Profinite methods in automata theory *

Jean-Éric Pin†

Invited lecture at STACS 2009

Abstract

This survey paper presents the success story of the topological approach to automata theory. It is based on profinite topologies, which are built from finite topogical spaces. The survey includes several concrete applications to automata theory.

In mathematics, p-adic analysis is a powerful tool of number theory. The p-adic topology is the emblematic example of a *profinite topology*, a topology that is in a certain sense built from finite topological spaces. The aim of this survey is to convince the reader that profinite topologies also play a key role in automata theory, confirming once again the following quote of Marshall Stone [38, p.814]:

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

Unfortunately, this topic is rather abstract and not really intuitive. In particular, the appropriate framework to present the whole theory, namely *uniform spaces*, is unlikely to be sufficiently familiar to the average participant to STACS. To thwart this "user unfriendly" aspect, I downgraded from uniform spaces to metric spaces in this survey. This is sufficient to address most of the theory and it certainly makes the presentation easier to follow. When uniform spaces are really needed, I simply include a short warning addressed to the more advanced readers, preceded by the sign ☡. More details can be found in specialized articles [1, 2, 3, 5, 27, 30, 40].

Profinite topologies for free groups were explored by M. Hall in [13]. However, the idea of profinite topologies goes back at least to Birkhoff [8, Section 13]. In this paper, Birkhoff introduces topologies defined by congruences on abstract algebras and states that, if each congruence has finite index, then the completion of the topological algebra is compact. Further, he explicitly mentions three examples: p-adic numbers, Stone's duality of Boolean algebras and topologization of free groups. The duality between Boolean algebras and Stone spaces also appears in [1], [2, Theorem 3.6.1] and [31]. It is also the main ingredient in [12], where the extended duality between lattices and Priestley spaces is used. This duality approach is so important that it would deserve a survey article on its own. But due to the lack of space, I forwent, with some regrets, from presenting it in the present paper. The interested reader will find duality proofs of the results of Sections 4 and 5 in [12].

The survey is organised as follows. Section 1 is a brief reminder on metric spaces. Profinite words are introduced in Section 2 and used to give equational descriptions of varieties of finite monoids in Section 3 and of lattices of regular languages in Sections 4 and 5. We discuss various extensions of the profinite metric in Section 6 and we conclude in Section 7.

---

*The author acknowledge support from the AutoMathA programme of the European Science Foundation.

†LIAFA, Université Paris-Diderot and CNRS, Case 7014, 75205 Paris Cedex 13, France.

1 Metric spaces

A *metric* d on a set E is a map $d : E \to \mathbb{R}_+$ from E into the set of nonnegative real numbers satisfying the three following conditions, for every $x, y, z \in E$:

(1) $d(x, y) = 0$ if and only if $x = y$,

(2) $d(y, x) = d(x, y)$,

(3) $d(x, z) \leqslant d(x, y) + d(y, z)$

An *ultrametric* satisfies the stronger property

(3′) $d(u, w) \leqslant \max\{d(u, v), d(v, w)\}$.

A *metric space* is a set E together with a metric d on E. The topology defined by d is obtained by taking as a basis the *open ε-balls* defined for $x \in E$ and $\varepsilon > 0$ by $B(x, \varepsilon) = \{y \in E \mid d(x, y) < \varepsilon\}$. In other words, an *open* set is a (possibly infinite) union of open balls. The complement of an open set is called a *closed set.* A set is *clopen* if it is both open and closed. Every metric space is *Hausdorff*, which means that any two distinct points can be separated by open sets.

A *Cauchy sequence* is a sequence $(x_n)_{n\geqslant 0}$ of elements of E such that for each $\varepsilon > 0$, there exists a integer k such that, for each $n \geqslant k$ and $m \geqslant k$, $d(x_n, x_m) < \varepsilon$.

Let (E, d) and (E', d') be two metric spaces. A function φ from E into E' is said to be *uniformly continuous* if for each $\varepsilon > 0$, there exists $\delta > 0$ such that the relation $d(x, y) < \delta$ implies $d'(\varphi(x), \varphi(y)) < \varepsilon$. If φ is uniformly continuous, the image under φ of a Cauchy sequence of E is a Cauchy sequence of E'. We say that φ is a *uniform isomorphism* if it is a uniformly continuous bijection and φ^{-1} is also uniformly continuous. Two metric spaces are *uniformly isomorphic* is there is a uniform isomorphism between them.

A metric space is *complete* if every Cauchy sequence is convergent. The *completion* of a metric space E is a complete metric space $\widehat{E}$ together with an isometric embedding of E as a dense subspace of $\widehat{E}$. One can prove that every metric space admits a completion, which is unique up to uniform isomorphism. Further, if φ is a uniformly continuous function from (E, d) in a metric space (E', d'), φ admits a uniformly continuous extension $\hat{\varphi} : \widehat{E} \to E'$ and this extension is unique.

The completion of E can be constructed as follows. Let $C(E)$ be the set of Cauchy sequences in E. Define an equivalence relation $\sim$ on $C(E)$ as follows. Two Cauchy sequences $x = (x_n)_{n\geqslant 0}$ and $y = (y_n)_{n\geqslant 0}$ are equivalent if the interleave sequence $x_0, y_0, x_1, y_1, \ldots$ is also a Cauchy sequence. The completion of E is defined to be the set $\widehat{E}$ of equivalence classes of $C(E)$. The metric d on E extends to a metric on $\widehat{E}$ defined by

$$d(x, y) = \lim_{n\to\infty} d(x_n, y_n)$$

where x and y are representative Cauchy sequences of elements in $\widehat{E}$. The definition of the equivalence insures that the above definition does not depend on the choice of x and y in their equivalence class and the fact that $\mathbb{R}$ is complete ensures that the limit exists.

2 Profinite words

In this section, A denotes a finite alphabet. The set of profinite words is defined as the completion of A^* for a certain metric. One can actually choose one of two natural metrics, which define the same uniform structure. One makes use of finite automata and the other one of finite monoids.

2.1 Separating words

A deterministic finite automaton (DFA) *separates* two words if it accepts one of them but not the other. Similarly, a finite monoid M *separates* two words u and v of A^* if there is a monoid morphism $\varphi : A^* \to M$ such that $\varphi(u) \neq \varphi(v)$.

Example 2.1

(1) *The words ababa and abaa can be separated by a group of order* 2. *Indeed, let* $\pi : A^* \to \mathbb{Z}/2\mathbb{Z}$ *be the morphism defined by* $\pi(x) = |x| \pmod 2$. *Then* $\pi(ababa) = 1$ *and* $\pi(abaa) = 0$ *and hence* π *separates* u *and* v.

(2) *More generally, two words* u *and* v *of unequal length can be separated by a finite cyclic group. Indeed, suppose that* $|u| < |v|$ *and let* $n = |v|$. *Let* $\pi : A^* \to \mathbb{Z}/n\mathbb{Z}$ *be the morphism defined by* $\pi(x) = |x| \pmod n$. *Then* $\pi(v) = 0$ *but* $\pi(u) \neq 0$. *A similar idea can be applied if the number of occurrences of some letter* a *is not the same in* u *and* v.

(3) *Let* U_2 *be the monoid defined on the set* $\{1, a, b\}$ *by the operation* $aa = ba = a$, $bb = ab = b$ *and* $1x = x1 = x$ *for all* $x \in \{1, a, b\}$. *Let* u *and* v *be words of* $\{a, b\}^*$. *Then the words* ua *and* vb *can be separated by the morphism* $\pi : A^* \to U_2$ *defined by* $\pi(a) = a$ *and* $\pi(b) = b$ *since* $\pi(ua) = a$ *and* $\pi(ub) = b$.

These examples are a particular case of a general result.

Proposition 2.1 *Any pair of distinct words of* A^* *can be separated by a finite monoid.*

Proof. Let u and v be two distinct words of A^*. Since the language $\{u\}$ is regular, there exists a morphism φ from A^* onto a finite monoid M which recognizes it, that is, such that $\varphi^{-1}(\varphi(u)) = \{u\}$. It follows that $\varphi(v) \neq \varphi(u)$ and thus φ separates u and v. $\square$

2.2 Profinite metrics

We now define two metrics on A^* with the following idea in mind: two words are close for d_1 [d_2] if a large DFA [monoid] is required to separate them. Let us denote by $|\mathcal{A}|$ the number of states of a DFA $\mathcal{A}$. Given two words $u, v \in A^*$, we set

$$r_1(u, v) = \min \{|\mathcal{A}| \mid \mathcal{A} \text{ is a DFA that separates } u \text{ and } v\}$$
$$r_2(u, v) = \min \{|M| \mid M \text{ is a monoid that separates } u \text{ and } v\}$$

We also set $d_1(u, v) = 2^{-r_1(u,v)}$ and $d_2(u, v) = 2^{-r_2(u,v)}$ with the usual conventions $\min \emptyset = +\infty$ and $2^{-\infty} = 0$.

Proposition 2.2 *Let* d *be one of the functions* d_1 *or* d_2. *Then* d *is an ultrametric and it satisfies the relations* $d(uw, vw) \leqslant d(u, v)$ *and* $d(wu, wv) \leqslant d(u, v)$ *for all* $u, v, w \in A^*$.

Note that the topology induced on A^* by d_1 or d_2 is discrete: every subset of A^* is clopen. Further, d_1 and d_2 define the same uniform structure.

Proposition 2.3 *The metrics* d_1 *and* d_2 *are uniformly equivalent. More precisely, the following relation holds:* $2^{-\frac{1}{d_1}} \leqslant d_2 \leqslant d_1$.

We let the reader verify that changing DFAs to NFAs in the definition of d_1 would also lead to a uniformly equivalent metric. Thus (A^*, d_1) and (A^*, d_2) are metric spaces, and their completion are uniformly isomorphic. In the sequel, we shall only use d_2 (rather than d_1) and simplify the notation to d.

The completion of (A^*, d), denoted by $\widehat{A^*}$, is the set of *profinite words* on the alphabet A. Let us state some useful properties.

Proposition 2.4

(1) *The concatenation product is a uniformly continuous from $A^* \times A^*$ to A^*.*

(2) *Every morphism φ from A^* into a discrete finite monoid M is uniformly continuous.*

It follows from Proposition 2.4 and from the density of A^* in $\widehat{A^*}$ that the product on A^* can be extended by continuity to $\widehat{A^*}$. This extended product makes $\widehat{A^*}$ a topological monoid, called the *free profinite monoid*.

By the same argument, every morphism φ from A^* onto a finite monoid M extends uniquely to a uniformly continuous morphism from $\widehat{A^*}$ onto M. However, there are some noncontinuous morphisms form $\widehat{A^*}$ onto a finite monoid. For instance, the morphism φ from $\widehat{A^*}$ to $\{0, 1\}$, defined by $\varphi(u) = 1$ if $u \in A^*$ and $\varphi(u) = 0$ otherwise, is not continuous since $\varphi^{-1}(1) = A^*$ is not closed. Now, the restriction of φ to A^*, which is continuous, has a continuous extension to $\widehat{A^*}$. But this extension maps every profinite word to 1 and is therefore not equal to φ.

Another useful example is the following. The set 2^A of subsets of A is a monoid under union and the function $c : A^* \to 2^A$ defined by $c(a) = \{a\}$ is a morphism. Thus $c(u)$ is the set of letters occurring in u. Now c extends into a uniformly continuous morphism from $\hat{A}^*$ onto 2^A, also denoted c and called the *content* mapping.

Since A^* embeds naturally in $\widehat{A^*}$, every finite word is a profinite word. However, it is relatively difficult to give "concrete" examples of profinite words which are not words. One such example is the profinite word x^ω, associated with every finite word x. The formal definition is

$$x^\omega = \lim_{n\to\infty} x^{n!}$$

and is justified by the fact that the sequence $x^{n!}$ has a limit in $\widehat{A^*}$.

Proposition 2.5 *For each word x, the sequence $(x^{n!})_{n\geqslant 0}$ is a Cauchy sequence. It converges to an idempotent element of $\widehat{A^*}$.*

Proof. For the first part of the statement, it suffices to show that for $p, q \geqslant n$, $x^{p!}$ and $x^{q!}$ cannot be separated by a monoid of size $\leqslant n$. Let indeed $\varphi : A^* \to M$ be a monoid morphism, with $|M| \leqslant n$, and put $s = \varphi(x)$. Since M is finite, s has an idempotent power $e = s^r$, with $r \leqslant n$. By the choice of p and q, the integer r divides simultaneously $p!$ and $q!$. Consequently, $s^{p!} = s^{q!} = e$, which shows that M cannot separate $x^{p!}$ and $x^{q!}$.

For n large enough, we also have $\varphi(x^{n!})\varphi(x^{n!}) = ee = e = \varphi(x^{n!})$. It follows that the limit of the sequence $(x^{n!})_{n\geqslant 0}$ is idempotent. □

Note that x^ω is simply a notation and one should resist the temptation to interpret it as an infinite word. To get the right intuition, let us compute the image of x^ω under a morphism onto in a finite monoid. Let M be a finite monoid, $\varphi : A^* \to M$ a morphism and let $s = \varphi(u)$. Then the sequence $s^{n!}$ is ultimately equal to s^ω, the unique idempotent of the subsemigroup of M generated by s. Consequently, we obtain the formula $\hat{\varphi}(x^\omega) = \varphi(x)^\omega$, which justifies the notation x^ω.

Another convenient way to define profinite words is to use projective systems (see [3] for more details). Suppose we are given, for each morphism φ from A^* onto a finite monoid M, an element x_φ of M. This system of elements is *projective* if for any surjective morphisms $\varphi : A^* \to M$ and $\pi : M \to N$, one has $x_{\pi\circ\varphi} = \pi(x_\varphi)$.

Proposition 2.6 *For each projective system of elements (x_φ), there is a unique profinite word x such that, for every morphism $\varphi : A^* \to M$, one has $\hat{\varphi}(x) = x_\varphi$. In particular, if two profinite words u and v satisfy $\hat{\varphi}(u) = \hat{\varphi}(v)$ for all morphisms φ onto a finite monoid, then they are equal.*

We now state the most important topological property of $\widehat{A^*}$.

Theorem 2.7 *The set of profinite words $\widehat{A^*}$ is compact.*

If A is infinite, a profinite uniform structure can also be defined on A^* and its completion is still a compact space. However, this space is not metrizable anymore.

What about sequences? First, every profinite word is the limit of a Cauchy sequence of words. Next, a sequence of profinite words $(u_n)_{n\geqslant 0}$ is converging to a profinite word u if and only if, for every morphism φ from A^* onto a finite monoid, $\hat{\varphi}(u_n)$ is ultimately equal to $\hat{\varphi}(u)$.

Here is another example. Recall that a nonempty subset I of a monoid M is an *ideal* if, for each $s \in I$ and $x, y \in M$, $xsy \in I$. One can show that any finite monoid and any compact monoid has a unique minimal ideal (for inclusion), called *the* minimal ideal of M.

Let us fix a total order on the alphabet A and let $u_0, u_1, \ldots$ be the ordered sequence of all words of A^* in the induced shortlex order. For instance, if $A = \{a, b\}$ with $a < b$, the first elements of this sequence would be

$$1, a, b, aa, ab, ba, bb, aaa, aab, aba, abb, baa, bab, bba, bbb, aaaa, \ldots$$

It is proved in [32, 4] that the sequence of words $(v_n)_{n\geqslant 0}$ defined by

$$v_0 = u_0,\ v_{n+1} = (v_n u_{n+1} v_n)^{(n+1)!}$$

converges to a profinite word ρ_A, which is idempotent and belongs to the minimal ideal of $\widehat{A^*}$. We shall meet again this profinite word at the end of Section 5.2.

3 Equational definitions of varieties

A *variety of monoids* is a class of monoids closed under taking submonoids, quotients and direct products. Similarly, a *variety of finite monoids* is a class of finite monoids closed under taking submonoids, quotients and finite direct products. For instance, finite groups form a variety of finite monoids (the trick is that a submonoid of a finite group is a group). Another famous example is the variety of finite aperiodic monoids. Recall that a finite monoid M is *aperiodic* if there exists a positive integer n such that, for all $x \in M$, $x^n = x^{n+1}$.

Formally, an *identity* is a pair (u, v) of words of A^*, for some finite alphabet A. A monoid M satisfies the identity $u = v$ if, for every morphism $\varphi : A^* \to M$, $\varphi(u) = \varphi(v)$. It is a well known theorem of Birkhoff that varieties can be defined by a set of identities. A variety that can be defined by a finite set of identities is said to be *finitely based*. For instance, the variety of commutative monoids is finitely based, since it is defined by the single identity $xy = yx$. But in general, a variety is not finitely based, even if it is generated by a finite monoid. Consider the monoid $M = \{1, a, b, ab, ba, 0\}$ defined by the relations $aa = bb = 0$, $aba = a$ and $bab = b$. It has been proved that the variety generated by M is not finitely based.

An interesting question is to know whether varieties of finite monoids can also be defined by identities. The problem was solved by several authors but the most satisfactory answer is due to Reiterman [33]. A *profinite identity* is a pair (u, v) of profinite words of $\widehat{A^*}$, for some finite alphabet A. A finite monoid M satisfies the profinite identity $u = v$ if, for every morphism $\varphi : A^* \to M$, $\hat{\varphi}(u) = \hat{\varphi}(v)$. Reiterman's theorem is now the exact counterpart of Birkhoff's theorem:

Theorem 3.1 *Every variety of finite monoids can be defined by a set of profinite identities.*

For instance the variety of finite aperiodic monoids is defined by the identity $x^\omega = x^{\omega+1}$ and the variety of finite groups is defined by the identity $x^\omega = 1$.

4 Recognizable languages and clopen sets

A series of results, mainly due to Almeida [1, 3], [2, Theorem 3.6.1] and Pippenger [31], establishes a strong connection between regular languages and clopen sets. This section gives a short overview of these results.

Recall that a subset P of a monoid M is *recognizable* if there exists a morphism φ from M onto a finite monoid F such that $P = \varphi^{-1}(\varphi(P))$. For instance, the recognizable subsets of a free monoid are the regular languages.

The *syntactic congruence* of P is the congruence $\sim_P$ defined on M by $u \sim_P v$ if and only if, for all $x, y \in M$, the conditions $xuy \in P$ and $xvy \in P$ are equivalent. The monoid $M/{\sim_P}$ is called the *syntactic monoid* of P.

In the context of uniform spaces, the morphisms are uniformly continuous. It is therefore natural to extend the notion of recognizable set as follows: A subset P of a compact monoid M is *recognizable* if there exists a *uniformly continuous* morphism φ from M onto a finite discrete monoid F such that $P = \varphi^{-1}(\varphi(P))$. When M is a free profinite monoid, the recognizable subsets have a nice topological characterization, due to Hunter [14, Lemma 4].

Proposition 4.1 *Let P be a subset of $\widehat{A^*}$. The following conditions are equivalent:*

(1) *P is clopen,*

(2) *the syntactic congruence of P is a clopen subset of $\widehat{A^*} \times \widehat{A^*}$,*

(3) *P is recognizable (in the topological sense).*

Proof. Let us denote by $\sim_P$ the syntactic congruence of P and by $\hat\eta : \widehat{A^*} \to M$ its syntactic morphism. Recall that $s \sim_P t$ if, for all $u, v \in \widehat{A^*}$, the conditions $usv \in P$ and $utv \in P$ are equivalent.

(1) implies (2). It follows from the definition of $\sim_P$ that

$$\sim_P = \bigcap_{u,v\in\widehat{A^*}} \big((u^{-1}Pv^{-1} \times u^{-1}Pv^{-1}) \cup (u^{-1}P^cv^{-1} \times u^{-1}P^cv^{-1})\big) \qquad (4.1)$$

If P is clopen, each set $u^{-1}Pv^{-1}$ is also clopen. Indeed, $u^{-1}Pv^{-1}$ is the inverse image of the clopen set P under the continuous function $x \mapsto uxy$. Now, Formula (4.1) shows that $\sim_P$ is closed.

In order to show that the complement of $\sim_P$ is closed, consider a sequence (s_n, t_n) of elements of $(\sim_P)^c$, converging to a limit (s, t). Since $s_n \not\sim_P t_n$, there exist some profinite words u_n, v_n such that $u_ns_nv_n \in P$ and $u_nt_nv_n \notin P$. Since $\widehat{A^*} \times \widehat{A^*}$ is compact, the sequence (u_n, v_n) has a convergent subsequence. Let (u, v) be its limit. Since both P and P^c are closed and since the multiplication in $\widehat{A^*}$ is continuous, one gets $usv \in P$ and $utv \notin P$. Therefore, $s \not\sim_P t$, which shows that $(\sim_P)^c$ is closed. Thus $\sim_P$ is clopen.

(2) implies (3). If $\sim_P$ is clopen, then for each $s \in \widehat{A^*}$, there exists an open neighbourhood U of s such that $U \times U \subseteq {\sim_P}$. Therefore U is contained in the $\sim_P$-class of s. This proves that the $\sim_P$-classes form an open partition of $\widehat{A^*}$. By compactness, this partition is finite and thus P is recognizable. Further, since each $\sim_P$-class is open, the syntactic morphism of P is continuous.

(3) implies (1). Let $\pi : \widehat{A^*} \to M$ be the syntactic morphism of P. Since P is recognizable, M is finite. One has $P = \pi^{-1}(\pi(P))$ and since M is finite, $\pi(P)$ is clopen in M. Finally, since π is continuous, P is clopen in $\widehat{A^*}$. $\square$

We now turn to languages of A^*.

Proposition 4.2 *If L be a language of A^*, then $L = \overline{L} \cap A^*$. Further, the following conditions are equivalent:*

(1) *L is recognizable,*

(2) *$L = K \cap A^*$ for some clopen subset K of $\widehat{A^*}$,*

(3) *$\overline{L}$ is clopen in $\widehat{A^*}$,*

(4) *$\overline{L}$ is recognizable in $\widehat{A^*}$ (in the topological sense).*

Proof. The inclusion $L \subseteq \overline{L} \cap A^*$ is obvious. Let $u \in \overline{L} \cap A^*$ and let M be the syntactic monoid of $\{u\}$. Since M separates u from any word v different from u, one gets $r(u, v) \leqslant |M|$ if $u \neq v$. Let $(u_n)_{n \in \mathbb{N}}$ be a sequence of words of L converging to u. If $d(u_n, u) < 2^{-|M|}$, one has necessarily $u = u_n$ and thus $u \in L$.

(1) implies (2). If L is recognizable, there is a morphism φ from A^* onto a finite monoid M such that $L = \varphi^{-1}(\varphi(L))$. Let $K = \hat{\varphi}^{-1}(\varphi(L))$. Since M is discrete, $\varphi(L)$ is a clopen subset of M and since $\hat{\varphi}^{-1}$ is continuous, K is also clopen. Further, φ and $\hat{\varphi}$ coincide on A^* and thus $L = \hat{\varphi}^{-1}(\varphi(L)) \cap A^* = K \cap A^*$.

(2) implies (3). Suppose that $L = K \cap A^*$ with K clopen. Since K is open and A^* is dense in $\widehat{A^*}$, $K \cap A^*$ is dense in K. Thus $\overline{L} = \overline{K \cap A^*} = \overline{K} = K$. Thus $\overline{L}$ is clopen in $\widehat{A^*}$.

(3) implies (4) follows from Proposition 4.1.

(4) implies (1). Let $\widehat{\eta} : \widehat{A^*} \to F$ be the syntactic morphism of $\overline{L}$ and let $P = \widehat{\eta}(\overline{L})$. Let η be the restriction of $\widehat{\eta}$ to A^*. Then we have $L = \overline{L} \cap A^* = \widehat{\eta}^{-1}(P) \cap A^* = \eta^{-1}(P)$. Thus L is recognizable. $\square$

We now describe the closure in $\widehat{A^*}$ of a recognizable language of A^*.

Proposition 4.3 *Let L be a regular language of A^* and let $u \in \widehat{A^*}$. The following conditions are equivalent:*

(1) *$u \in \overline{L}$,*

(2) *$\hat{\varphi}(u) \in \varphi(L)$, for all morphisms φ from A^* onto a finite monoid,*

(3) *$\hat{\varphi}(u) \in \varphi(L)$, for some morphism φ from A^* onto a finite monoid that recognizes L,*

(4) *$\hat{\eta}(u) \in \eta(L)$, where η is the syntactic morphism of L.*

Proof. (1) implies (2). Let φ be a morphism from A^* onto a finite monoid F and let $\hat{\varphi}$ be its continuous extension to $\widehat{A^*}$. Then $\hat{\varphi}(\overline{L}) \subset \overline{\hat{\varphi}(L)}$ since $\hat{\varphi}$ is continuous, and $\overline{\hat{\varphi}(L)} = \hat{\varphi}(L) = \varphi(L)$ since F is discrete. Thus if $u \in \overline{L}$, then $\hat{\varphi}(u) \in \varphi(L)$.

(2) implies (4) and (4) implies (3) are trivial.

(3) implies (1). Let φ be a morphism from A^* onto a finite monoid F. Let u_n be a sequence of words of A^* converging to u. Since $\hat{\varphi}$ is continuous, $\hat{\varphi}(u_n)$ converges to $\hat{\varphi}(u)$. But since F is discrete, $\hat{\varphi}(u_n)$ is actually ultimately equal to $\hat{\varphi}(u_n)$. Thus for n large enough, one has $\hat{\varphi}(u_n) = \hat{\varphi}(u)$. It follows by (3) that $\varphi(u_n) = \hat{\varphi}(u_n) \in \varphi(L)$ and since φ recognizes L, we finally get $u_n \in \varphi^{-1}(\varphi(L)) = L$. Therefore $u \in \overline{L}$. $\square$

Let us denote by $\mathrm{Clopen}(\widehat{A^*})$ the Boolean algebra of all clopen sets of $\widehat{A^*}$.

Theorem 4.4 *The maps $L \mapsto \overline{L}$ and $K \mapsto K \cap A^*$ define mutually inverse isomorphism between the Boolean algebras $\mathrm{Reg}(A^*)$ and $\mathrm{Clopen}(\widehat{A^*})$. In particular, the following formulas hold, for all $L, L_1, L_2 \in \mathrm{Reg}(A^*)$:*

(1) $\overline{L^c} = (\overline{L})^c$,
(2) $\overline{L_1 \cup L_2} = \overline{L}_1 \cup \overline{L}_2$,
(3) $\overline{L_1 \cap L_2} = \overline{L}_1 \cap \overline{L}_2$.

Proof. Property (1) follows from Proposition 4.3. Indeed, let η be the syntactic morphism of L. Then since $L = \eta^{-1}(\eta(L))$ and $L^c = \eta^{-1}(\eta(L)^c)$, one has $\eta(L^c) = \eta(L)^c$. Therefore, one gets the following sequence of equalities:

$$\overline{L^c} = \hat{\eta}^{-1}(\eta(L^c)) = \hat{\eta}^{-1}(\eta(L)^c) = [\hat{\eta}^{-1}(\eta(L))]^c = (\overline{L})^c$$

Property (2) is a general result of topology and (3) is a consequence of (1) and (2). $\square$

Theorem 4.4 shows that the closure operator behaves nicely with respect to Boolean operations. It also behaves nicely for the left and right quotients and for inverse of morphisms.

Proposition 4.5 *Let L be a regular language of A^* and let $x, y \in A^*$. Then $\overline{x^{-1}Ly^{-1}} = x^{-1}\overline{L}y^{-1}$.*

Proposition 4.6 *Let $\varphi : A^* \to B^*$ be a morphism of monoids and L be a regular language of B^*. Then $\hat{\varphi}^{-1}(\overline{L}) = \overline{\varphi^{-1}(L)}$.*

5 Equational characterization of languages

A *lattice of languages* of A^* is a set of regular languages of A^* containing the empty language $\emptyset$, the full language A^* and which is closed under finite union and finite intersection. The aim of this section is to show that each lattice can be, in a certain sense, defined by a set of profinite equations. These results were obtained jointly with Mai Gehrke and Serge Grigorieff and first presented at ICALP'08 [12].

5.1 Lattices of languages

Formally, an *explicit equation* is a pair (u, v) of words of A^* and a *profinite equation* is a pair (u, v) of profinite words of $\widehat{A^*}$. We say that a language L of A^* *satisfies the explicit equation* $u \to v$ if the condition $u \in L$ implies $v \in L$ and that it *satisfies the profinite equation* $u \to v$ if the condition $u \in \overline{L}$ implies $v \in \overline{L}$. Since $\overline{L} \cap A^* = L$, the two definitions are consistent, that is, one can really consider explicit equations as a special case of profinite equations. Proposition 4.3 leads immediately to some equivalent definitions:

Corollary 5.1 *Let L be a regular language of A^*, let η be its syntactic morphism and let φ be any morphism onto a finite monoid recognizing L. The following conditions are equivalent:*

(1) *L satisfies the equation $u \to v$,*
(2) *$\hat{\eta}(u) \in \eta(L)$ implies $\hat{\eta}(v) \in \eta(L)$,*
(3) *$\hat{\varphi}(u) \in \varphi(L)$ implies $\hat{\varphi}(v) \in \varphi(L)$.*

Given a set E of equations of the form $u \to v$, the subset of $\mathrm{Reg}(A^*)$ *defined by E* is the set of all regular languages of A^* satisfying all the equations of E. It is easy to see that it is a lattice of languages.

Our aim is now to show that the converse also holds. We start with a result on languages interesting on its own right. Note in particular that there is no regularity assumption in this proposition.

Proposition 5.2 *Let L, $L_1, \ldots, L_n$ be languages. If L satisfies all the explicit equations satisfied by $L_1, \ldots, L_n$, then L belongs to the lattice of languages generated by $L_1, \ldots, L_n$.*

Proof. We claim that
$$L = \bigcup_{I \in \mathcal{I}} \bigcap_{i \in I} L_i \tag{5.2}$$
where $\mathcal{I}$ is the set of all subsets of $\{1, \ldots, n\}$ for which there exists a word $v \in L$ such that $v \in L_i$ if and only if $i \in I$. Let R be the right member of (5.2). If $u \in L$, let $I = \{i \mid u \in L_i\}$. By construction, $I \in \mathcal{I}$ and $u \in \cap_{i \in I} L_i$. Thus $u \in R$. This proves the inclusion $L \subseteq R$.

To prove the opposite direction, consider a word $u \in R$. By definition, there exists a set $I \in \mathcal{I}$ such that $u \in \cap_{i \in I} L_i$ and a word $v \in L$ such that $v \in L_i$ if and only if $i \in I$. We claim that the equation $v \to u$ is satisfied by each language L_i. Indeed, if $i \in I$, then $u \in L_i$ by definition. If $i \notin I$, then $v \notin L_i$ by definition of I, which proves the claim. It follows that $v \to u$ is also satisfied by L. Since $v \in L$, it follows that $u \in L$. This concludes the proof of (5.2) and shows that L belongs to the lattice of languages generated by $L_1, \ldots, L_n$. $\square$

It follows that finite lattices of languages can be defined by explicit equations.

Corollary 5.3 *A finite set of languages of A^* is a lattice of languages if and only if it can be defined by a set of explicit equations of the form $u \to v$, where $u, v \in A^*$.*

Proof. Consider a finite lattice $\mathcal{L}$ of languages and let E be the set of explicit equations satisfied by all the languages of $\mathcal{L}$. Proposition 5.2 shows that any language L that satisfies the equations of E belongs to $\mathcal{L}$. Thus $\mathcal{L}$ is defined by E. $\square$

We now are now ready for the main result.

Theorem 5.4 *A set of regular languages of A^* is a lattice of languages if and only if it can be defined by a set of equations of the form $u \to v$, where $u, v \in \widehat{A^*}$.*

Proof. For each regular language L, set
$$E_L = \{(u, v) \in \widehat{A^*} \times \widehat{A^*} \mid L \text{ satisfies } u \to v\}$$

Lemma 5.5 *For each regular language L, E_L is a clopen subset of $\widehat{A^*} \times \widehat{A^*}$.*

Proof. One has
$$\begin{aligned}
E_L &= \{(u, v) \in \widehat{A^*} \times \widehat{A^*} \mid L \text{ satisfies } u \to v\} \\
&= \{(u, v) \in \widehat{A^*} \times \widehat{A^*} \mid u \in \overline{L} \text{ implies } v \in \overline{L}\} \\
&= \{(u, v) \in \widehat{A^*} \times \widehat{A^*} \mid v \in \overline{L} \text{ or } u \notin \overline{L}\} \\
&= (\overline{L}^c \times \widehat{A^*}) \cup (\widehat{A^*} \times \overline{L})
\end{aligned}$$
The result follows since, by Proposition 4.2, $\overline{L}$ is clopen. $\square$

Let $\mathcal{L}$ be a lattice of languages and let E be the set of profinite equations satisfied by all languages of $\mathcal{L}$. We claim that E defines $\mathcal{L}$. First, by definition, every language of $\mathcal{L}$ satisfies the equations of E. It just remains to proving that if a language L satisfies the equations of E, then L belongs to $\mathcal{L}$.

First observe that the set
$$E_L \cup \{E_K^c \mid K \in \mathcal{L}\}$$

is a covering of $\widehat{A^*} \times \widehat{A^*}$. Indeed, if $(u,v) \notin \cup_{K\in\mathcal{L}} E_K^c$, then $(u,v) \in \cap_{K\in\mathcal{L}} E_K$, which means by definition that all the languages of $\mathcal{L}$ satisfy $u \to v$. It follows that L also satisfies this equation, and thus $(u,v) \in E_L$. Further, Proposition 5.5 shows that the elements of this covering are open sets. Since $\widehat{A^*} \times \widehat{A^*}$ is compact, it admits a finite subcovering, and we may assume that this covering contains E_L and is equal to

$$E_L \cup \{E_{L_1}^c, \ldots, E_{L_n}^c\}$$

for some languages $L_1, \ldots, L_n$ of $\mathcal{L}$. By the same argument as above, it follows that if an equation $u \to v$ is satisfied by $L_1, \ldots, L_n$, then it is satisfied by L. By Proposition 5.2, L belongs to the lattice of languages generated by $L_1, \ldots, L_n$ and hence belongs to $\mathcal{L}$. $\square$

Writing $u \leftrightarrow v$ for ($u \to v$ and $v \to u$), we get an equational description of the Boolean algebras of languages.

Corollary 5.6 *A set of regular languages of A^* is a Boolean algebra of languages if and only if it can be defined by a set of profinite equations of the form $u \leftrightarrow v$, where $u, v \in \widehat{A^*}$.*

These results apply in particular to any class of regular languages defined by a fragment of logic closed under conjunctions and disjunctions (first order, monadic second order, temporal, etc.). Consider for instance Büchi's *sequential calculus*, which comprises the relation symbols S and $<$ and a predicate $\mathbf{a}$ for each letter a. To each word nonempty word $u \in A^*$ is associated a structure

$$\mathcal{M}_u = (\{1, 2, \ldots, |u|\}, S, (\mathbf{a})_{a\in A})$$

where S denotes the successor relation on $\{1, 2, \ldots, |u|\}$, $<$ is the usual order and $\mathbf{a}$ is set of all positions i such that the i-th letter of u is an a. For instance, if $A = \{a, b\}$ and $u = abaab$, then $\mathbf{a} = \{1, 3, 4\}$ and $\mathbf{b} = \{2, 5\}$. The language defined by a sentence φ is the set of words which satisfy φ. Several fragments will be considered in this survey. We use a transparent notation of the form *Type*[Signature] to designate these fragments. For instance $FO[<]$ denotes the set of first order formulas in the signature $<$ and $\mathcal{B}\Sigma_1[S]$ consists of the Boolean combinations of existential first order formulas in the signature S.

This latter fragment allows to specify some combinatorial properties of words, like "the factor aa occurs at least twice", which defines the language $A^*aaA^*aaA^* \cup A^*aaaA^*$. Indeed, this language is described by the formula

$$\varphi = \exists x_1\ \exists x_2\ \exists y_1\ \exists y_2\ (\neg(x_1 = y_1) \wedge Sx_1x_2 \wedge Sy_1y_2 \wedge \mathbf{a}x_1 \wedge \mathbf{a}x_2 \wedge \mathbf{a}y_1 \wedge \mathbf{a}y_2$$

The $\mathcal{B}\Sigma_1(S)$-definable languages form a lattice of languages. An equational description of these languages can be derived from the results of [25]: for all $r, s, u, v, x, y \in A^*$,

$$\begin{aligned} &ux^\omega y \leftrightarrow ux^{\omega+1}v && ux^\omega ry^\omega sx^\omega ty^\omega v \leftrightarrow ux^\omega ty^\omega sx^\omega ry^\omega y \\ &x^\omega uy^\omega vx^\omega \leftrightarrow y^\omega vx^\omega uy^\omega && y(xy)^\omega \leftrightarrow (xy)^\omega \leftrightarrow (xy)^\omega x \end{aligned}$$

Note that this example does not enter in the category considered in the next section since the correspondence lattice of languages is not closed under quotient.

We now specialize Theorems 5.4 and Corollary 5.6 to lattices of languages closed under quotient in Section 5.2 and to varieties and $\mathcal{C}$-varieties of languages in Section 5.3.

5.2 Lattices of languages closed under quotient

We say that a lattice of regular languages $\mathcal{L}$ is *closed under quotient* if for every $L \in \mathcal{L}$ and $u \in A^*$, $u^{-1}L$ and Lu^{-1} are also in $\mathcal{L}$. The equational description of such lattices can be simplified by introducing a convenient definition.

Let u and v be two profinite words of $\widehat{A^*}$. We say that L *satisfies the equation* $u \leqslant v$ if, for all $x, y \in \widehat{A^*}$, it satisfies the equation $xvy \rightarrow xuy$. Since A^* is dense in $\widehat{A^*}$, it is equivalent to state that L satisfies these equations only for all $x, y \in A^*$. But there is a much more convenient characterization using the syntactic ordered monoid of L.

Recall that the *syntactic preorder* of a language L is the relation $\leqslant_L$ over A^* defined by $u \leqslant_L v$ if and only if, for every $x, y \in M$,

$$xvy \in L \Rightarrow xuy \in L$$

It is easy to see that $\leqslant_L$ is a partial preorder on A^*, whose associated equivalence relation is the *syntactic congruence* of L. Therefore, $\leqslant_L$ induces a partial order on the syntactic monoid M of L, called the *syntactic order* of L. The ordered monoid $(M, \leqslant_L)$ is called the *syntactic ordered monoid* of L.

Proposition 5.7 *Let L be a regular language of A^*, let $(M, \leqslant_L)$ be its syntactic ordered monoid and let $\eta : A^* \to M$ be its syntactic morphism. Then L satisfies the equation $u \leqslant v$ if and only if $\hat{\eta}(u) \leqslant_L \hat{\eta}(v)$.*

Proof. Corollary 5.1 shows that L satisfies the equation $u \leqslant v$ if and only if, for every $x, y \in A^*$, $\hat{\eta}(xvy) \in \eta(L)$ implies $\hat{\eta}(xuy) \in \eta(L)$. Since $\hat{\eta}(xvy) = \hat{\eta}(x)\hat{\eta}(v)\hat{\eta}(y) = \eta(x)\hat{\eta}(v)\eta(y)$ and since η is surjective, this is equivalent to saying that, for all $s, t \in M$, $s\hat{\eta}(v)t \in \eta(L)$ implies $s\hat{\eta}(u)t \in \eta(L)$, which exactly means that $\hat{\eta}(u) \leqslant_L \hat{\eta}(v)$. □

We can now state the equational characterization of lattices of languages closed under quotients.

Theorem 5.8 *A set of regular languages of A^* is a lattice of languages closed under quotients if and only if it can be defined by a set of equations of the form $u \leqslant v$, where $u, v \in \widehat{A^*}$.*

Theorem 5.8 can be readily extended to Boolean algebras. Let u and v be two profinite words. We say that a regular language L *satisfies the equation* $u = v$ if it satisfies the equations $u \leqslant v$ and $v \leqslant u$. Proposition 5.7 now gives immediately:

Proposition 5.9 *Let L be a regular language of A^* and let η be its syntactic morphism. Then L satisfies the equation $u = v$ if and only if $\hat{\eta}(u) = \hat{\eta}(v)$.*

This leads to the following equational description of the Boolean algebras of languages closed under quotients.

Corollary 5.10 *A set of regular languages of A^* is a Boolean algebra of languages closed under quotients if and only if it can be defined by a set of equations of the form $u = v$, where $u, v \in \widehat{A^*}$.*

Let us illustrate these results by three examples taken from [12].

(1) A *language with zero* is a language whose syntactic monoid has a zero. Languages with zero form a lattice of languages closed under quotient. They are characterized by the equations $x\rho_A = \rho_A = \rho_A x$ for all $x \in A^*$, where ρ_A is the profinite word defined at the end of Section 2.

(2) A language L of A^* is *dense* if, for every word $u \in A^*$, $L \cap A^*uA^* \neq \emptyset$. One can show that regular nondense or full languages form a lattice of languages closed under quotients. They are characterized by the equations $x \leqslant \rho_A$ and $x\rho_A = \rho_A = \rho_A x$ for all $x \in A^*$.

(3) Recall that a language L is *sparse* if it has a polynomial density, that is, if $|L \cap A^n| = O(n^k)$ for some $k > 0$. Equivalently, a language is sparse if it is a finite union of languages of the form $u_0 v_1^* u_1 \cdots v_n^* u_n$, where u_0, v_1, …, v_n, u_n are words. Sparse or full languages form a lattice of languages closed under quotient and thus admit an equational description. On a one letter alphabet, every recognizable language is sparse and the result is trivial. If $|A| \geqslant 2$, one can take the following set of equations: $x\rho_A = \rho_A = \rho_A x$ for all $x \in A^*$ and $(x^\omega y^\omega)^\omega = \rho_A$ for each $x, y \in A^+$ such that the first letter of x is different from the first letter of y.

5.3 Varieties of languages

A *class of languages* $\mathcal{F}$ associates with each alphabet A a set $\mathcal{F}(A^*)$ of regular languages of A^*. A *positive variety of languages* is a class of languages $\mathcal{V}$ such that

(1) for each alphabet A, $\mathcal{V}(A^*)$ is a lattice of languages closed under quotient,

(2) for each morphism of monoid $\varphi : A^* \to B^*$, $X \in \mathcal{V}(B^*)$ implies $\varphi^{-1}(X) \in \mathcal{V}(A^*)$,

A *variety* of languages is a positive variety of languages closed under complement.

For [positive] varieties, is is wise to use *identities* as we did for Reiterman's theorem. Intuitively, an identity is an equation in which one can substitute a word for each letter. More formally, let u and v be two profinite words of $\widehat{B}^*$ and let L be a regular language of A^*. One says that L *satisfies the profinite identity* $u \leqslant v$ $[u = v]$ if, for all morphisms $\gamma : B^* \to A^*$, L satisfies the equation $\hat{\gamma}(u) \leqslant \hat{\gamma}(v)$ $[\hat{\gamma}(u) = \hat{\gamma}(v)]$.

Theorem 5.11 *A class of languages is a positive variety of languages if and only if it can be defined by a set of profinite identities of the form $u \leqslant v$. It is a variety of languages if and only if it can be defined by a set of profinite identities of the form $u = v$.*

Theorem 5.11 and Reiterman's theorem allows one to recover Eilenberg's variety theorem. Let us first recall this important result.

If $\mathbf{V}$ is a variety of finite monoids, denote by $\mathcal{V}(A^*)$ the set of regular languages of A^* whose syntactic monoid belongs to $\mathbf{V}$. The correspondence $\mathbf{V} \to \mathcal{V}$ associates with each variety of finite monoids a variety of languages. Conversely, to each variety of languages $\mathcal{V}$, we associate the variety of monoids $\mathbf{V}$ generated by the monoids of the form $M(L)$ where $L \in \mathcal{V}(A^*)$ for a certain alphabet A. Eilenberg's variety theorem [10] states that the correspondences $\mathbf{V} \to \mathcal{V}$ and $\mathcal{V} \to \mathbf{V}$ define mutually inverse bijective correspondences between varieties of finite monoids and varieties of languages.

Now, it follows from Theorem 5.11 that any variety of languages can be defined by a set of profinite identities. And Reiterman's theorem states that varieties of finite monoids can also be defined by profinite identities. This gives the variety theorem.

There is an analogous result for ordered monoids [22], which gives mutually inverse bijective correspondences between the varieties of finite ordered monoids and the positive varieties of languages.

Equational descriptions are known for a large number of [positive] varieties of languages. We just give a few emblematic examples in elliptic style below, but many more can be found in the survey articles [9, 24].

(1) Finite or full languages: $yx^\omega = x^\omega = x^\omega y$ and $y \leqslant x^\omega$.

(2) Star-free languages (closure of finite languages under Boolean operations and product): $x^{\omega+1} = x^\omega$. These languages are also captured by the logical fragment $FO[<]$.
(3) Shuffle ideals (finite unions of languages of the form $A^*a_1A^*a_2A^* \cdots a_kA^*$, where $a_1, \ldots, a_k$ are letters): $x \leqslant 1$. These languages are also captured by the logical fragment $\Sigma_1[<]$.
(4) Piecewise testable languages (the Boolean closure of shuffle ideals): $x^{\omega+1} = x^\omega$ and $(xy)^\omega = (yx)^\omega$. These languages are also captured by the logical fragment $\mathcal{B}\Sigma_1[<]$.
(5) Unambiguous star-free languages (closure of finite languages under Boolean operations and unambiguous product): $x^{\omega+1} = x^\omega$ and $(xy)^\omega(yx)^\omega(xy)^\omega = (xy)^\omega$. These languages are also captured by the logical fragments $FO_2[<]$ (first order with two variables), by $\Delta_2[<]$ or by unary temporal logic (based on the operators *eventually in the future* and *eventually in the past*). Finally, they are disjoint unions of unambiguous products of the form $A_0^*a_1A_1^* \cdots a_kA_k^*$, where $a_1, \ldots, a_k$ are letters and $A_0, \ldots, A_k$ are subsets of A.

A more general notion was introduced by Straubing [39] (see also Ésik and Ito [11]). Let $\mathcal{C}$ be a class of morphisms between finitely generated free monoids that is closed under composition and contains all length-preserving morphisms. Examples include the classes of *length-preserving* morphisms, of *length-multiplying* morphisms, of *non-erasing* morphisms, of *length-decreasing* morphisms and of course the class of all morphisms.
A *positive $\mathcal{C}$-variety of languages* is a class of languages $\mathcal{V}$ such that

(1) for every alphabet A, $\mathcal{V}(A^*)$ is a lattice of languages closed under quotient,
(2) if $\varphi\colon A^* \to B^*$ is a morphism of $\mathcal{C}$, $L \in \mathcal{V}(B^*)$ implies $\varphi^{-1}(L) \in \mathcal{V}(A^*)$,

A *$\mathcal{C}$-variety of languages* is a positive $\mathcal{C}$-variety of languages closed under complement.

It is easy to extend Theorem 5.11 to $\mathcal{C}$-varieties by using $\mathcal{C}$-identities. Let us say that a language L *satisfies the profinite $\mathcal{C}$-identity* $u \leqslant v$ $[u = v]$ if, for all $\mathcal{C}$-morphisms $\gamma : B^* \to A^*$, L satisfies the equation $\hat\gamma(u) \leqslant \hat\gamma(v)$ $[\hat\gamma(u) = \hat\gamma(v)]$. Then we can state

Theorem 5.12 *A class of languages of A^* is a positive $\mathcal{C}$-variety of languages if and only if it can be defined by a set of profinite $\mathcal{C}$-identities of the form $u \leqslant v$. It is a $\mathcal{C}$-variety of languages if and only if it can be defined by a set of profinite $\mathcal{C}$-identities of the form $u = v$.*

5.4 Summary

We summarize below the various types of equations...

Closed under	**Equations**	**Definition**
$\cup, \cap$	$u \to v$	$\hat\eta(u) \in \hat\eta(L) \Rightarrow \hat\eta(v) \in \hat\eta(L)$
quotient	$u \leqslant v$	$xvy \to xuy$
complement	$u \leftrightarrow v$	$u \to v$ and $v \to u$
quotient and complement	$u = v$	$xvy \leftrightarrow xuy$

...and the various types of $\mathcal{C}$-identities.

Class of morphisms $\mathcal{C}$	Interpretation of variables
all morphisms	words
nonerasing morphisms	nonempty words
length multiplying morphisms	words of equal length
length preserving morphisms	letters

6 Pro-V uniformities

The profinite uniformities defined in Section 2 can be generalized in various ways using varieties of finite ordered monoids [30] or even lattices of languages [12]. In this section, we only consider the case of a variety of finite monoids $\mathbf{V}$.

The idea is to generalize the metric d on A^* by defining $d_{\mathbf{V}}$ as follows:

$$\begin{aligned} r_{\mathbf{V}}(u,v) &= \min\{|M| \mid M \text{ is a monoid of } \mathbf{V} \text{ that separates } u \text{ and } v\} \\ d_{\mathbf{V}}(u,v) &= 2^{-r_{\mathbf{V}}(u,v)} \end{aligned}$$

Unfortunately, $d_{\mathbf{V}}$ is not always a metric since the monoids of $\mathbf{V}$ may not suffice to separate two distinct words. For instance, if $\mathbf{V}$ is the variety $\mathbf{Com}$ of finite commutative monoids, there is no way to separate the words ab and ba.

There are two possibilities to overcome this difficulty. The first solution is algebraic: the relation $\sim_{\mathbf{V}}$ defined on A^* by $u \sim_{\mathbf{V}} v$ if and only if $d_{\mathbf{V}}(u,v) = 0$ is a congruence on A^* and $d_{\mathbf{V}}$ induces a metric on the quotient monoid $A^*/{\sim_{\mathbf{V}}}$ and one can take the completion of the metric space $(A^*/{\sim_{\mathbf{V}}}, d_{\mathbf{V}})$. For instance, if $\mathbf{V} = \mathbf{Com}$, $A^*/{\sim_{\mathbf{V}}}$ is the free commutative monoid $\mathbb{N}^A$.

The second solution is topological. Even if $d_{\mathbf{V}}$ fails to be a metric, it still satisfies conditions (2) and (3) of the definition of a metric and this suffices to define an Hausdorff completion. A systematic study of the corresponding uniform spaces can be found in [30].

The two methods lead to the same object, called the free pro-$\mathbf{V}$ monoid on A, denoted by $\widehat{F}_{\mathbf{V}}(A)$. This monoid is compact and can be alternatively defined as the quotient of $\widehat{A^*}$ by the congruence induced by the profinite identities defining $\mathbf{V}$.

The study of these free profinite monoids, for various varieties $\mathbf{V}$, is a central topic of the theory of finite monoids and regular languages. It is not possible to describe in detail the numerous results obtained in this area, and we refer the interested reader to the survey articles [3, 5] and to the articles of Almeida, Auinger, Margolis, Steinberg, Weil, Zeitoun or the author for more information.

We limit ourselves in this survey to direct consequences in automata theory. First, uniform continuous functions have a very concrete characterization.

Theorem 6.1 *A function $f : A^* \to B^*$ is uniformly continuous for $d_{\mathbf{V}}$ if and only if, for every language L of B^* recognized by a monoid of $\mathbf{V}$, the language $f^{-1}(L)$ is also recognized by a monoid of $\mathbf{V}$.*

It is worth considering separately the case where $\mathbf{V}$ is the variety of all finite monoids.

Corollary 6.2 *A function $f : A^* \to B^*$ is uniformly continuous for d if and only if, for every regular language L of B^*, the language $f^{-1}(L)$ is also regular.*

We illustrate the power of this approach by solving a standard exercise in automata theory: Show that the *square root* of a regular language is regular.

Proof. Since the concatenation product is uniformly continuous, the product h of two uniformly continuous functions f and g, defined by $h(u) = f(u)g(u)$, is also uniformly continuous. In particular, the function $u \to u^2$, from A^* into itself, is uniformly continuous. It follows that, if L is regular, the set $\{u \in A^* \mid u^2 \in L\}$ is also regular. $\square$

Here are two more advanced results. Given a class $\mathcal{L}$ of regular languages, the *polynomial closure* $\mathrm{Pol}(\mathcal{L})$ of $\mathcal{L}$ is the set of all languages which are finite unions of languages of the form $L_0a_1L_1 \cdots a_kL_k$ where $a_1, \ldots, a_k$ are letters and $L_0, \ldots, L_k$ are languages of $\mathcal{L}$. It can be shown that if $\mathcal{V}$ is a variety of languages, then $\mathrm{Pol}(\mathcal{V})$ is a positive variety of languages. Further, it follows from the results of [29] that, given the profinite equations satisfied by $\mathcal{V}$, one can find, at least implicitly, the profinite equations defining $\mathrm{Pol}(\mathcal{V})$.

Theorem 6.3 *The positive variety $Pol(\mathcal{V})$ is defined by the profinite identities of the form $x^\omega y x^\omega \leqslant x^\omega$ where x and y are profinite words such that the identities $x = y = x^2$ hold in $\mathcal{V}$.*

Similar results hold for the unambiguous polynomial closure (the identities are now of the form $x^\omega y x^\omega = x^\omega$) and for the closure under Boolean operations and product.

Theorem 6.4 *The closure under Boolean operations and product of a variety of languages $\mathcal{V}$ is defined by the profinite identities of the form $x^{\omega+1} = x^\omega$ where x and y are profinite words such that the identities $x = y = x^2$ hold in $\mathcal{V}$.*

We now concentrate on two important particular cases which proved to have unexpected connections with automata theory: the variety of finite groups $\mathbf{G}$ and the variety of finite p-groups $\mathbf{G}_p$, where p is a prime number. Recall that a *p-group* is a finite group whose order is a power of p.

6.1 Progroup topology

The pro-group topology was originally introduced by M. Hall in group theory [13]. It was first considered for the free monoid by Reutenauer [34, 35] and studied in full details in [19, 26, 23], notably in connection with a celebrated problem of Rhodes in finite semigroup theory. This problem was ultimately solved by Ash using a combinatorial approach [6] and by Ribes and Zalesskii using profinite methods [36].

The definition of the *pro-group metric* is obtained by taking $\mathbf{V} = \mathbf{G}$ in the definition of $d_\mathbf{V}$. One can show that $d_\mathbf{G}$ is an ultrametric, but contrary to the profinite metric d, the topology induced by $d_\mathbf{G}$ on A^* is not discrete and it is an interesting question to decide whether a given regular language is open, closed or clopen for this topology.

Recall that a *group language* is a language whose syntactic monoid is a group, or, equivalently, is recognized by a finite deterministic automaton in which each letter defines a permutation of the set of states. According to the definition of a polynomial closure, a *polynomial of group languages* is a finite union of languages of the form $L_0a_1L_1 \cdots a_kL_k$ where $a_1, \ldots, a_k$ are letters and $L_0, \ldots, L_k$ are group languages. It can be shown that a regular language is clopen if and only if it is a group language. For the open sets, the following characterization holds.

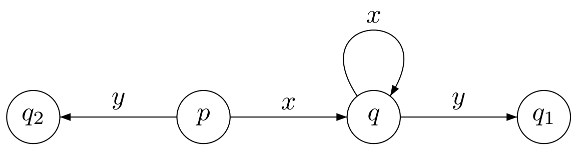
Theorem 6.5 *Let L be a regular language. The following conditions are equivalent:*

(1) *L is a polynomial of group languages,*

(2) *L is open in the group topology,*

(3) *L satisfies the identity $x^\omega \leqslant 1$,*

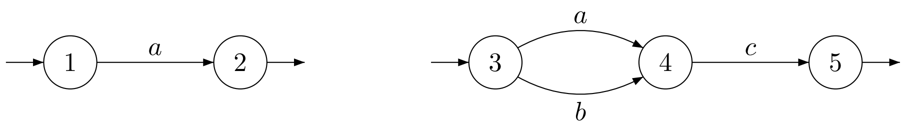
(4) *the minimal deterministic automaton of L contains no configuration of the form*



where $x, y \in A^*$, q_1 is final and q_2 is nonfinal.

One can show also that the closure of regular language for $d_{\mathbf{G}}$ is again regular and can be effectively computed [26, 36].

These results also permit to study another class of finite automata. A *reversible automaton* is a finite automaton whose transitions are both deterministic and co-deterministic. In other words, each letter a induces a partial one-to-one map from the set of states into itself. However, we make no assumption on the set of initial states and the set of final states, which can be arbitrary. It is not difficult to see that any finite language can be accepted by a reversible automaton. For instance, a reversible automaton accepting $\{a, ac, bc\}$ is represented below:



It is tempting to guess that a language is accepted by some reversible automaton if and only if its minimal DFA is reversible, but this is not the case and the characterization of these languages [20] is more involved.

Theorem 6.6 *Let L be a regular language and let M be its syntactic monoid. The following conditions are equivalent:*

(1) *L is accepted by a reversible automaton,*

(2) *the idempotents of M commute and L is closed in the profinite group topology of A^*.*

(3) *L satisfies the identities $x^\omega y^\omega = y^\omega x^\omega$ and $1 \leqslant x^\omega$,*

6.2 Pro-p topology

The definition of the *pro-p metric* is obtained by taking $\mathbf{V} = \mathbf{G}_p$ in the definition of $d_{\mathbf{V}}$. The resulting ultrametric d_p defines the *p-adic topology* on A^*. When A is a one letter alphabet, the free monoid A^* is isomorphic to the additive monoid $\mathbb{N}$ and its pro-p completion is the additive group of p-adic numbers.

As for $d_{\mathbf{G}}$, the closure of regular language for d_p is again regular and can be effectively computed [37, 18], but this is a difficult result.

There is also a nice connection [21] between this topology and a generalization of the *binomial coefficients*. Let u and v be two words of A^*. Let $u = a_1 \cdots a_n$, with $a_1, \ldots, a_n \in A$. Then u is a *subword* of v if there exist $v_0, \ldots, v_n \in A^*$ such that $v = v_0 a_1 v_1 \ldots a_n v_n$. Following [10, 15], we define the binomial coefficient of u and v by setting

$$\binom{v}{u} = |\{(v_0, \ldots, v_n) \mid v = v_0 a_1 v_1 \ldots a_n v_n\}|$$

Observe that if a is a letter, then $\binom{v}{a}$ is simply the number of occurrences of a in v. Further, if $u = a^n$ and $v = a^m$, then $\binom{v}{u} = \binom{m}{n}$ and hence these numbers constitute a generalization of the classical binomial coefficients. Let us set now

$$r'_p(u,v) = \min \left\{ |x| \mid x \in A^* \text{ and } \binom{u}{x} \not\equiv \binom{v}{x} \pmod p \right\}$$
$$d'_p(u,v) = p^{-r'_p(u,v)}.$$

It is proved in [21, Theorem 4.4] that d'_p is an ultrametric uniformly equivalent to d_p. The next proposition should be compared with Proposition 2.5.

Proposition 6.7 *For every word $u \in A^*$, one has $\lim_{n\to\infty} u^{p^n} = 1$ for the metric d_p.*

Proof. By the definition of the topology, it suffices to show that if $\varphi : A^* \to G$ is a monoid morphism onto a discrete p-group G, then $\lim_{n\to\infty} \varphi(g^{p^n}) = 1$. But if $|G| = p^k$, then for $n \geqslant k$, $\varphi(g^{p^n}) = 1$ since the order of $\varphi(g)$ divides p^k. □

There is another nice example of converging sequence, related to the Prouhet-Thue-Morse word $t = abbabaabbaababba\cdots$. Recall that this infinite word on the alphabet $\{a, b\}$ is obtained from a by iterating the morphism τ defined by $\tau(a) = ab$ and $\tau(b) = ba$.

Denoting by $t[m]$ the prefix of length m of t, one has:

Theorem 6.8 (See [7]) *For every prime number p, there exists a strictly increasing sequence $m_1 < m_2 < \cdots$ such that $\lim_{n\to\infty} t[m_n] = 1$.*

The sequence m_n depends on p but can be explicitly given. For $p \neq 2$, one can choose $m_n = 2^n p^{1+\lfloor log_p n\rfloor}$, but as often in mathematics, the case $p = 2$ is singular. In this case, one can take $m_n = 2^k$ if $F_{k-1} \leqslant n < F_k$, where F denotes the Fibonacci sequence defined by $F_0 = 0$, $F_1 = 1$ and $F_{n+2} = F_{n+1} + F_n$ for every $n \geqslant 0$.

The connection between the pro-p topology and the binomial coefficients comes from the characterization of the languages recognized by a p-group given by Eilenberg and Schützenberger (see [10, Theorem 10.1, p. 239]). Let us call a p-group language a language recognized by a p-group.

Proposition 6.9 *A language of A^* is a p-group language if and only if it is a Boolean combination of the languages*

$$L(x, r, p) = \{u \in A^* \mid \binom{u}{x} \equiv r \bmod p\},$$

for $0 \leqslant r < p$ and $x \in A^*$.

We conclude this section with a result presented at STACS last year [28], which extends a classical result of Mahler [16, 17].

Let $f : A^* \to \mathbb{Z}$ be a function. For each letter a, we define the difference operator Δ^a by $(\Delta^a f)(u) = f(ua) - f(u)$. One can now define inductively an operator Δ^w for each word $w \in A^*$ by setting $(\Delta^1 f)(u) = f(u)$, and for each letter $a \in A$, $(\Delta^{aw} f)(u) = (\Delta^a(\Delta^w f))(u)$. It is easy to see that these operators can also be defined directly by setting

$$\Delta^w f(u) = \sum_{0\leqslant|x|\leqslant|w|} (-1)^{|w|+|x|} \binom{w}{x} f(ux)$$

For instance, $\Delta^{aab} f(u) = -f(u) + 2f(ua) + f(ub) - f(uaa) - 2f(uab) + f(uaab)$.

One can show that for each function $f : A^* \to \mathbb{Z}$, there exists a unique family $\langle f, v\rangle_{v\in A^*}$ of integers such that, for all $u \in A^*$, $f(u) = \sum_{v\in A^*} \langle f, v\rangle \binom{u}{v}$. These coefficients are given by

$$\langle f, v\rangle = (\Delta^v f)(1) = \sum_{0\leqslant|x|\leqslant|v|} (-1)^{|v|+|x|} \binom{v}{x} f(x)$$

If n is a non-zero integer, we denote by $|n|_p$ the *p-adic norm* of n, which is the real number p^{-k}, where k is the largest integer such that p^k divides n. By convention, $|0|_p = 0$. The main result of [28] gives a simple description of the uniformly continuous functions for d_p.

Theorem 6.10 *Let $f(u) = \sum_{v \in A^*} \langle f, v \rangle \binom{u}{v}$ be the Mahler's expansion of a function from A^* to $\mathbb{Z}$. The following conditions are equivalent:*

(1) *f is uniformly continuous for d_p,*

(2) *the partial sums $\sum_{0 \leqslant |v| \leqslant n} \langle f, v \rangle \binom{u}{v}$ converge uniformly to f,*

(3) $\lim_{|v| \to \infty} |\langle f, v \rangle|_p = 0$.

7 Conclusion

Profinite topologies are a powerful tool to solve decidability problems on regular languages. In particular, they lead to equational definitions of lattices of languages which can sometimes be used to obtain decidability results. For instance, it follows from the deep results of McNaughton and Schützenberger that $FO[<]$-definable languages are defined by the identity $x^\omega = x^{\omega+1}$. Since Büchi has shown that monadic second order $MSO[<]$ captures all regular languages, it follows that one can effectively decide whether a monadic second order formula is equivalent to a first order formula on finite words (or, in the language of model theory, on finite coloured linear orders).

There are however two problems to extend this type of arguments to a given lattice of languages. First, one needs to find effectively the equations foretold by Theorem 5.4. This step can be extremely difficult. For instance, it is conjectured that the languages captured by the logical fragment $\mathcal{B}\Sigma_2[<]$ are defined by the identities

$$(x^\omega p y^\omega q x^\omega)^\omega x^\omega p y^\omega s x^\omega (x^\omega r y^\omega s x^\omega)^\omega = (x^\omega p y^\omega q x^\omega)^\omega (x^\omega r y^\omega s x^\omega)^\omega$$

where $x, y, p, q, r, s \in \widehat{A^*}$ are profinite words with the same content, but this conjecture is still open.

If some set of equations have been found, one still needs to decide whether a given regular language satisfies these equations. This second problem might also be difficult to solve. For instance, it is not clear whether the implicit descriptions given in Theorems 6.3 ant 6.4 lead to effective decision criteria. A lot of work has been done, notably by Almeida and Steinberg, to address this type of questions. A key idea is that certain varieties can be defined by identities involving words and simple profinite operations, like the ω operation. When a basis of such identities can be found, the second problem becomes generally easy.

To conclude, we would like to suggest a new path of research. The starting point is the following observation: the hierarchies considered in computability, in complexity theory and in descriptive set theory are defined in terms of appropriate reductions. In each case, the definition of a reduction follows the same pattern: given two sets X and Y, Y reduces to X if there exists a function f such that $X = f^{-1}(Y)$. In complexity theory, f is required to be computable in polynomial time. In descriptive set theory, f is continuous. We propose to study the reductions between regular languages based on uniformly continuous functions (for instance for some metric $d_{\mathbf{V}}$). One could then explore the corresponding hierarchies, as it has been done in descriptive set theory and in computability theory.

References

[1] J. Almeida, Residually finite congruences and quasi-regular subsets in uniform algebras, *Partugaliæ Mathematica* **46** (1989), 313–328.

[2] J. Almeida, *Finite semigroups and universal algebra*, World Scientific Publishing Co. Inc., River Edge, NJ, 1994. Translated from the 1992 Portuguese original and revised by the author.

[3] J. Almeida, Profinite semigroups and applications, in *Structural theory of automata, semigroups, and universal algebra*, pp. 1–45, *NATO Sci. Ser. II Math. Phys. Chem.* vol. 207, Springer, Dordrecht, 2005. Notes taken by Alfredo Costa.

[4] J. Almeida and M. V. Volkov, Profinite identities for finite semigroups whose subgroups belong to a given pseudovariety, *J. Algebra Appl.* **2**,2 (2003), 137–163.

[5] J. Almeida and P. Weil, Relatively free profinite monoids: an introduction and examples, in *NATO Advanced Study Institute Semigroups, Formal Languages and Groups*, J. Fountain (ed.), vol. 466, pp. 73–117, Kluwer Academic Publishers, 1995.

[6] C. J. Ash, Inevitable graphs: a proof of the type II conjecture and some related decision procedures, *Internat. J. Algebra Comput.* **1**,1 (1991), 127–146.

[7] J. Berstel, M. Crochemore and J.-E. Pin, Thue sequence and p-adic topology of the free monoid, *Discrete Mathematics* **76** (1989), 89–94.

[8] G. Birkhoff, Moore-Smith convergence in general topology, *Ann. of Math. (2)* **38**,1 (1937), 39–56.

[9] M. J. J. Branco, Varieties of languages, in *Semigroups, algorithms, automata and languages (Coimbra, 2001)*, pp. 91–132, World Sci. Publ., River Edge, NJ, 2002.

[10] S. Eilenberg, *Automata, languages, and machines. Vol. B*, Academic Press [Harcourt Brace Jovanovich Publishers], New York, 1976.

[11] Z. Ésik and M. Ito, Temporal logic with cyclic counting and the degree of aperiodicity of finite automata, *Acta Cybernetica* **16** (2003), 1–28.

[12] M. Gehrke, S. Grigorieff and J.-E. Pin, Duality and equational theory of regular languages, in *ICALP 2008, Part II*, L. Aceto and al. (ed.), Berlin, 2008, pp. 246–257, *Lect. Notes Comp. Sci.* vol. 5126, Springer.

[13] M. Hall, Jr., A topology for free groups and related groups, *Ann. of Math. (2)* **52** (1950), 127–139.

[14] R. Hunter, Certain finitely generated compact zero-dimensional semigroups, *J. Austral. Math. Soc. (Series A)* **44** (1988), 265–270.

[15] M. Lothaire, *Combinatorics on words*, *Cambridge Mathematical Library*, Cambridge University Press, Cambridge, 1997. With a foreword by Roger Lyndon and a preface by Dominique Perrin, Corrected reprint of the 1983 original, with a new preface by Perrin.

[16] K. Mahler, An interpolation series for continuous functions of a p-adic variable., *J. Reine Angew. Math.* **199** (1958), 23–34. Correction **208** (1961), 70–72.

[17] K. Mahler, A correction to the paper "An interpolation series for continuous functions of a p-adic variable.", *J. Reine Angew. Math.* **208** (1961), 70–72.

[18] S. Margolis, M. Sapir and P. Weil, Closed subgroups in pro-**V** topologies and the extension problem for inverse automata, *Internat. J. Algebra Comput.* **11**,4 (2001), 405–445.

[19] J.-E. Pin, Topologies for the free monoid, *J. of Algebra* **137** (1991), 297–337.

[20] J.-E. Pin, On reversible automata, in *Proceedings of the first LATIN conference*, Saõ-Paulo, 1992, pp. 401–416, *Lect. Notes Comp. Sci.* n° 583, Springer.

[21] J.-E. Pin, Topologie p-adique sur les mots, *Journal de théorie des nombres de Bordeaux* **5** (1993), 263–281.

[22] J.-E. Pin, A variety theorem without complementation, *Russian Mathematics (Iz. VUZ)* **39** (1995), 80–90.

[23] J.-E. Pin, Polynomial closure of group languages and open sets of the Hall topology, *Theoret. Comput. Sci.* **169** (1996), 185–200.

[24] J.-E. Pin, Syntactic semigroups, in *Handbook of formal languages*, G. Rozenberg and A. Salomaa (ed.), vol. 1, ch. 10, pp. 679–746, Springer, 1997.

[25] J.-E. Pin, The expressive power of existential first order sentences of Büchi's sequential calculus, *Discrete Mathematics* **291** (2005), 155–174.

[26] J.-E. Pin and C. Reutenauer, A conjecture on the Hall topology for the free group, *Bull. London Math. Soc.* **23** (1991), 356–362.

[27] J.-E. Pin and P. V. Silva, A topological approach to transductions, *Theoret. Comput. Sci.* **340** (2005), 443–456.

[28] J.-E. Pin and P. V. Silva, A Mahler's theorem for functions from words to integers, in *25th International Symposium on Theoretical Aspects of Computer Science (STACS 2008)*, S. Albers and P. Weil (ed.), Dagstuhl, Germany, 2008, pp. 585–596, Internationales Begegnungs- Und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany.

[29] J.-E. Pin and P. Weil, Polynomial closure and unambiguous product, *Theory Comput. Systems* **30** (1997), 1–39.

[30] J.-E. Pin and P. Weil, Uniformities on free semigroups, *International Journal of Algebra and Computation* **9** (1999), 431–453.

[31] N. Pippenger, Regular languages and Stone duality, *Theory Comput. Syst.* **30**,2 (1997), 121–134.

[32] N. R. Reilly and S. Zhang, Decomposition of the lattice of pseudovarieties of finite semigroups induced by bands, *Algebra Universalis* **44**,3-4 (2000), 217–239.

[33] J. Reiterman, The Birkhoff theorem for finite algebras, *Algebra Universalis* **14**,1 (1982), 1–10.

[34] C. Reutenauer, Une topologie du monoïde libre, *Semigroup Forum* **18**,1 (1979), 33–49.

[35] C. Reutenauer, Sur mon article: "Une topologie du monoïde libre" [Semigroup Forum **18** (1979), no. 1, 33–49; MR 80j:20075], *Semigroup Forum* **22**,1 (1981), 93–95.

[36] L. Ribes and P. A. Zalesskii, On the profinite topology on a free group, *Bull. London Math. Soc.* **25**,1 (1993), 37–43.

[37] L. Ribes and P. A. Zalesskii, The pro-p topology of a free group and algorithmic problems in semigroups, *Internat. J. Algebra Comput.* **4**,3 (1994), 359–374.

[38] M. H. Stone, The representation of Boolean algebras, *Bull. Amer. Math. Soc.* **44**,12 (1938), 807–816.

[39] H. Straubing, On logical descriptions of regular languages, in *LATIN 2002*, Berlin, 2002, pp. 528–538, *Lect. Notes Comp. Sci.* n° 2286, Springer.

[40] P. Weil, Profinite methods in semigroup theory, *Int. J. Alg. Comput.* **12** (2002), 137–178.