

Formal Languages and Number Theory

Jeffrey Shallit

Department of Computer Science

University of Waterloo

Waterloo, Ontario N2L 3G1

Canada

`shallit@graceland.uwaterloo.ca`

`http://www.math.uwaterloo.ca/~shallit`

Outline of the Talk

This talk is about application of number theory to problems of formal language theory and vice versa.

Table of Contents

1. State complexity of the intersection of unary languages.
2. Descriptive complexity of unary context-free grammars representing regular languages.
3. Proof that the primitive words are not unambiguously context-free.
4. Transcendence in finite characteristic.

Languages

- If Σ is a finite set of symbols, then by Σ^* we mean the free monoid over Σ (set of all finite strings of symbols chosen from Σ);
- A *language* is a subset of Σ^* .
- Example: $PALL = \{x \in \{a, b\}^* : x = x^R\}$
- The basic operations on languages are union, concatenation, and Kleene closure.
- concatenation:

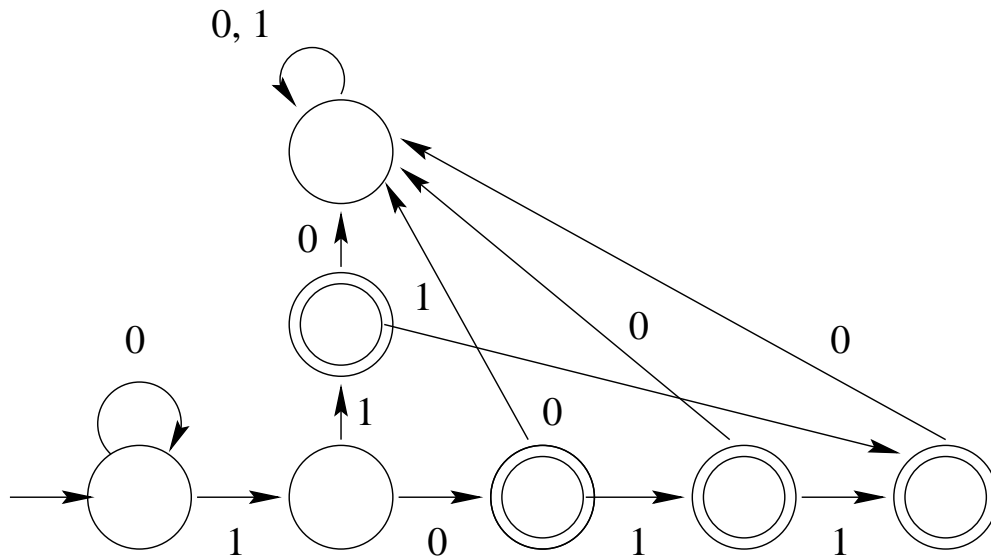
$$L_1L_2 := \{xy : x \in L_1, y \in L_2\}.$$

- Kleene closure:

$$L^* := \bigcup_{i \geq 0} L^i.$$

Finite Automata

- A *deterministic finite automaton* (DFA) is a simple model of a computer.



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

Basics of Finite Automata

- 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 a finite set of symbols, called the *input alphabet*;
 - $q_0 \in Q$ is the *start state*;
 - $F \subseteq Q$ is the set of *final states*;
 - $\delta : Q \times \Sigma \rightarrow Q$ is the *transition function*, which is extended to $\delta : Q \times \Sigma^*$ in the obvious way
- The *language accepted by* M is denoted by $L(M)$ and is given by $\{w \in \Sigma^* \mid \delta(q_0, w) \in F\}$.
- A language L is said to be *regular* if it is accepted by some DFA M .

State Complexity

The *state complexity* of a regular language L , $sc(L)$, is the minimum number of states needed to accept it by a DFA.

The problem:

Given languages L, L' with state complexity n, n' respectively, what are good bounds on the state complexity of $L \cup L', LL', L^*$, etc.?

For the state complexity of intersection, we have the following bound:

Proposition. *We have*

$$sc(L \cap L') \leq sc(L)sc(L').$$

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
- $\delta''([p, q], a) = [\delta(p, a), \delta'(q, a)]$. ■

State Complexity of Intersection

The upper bound of $sc(L)sc(L')$ can be achieved if L, L' are over an alphabet of size at least 2:

Proposition. (S. YU.) *Define*

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

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

Then

$$sc(L \cap L') = nn'.$$

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 $\gcd(n, n') > 1$?

State Complexity of Intersection for Unary Languages

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 that

Theorem. *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*

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

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

Two New Number-Theoretic Functions

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

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

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

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

- The asymptotic behavior of F and G is still not known precisely
- There is a relation to JACOBSTHAL'S function $g(n)$, which is the *least integer r such that every set of r consecutive integers contains at least one integer relatively prime to n* .
- IWANIEC proved [1978] using the linear sieve that $g(n) = O((\log n)^2)$.
- It then follows that if $n \leq n'$, we have $F(n, n') \geq G(n, n') \geq nn' - c(\log n)^2 n$ for some constant c .

Context-Free Grammars

- A method for generating languages
- Modern mathematical formulation due to CHOMSKY [1956], although idea goes back to Indian philologist PANINI, c. 400 B.C.E.
- Consists of a start symbol and *rewriting rules*, e.g.:

$$S \rightarrow aSa$$

$$S \rightarrow bSb$$

$$S \rightarrow a$$

$$S \rightarrow b$$

$$S \rightarrow \epsilon$$

which generates the palindromes over $\{a, b\}$.

- 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 $A \rightarrow \gamma$, where $A \in V$ and $\gamma \in (V \cup \Sigma)^*$
 - S is the start symbol

Context-Free Grammars

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

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

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

Descriptive Complexity of Context-Free Grammars

- 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 proved [1971] 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?
 - It is possible to show that 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?

An Example Exhibiting 2^{cn^2} Blowup

$$A_0 \rightarrow a$$

$$A_{i+1} \rightarrow A_i A_i \quad (i \geq 0)$$

$$\text{so } A_i \implies^* \{a^{2^i}\}$$

$$B_i \rightarrow aA_i$$

$$\text{so } B_i \implies^* \{a^{2^i+1}\}$$

$$C_0 \rightarrow a$$

$$C_{i+1} \rightarrow a \mid C_i C_i \quad (i \geq 0)$$

$$\text{so } C_i \implies^* \{a, a^2, a^3, \dots, a^{2^i}\}$$

$$D_i \rightarrow D_i B_i \mid C_i \quad (i \geq 0)$$

$$\begin{aligned} \text{so } D_i &\implies^* \{a, a^2, a^3, \dots, a^{2^i}\} \{a^{2^i+1}\}^* \\ &= \{a^j : j \not\equiv 0 \pmod{2^i + 1}\}. \end{aligned}$$

An Example Exhibiting 2^{cn^2} Blowup

And finally, let

$$S_i \rightarrow \epsilon \mid D_0 \mid D_1 \mid D_2 \mid \cdots \mid D_i$$

so $S_i \Longrightarrow^* \{\epsilon\} \cup \{a^k : k \not\equiv 0 \pmod{\text{lcm}(2^0 + 1, 2^1 + 1, \dots, 2^i + 1)}\}$.

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

$$\text{lcm}(2^0 + 1, 2^1 + 1, \dots, 2^n + 1)$$

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

It remains to estimate

$$\text{lcm}(2^0 + 1, 2^1 + 1, \dots, 2^n + 1)$$

An Example Exhibiting 2^{cn^2} Blowup

We use the following theorem of BÉZIVIN [1989]:

Theorem. *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

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

where

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

and

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

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

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

An Example Exhibiting 2^{cn^2} Blowup

On the other hand, it is easy to see that

$$\lim_{n \rightarrow \infty} \frac{(2^0 + 1)(2^1 + 1) \cdots (2^n + 1)}{2^0 \cdot 2^1 \cdots 2^n} = c_1,$$

where $c_1 \doteq 4.768$, so it follows that

$$\log((2^0 + 1)(2^1 + 1) \cdots (2^n + 1)) \sim \log c_1 + \frac{n(n + 1)}{2} \log 2.$$

Putting this together with the BÉZIVIN result, we get

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

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 $k \geq 2$.
- Open problem in formal languages: is the language P of primitive words context-free?
- Answer is almost certainly no, but nobody knows how to prove this.
- PETERSEN [1996] proved the weaker result that P is not unambiguously context-free.
- He did this using the CHOMSKY-SCHÜTZENBERGER theorem, 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 was found by ALLOUCHE using the theory of automatic sequences.

Example of Chomsky-Schützenberger Theorem

Consider the unambiguous grammar

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

which represents strings of “if-then-else” clauses.

Then this has the following commutative image:

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

This system has the following power series solutions:

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

By the CHOMSKY-SCHÜTZENBERGER theorem, each variable satisfies an algebraic equation over $\mathbb{Q}(x)$.

For example, we have

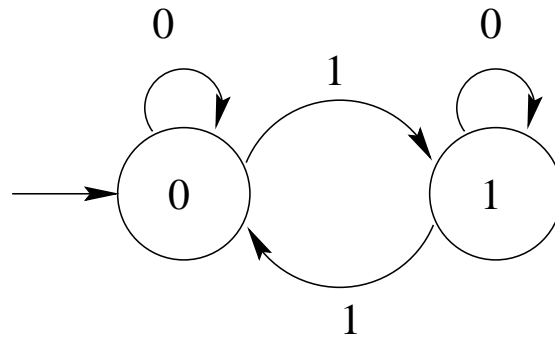
$$x(2x - 1)S^2 + (2x - 1)S + 1 = 0$$

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 $\tau : Q \rightarrow \Delta$ 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 $\Sigma = \{0, 1, 2, \dots, k - 1\}$.
- 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.

Example: The Thue-Morse sequence

- 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 .
- $(t_n)_{n \geq 0} = 01101001 \dots$
- We have $t_0 = 0$; $t_{2n} = t_n$, and $t_{2n+1} = 1 - t_n$ for $n \geq 0$.
- 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:



Robustness of the Notion of Automatic Sequence

- the order in which the base- k digits are fed into the automaton 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.
- if a symbol in an automatic sequence occurs with well-defined frequency r , then r is rational.

The Theorem of Christol

Theorem. (CHRISTOL [1980]). Let $(u_n)_{n \geq 0}$ be a sequence over

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

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.

Example.

Let, as before, $(t_n)_{n \geq 0}$ denote the THUE-MORSE sequence, i.e., $t_n = \text{sum of the bits in the binary expansion of } n, \text{ mod } 2$. 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

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

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

Back to Primitive Words

- 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

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

- 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

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

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

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

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

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

- But

$$\begin{aligned}
\frac{\psi_k(n)}{k} &= \sum_{d|n} \mu(d) k^{n/d-1} \\
&= \mu(n) + \sum_{\substack{d|n \\ d \neq n}} \mu(d) k^{n/d-1} \\
&\equiv \mu(n) \pmod{p}.
\end{aligned}$$

- It follows that

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

and so the sequence $(\mu(n) \pmod{p})_{n \geq 0}$ must be p -automatic.

- But then $(\mu(n)^2 \pmod{p})_{n \geq 0}$ would be p -automatic.
- However, $\mu(n)^2 \equiv 1 \pmod{p}$ if and only if n is square-free.
- By a classical theorem, the density of the squarefree numbers exists and is equal to $6/\pi^2$, an irrational number.
- But 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.

Transcendence in Finite Characteristic

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,

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

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

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

Transcendence in Finite Characteristic

WADE proved that π_q is transcendental. Here is another proof, due to ALLOUCHE, using automatic sequences and CHRISTOL'S theorem.

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 $[k] := X^{q^k} - X$.

Transcendence in Finite Characteristic

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

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

Transcendence in Finite Characteristic

Hence

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

where

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

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 .

Transcendence in Finite Characteristic

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

$$\begin{aligned} c(q^n - 1) &= \sum_{\substack{k \geq 1 \\ q^k - 1 \mid q^n - 1}} 1 \\ &= \sum_{\substack{k \geq 1 \\ k \mid n}} 1 \\ &= d(n), \end{aligned}$$

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. ■