DIMACS Series in Discrete Mathematics
and Theoretical Computer Science
Volume 64, 2004

Formal Languages and Number Theory

Jeffrey Shallit

ABSTRACT. I illustrate the connection between formal language theory and number theory. I give three examples (dealing with unary regular languages, unary context-free languages, and primitive words) where number theory may be applied to solve problems in formal language theory. I also give one example (transcendence in finite characteristic) where formal language theory may be applied to solve a problem in number theory.

1. Introduction.

An *alphabet* is a (usually) finite set of symbols Σ. A *word* or *string* — the terms are synonymous — is a (usually) finite sequence of symbols chosen from Σ. By Σ^* we mean the free monoid over Σ (the set of all finite strings with symbols chosen from Σ). A *language* is a (finite or infinite) subset of Σ^*. For example, if x^R denotes the reversal of the string x, then the language PAL of palindromes over $\{a, b\}$ can be defined as follows:

$$\begin{aligned} \texttt{PAL} &= \{x \in \{a,b\}^* \ : \ x = x^R\} \\ &= \{\epsilon, a, b, aa, bb, aaa, aba, bab, bbb, \ldots\}. \end{aligned}$$

Note that ϵ denotes the empty string.

The basic operations on languages include union, intersection, complementation, concatenation (defined by $L_1 L_2 := \{xy \ : x \in L_1, \ y \in L_2\}$) and Kleene closure (defined by $L^* := \bigcup_{i \geq 0} L^i$.)

Formal language theory is the study of the properties of languages. By contrast, number theory is the study of the properties of integers. Nevertheless, these two areas have many interesting intersections. For a survey, see a previous paper of mine [**23**] and the new book [**7**].

In this paper, I illustrate the connections between formal language theory and number theory by discussing four examples. In the first three examples, given in sections 2, 3, and 4, I show how number theory may be used to solve problems in formal language theory. These examples are surely in the spirit of the theme of the conference, entitled "Unusual Applications of Number Theory". In the last

2000 *Mathematics Subject Classification.* Primary 11B85; Secondary 68Q42, 68Q45, 11T99.
Key words and phrases. Finite automata; context-free grammar; transcendence in finite characteristic.
Research supported in part by a grant from NSERC.

example, given in section 5, I show how formal language theory may be applied to solve a problem in number theory. (This example would be suitable for a "co-conference".) I do not make any attempt to be comprehensive or provide all details. Further, while the sections 2 and 3 discuss original results, sections 4 and 5 are based on results of Jean-Paul Allouche.

2. State complexity of the intersection of unary languages

A *deterministic finite automaton* (DFA) is a simple model of a computer. It consists of a finite nonzero number of states. One state, called the *start state*, is drawn with a single arrow entering. Other states, called *accepting* or *final* states, are drawn with two concentric circles. A string is said to be *accepted* if it is the label of a path beginning with the start state and ending at some final state. The *language accepted by a finite automaton* M, written $L(M)$, is the set of all strings accepted by M. For example, in Figure 1 below, the given automaton accepts the base-2 representations of the prime numbers ≤ 11.

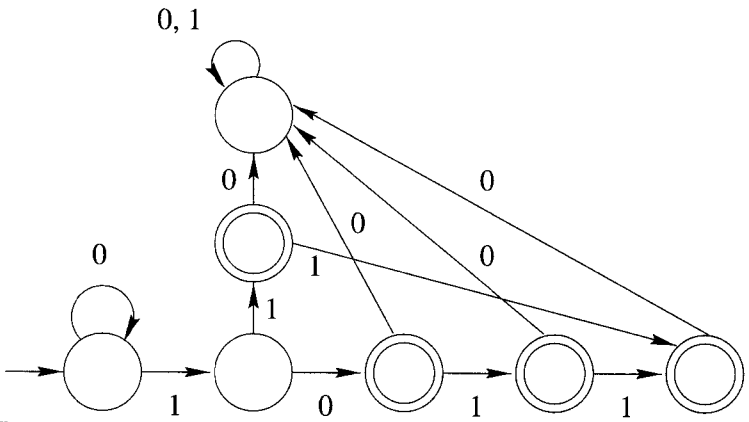

FIGURE 1. Transition diagram for automaton accepting the base-2 representations of the primes $p \leq 11$

More formally, a DFA is a quintuple: $M = (Q, \Sigma, \delta, q_0, F)$ where

- Q is a finite set of states;
- the *size* of M is $|M| := |Q|$, the number of states;
- Σ is the input alphabet;
- $q_0 \in Q$ is the start state;
- $F \subseteq Q$ is the set of final states;
- $\delta : Q \times \Sigma \to Q$ is the *transition function*, which is extended to $\delta : Q \times \Sigma^* \to Q$ in the obvious way.

The *language accepted by* M is denoted by $L(M)$ and is given by $\{w \in \Sigma^* : \delta(q_0, w) \in F\}$. A language L is said to be *regular* if it is accepted by some DFA M.

The *state complexity* of a regular language L, $\mathrm{sc}(L)$, is the minimum number of states required by any DFA which accepts it; see, for example, [**28**].

The state complexity problem is the following: given regular languages L, L' with state complexity n, n' respectively, what are good bounds on the state complexity of $L \cap L'$, $L \cup L'$, LL', L^*, etc.?

For the state complexity of intersection, we have the following upper bound, which is well-known.

PROPOSITION 1. *We have*

$$\mathrm{sc}(L \cap L') \leq \mathrm{sc}(L)\mathrm{sc}(L').$$

PROOF. Let L be accepted by the DFA $(Q, \Sigma, \delta, q_0, F)$ and L' be accepted by the DFA $(Q', \Sigma, \delta', q_0', F')$. Then $L \cap L'$ can be accepted by a DFA $(Q'', \Sigma, \delta'', q_0'', F'')$ where

- $Q'' := Q \times Q'$;
- $q_0'' := [q_0, q_0']$;
- $F'' := F \times F'$; and
- $\delta''([p, q], a) = [\delta(p, a), \delta(q, a)]$.

□

As Yu and Zhuang observed [**27**], the upper bound of $\mathrm{sc}(L)\mathrm{sc}(L')$ can be achieved if L, L' are over an alphabet of size at least 2:

PROPOSITION 2. *Define*

$$L := \{x \in \{a, b\}^* \ : \ |x|_a \equiv 0 \pmod{n}\};$$
$$L' := \{y \in \{a, b\}^* \ : \ |y|_b \equiv 0 \pmod{n'}\}.$$

Then

$$\mathrm{sc}(L \cap L') = nn'.$$

But what if L, L' are unary, that is, defined over an alphabet of one symbol? Clearly if $\gcd(n, n') = 1$ then the bound nn' can again be achieved, by taking $L = (a^n)^*$ and $L' = (a^{n'})^*$. But what if $\gcd(n, n') > 1$? This problem was stated as an unsolved problem by Yu [**26**].

To obtain a lower bound on the state complexity of intersection of unary regular languages, we examine the topology of unary DFA's. A connected unary DFA has the property that its transition diagram consists of

- a *tail* of $t \geq 0$ states and
- a *cycle* of $c \geq 1$ states.

It is then not hard to prove the following:

THEOREM 3. *Let M, M' be unary DFA's with tails of size t, t' and cycles of size c, c', respectively. If L, L' are the corresponding languages, we have*

$$\mathrm{sc}(L \cap L') \leq \max(t, t') + \mathrm{lcm}(c, c'). \tag{1}$$

Furthermore, for all $t, t' \geq 0$ and $c, c' \geq 1$ there exist unary languages for which the bound (1) is achieved.

(This theorem was obtained independently by the author [**24**] and G. Pighizzini [**20**]; also see [**22**].)

Thus, to estimate the worst-case behavior for the state complexity of intersection of unary languages with n and n' states, respectively, we must estimate the function

$$F(n, n') = \max_{\substack{1 \leq c \leq n \\ 1 \leq c' \leq n'}} \left(\max(n - c, n' - c') + \mathrm{lcm}(c, c') \right).$$

This in turn suggests studying the somewhat simpler and more natural function

$$G(n,n') = \max_{\substack{1\le c\le n\\ 1\le c'\le n'}} \operatorname{lcm}(c,c').$$

Although the asymptotic behavior of F and G is still not known precisely, there is a relation to Jacobsthal's function $g(n)$, which is defined to be the least integer r such that every set of r consecutive integers contains at least one integer relatively prime to n [**14**].

Iwaniec [**13**] proved using the linear sieve that $g(n) = O((\log n)^2)$. The following lower bound for our problem then follows:

THEOREM 4. *If $n \le n'$, we have $F(n,n') \ge G(n,n') \ge nn' - c_1(\log n)^2 n$ for some constant c_1.*

Note: results on the average state complexity of operations on unary DFA's were obtained by Nicaud [**18**].

3. Grammatical complexity of unary context-free grammars

Context-free grammars are a method for generating languages. The modern mathematical formulation is due to Chomsky [**8**], although the basic idea goes back to Indian philologist Panini, c. 400 B.C.E.

A context-free grammar consists of a start symbol and *rewriting* or *production rules*, e.g.:

$$\begin{aligned} S &\to aSa \\ S &\to bSb \\ S &\to a \\ S &\to b \\ S &\to \epsilon \end{aligned}$$

which generates the palindromes over $\{a,b\}$. Often multiple productions are abbreviated using |, e.g.,

$$S \to aSa \mid bSb \mid a \mid b \mid \epsilon.$$

To derive a word, one starts with the start symbol S and then successively replaces an occurrence of a variable with a right-hand-side of a production, until a string of symbols (without variables) is obtained.

More formally, a context-free grammar (CFG) is a 4-tuple $G = (V, \Sigma, P, S)$ where

- V is a finite set of *variables*;
- Σ is a finite alphabet;
- P is a set of *production rules* of the form $A \to \gamma$, where $A \in V$ and $\gamma \in (V \cup \Sigma)^*$; and
- S is the *start symbol.*

We write $\alpha \Longrightarrow \beta$ if β can be obtained from α by the use of one production rule. We write $\Longrightarrow^*$ for the reflexive, transitive closure of $\Longrightarrow$. Then $L(G)$, the language generated by G is formally defined as

$$L(G) := \{x \in \Sigma^* \ : \ S \Longrightarrow^* x\}.$$

Context-free grammars generate a class of languages, the context-free languages, which are a strict superset of the class of regular languages.

A *leftmost derivation* in a grammar is a series of steps deriving a terminal string, in which the leftmost variable is replaced at each step. A grammar G is said to be *ambiguous* if there exists at least one word in $L(G)$ possessing at least two distinct leftmost derivations.

We can measure the size of a context-free grammar as the number of symbols needed to write down its description. Suppose a CFG G generates a regular language. How big can the corresponding DFA be, in terms of the size of G?

If the CFG is over an alphabet with at least 2 symbols, the answer is, there is no recursive bound. More precisely, Meyer and Fischer [**17**] proved that given any recursive function f, for arbitrarily large integers n there exists a CFG of size n describing a regular language L such that any DFA accepting L has at least $f(n)$ states.

But how about the unary case? As Ginsburg and Rice [**12**] proved, any unary CFG generates a regular language. Further, the author together with M.-w. Wang (and independently, G. Pighizzini [**21**]) have shown there exists a constant such that any unary CFG of size n describing a regular language can be accepted by a DFA with at most $O(2^{cn^2})$ states.

But is this bound achievable? We give an example based on number theory that achieves this bound. Consider the following productions:

$$\begin{aligned}
A_0 &\to a \\
A_{i+1} &\to A_i A_i \quad (i \geq 0) \\
B_i &\to a A_i \\
C_0 &\to a \\
C_{i+1} &\to a \mid C_i C_i \quad (i \geq 0) \\
D_i &\to D_i B_i \mid C_i \quad (i \geq 0) \\
S_i &\to \epsilon \mid D_0 \mid D_1 \mid D_2 \mid \cdots \mid D_i \quad (i \geq 0)
\end{aligned}$$

It follows that, for $i \geq 0$,

$$\begin{aligned}
A_i &\Longrightarrow^* \{a^{2^i}\} \\
B_i &\Longrightarrow^* \{a^{2^i+1}\} \\
C_i &\Longrightarrow^* \{a, a^2, a^3, \ldots, a^{2^i}\} \\
D_i &\Longrightarrow^* \{a, a^2, a^3, \ldots, a^{2^i}\}\{a^{2^i+1}\}^* = \{a^j \ : \ j \not\equiv 0 \pmod{2^i+1}\}. \\
S_i &\Longrightarrow^* \{\epsilon\} \cup \{a^k \ : \ k \not\equiv 0 \pmod{\operatorname{lcm}(2^0+1, 2^1+1, \ldots, 2^i+1)}\}.
\end{aligned}$$

Now let $G_n = (V_n, \{a\}, P_n, S_n)$, where

$$V_n = \{A_i, B_i, C_i, D_i, S_i \ : \ 0 \leq i \leq n\}$$

and P_n is the set of $O(n)$ productions given above involving these variables. It is clear that $L(G_n)$ is regular. The shortest string not generated by G_n is of length

$$\operatorname{lcm}(2^0+1, 2^1+1, \ldots, 2^n+1)$$

and so any DFA accepting $L(G_n)$ must have at least this many states.

It remains to estimate

$$\operatorname{lcm}(2^0+1, 2^1+1, \ldots, 2^n+1).$$

We use the following theorem of Bézivin [1989]:

THEOREM 5. *Let a, b be integers with $b \neq 0$ and $\gcd(a,b) = 1$. Let α, β be zeroes of the polynomial $X^2 - aX - b$. For $m \geq 2$ define*

$$u_m(n) = \frac{\alpha^{mn} - \beta^{mn}}{\alpha^n - \beta^n}.$$

Then

$$\lim_{n\to\infty} \frac{\log(u_m(1)u_m(2)\cdots u_m(n))}{\log \operatorname{lcm}(u_m(1), u_m(2), \ldots, u_m(n))} = \frac{(m-1)L(m)\pi^2}{6H(m)},$$

where

$$L(m) = \prod_{p \mid m} \left(1 - \frac{1}{p^2}\right)$$

and

$$H(m) = \sum_{\substack{d \mid m \\ d>1}} \frac{\varphi(d)\varphi(m/d)d}{m}.$$

Now take $a = 3, b = -2, m = 2$ in Bézivin's theorem. Then $\alpha = 2$ and $\beta = 1$, and we obtain

$$\lim_{n\to\infty} \frac{\log((2^0+1)(2^1+1)\cdots(2^n+1))}{\log \operatorname{lcm}(2^0+1, 2^1+1, \ldots, 2^n+1))} = \frac{\pi^2}{8}.$$

On the other hand, it is easy to see that

$$\lim_{n\to\infty} \frac{(2^0+1)(2^1+1)\cdots(2^n+1)}{2^0 \cdot 2^1 \cdots 2^n} = c_2,$$

where $c_2 \doteq 4.768$, so it follows that

$$\log((2^0+1)(2^1+1)\cdots(2^n+1)) \sim \log c_2 + \frac{n(n+1)}{2}\log 2.$$

Putting this together with the Bézivin result, we get

$$\log \operatorname{lcm}(2^0+1, 2^1+1, \ldots, 2^n+1) \sim \frac{4\log 2}{\pi^2} n^2.$$

We remark that other constructions are possible which achieve the 2^{cn^2} bound. For example, instead of $\operatorname{lcm}(2^0+1, 2^1+1, \ldots, 2^n+1)$, we could instead use a result of Szymiczek [**25**] and consider $\operatorname{lcm}(2^1-1, 2^2-1, \ldots, 2^n-1)$. Or we could use a result of Matiyasevich and Guy [**16**] and consider $\operatorname{lcm}(F_1, F_2, \ldots, F_n)$, where F_i is the ith Fibonacci number. For related results, see [**15, 1, 2**].

4. The primitive words problem

Let Σ be a finite alphabet with at least two letters. A word $w \in \Sigma^*$ is said to be *primitive* if it cannot be expressed in the form x^k with $k \geq 2$. For example, *abaab* is primitive, but *abaaba* $= (aba)^2$ is not. A major open problem in formal languages is the following: is the language P of primitive words over $\{a, b\}$ context-free? The answer is almost certainly no, but nobody currently knows how to prove this. Petersen [**19**] proved the following weaker result:

THEOREM 6. *P is not unambiguously context-free (i.e., if P is a CFL, then any grammar for it is ambiguous).*

Petersen used the Chomsky-Schützenberger theorem [**9**], which states that if L is a context-free language having an unambiguous grammar, and $a_n := |L \cap \Sigma^n|$, then $\sum_{n\geq 0} a_n X^n$ is a formal power series in $\mathbb{Z}[[X]]$ which is algebraic over $\mathbb{Q}(X)$.

Recently a remarkably simple proof of Petersen's result was found by Allouche [**5, 6**] using the theory of automatic sequences. In this section we explain Allouche's proof.

First, we give an example of Chomsky-Schützenberger theorem. Consider the unambiguous grammar

$$\begin{aligned} S &\to M \mid U \\ M &\to 0M1M \mid \epsilon \\ U &\to 0S \mid 0M1U \end{aligned}$$

which represents strings of "if-then-else" clauses. Then this grammar has the following commutative image:

$$\begin{aligned} S &= M + U \\ M &= x^2M^2 + 1 \\ U &= Sx + x^2MU \end{aligned}$$

This system of equations has the following power series solutions:

$$\begin{aligned} M &= 1 + x^2 + 2x^4 + 5x^6 + 14x^8 + 42x^10 + \cdots \\ U &= x + x^2 + 3x^3 + 4x^4 + 10x^5 + 15x^6 + 35x^7 + 56x^8 + \cdots \\ S &= 1 + x + 2x^2 + 3x^3 + 6x^4 + 10x^5 + 20x^6 + 35x^7 + \cdots \end{aligned}$$

By the Chomsky-Schützenberger theorem, each variable satisfies an algebraic equation over $\mathbb{Q}(x)$. For example, we have

$$x(2x-1)S^2 + (2x-1)S + 1 = 0$$

Now we digress a moment to discuss automata as computers of sequences.

We can generalize our notion of automaton to provide an output, not simply accept/reject. Formally, we define a *deterministic finite automaton with output* (DFAO) as a sextuple: $(Q, \Sigma, \delta, q_0, \Delta, \tau)$, where Δ is the finite *output alphabet* and $\tau : Q \to \Delta$ is the *output mapping*.

Next, we decide on a integer base $k \geq 2$ and represent n as a string of symbols over the alphabet $\Sigma = \{0, 1, 2, \ldots, k-1\}$. To compute f_n, given an automaton M, express n in base-k, say, $a_r a_{r-1} \cdots a_1 a_0$, and compute $f_n = \tau(\delta(q_0, a_r a_{r-1} \cdots a_1 a_0))$. Any sequence that can be computed in this way is said to be *k-automatic* [**11**].

The Thue-Morse sequence $(t_n)_{n\geq 0}$ is defined as follows: t_n is the parity of the number of 1's in the binary expansion of n. We have

$$(t_n)_{n\geq 0} = 0\,1\,1\,0\,1\,0\,0\,1\cdots$$

Note that $t_0 = 0$; $t_{2n} = t_n$, and $t_{2n+1} = 1 - t_n$ for $n \geq 0$.

Axel Thue (c. 1906) studied this sequence because it is *cubefree*: it contains no subword of the form www, where w is a nonempty word. It is computed by the following DFAO:

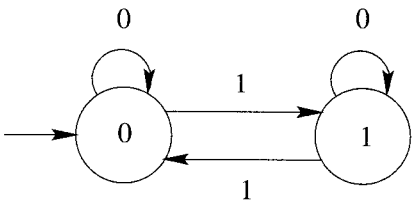

FIGURE 2. Automaton computing the Thue-Morse sequence

We observe that the notion of automatic sequence is *robust*: it does not change under small changes to the computational model. For example, the order in which the base-k digits are fed into the automaton in does not matter (provided it is fixed for all n); other representations also work (such as expansion in base-$(-k)$); automatic sequences are closed under many operations, such as shift, periodic deletion, q-block compression, and q-block substitution; and if a symbol in an automatic sequence occurs with well-defined frequency r, then r is rational [**11**].

The theorem of Christol [**10**] is the most important in the area:

THEOREM 7 (Christol, 1979). *Let $(u_n)_{n\geq 0}$ be a sequence over*

$$\Sigma = \{0, 1, \ldots, p-1\},$$

where p is a prime. Then the formal power series $U(X) = \sum_{n\geq 0} u_n X^n$ is algebraic over $GF(p)[X]$ if and only if $(u_n)_{n\geq 0}$ is p-automatic.

Let us consider an example. Let, as before, $(t_n)_{n\geq 0}$ denote the Thue-Morse sequence, i.e., t_n = sum of the bits in the binary expansion of n, mod 2. Then $t_{2n} \equiv t_n$ and $t_{2n+1} \equiv t_n + 1$. If we set $A(X) = \sum_{n\geq 0} t_n X^n$, then

$$\begin{aligned} A(X) &= \sum_{n\geq 0} t_{2n} X^{2n} + \sum_{n\geq 0} t_{2n+1} X^{2n+1} \\ &= \sum_{n\geq 0} t_n X^{2n} + X \sum_{n\geq 0} t_n X^{2n} + X \sum_{n\geq 0} X^{2n} \\ &= A(X^2) + XA(X^2) + X/(1-X^2) \\ &= A(X)^2(1+X) + X/(1+X)^2. \end{aligned}$$

Hence $(1+X)^3 A^2 + (1+X)^2 A + X = 0$.

We can now return to Allouche's proof of Petersen's result.

PROOF. Let $\psi_k(n)$ be the number of primitive words of length n over a k-letter alphabet. Then it is easy to see (using Möbius inversion) that

$$\psi_k(n) = \sum_{d \mid n} \mu(d) k^{n/d}.$$

If P_k were unambiguously context-free then by the Chomsky-Schützenberger theorem

$$R(X) = \sum_{n\geq 1} \psi_k(n) X^n$$

would be algebraic over $\mathbb{Q}(X)$. Then

$$\hat{R}(X) = \sum_{n \geq 1} \frac{\psi_k(n)}{k} X^n$$

would also be algebraic over $\mathbb{Q}(X)$.

Let p be a prime dividing k. Then it is not hard to see that

$$\hat{R}_p(X) = \sum_{n \geq 1} (\frac{\psi_k(n)}{k} \bmod p) X^n$$

would also be algebraic over $GF(p)(X)$. But

$$\frac{\psi_k(n)}{k} = \sum_{d \mid n} \mu(d) k^{n/d-1} = \mu(n) + \sum_{d \mid n d \neq n} \mu(d) k^{n/d-1} \equiv \mu(n) \pmod{p}.$$

It follows that

$$\hat{R}_p(X) = \sum_{n \geq 1} \mu(n) X^n$$

and so the sequence $(\mu(n) \bmod p)_{n \geq 0}$ must be p-automatic. But then $(\mu(n)^2 \bmod p)_{n \geq 0}$ would be p-automatic.

However, $\mu(n)^2 \equiv 1 \pmod{p}$ if and only if n is squarefree. By a classical theorem, the density of the squarefree numbers exists and is equal to $6/\pi^2$, an irrational number. But, as remarked earlier, the density of symbols in automatic sequences (if it exists) must be rational, a contradiction. It follows that $R(X)$ is not algebraic over $\mathbb{Q}(X)$ and so P_k is not unambiguously context-free. This completes the proof. □

5. Transcendence in Finite Characteristic

In this section, we turn the tables and illustrate an application of formal language theory to number theory.

Define for $n \geq 1$

$$\zeta_q(n) = \sum_{\substack{P \text{ monic} \\ P \in GF(q)[X]}} \frac{1}{P^n}$$

Thus, for example,

$$\begin{aligned} \zeta_2(1) &= \frac{1}{1} + \frac{1}{X} + \frac{1}{X+1} + \frac{1}{X^2} + \frac{1}{X^2+1} + \frac{1}{X^2+X} + \cdots \\ &= 1 + X^{-2} + X^{-3} + X^{-4} + X^{-5} + X^{-9} + X^{-10} + \cdots \\ &\in GF(2)[[X^{-1}]]. \end{aligned}$$

This function ζ_q, now called the Carlitz zeta function, has many properties similar to those of the Riemann zeta function. For example, it admits the following Euler product:

$$\zeta_q(n) = \prod_{\substack{P \text{ irreducible} \\ P \in GF(q)[X]}} \frac{1}{1 - \frac{1}{P^n}}.$$

Carlitz also showed that if $q-1 \mid n$, then $\zeta_q(n) = \pi_q^n \cdot r$ where r is a rational function and

$$\pi_q := \prod_{k \geq 1} \left(1 - \frac{X^{q^k} - X}{X^{q^{k+1}} - X}\right).$$

Wade proved the following theorem:

THEOREM 8. *π_q is transcendental.*

Here is another proof of Wade's result, due to Allouche [**4**], using automatic sequences and Christol's theorem (also see [**3**].

PROOF. Taking the logarithmic derivative, we get

$$\begin{aligned}
\frac{\pi_q'}{\pi_q} &= \sum_{k \geq 1} \left(\frac{1}{1 - \frac{X^{q^k} - X}{X^{q^{k+1}} - X}}\right) \left(\frac{(X^{q^{k+1}} - X) - (X^{q^k} - X)}{(X^{q^{k+1}} - X)^2}\right) \\
&= \sum_{k \geq 1} \frac{1}{X^{q^{k+1}} - X} \\
&= \left(\sum_{k \geq 1} \frac{1}{X^{q^k} - X}\right) - \frac{1}{X^q - X}.
\end{aligned}$$

Now suppose that π_q is algebraic over $GF(q)(X)$. Then so is the formal derivative π_q'. Hence so is π_q'/π_q. But then so is

$$\sum_{k \geq 1} \frac{1}{X^{q^k} - X} = \sum_{k \geq 1} \frac{1}{[k]},$$

the so-called "bracket series" introduced by Wade, who defined $[k] := X^{q^k} - X$.

Thus to prove π_q transcendental, it suffices to show that

$$\sum_{k \geq 1} \frac{1}{X^{q^k} - X}$$

is transcendental. We now have

$$\begin{aligned}\sum_{k\geq 1}\frac{1}{X^{q^k}-X} &= \sum_{k\geq 1}\frac{1}{X^{q^k}\left(1-\left(\frac{1}{X}\right)^{q^k-1}\right)}\\ &= \sum_{k\geq 1}\frac{1}{X^{q^k}}\sum_{n\geq 0}\left(\frac{1}{X}\right)^{n(q^k-1)}\\ &= \frac{1}{X}\sum_{k\geq 1}\frac{1}{X^{q^k-1}}\sum_{n\geq 0}\left(\frac{1}{X}\right)^{n(q^k-1)}\\ &= \frac{1}{X}\sum_{\substack{k\geq 1\\ n\geq 0}}\left(\frac{1}{X}\right)^{(n+1)(q^k-1)}\\ &= \frac{1}{X}\sum_{\substack{k\geq 1\\ n\geq 1}}\left(\frac{1}{X}\right)^{n(q^k-1)}.\end{aligned}$$

Hence

$$\begin{aligned}\sum_{k\geq 1}\frac{1}{X^{q^k}-X} &= \frac{1}{X}\sum_{m\geq 1}\left(\frac{1}{X}\right)^m \sum_{\substack{k,n\geq 1\\ n(q^k-1)=m}} 1\\ &= \frac{1}{X}\sum_{k\geq 1}\left(\frac{1}{X}\right)^m \sum_{\substack{k\geq 1\\ q^k-1\,|\,m}} 1\\ &= \frac{1}{X}\sum_{m\geq 1}\left(\frac{1}{X}\right)^m c(m),\end{aligned}$$

where

$$c(m) := \sum_{\substack{k\geq 1\\ q^k-1\,|\,m}} 1.$$

Now, by Christol's theorem, in order to show that $\sum_{k\geq 1}\frac{1}{X^{q^k}-X}$ is transcendental over $GF(q)(X)$, it suffices to prove that $(c(m) \bmod p)_{m\geq 1}$ is not q-automatic, where $q = p^e$ for some e.

If the sequence $(c(m) \bmod p)_{m\geq 1}$ were q-automatic, then the subsequence $(c(q^n-1) \bmod p)_{n\geq 0}$ would be ultimately periodic. But

$$c(q^n-1) = \sum_{\substack{k\geq 1\\ q^k-1\,|\,q^n-1}} 1 = \sum_{\substack{k\geq 1\\ k\,|\,n}} 1 = d(n),$$

where $d(n)$ is the number of positive integral divisors of n.

It now suffices to show that $(d(n) \bmod p)_{n\geq 1}$ is not ultimately periodic. This can be done by a simple argument using Dirichlet's theorem. This contradiction completes the proof. □

6. Acknowledgments

I thank Gerry Myerson for pointing out the paper of Szymiczek [**25**], and Jean-Eric Pin for pointing out the work of Nicaud [**18**].

References

[1] S. Akiyama. Lehmer numbers and an asymptotic formula for π. *J. Number Theory* **36** (1990), 328–331.

[2] S. Akiyama. A new type of inclusion exclusion principle for sequences and asymptotic formulas for $\zeta(k)$. *J. Number Theory* **45** (1993), 200–214.

[3] J.-P. Allouche. Sur la transcendance de la série formelle π. *Séminaire de Théorie des Nombres de Bordeaux* **2** (1990), 103–117.

[4] J.-P. Allouche. Finite automata and arithmetic. *Sém. Lotharingien de Combinatoire* **B30c** (1993), 1–23.

[5] J.-P. Allouche. Note on the transcendence of a generating function. In A. Laurincikas, E. Manstavicius, and V. Stakenas, editors, *Proceedings of the 2nd International Conference in Honour of J. Kubilius (Palanaga, Lithuania, September 23–27, 1996)*, Vol. 4 of *New Trends in Probability and Statistics*, pp. 461–465. VSP, Utrecht, 1997.

[6] J.-P. Allouche. Transcendence of formal power series with rational coefficients. *Theoret. Comput. Sci.* **218** (1999), 143–160.

[7] J.-P. Allouche and J. O. Shallit. *Automatic Sequences: Theory, Applications, Generalizations.* Cambridge University Press, 2003.

[8] N. Chomsky. Three models for the description of language. *IRE Trans. Info. Theory* **2** (1956), 113–124.

[9] N. Chomsky and M. P. Schützenberger. The algebraic theory of context-free languages. In P. Braffort and D. Hirschberg, editors, *Computer Programming and Formal Systems*, pp. 118–161. North Holland, Amsterdam, 1963.

[10] G. Christol. Ensembles presque périodiques k-reconnaissables. *Theoret. Comput. Sci.* **9** (1979), 141–145.

[11] A. Cobham. Uniform tag sequences. *Math. Systems Theory* **6** (1972), 164–192.

[12] S. Ginsburg and H. G. Rice. Two families of languages related to ALGOL. *J. Assoc. Comput. Mach.* **9** (1962), 350–371.

[13] H. Iwaniec. On the problem of Jacobsthal. *Dem. Math.* **11** (1978), 225–231.

[14] E. Jacobsthal. Über Sequenzen ganzer Zahlen. von denen keine zu n teilerfremd ist. I-III. *Norske Vid. Selsk. Forh. Trondheim* **33** (1960), 117–139.

[15] P. Kiss and F. Mátyás. An asymptotic formula for π. *J. Number Theory* **31** (1989), 255–259.

[16] Yu. V. Matiyasevich and R. K. Guy. A new formula for π. *Amer. Math. Monthly* **93** (1986), 631–635.

[17] A. R. Meyer and M. J. Fischer. Economy of description by automata, grammars, and formal systems. In *Proc. 12th Annual Symposium on Switching and Automata Theory*, pp. 188–191, 1971.

[18] C. Nicaud. Average state complexity of operations on unary automata. In *Proc. 24th Symposium, Mathematical Foundations of Computer Science 1999*, Vol. 1672 of *Lecture Notes in Computer Science*, pp. 231–240. Springer-Verlag, 1999.

[19] H. Petersen. On the language of primitive words. *Theoret. Comput. Sci.* **161** (1996), 141–156.

[20] G. Pighizzini. Unary language concatenation and its state complexity, In S. Yu and A. Păun, eds., *Proc. 5th International Conference on Implementation and Application of Automata (CIAA 2000)*, Lecture Notes in Computer Science Vol. 2088, Springer-Verlag, 2001, pp. 252–262.

[21] G. Pighizzini. In M. Nielsen and B. Rovan, editors, *Proc. 25th International Symposium on Mathematical Foundations of Computer Science (MFCS 2000)*, Vol. 1893 of *Lecture Notes in Computer Science*, pp. 599–608. Springer-Verlag, 2000.

[22] G. Pighizzini and J. Shallit. Unary language operations, state complexity, and Jacobsthal's function. *Int'l. J. Found. Comput. Sci.* **13** (2002), 145–149.

[23] J. Shallit. Number theory and formal languages. In D. A. Hejhal, J. Friedman, M. C. Gutzwiller, and A. M. Odlyzko, editors, *Emerging Applications of Number Theory*, Vol. 109 of *IMA Volumes in Mathematics and Its Applications*, pp. 547–570. Springer-Verlag, 1999.

[24] J. Shallit. State complexity and Jacobsthal's function. In S. Yu and A. Păun, eds., *Proc. 5th International Conference on Implementation and Application of Automata (CIAA 2000)*, Lecture Notes in Computer Science Vol. 2088, Springer-Verlag, 2001, pp. 272–278.

[25] K. Szymiczek. On the distribution of prime factors of Mersenne numbers. *Prace Math.* **13** (1969), 33–49.

[26] S. Yu. State complexity of regular languages. In *International Workshop on Descriptional Complexity of Automata, Grammars and Related Structures, Preproceedings*, pp. 77–88. Department of Computer Science, Otto-von-Guericke University of Magdeburg, July 1999.

[27] S. Yu and Q. Zhuang. On the state complexity of intersection of regular languages. *SIGACT News* **22**(3) (1991), 52–54.

[28] S. Yu, Q. Zhuang, and K. Salomaa. The state complexity of some basic operations on regular languages. *Theoret. Comput. Sci.* **125** (1994), 315–328.

School of Computer Science, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada

E-mail address: shallit@graceland.uwaterloo.ca