\emptyset^c \cup \emptyset^c a \cup \emptyset^c aa\emptyset^c \cup \emptyset^c bb\emptyset^c\big)^c$ is star-free.

(4) $(aa)^*$ is not star-free since the syntactic monoid of a^2 is not aperiodic.

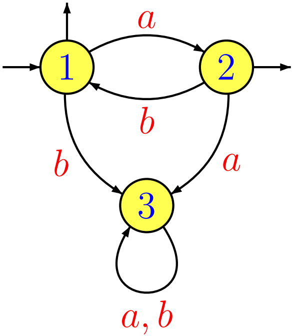


1	1	2
a	2	1
b	–	–

$$a^2 = 1$$
$$b = 0$$

The syntactic monoid of $(ab)^*$.

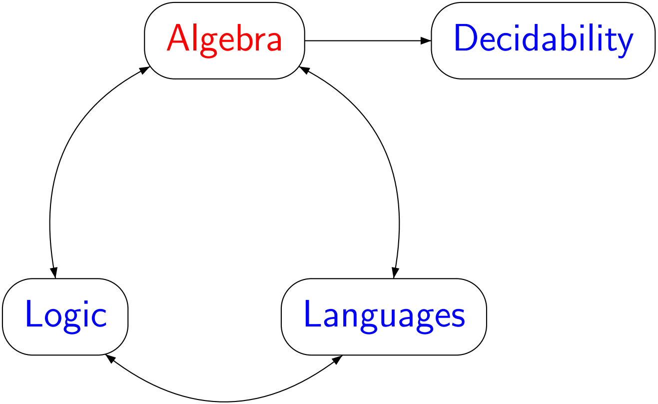
One has $M = \{1, a, b, ab, ba, aa\}$. It is aperiodic since $1^2 = 1$, $a^2 = a^3$, $b^2 = b^3$, $(ab)^2 = ab$, $(ba)^2 = ba$, $(aa)^2 = (aa)^3$. Thus $(ab)^*$ is star-free.



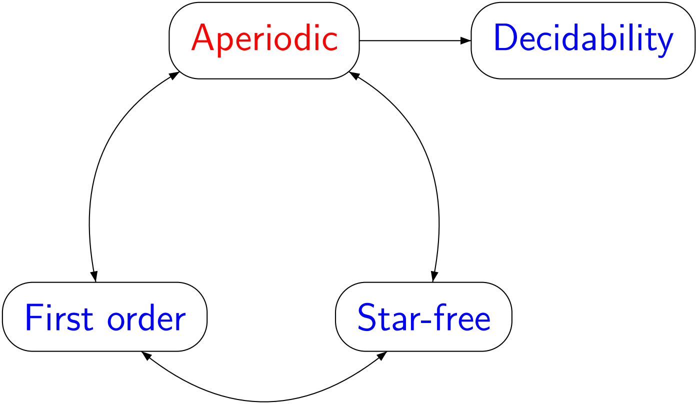
1	1	2	3
a	2	3	3
b	3	1	3
aa	3	3	3
ab	1	3	3
ba	3	2	3

$$bb = aa = 0$$
$$aba = a$$
$$bab = b$$

The first virtuous circle



An instance of the first virtuous circle



Part II

The profinite world

Quotation (M. Stone)

A cardinal principle of modern mathematical research may be stated as a maxim: One must always topologize.

Varieties

A Birkhoff variety of monoids is a class of monoids closed under taking submonoids, quotients (= homomorphic images) and direct products.

A variety of finite monoids is a class of finite monoids closed under taking submonoids, quotients and finite direct products.

Groups do not form a Birkhoff variety of monoids, but finite groups form a variety of finite monoids.

Birkhoff's theorem

Theorem (Birkhoff 1935)

A class of monoids is a Birkhoff variety iff it is defined by a set of identities.

For instance, commutative monoids are defined by the identity $xy = yx$.

What happens for finite monoids?

Separating words

A monoid M separates two words u and v of A^* if there exists a monoid morphism $\varphi : A^* \to M$ such that $\varphi(u) \neq \varphi(v)$.

For instance, the morphism which maps each word onto its length modulo 2 is a morphism from $\{a, b\}^*$ onto $\mathbb{Z}/2\mathbb{Z}$ which separates $abaaba$ and $abaabab$.

The profinite metric

Let u and v be two words. Put

$$r(u,v) = \min\{|M| \mid M \text{ is a finite monoid that separates } u \text{ and } v\}$$

$$d(u,v) = 2^{-r(u,v)}$$

Intuitively, two words are close for d if one needs a large monoid to separate them.

Then d is an ultrametric, for which the product of words is uniformly continuous.

Main properties of d

A sequence of words u_n is a Cauchy sequence iff, for every monoid morphism φ from A^* to a finite monoid, the sequence $\varphi(u_n)$ is ultimately constant.

A sequence of words u_n is converging to a word u iff, for every monoid morphism φ from A^* to a finite monoid, the sequence $\varphi(u_n)$ is ultimately equal to $\varphi(u)$.

The free profinite monoid

The completion of the metric space (A^*, d) is the free profinite monoid on A and is denoted by $\widehat{A^*}$. Its elements are called profinite words.

The product is uniformly continuous on A^* and hence can be extended to $\widehat{A^*}$. Further, if A is finite, $\widehat{A^*}$ is compact.

Any morphism $\varphi : A^* \to M$, where M is a (discrete) finite monoid is uniformly continuous. Since A^* is dense in $\widehat{A^*}$, such a morphism extends in a unique way to a uniformly continuous morphism $\hat{\varphi} : \widehat{A^*} \to M$.

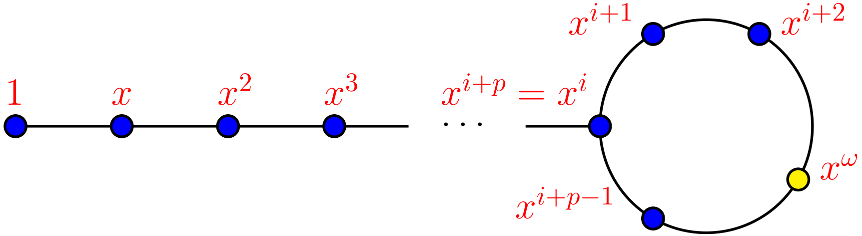
Profinite as projective limit

The profinite monoid can be viewed as the projective limit of the directed system formed by the surjective morphisms between finite monoids.

In particular, a profinite word ρ is completely determined by its images $\varphi(\rho)$, where φ runs over the class of morphisms from A^* onto a finite monoid.

A nonfinite profinite word

For each $u \in A^*$, the sequence $u^{n!}$ is a Cauchy sequence and hence converges in $\widehat{A^*}$ to a limit, denoted by u^ω. If φ is a morphism from A^* onto a finite monoid, $\varphi(u^\omega)$ is the unique idempotent of the semigroup generated by $x = \varphi(u)$.



Reiterman's theorem

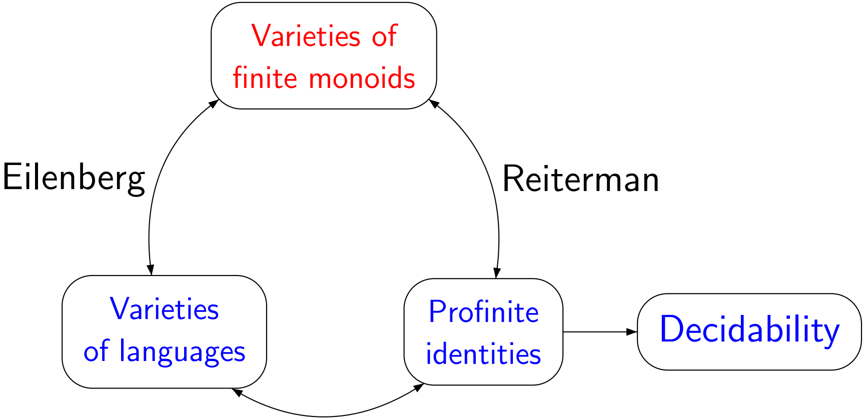
Define a profinite identity as a formal equality of the form $u = v$, where u and v are elements of a free profinite monoid.

Theorem (Reiterman 1982)

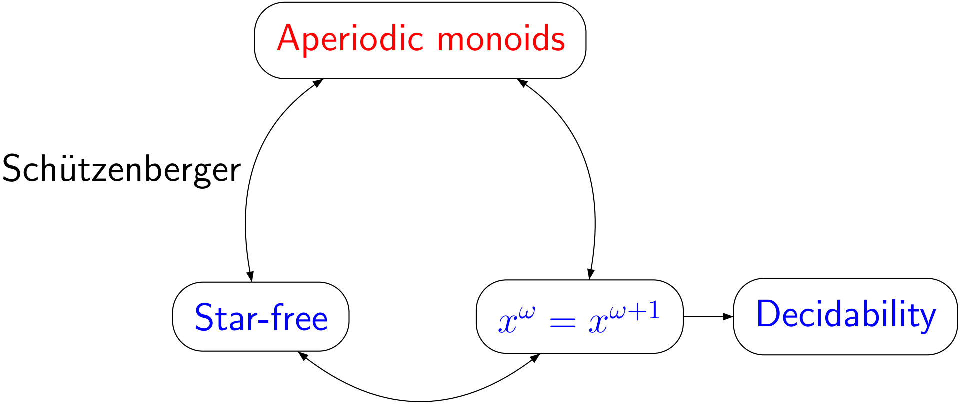
A class of finite semigroups is a variety iff it is defined by a set of profinite identities.

The variety of finite groups is defined by the single identity $x^\omega = 1$ since, in a finite group, the unique idempotent is the identity.

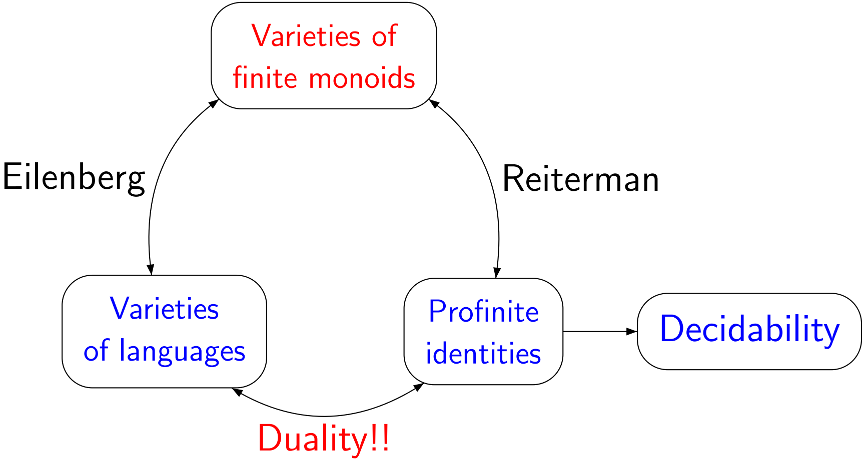
The second virtuous circle



An instance of the second virtuous circle



Duality pops up...



Part III

Duality

The dual space of $\mathrm{Rec}(A^*)$

Let $\mathrm{Rec}(A^*)$ be the distributive lattice of recognizable languages of A^*.

Its dual space X_A is the set of prime filters of $\mathrm{Rec}(A^*)$ or, equivalently, the set of lattice valuations

$$v : \mathrm{Rec}(A^*) \to \{0, 1\}$$

What are the prime filters?

The prime filters

Prime filters, valuations and profinite words can be identified. Indeed, if ρ is a profinite word, the set

$$\{\varphi^{-1}(\varphi(\rho)) \mid \varphi \text{ is a morphism from } A^* \text{ onto a finite monoid}\}$$

is a prime filter.

Conversely, if v is a valuation, one can define a profinite word ρ by the condition $\varphi(\rho) = m$ if $v(\varphi^{-1}(m)) = 1$ for each morphism φ from A^* onto a finite monoid.

Priestley duality

Thus X_A is the set of profinite words. By Priestley duality, there is an injective morphism of distributive lattices from $\mathrm{Rec}(A^*)$ into $\mathcal{P}(X_A)$:

$$\begin{aligned} L &\to \{\text{prime filters containing } L\} \\ &\to \{\text{valuations such that } v(L) = 1\} \\ &\to \{\text{profinite words } \rho \text{ such that } \varphi(\rho) \in \varphi(L)\} \end{aligned}$$

These sets are exactly the clopen sets of X_A. Further, since each singleton $\{u\}$ is a recognizable language, A^* embeds into $\mathcal{P}(X_A)$.

Residuals

The right and left residuals of L by K are defined by:

$$K\backslash L = \{u \in A^* \mid Ku \subseteq L\}$$
$$L/K = \{u \in A^* \mid uK \subseteq L\}$$

The unary versions, given by taking K to be a singleton, are the most commonly used and are called quotients. If v is a word and L is a language

$$v^{-1}L = \{u \in A^* \mid vu \in L\}$$
$$Lv^{-1} = \{u \in A^* \mid uv \in L\}$$

Residuals and product

It is easy to see that $\mathrm{Rec}(A^*)$ is closed under residual. In fact $(\mathrm{Rec}(A^*), \cdot, /, \backslash)$ is a residuated Boolean algebra and

$$\backslash, / \;:\; \mathrm{Rec}(A^*) \times \mathrm{Rec}(A^*) \to \mathrm{Rec}(A^*)$$

are residuals of the concatenation product, that is,

$$HK \subseteq L \Leftrightarrow K \subseteq H\backslash L \Leftrightarrow H \subseteq L/K$$

It follows that the product is a continous open map on X_A.

Some other consequences of duality

The following properties hold:

(1) X_A is the space of profinite words and each of them induces a term function of arity $|A|$ on any finite monoid.

(2) The identity $(H\backslash L)/K = H\backslash(L/K)$ in $Rec(A^*)$ is equivalent to stating that the product is associative on X_A.

(3) The map $u \to p_u = \{L \mid u \in L\}$ embeds $(A^*, \cdot, 1)$ into $(X_A, \cdot, p_1)$ as a discrete submonoid.

Part IV

Back to the future

Reiterman revisited

For $L \in \mathrm{Rec}(A^*)$, the syntactic monoid of L is the dual space of the quotient subalgebra of $\mathrm{Rec}(A^*)$ generated by L.

Any quotient subalgebra of $\mathrm{Rec}(A^*)$ corresponds dually to a topological monoid quotient of X_A and is thus given by a congruence on the profinite words (Reiterman's identities).

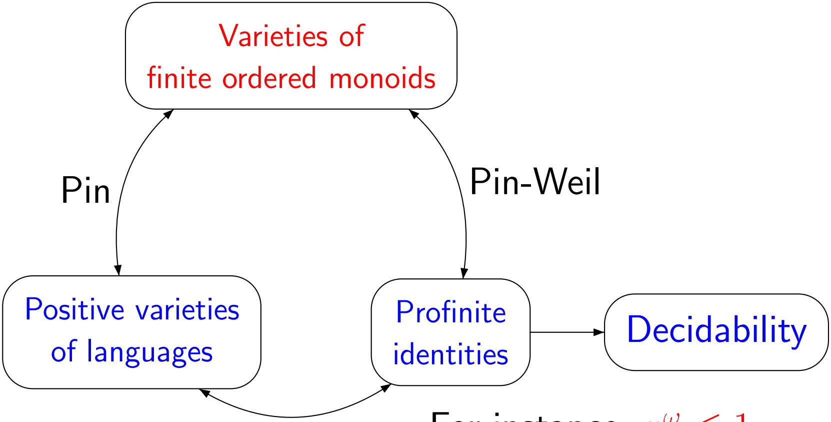
Extensions of Eilenberg's theorem

Eilenberg's definition of varieties of languages requires three conditions:

(1) closure under Boolean operations,
(2) closure under quotients,
(3) closure under inverse of morphisms between free monoids

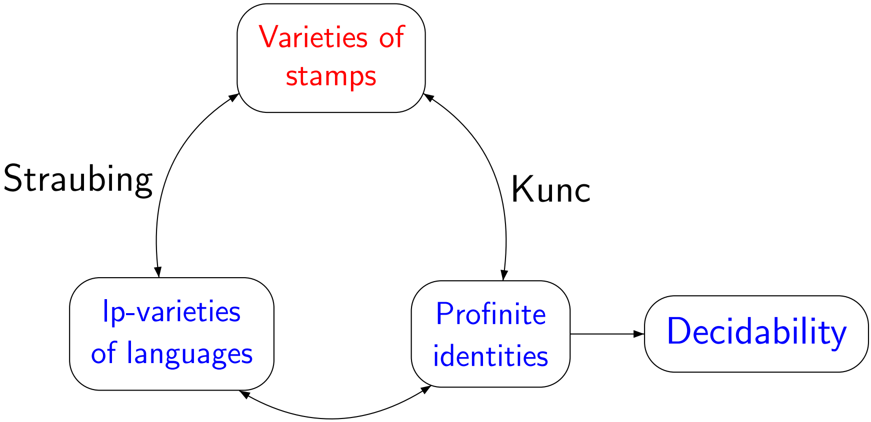
One can relax these conditions by changing the algebraic counterpart.

No complement (union and intersection only)

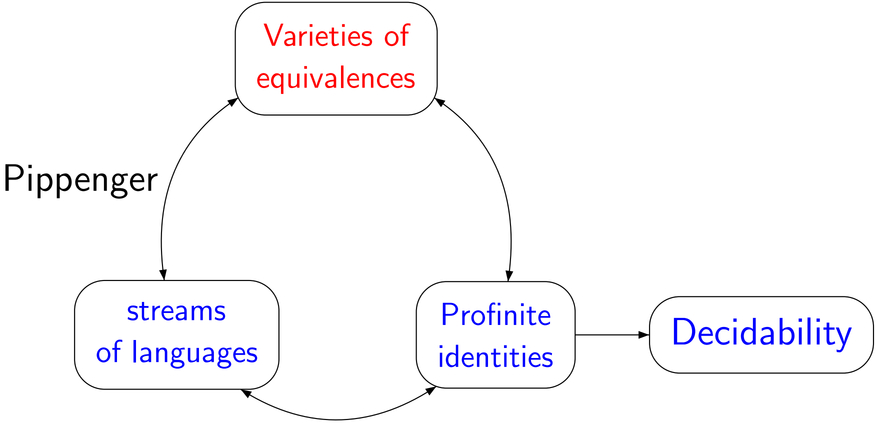


For instance, $x^\omega \leqslant 1$

Inverse of length-preserving morphims only

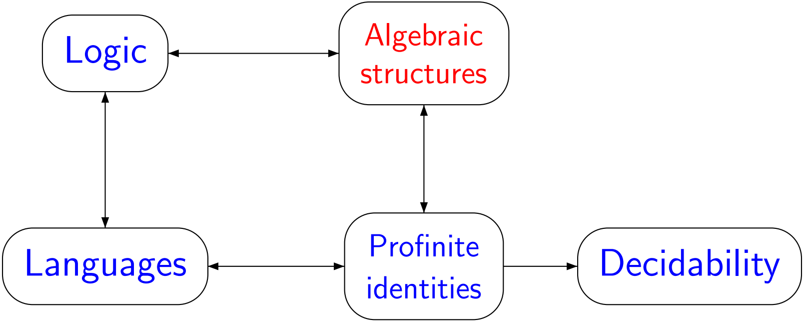


No residuals



Hope for the future

All these extensions are particular cases of our results. Further one can hope to merge the two virtuous circles



A case study

Let $\mathcal{C}$ be the lattice generated by the languages

$$\langle u \rangle = A^* u A^*, \quad u \in A^*.$$

The languages of $\mathcal{C}$ are called positively strongly locally testable (PSLT) languages.

Fact. A language is PSLT iff it can be expressed by a Σ_1-formula in Büchi's logic with the successor relation (instead of $<$).

Equations for PSLT languages

$$x^\omega y x^\omega = x^\omega y x^\omega y x^\omega$$

$$x^\omega y x^\omega z x^\omega = x^\omega z x^\omega y x^\omega$$

$$x^\omega y x^\omega \leqslant x^\omega$$

$$x^\omega u y^\omega v x^\omega \in P \Leftrightarrow y^\omega v x^\omega u y^\omega \in P$$

$$y(xy)^\omega \in P \Leftrightarrow (xy)^\omega \in P \Leftrightarrow (xy)^\omega x \in P$$

Connection with symbolic dynamics

An element of $A^{\mathbb{Z}}$ is a two-sided infinite word

$$u = \cdots\ u_{-2}u_{-1}u_0u_1u_2\cdots$$

In symbolic dynamics, a subshift is a subset of $A^{\mathbb{Z}}$ that is closed (for the product topology) and shift invariant.

We are currently working on a representation of the profinite quotient for PSLT languages using subshifts.