Equations defining the polynomial closure of a lattice of regular languages

Mário J. J. Branco[*] Jean-Éric Pin[‡]

April 28, 2009

The *polynomial closure* Pol($\mathcal{L}$) of a class of languages $\mathcal{L}$ of A^* is the set of languages that are finite unions of marked products of the form $L_0a_1L_1 \cdots a_nL_n$, where the a_i are letters and the L_i are elements of $\mathcal{L}$.

The main result of this paper gives an equational description of Pol($\mathcal{L}$), given an equational description of $\mathcal{L}$, when $\mathcal{L}$ is a lattice of regular languages closed under quotients, or a *quotienting algebra of languages*, as we call it in the sequel. The term "equational description" refers to a recent paper [5], where it was shown that any lattice of regular languages can be defined by a set of profinite equations. More formally, our main result can be stated as follows:

> *If $\mathcal{L}$ is a quotienting algebra of languages, then Pol($\mathcal{L}$) is defined by the set of equations of the form $x^\omega y x^\omega \leqslant x^\omega$, where x, y are profinite words such that the equations $x = x^2$ and $y \leqslant x$ are satisfied by $\mathcal{L}$.*

As an application of this result, we establish a set of profinite equations defining the class of languages of the form $L_0a_1L_1 \cdots a_nL_n$, where each language L_i is either of the form u^* (where u is a word) or A^* (where A is the alphabet) and we prove that this class is decidable. Let us now give the motivations of our work and a brief survey of the previously known results.

Motivations. The polynomial closure occurs in several difficult problems on regular languages. For instance, a language has *star-height one* if and only if it belongs to the polynomial closure of the set of languages of the form F or F^*, where F is a finite language. Although this class is known to be decidable, it is still an open problem to find profinite equations for this class. Such a result could serve, in turn, to discover a language of generalized star-height > 1, a widely open problem.

The polynomial closure is also one of the two operations appearing in the definition of the *concatenation hierarchy* over a given set $\mathcal{L}$ of regular languages, defined by induction on n as follows. The level 0 is $\mathcal{L}$ and, for each $n \geqslant 0$, the level $2n + 1$ is the polynomial closure of the level $2n$ and the level $2n + 2$ is the Boolean closure of the level $2n + 1$. The simplest hierarchy is built on the initial set $\mathcal{L} = \{\emptyset, A^*\}$. A nice result of Thomas [15] shows that a regular

[*]CAUL and Dep. Matemática da Faculdade de Ciências, Universidade de Lisboa, Av. Prof. Gama Pinto, 2, 1649-003 Lisboa, Portugal

[†]LIAFA, Université Paris 7 and CNRS, Case 7014, 75205 Paris Cedex 13, France.

[‡]The authors acknowledge support from the AutoMathA programme of the European Science Foundation and the projects ISFL-1-143 and PTDC/MAT/69514/2006 of CAUL, financed by FCT and FEDER.

language is of level $2n+1$ in this hierarchy if and only if it is definable by a Σ_{n+1}-sentence of first order logic in the signature $\{<, (\mathbf{a})_{a \in A}\}$, where $\mathbf{a}$ is a predicate giving the positions of the letter a. Similar logical interpretations hold for other hierarchies, but unfortunately, only the very low levels of such hierarchies are known to be decidable and in general, this type of decidability problems is considered to be difficult. Our result certainly does not solve the problem in general, but it gives an algebraic approach that can be successful in some particular cases, like the one considered in Section 3, which does not follow from the results of [9].

Known results. A similar result was known when $\mathcal{L}$ is a variety of languages, that is, a class of regular languages closed under Boolean operations, quotients and inverse of morphisms, but depended on the conjunction of two theorems. The first theorem [11] relied on Eilenberg's theory of varieties, which gives a bijective correspondence between varieties of languages and varieties of finite monoids. It stated, in essence, that the polynomial closure corresponds, on the monoid level, to a certain Mal'cev product of varieties. The second result [10] gave identities for the Mal'cev product of two varieties of finite monoids. These results have been extended in [7] to positive varieties of languages and in [9] to quotienting algebras closed under inverse of length-preserving morphisms. However, all these proofs relied on the original proof of [11] and required the use of Mal'cev products and relational morphisms.

In summary, our new result is more general than all the previously known results. Further, our new proof combines various ideas from the above-mentioned papers, but avoids the use of Mal'cev products, a major difference with the original proof, although the experienced reader will still recognize their ghost in this paper. This could be a decisive advantage for potential extensions to other structures, like words over linear orders or finite trees.

1 Definitions and background

1.1 Languages, monoids and syntactic order

Let A be a finite alphabet. A *lattice of languages* is a set of regular languages of A^* containing the empty language, the full language A^* and closed under finite intersection and finite union. We denote by L^c the complement of a language L of A^*.

Let L be a language of A^* and let u be a word. The *left quotient* of L by u is the language $u^{-1}L = \{v \in A^* \mid uv \in L\}$. The *right quotient* Lu^{-1} is defined in a symmetrical way. A *quotienting algebra of languages* is a lattice of languages closed under the operations $L \mapsto u^{-1}L$ and $L \mapsto Lu^{-1}$, for any word u.

An *ordered monoid* is a monoid M equipped with a partial order $\leqslant$ compatible with the product on M: for all $x, y, z \in M$, if $x \leqslant y$ then $zx \leqslant zy$ and $xz \leqslant yz$. For each $x \in M$, we set $\downarrow x = \{y \in M \mid y \leqslant x\}$. A *morphism of ordered monoids* is an order-preserving monoid morphism.

The *syntactic congruence* of a language L of A^* is the equivalence relation on A^* defined by $u \sim_L v$ if and only if, for every $x, y \in A^*$,

$$xvy \in L \iff xuy \in L$$

The monoid $M = A^*/{\sim_L}$ is the *syntactic monoid* of L and the natural morphism $\eta : A^* \to M$ is called the *syntactic morphism* of L. It is a well-known fact that a language is regular if and only if its syntactic monoid is finite.

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 $x, y \in A^*$, $xvy \in L$ implies $xuy \in L$. The associated equivalence relation is the syntactic congruence $\sim_L$. Further, $\leqslant_L$ induces a partial order on the syntactic monoid M of L. This partial order $\leqslant$ is compatible with the product and can also be defined directly on M as follows: given $u, v \in M$, one has $u \leqslant v$ if and only if, for all $x, y \in M$, $xvy \in \eta(L)$ implies $xuy \in \eta(L)$. The ordered monoid $(M, \leqslant)$ is called the *syntactic ordered monoid* of L.

1.2 Factorization forests

We review in this section an important combinatorial result of I. Simon on finite semigroups. A *factorization forest* is a function F that associates with every word x of A^2A^* a factorization $F(x) = (x_1, \ldots, x_n)$ of x such that $n \geqslant 2$ and $x_1, \ldots, x_n \in A^+$. The integer n is the *degree* of the factorization $F(x)$. Given a factorization forest F, the *height function* of F is the function $h : A^* \to \mathbb{N}$ defined recursively by

$$h(x) = \begin{cases} 0 & \text{if } |x| \leqslant 1 \\ 1 + \max\{h(x_i) \mid 1 \leqslant i \leqslant n\} & \text{if } F(x) = (x_1, \ldots, x_n) \end{cases}$$

The *height* of F is the least upper bound of the heights of the words of A^*.

Let M be a finite monoid and let $\varphi : A^* \to M$ be a morphism. A factorization forest F is *Ramseyan modulo* φ if, for every word x of A^2A^*, $F(x)$ is either of degree 2 or there exists an idempotent e of M such that $F(x) = (x_1, \ldots, x_n)$ and $\varphi(x_1) = \varphi(x_2) = \cdots = \varphi(x_n) = e$ for $1 \leqslant i \leqslant n$. The factorization forest theorem was first proved by I. Simon in [12, 13, 14] and later improved in [2, 3, 4, 6]:

Theorem 1.1 *Let φ be a morphism from A^* into a finite monoid M. There exists a factorization forest of height $\leqslant 3|M| - 1$ which is Ramseyan modulo φ.*

1.3 Profinite monoids and equations

We briefly recall the definition of a free profinite monoid. More details can be found in [1, 8]. A finite monoid M *separates* two words u and v of A^* if there is a morphism $\varphi : A^* \to M$ such that $\varphi(u) \neq \varphi(v)$. We set

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

and $d(u, v) = 2^{-r(u,v)}$, with the usual conventions $\min \emptyset = +\infty$ and $2^{-\infty} = 0$. Then d is a *metric* on A^* and the completion of A^* for this metric is denoted by $\widehat{A^*}$. The product on A^* can be extended by continuity to $\widehat{A^*}$. This extended product makes $\widehat{A^*}$ a compact topological monoid, called the *free profinite monoid*. Its elements are called *profinite words*.

Every finite monoid M can be considered as a discrete metric space for the discrete metric d, defined by $d(x, y) = 0$ if $x = y$, and $d(x, y) = 1$ otherwise.

Now, every morphism φ from A^* into a finite monoid is uniformly continuous and therefore can be extended (in a unique way) into a uniformly continuous morphism $\hat{\varphi}$ from $\widehat{A^*}$ to M.

Since A^* embeds naturally in $\widehat{A^*}$, every finite word is a profinite word. We shall also use the operator $x \mapsto x^\omega$ in $\widehat{A^*}$, which is formally defined by the formula $x^\omega = \lim_{n\to\infty} x^{n!}$ and is justified by the fact that the sequence $(x^{n!})_{n\geqslant 0}$ is a Cauchy sequence in $\widehat{A^*}$ and hence has a limit in $\widehat{A^*}$. Let φ be a morphism from A^* onto a finite monoid M and let $s = \hat{\varphi}(x)$. Then the sequence $(s^{n!})_{n\geqslant 0}$ is ultimately equal to s^ω, where ω is the least integer k such that for all $t \in M$, t^k is idempotent. Consequently, we obtain the formula $\hat{\varphi}(x^\omega) = \hat{\varphi}(x)^\omega$, which gives ground to the notation x^ω.

Let L be a regular language of A^*, let $(M, \leqslant)$ be its syntactic ordered monoid and let $\eta : A^* \to M$ its syntactic morphism. Given two profinite words $u, v \in \widehat{A^*}$, we say that L *satisfies the (profinite) equation* $u \leqslant v$ (resp. $u = v$) if $\hat{\eta}(u) \leqslant \hat{\eta}(v)$ (resp. $\hat{\eta}(u) = \hat{\eta}(v)$). By extension, we say that a set of languages $\mathcal{L}$ *satisfies a set of equations* Σ if every language of $\mathcal{L}$ satisfies every equation of Σ.

2 Polynomial closure of lattices of languages

Let $\mathcal{L}$ be a set of languages of A^*. An *$\mathcal{L}$-monomial of degree n* is a language of the form $L_0a_1L_1 \cdots a_nL_n$, where each a_i is a letter of A and each L_i is a language of $\mathcal{L}$. An *$\mathcal{L}$-polynomial* is a finite union of $\mathcal{L}$-monomials. Its *degree* is the maximum of the degrees of its monomials. The *polynomial closure* of $\mathcal{L}$, denoted by Pol($\mathcal{L}$), is the set of all $\mathcal{L}$-polynomials.

Our main result gives an equational description of Pol($\mathcal{L}$), given an equational description of $\mathcal{L}$, when $\mathcal{L}$ is a quotienting algebra of languages. To state this result in a concise way, let us introduce a convenient notation. Given a set $\mathcal{R}$ of regular languages, denote by $\Sigma(\mathcal{R})$ the set of equations of the form $x^\omega y x^\omega \leqslant x^\omega$, where x, y are profinite words of $\widehat{A^*}$ such that the equations $x = x^2$ and $y \leqslant x$ are satisfied by $\mathcal{R}$. Note that the function mapping $\mathcal{R}$ to the class of languages satisfying $\Sigma(\mathcal{R})$ is monotonic for the inclusion. We can now state our main result:

Theorem 2.1 *If $\mathcal{L}$ is a quotienting algebra of languages, then Pol($\mathcal{L}$) is defined by the set of equations $\Sigma(\mathcal{L})$.*

The proof is divided into several parts. We first prove in Proposition 2.2 that Pol($\mathcal{L}$) satisfies the equations of $\Sigma(\mathcal{L})$. To establish the converse of this property, we consider a language K satisfying all the equations of $\Sigma(\mathcal{L})$. We convert this property into a topological property (Proposition 2.4) and then use a compactness argument to show that K satisfies the equations of $\Sigma(\mathcal{F})$, where $\mathcal{F}$ is a finite sublattice of $\mathcal{L}$ (Proposition 2.5). The final part of the proof consists in proving that K belongs to Pol($\mathcal{F}$). This is where the factorization forest theorem arises, but a series of lemmas (Lemmas 2.6 to 2.11) are still necessary to find explicitly a polynomial expression for K.

Proposition 2.2 *If $\mathcal{L}$ is a lattice of languages, then Pol($\mathcal{L}$) satisfies the equations of $\Sigma(\mathcal{L})$.*

Proof. Since, by [5, Theorem 7.2] the set of languages satisfying $\Sigma(\mathcal{L})$ is a lattice of languages, it suffices to prove the result for any $\mathcal{L}$-monomial. Let $L = L_0a_1L_1 \cdots a_nL_n$ be an $\mathcal{L}$-monomial and let $\eta\colon A^* \to M$ be its syntactic morphism. Let, for $0 \leqslant i \leqslant n$, $\eta_i\colon A^* \to M_i$ be the syntactic morphism of L_i. Let x and y be two profinite words such that each L_i satisfies the two equations $x = x^2$ and $y \leqslant x$.

Since A^* is dense in $\widehat{A^*}$, one can find a word $x' \in A^*$ such that $r(x', x) > \max\{|M_0|, \ldots, |M_n|, |M|\}$. It follows that $\eta(x') = \hat{\eta}(x)$ and, for $0 \leqslant i \leqslant n$, $\eta_i(x') = \hat{\eta}_i(x)$. Similarly, one can associate with y a word $y' \in A^*$ such that $\eta(y') = \hat{\eta}(y)$ and, for $0 \leqslant i \leqslant n$, $\eta_i(y') = \hat{\eta}_i(y)$. It follows that each L_i satisfies the equations $x' = {x'}^2$ and $y' \leqslant x'$ and that L satisfies the equation $x^\omega y x^\omega \leqslant x^\omega$ if and only if it satisfies the equations ${x'}^\omega y' {x'}^\omega \leqslant {x'}^\omega$. In other terms, it suffices to prove the result when x and y are words.

We need to establish the relation $(*) : \hat{\eta}(x^\omega y x^\omega) \leqslant \hat{\eta}(x^\omega)$. Let k be an integer such that $k > n$ and $\hat{\eta}(x^\omega) = \eta(x^k)$. Since $\hat{\eta}(x^\omega y x^\omega) = \eta(x^k y x^k)$, proving $(*)$ amounts to showing that $x^k y x^k \leqslant_L x^k$. Let $u, v \in A^*$ and suppose that $ux^kv \in L$. Thus $ux^kv = u_0a_1u_1 \cdots a_nu_n$, where, for $0 \leqslant i \leqslant n$, $u_i \in L_i$. Since $k > n$, one can find $h \in \{0, \ldots, n\}$, $j \in \{1, \ldots, k\}$ and $u'_h, u''_h \in A^*$ such that $u_h = u'_h x u''_h$, $ux^{j-1} = u_0a_1u_1 \cdots a_hu'_h$ and $x^{k-j}v = u''_h a_{h+1}u_{h+1} \cdots a_nu_n$. Since $u_h \in L_h$ and L_h satisfies the equations $x = x^2$ and $y \leqslant x$, one has $u'_h x^{k-j+1} y x^j u''_h \in L_h$, and since $ux^kyx^kv = u_0a_1u_1 \cdots a_h(u'_h x^{k-j+1} y x^j u''_h)a_{h+1}u_{h+1} \cdots a_nu_n$, one gets $ux^kyx^kv \in L$. Thus $x^kyx^k \leqslant_L x^k$, which completes the proof. □

The rest of this section is devoted to showing the converse implication in Theorem 2.1. Let us introduce, for each regular language L of A^*, the sets

$$E_L = \Big\{(x, y) \in \widehat{A^*} \times \widehat{A^*} \mid L \text{ satisfies } x = x^2 \text{ and } y \leqslant x\Big\}$$
$$F_L = \Big\{(x, y) \in \widehat{A^*} \times \widehat{A^*} \mid L \text{ satisfies } x^\omega y x^\omega \leqslant x^\omega\Big\}.$$

Recall that a subset of a topological space is *clopen* if it is both open and closed.

Lemma 2.3 *For each regular language L of A^*, the sets E_L and F_L are clopen in $\widehat{A^*} \times \widehat{A^*}$.*

Proof. Let $\eta\colon A^* \to M$ be the syntactic morphism of L. The formula $\alpha(x, y) = \big(\hat{\eta}(x), \hat{\eta}(x^2), \hat{\eta}(y)\big)$ defines a continuous map α from $\widehat{A^*} \times \widehat{A^*}$ into M^3, considered as a discrete space. Setting $\Delta = \{(s, t, u) \in M^3 \mid s = t \text{ and } u \leqslant s\}$, we get

$$E_L = \Big\{(x, y) \in \widehat{A^*} \times \widehat{A^*} \mid \hat{\eta}(x) = \hat{\eta}(x^2) \text{ and } \hat{\eta}(y) \leqslant \hat{\eta}(x)\Big\} = \alpha^{-1}(\Delta)$$

Now, since M is a discrete topological space, Δ is clopen in M^3 and thus E_L is a clopen subset of $\widehat{A^*} \times \widehat{A^*}$.

A similar argument, using the continuous map $\beta : \widehat{A^*} \times \widehat{A^*} \to M^2$ defined by $\beta(x, y) = \big(\hat{\eta}(x^\omega y x^\omega), \hat{\eta}(x^\omega)\big)$, would show that F_L is clopen. □

We now convert our equational conditions into a topological property. Recall that a *cover* [*open cover*] of a topological space X is a collection of subsets [open subsets] of X whose union is X.

Proposition 2.4 *Let $\mathcal{F}$ be a set of regular languages of A^* and let K be a regular language of A^*. The following conditions are equivalent:*

(1) *K satisfies the profinite equations of $\Sigma(\mathcal{F})$,*

(2) *the set $\{F_K\} \cup \{E_L^c \mid L \in \mathcal{F}\}$ is an open cover of $\widehat{A^*} \times \widehat{A^*}$.*

Proof. Indeed $\mathcal{F}$ satisfies the two profinite equations $x = x^2$ and $y \leqslant x$ if and only if $(x, y) \in \bigcap_{L\in\mathcal{F}} E_L$ or, equivalently, $(x, y) \notin \bigcup_{L\in\mathcal{F}} E_L^c$. Similarly, K satisfies the equation $x^\omega y x^\omega \leqslant x^\omega$ if and only if $(x, y) \in F_K$. Now, condition (1) is equivalent to saying that $(x, y) \notin \bigcup_{L\in\mathcal{F}} E_L^c$ implies $(x, y) \in F_K$, which is another way to say that $\{F_K\} \cup \{E_L^c \mid L \in \mathcal{F}\}$ is a cover of $\widehat{A^*} \times \widehat{A^*}$. Further, Proposition 2.3 shows that it is an open cover. □

Proposition 2.5 *If K satisfies the equations of $\Sigma(\mathcal{L})$, there is a finite subset $\mathcal{F}$ of $\mathcal{L}$ such that K satisfies the equations of $\Sigma(\mathcal{F})$.*

Proof. Proposition 2.4 shows that $\{F_K\} \cup \{E_L^c \mid L \in \mathcal{L}\}$ is a cover of $\widehat{A^*} \times \widehat{A^*}$. Since $\widehat{A^*}$ is compact, one can extract from this cover a finite cover, say $\{F_K\} \cup \{E_L^c \mid L \in \mathcal{F}\}$. By Proposition 2.4 again, K satisfies the profinite equations of the form $x^\omega y x^\omega \leqslant x^\omega$ such that all the languages of $\mathcal{F}$ satisfy the equations $x = x^2$ and $y \leqslant x$. □

Let K be a regular language satisfying all the equations of $\Sigma(\mathcal{L})$ and let $\eta : A^* \to M$ be its syntactic morphism. Let also $\mathcal{F} = \{L_1, \ldots, L_n\}$ be a finite subset of $\mathcal{L}$ as given by Proposition 2.5. For $1 \leqslant i \leqslant n$, let $\eta_i\colon A^* \to M_i$ be the syntactic morphism of L_i. Let $\mu\colon A^* \to M_1 \times \cdots \times M_n$ be the morphism defined by $\mu(u) = (\eta_1(u), \ldots, \eta_n(u))$. Finally, let $V = \mu(A^*)$ and, for $1 \leqslant i \leqslant n$, let $\pi_i : V \to M_i$ be the natural projection. We set $S = \{(\eta(u), \mu(u)) \mid u \in A^*\}$. Then S is a submonoid of $M \times V$ and the two morphisms $\alpha : S \to M$ and $\beta : S \to V$ defined by $\alpha(m, v) = m$ and $\beta(m, v) = v$ are surjective. Further, the morphism $\delta : A^* \to S$ defined by $\delta(u) = (\eta(u), \mu(u))$ satisfies $\eta = \alpha \circ \delta$ and $\mu = \beta \circ \delta$. The situation is summarized in the following diagram:

$$\begin{tikzcd}
 & A^* \arrow[dl, "\eta"'] \arrow[d, "\delta"] \arrow[dr, "\mu"] \arrow[drr, "\eta_i"] & & \\
M & S \arrow[l, "\alpha"'] \arrow[r, "\beta"] & V \arrow[r, "\pi_i"] & M_i
\end{tikzcd}$$

We now arrive at the last step of the proof of Theorem 2.1, which consists in proving that K belongs to $\mathrm{Pol}(\mathcal{F})$.

We start with three auxiliary lemmas. The first one states that every downward closed language recognized by μ belongs to $\mathcal{L}$ and relies on the fact that $\mathcal{L}$ is a quotienting algebra of languages. The second one gives a key property of S and this is the only place in the proof where we use the equations of $\Sigma(\mathcal{L})$. The third one is an elementary, but useful, observation.

Lemma 2.6 *Let $t \in V$. Then the language $\mu^{-1}(\downarrow t)$ belongs to $\mathcal{L}$.*

Proof. Let $t = (t_1, \ldots, t_n)$ and let z be a word such that $\mu(z) = t$. Then $t_i = \eta_i(z)$ and $\mu^{-1}(\downarrow t) = \bigcap_{1 \leqslant i \leqslant n} \eta_i^{-1}(\downarrow t_i)$. Moreover, one gets for each $i \in \{1, \ldots, n\}$,

$$\eta_i^{-1}(\downarrow t_i) = \{x \in A^* \mid \eta_i(x) \leqslant \eta_i(z)\} = \{x \in A^* \mid x \leqslant_{L_i} z\} = \bigcap_{(u,v) \in E_i} u^{-1} L_i v^{-1}$$

where $E_i = \{(u, v) \in A^* \times A^* \mid uzv \in L_i\}$. Since L_i is regular, there are only finitely many quotients of the form $u^{-1}L_iv^{-1}$ and hence the intersection is finite. The result follows, since $\mathcal{L}$ is a quotienting algebra of languages. □

Lemma 2.7 *For every idempotent* $(e, f) \in S$ *and for every* $(s, t) \in S$ *such that* $t \leqslant f$*, one has* $ese \leqslant e$*.*

Proof. Let x and y be two words such that $\delta(x) = (e, f)$ and $\delta(y) = (s, t)$. Then $\eta(x) = e$, $\mu(x) = f$, $\eta(y) = s$ and $\mu(y) = t$ and since f is idempotent and $t \leqslant f$, $\mathcal{F}$ satisfies the equations $x = x^2$ and $y \leqslant x$. Therefore K satisfies the equation $x^\omega y x^\omega \leqslant x^\omega$. It follows that $\hat{\eta}(x^\omega y x^\omega) \leqslant \hat{\eta}(x^\omega)$, that is $ese \leqslant e$. □

Before we continue, let us point out a subtlety in the proof of Lemma 2.7. It looks like we have used words instead of profinite words in this proof and the reader may wonder whether one could change "profinite" to "finite" in the statement of our main result. The answer is negative for the following reason: if $\mathcal{F}$ satisfies the equations $x = x^2$ and $y \leqslant x$, it does not necessarily imply that $\mathcal{L}$ satisfies the same equations. In fact, the choice of $\mathcal{F}$ comes from the extraction of the finite cover and hence is bound to K.

We now set, for each idempotent f of V, $L(f) = \mu^{-1}(\downarrow f)$.

Lemma 2.8 *For each idempotent* f *of* V*, one has* $L(1)L(f)L(1) = L(f)$*.*

Proof. Since $1 \in L(1)$, one gets the inclusion $L(f) = 1L(f)1 \subseteq L(1)L(f)L(1)$. Let now $s, t \in L(1)$ and $x \in L(f)$. Then by definition, $\mu(s) \leqslant 1$, $\mu(x) \leqslant f$ and $\mu(t) \leqslant 1$. It follows that $\mu(sxt) = \mu(s)\mu(x)\mu(t) \leqslant 1f1 = f$, whence $sxt \in L(f)$. This gives the opposite inclusion $L(1)L(f)L(1) \subseteq L(f)$. □

We now come to the combinatorial argument of the proof. By Theorem 1.1, there exists a factorization forest F of height $\leqslant 3|S| - 1$ which is Ramseyan modulo δ. We use this fact to associate with each word x a certain language $R(x)$, defined recursively as follows:

$$R(x) = \begin{cases} L(1)xL(1) & \text{if } |x| \leqslant 1 \\ R(x_1)R(x_2) & \text{if } F(x) = (x_1, x_2) \\ R(x_1)L(f)R(x_k) & \text{if } \; F(x) = (x_1, \ldots, x_k), \text{ with } k \geqslant 3 \text{ and} \\ & \qquad \delta(x_1) = \cdots = \delta(x_k) = (e, f) \end{cases}$$

In particular $R(1) = L(1)$, since $L(1)$ is a submonoid of A^*.

Denote by $\mathcal{E}$ the finite set of languages of the form $L(f)$, where f is an idempotent of V. We know by Lemma 2.6 that $\mathcal{E}$ is a subset of $\mathcal{L}$. Let us say that an $\mathcal{E}$-monomial is in *normal form* if it is of the form $L(1)a_0L(f_1)a_1 \cdots L(f_k)a_kL(1)$ where $f_1, \ldots, f_k$ are idempotents of V.

Lemma 2.9 *For each $x \in A^*$, $R(x)$ is equal to an $\mathcal{E}$-monomial in normal form of degree $\leqslant 2^{h(x)}$.*

Proof. We prove the result by induction on the length of x. The result is true if $|x| \leqslant 1$. Suppose that $|x| \geqslant 2$. If $F(x) = (x_1, x_2)$, then $R(x) = R(x_1)R(x_2)$ otherwise $R(x) = R(x_1)L(f)R(x_k)$. We treat only the latter case, since the first one is similar. By the induction hypothesis, $R(x_1)$ and $R(x_k)$ are equal to $\mathcal{E}$-monomials in normal form. It follows by Lemma 2.8 that $R(x)$ is equal to an $\mathcal{E}$-monomial in normal form, whose degree is lesser than or equal to the sum of the degrees of $R(x_1)$ and $R(x_k)$. The result now follows from the induction hypothesis, since $2^{h(x_1)} + 2^{h(x_k)} \leqslant 2^{1+\max\{h(x_1),\ldots,h(x_k)\}} \leqslant 2^{h(x)}$. $\square$

Lemma 2.10 *For each $x \in A^*$, one has $x \in R(x)$.*

Proof. We prove the result by induction on the length of x. The result is trivial if $|x| \leqslant 1$. Suppose that $|x| \geqslant 2$. If $F(x) = (x_1, x_2)$, one has $x_1 \in R(x_1)$ and $x_2 \in R(x_2)$ by the induction hypothesis and hence $x \in R(x)$ since $R(x) = R(x_1)R(x_2)$. Suppose now that $F(x) = (x_1, \ldots, x_k)$ with $k \geqslant 3$ and $\delta(x_1) = \cdots = \delta(x_k) = (e, f)$. Then $R(x) = R(x_1)L(f)R(x_k)$. Since $x_1 \in R(x_1)$ and $x_k \in R(x_k)$ by the induction hypothesis and $\mu(x_2 \cdots x_{k-1}) = f$, one gets $x_2 \cdots x_{k-1} \in L(f)$ and finally $x \in R(x_1)L(f)R(x_k)$, that is, $x \in R(x)$. $\square$

If R is a language, let us write $\eta(R) \leqslant \eta(x)$ if, for each $u \in R$, $\eta(u) \leqslant \eta(x)$.

Lemma 2.11 *For each $x \in A^*$, one has $\eta(R(x)) \leqslant \eta(x)$.*

Proof. We prove the result by induction on the length of x. First, applying Lemma 2.7 with $e = f = 1$ shows that if $(s, t) \in S$ and $t \leqslant 1$, then $s \leqslant 1$. It follows that $\eta(R(1)) = \eta(L(1)) = \eta(\mu^{-1}(\downarrow 1)) \leqslant 1$.

If $|x| \leqslant 1$, one gets $R(x) = L(1)xL(1)$ and $\eta(R(x)) = \eta(L(1))\eta(x)\eta(L(1)) \leqslant \eta(x)$ since $\eta(L(1)) \leqslant 1$. Suppose now that $|x| \geqslant 2$. If $F(x) = (x_1, x_2)$, then $R(x) = R(x_1)R(x_2)$ and by the induction hypothesis, $\eta(R(x_1)) \leqslant \eta(x_1)$ and $\eta(R(x_2)) \leqslant \eta(x_2)$. Therefore, $\eta(R(x)) = \eta(R(x_1))\eta(R(x_2)) \leqslant \eta(x_1)\eta(x_2) = \eta(x)$. Finally, suppose that $F(x) = (x_1, \ldots, x_k)$ with $k \geqslant 3$ and $\delta(x_1) = \cdots = \delta(x_k) = (e, f)$. Then $R(x) = R(x_1)L(f)R(x_k)$. By the induction hypothesis, $\eta(R(x_1)) \leqslant e$ and $\eta(R(x_k)) \leqslant e$. Now, if $u \in L(f)$, one gets $\mu(u) \leqslant f$. Since $(\eta(u), \mu(u)) \in S$, it follows from Lemma 2.7 that the relation $e\eta(u)e \leqslant e$ holds in M. Finally, we get $\eta(R(x)) = \eta(R(x_1))\eta(L(f))\eta(R(x_k)) \leqslant e\eta(L(f))e \leqslant e = \eta(x)$. $\square$

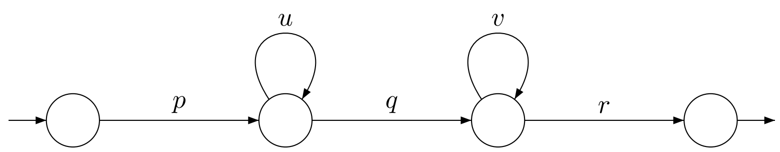
We can now conclude the proof of Theorem 2.1. We claim that $K = \bigcup_{x \in K} R(x)$. The inclusion $K \subseteq \bigcup_{x \in K} R(x)$ is an immediate consequence of Lemma 2.10. To prove the opposite inclusion, consider a word $u \in R(x)$ for some $x \in K$. It follows from Lemma 2.11 that $\eta(u) \leqslant \eta(x)$. Since $\eta(x) \in \eta(K)$, one gets $\eta(u) \in \eta(K)$ and finally $u \in K$. Now, by Lemma 2.9, each language $R(x)$ is an $\mathcal{E}$-monomial of degree $\leqslant 2^{h(x)}$. Since $h(x) \leqslant 3|S| - 1$ for all x, and since $\mathcal{E}$ is finite, there are only finitely many such monomials. Therefore K is equal to an $\mathcal{E}$-polynomial. Finally, Lemma 2.6 shows that each $\mathcal{E}$-polynomial belongs to $\mathrm{Pol}(\mathcal{L})$, and thus $K \in \mathrm{Pol}(\mathcal{L})$. $\square$

3 A case study

The *density* of a language $L \subseteq A^*$ is the function which counts the number of words of length n in L. More formally, it is the function $d_L : \mathbb{N} \to \mathbb{N}$ defined by $d_L(n) = |L \cap A^n|$. See [16] for a general reference. If $d_L(n) = O(1)$, then L is called a *slender language*. A regular language of A^* is *slender* if and only if it is a finite union of languages of the form $u_0 v^* u_1$, where $u_0, v, u_1 \in A^*$ (see [16, Theorem 3.6]). A language is *sparse* if it is of polynomial density. One can show that a regular language is sparse if and only 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, $\ldots$, v_n, u_n are words.

We shall also use the following characterization of regular nonslender languages, in which $i(u)$ denotes the first letter (or *initial*) of a word u.

Proposition 3.1 *A regular language L is nonslender if and only if there exist words $p, q, r \in A^*$ and $u, v \in A^+$ such that $i(u) \neq i(qv)$ and $pu^*qv^*r \subseteq L$.*



If $|A| \leqslant 1$, every regular language is slender, but if $|A| \geqslant 2$, the full language A^* is not slender and thus regular slender languages do not form a lattice of languages. However, the regular languages that are either slender or full form a quotienting algebra of languages, denoted by $\mathcal{S}$ in the sequel. Two sets of profinite equations for $\mathcal{S}$ were given in [5]. We shall just mention the second one, which requires a convenient writing convention. Let L be a regular language of A^* and let $\eta : A^* \to M$ be its syntactic morphism. If x is a profinite word of $\widehat{A^*}$, we say that L satisfies the equation $x \leqslant 0$ [$x = 0$], if the monoid M has a zero, denoted by 0, and if $\hat{\eta}(x) \leqslant 0$ [$\hat{\eta}(x) = 0$].

Proposition 3.2 *Suppose that $|A| \geqslant 2$. A regular language of A^* is slender or full if and only if it satisfies the equations $x \leqslant 0$ for all $x \in A^*$ and the equations $x^\omega u y^\omega = 0$ for each $x, y \in A^+$, $u \in A^*$ such that $i(x) \neq i(uy)$.*

We are interested in the polynomial closure of $\mathcal{S}$. The languages of Pol($\mathcal{S}$) are finite unions of languages of the form $L_0 a_1 L_1 \cdots a_n L_n$, where the a_i are letters and the L_i are languages of the form A^* or u^* for some word u.[1] In particular, Pol($\mathcal{S}$) contains all regular sparse languages but it also contains the nonsparse language A^* if $|A| \geqslant 2$.

The main result of this section is an equational description of Pol($\mathcal{S}$). Let us denote by $\Sigma'(\mathcal{S})$ the set of equations of the form

$$(x^\omega y^\omega)^\omega z (x^\omega y^\omega)^\omega \leqslant (x^\omega y^\omega)^\omega$$

where $z \in A^*$ and $x, y \in A^+$ and $i(x) \neq i(y)$.

Theorem 3.3 *A regular language of A^* belongs to Pol($\mathcal{S}$) if and only if it satisfies the equations of $\Sigma'(\mathcal{S})$.*

[1] To see this, it suffices to replace each word $u_i = a_1 \cdots a_k$ by $1^* a_1 1^* a_2 1^* \cdots 1^* a_k 1^*$ in each monomial of the form $u_0 v_1^* u_1 \cdots v_n^* u_n$.

Proof. Let us first settle a trivial case. If $|A| \leqslant 1$, every regular language belongs to $\mathrm{Pol}(\mathcal{S})$, but on the other hand, the set $\Sigma'(\mathcal{S})$ is empty because the condition $i(x) \neq i(y)$ is never satisfied! We suppose now that $|A| \geqslant 2$.

We show that every language of $\mathrm{Pol}(\mathcal{S})$ satisfies the equations of $\Sigma'(\mathcal{S})$ by applying Theorem 2.1. It suffices to verify that, if $i(x) \neq i(y)$, $\mathcal{S}$ satisfies the equations $(x^\omega y^\omega)^\omega = \big((x^\omega y^\omega)^\omega\big)^2$ and $z \leqslant (x^\omega y^\omega)^\omega$. But this is trivial, since we know by Proposition 3.2 that $\mathcal{S}$ satisfies the equations $x^\omega y^\omega = 0$ (take $u = 1$ in the equation $x^\omega u y^\omega = 0$) and $z \leqslant 0$.

Let K be a regular language satisfying the equations of $\Sigma'(\mathcal{S})$ and let $\eta : A^* \to M$ be its syntactic morphism. We immediately derive from $\Sigma'(\mathcal{S})$ a more comprehensible property, which is the counterpart of Lemma 2.7 in the proof of Theorem 2.1.

Lemma 3.4 *Let e be an idempotent of M. Then either $\eta^{-1}(e)$ is slender, or for all $s \in M$, $ese \leqslant e$.*

Proof. Let $L = \eta^{-1}(e)$. Since L is not slender, Proposition 3.1 tells us that one can find words $p, q, r \in A^*$ and $u, v \in A^+$ such that $i(u) \neq i(qv)$ and $pu^*qv^*r \subseteq L$. Further, since e is idempotent, L is a semigroup and we have in fact $(pu^*qv^*r)^+ \subseteq L$. It follows that

$$p(u^*(qvrp)^*)^*qr \subseteq p(u^*(qv^*rp)^*)^*u^*qv^*r \subseteq (pu^*qv^*r)^+ \subseteq L$$

Setting $x = u$, $y = qvrp$ and $t = qr$, we get $i(x) \neq i(y)$ and $p(x^*y^*)^*t \subseteq L$. It follows in particular that

$$\hat\eta\big(p(x^\omega y^\omega)^\omega t\big) = e \qquad (1)$$

Let $s \in M$ and let w be a word such that $\eta(w) = s$. Since L satisfies the equations of $\Sigma'(\mathcal{S})$, it satisfies in particular the equation $(x^\omega y^\omega)^\omega twp(x^\omega y^\omega)^\omega \leqslant (x^\omega y^\omega)^\omega$ and hence also $p(x^\omega y^\omega)^\omega twp(x^\omega y^\omega)^\omega t \leqslant p(x^\omega y^\omega)^\omega t$. This means that

$$\hat\eta\big(p(x^\omega y^\omega)^\omega twp(x^\omega y^\omega)^\omega t\big) \leqslant \hat\eta\big(p(x^\omega y^\omega)^\omega t\big) \qquad (2)$$

Now, using (1), (2) and the relation $\eta(w) = s$, we get $ese \leqslant e$. □

The end of the proof is similar to that of Theorem 2.1, with the major difference that we do not use the morphisms δ and μ anymore. By Theorem 1.1, there exists a factorization forest F of height $\leqslant 3|M| - 1$ which is Ramseyan modulo η. We associate with each idempotent $e \in M$ the language $L(e)$ equal to $\eta^{-1}(e)$ if this language is slender, and to A^* otherwise. Let us denote by $\mathcal{E}$ the set of languages of the form $L(e)$. By definition, every language of $\mathcal{E}$ is slender or full. We also associate with each word x a language $R(x)$, defined as follows:

$$R(x) = \begin{cases} L(1)xL(1) & \text{if } |x| \leqslant 1 \\ R(x_1)R(x_2) & \text{if } F(x) = (x_1, x_2) \\ R(x_1)L(e)R(x_k) & \text{if } \ F(x) = (x_1, \ldots, x_k), \text{ with } k \geqslant 3 \text{ and} \\ & \qquad \eta(x_1) = \cdots = \eta(x_k) = e \end{cases}$$

The proof now consists in adapting Lemmas 2.8, 2.9, 2.10 and 2.11 to our new definitions. We just give here a sketch of these proofs (detailed proofs can be found in the Appendix). For Lemma 2.8, one needs to prove that, for each idempotent $e \in M$, $L(1)L(e)L(1) = L(e)$. The key observation is that if $\eta^{-1}(e)$ is slender, then $\eta^{-1}(1)$ is slender: indeed $\eta^{-1}(1)\eta^{-1}(e) \subseteq \eta^{-1}(e)$ and if the density of $\eta^{-1}(1)$ is not bounded, the density of $\eta^{-1}(e)$ cannot be bounded. Therefore, if $\eta^{-1}(e)$ is slender, one can follow the original proof. If $\eta^{-1}(e)$ is not slender, then $L(e) = A^*$ and the result is trivial, since $1 \in L(1)$.

The proofs of Lemmas 2.9 and 2.10 are unchanged. The proof of Lemma 2.11 requires a slight modification in the case where $F(x) = (x_1, \ldots, x_k)$ with $k \geqslant 3$, $\eta(x_1) = \cdots = \eta(x_k) = e$ and $\eta^{-1}(e)$ nonslender. Then $R(x) = R(x_1)A^*R(x_k)$ and by the induction hypothesis $\eta(R(x_1)) \leqslant e$ and $\eta(R(x_k)) \leqslant e$. Further, Lemma 3.4 shows that, for all $s \in M$, $ese \leqslant e$. Therefore, for each $s_1 \in \eta(R(x_1))$, $s_k \in \eta(R(x_k))$ and $s \in M$, one gets $s_1 s s_k \leqslant ese \leqslant e$. It follows that $\eta(R(x)) \leqslant e$, which completes the proof, since $\eta(x) = e$.

The rest of the proof is unchanged and shows that K is equal to an $\mathcal{E}$-polynomial. Since each $\mathcal{E}$-monomial is itself in $\mathrm{Pol}(\mathcal{S})$, it follows that $K \in \mathrm{Pol}(\mathcal{S})$. $\square$

Corollary 3.5 *There is an algorithm to decide whether a given regular language belongs to Pol($\mathcal{S}$).*

Proof. Let L be a regular language and let $\eta : A^* \to M$ be its syntactic morphism. By Theorem 3.3, L belongs to $\mathrm{Pol}(\mathcal{S})$ if and only if it satisfies the equations of $\Sigma'(\mathcal{S})$. Setting

$$F = \bigcup_{\substack{a,b \in A \\ a \neq b}} \eta(a)M \times \eta(b)M$$

it suffices to verify that the property $(x^\omega y^\omega)^\omega z (x^\omega y^\omega)^\omega \leqslant (x^\omega y^\omega)^\omega$ holds for all $(x, y) \in F$ and all $z \in M$. Since M and F are finite, this property is decidable. $\square$

References

[1] 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.

[2] J. Chalopin and H. Leung, On factorization forests of finite height, *Theoret. Comput. Sci.* **310**,1-3 (2004), 489–499.

[3] T. Colcombet, A combinatorial theorem for trees: applications to monadic logic and infinite structures, in *Automata, languages and programming*, pp. 901–912, *Lect. Notes Comp. Sci.* vol. 4596, Springer, Berlin, 2007.

[4] T. Colcombet, Factorisation Forests for Infinite Words, in *Fundamentals of Computation Theory, 16th International Symposium, FCT 2007, Budapest, Hungary, August 27-30, 2007, Proceedings*, E. Csuhaj-Varjú and Z. Ésik (ed.), pp. 226–237, *Lect. Notes Comp. Sci.* vol. 4639, Springer, Berlin, 2007.

[5] 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.), pp. 246–257, *Lect. Notes Comp. Sci.* vol. 5126, Springer, Berlin, 2008.

[6] M. Kufleitner, The Height of Factorization Forests, in *Mathematical Foundations of Computer Science 2008, 33rd International Symposium, MFCS 2008, Torun, Poland, August 25-29, 2008, Proceedings*, E. Ochmanski and J. Tyszkiewicz (ed.), pp. 443–454, *Lect. Notes Comp. Sci.* vol. 5162, Springer, Berlin, 2008.

[7] J.-E. Pin, Algebraic tools for the concatenation product, *Theoret. Comput. Sci.* **292** (2003), 317–342.

[8] J.-E. Pin, Profinite methods in automata theory, in *26th International Symposium on Theoretical Aspects of Computer Science (STACS 2009)*, S. Albers (ed.), pp. 31–50, Internationales Begegnungs- Und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany, Dagstuhl, Germany, 2009.

[9] J.-E. Pin and H. Straubing, Some results on $\mathcal{C}$-varieties, *Theoret. Informatics Appl.* **39** (2005), 239–262.

[10] J.-E. Pin and P. Weil, Profinite semigroups, Mal'cev products and identities, *J. of Algebra* **182** (1996), 604–626.

[11] J.-E. Pin and P. Weil, Polynomial closure and unambiguous product, *Theory Comput. Systems* **30** (1997), 383–422.

[12] I. Simon, Properties of factorization forests, in *Formal properties of finite automata and applications, Ramatuelle, France, May 23-27, 1988, Proceedings*, pp. 65–72, *Lect. Notes Comp. Sci.* vol. 386, Springer, Berlin, 1989.

[13] I. Simon, Factorization forests of finite height, *Theoret. Comput. Sci.* **72**,1 (1990), 65–94.

[14] I. Simon, A short proof of the factorization forest theorem, in *Tree automata and languages (Le Touquet, 1990)*, pp. 433–438, *Stud. Comput. Sci. Artificial Intelligence* vol. 10, North-Holland, Amsterdam, 1992.

[15] W. Thomas, Classifying regular events in symbolic logic, *J. Comput. System Sci.* **25**,3 (1982), 360–376.

[16] S. Yu, Regular languages, in *Handbook of formal languages*, G. Rozenberg and A. Salomaa (ed.), vol. 1, ch. 2, pp. 45–110, Springer, 1997.