Formal Model Research Group Seminar
Brno University of Technology

Brno, 15 june 2015

Finite automata and number theory

Christian MAUDUIT

Université Aix-Marseille et Institut Universitaire de France,

Institut de Mathématiques de Marseille UMR 7373 CNRS, France.

Motivation

The difficulty of the transition from the representation of an integer in a number system (e.g. $rep_{10}(n) = 19605131$) to its multiplicative representation (as a product of prime factors : $n = 7 \times 13 \times 17 \times 19 \times 23 \times 29$) is at the origin of many important open problems in mathematics and computer science.

Our talk concerns the study of independence between multiplicative functions and functions defined using a simple algorithm or ”deterministic” functions.

A simple remark

The study of

subsets E_i of $\mathbb{N}, (i \in \{1, ..., g\})$

is linked to the study of

infinite sequences of symbols over $A = \{a_1, ..., a_g\}$,

$w = w_0, w_1, ..., w_n, ...$ with $w_n \in A$.

Example 1. $E_1 = 2\mathbb{N}, E_2 = 2\mathbb{N} + 1$ and $w = (n \mod 2)_{n\in\mathbb{N}}$.

Representation of integers in base q

If q is an integer greater than or equal to 2, any positive integer n can be written in a unique way in base q in the form $n = \sum_{j=0}^{\ell} n_j q^j, \quad n_j \in \{0, \ldots, q-1\}, \quad n_\ell \geq 1$ and we denote by $rep_q(n) = n_\ell \ldots n_0 \in \{0, \ldots, q-1\}^*$ the representation of n in base q (for any finite alphabet A, we denote by A^* the set of finite words over A).

To any $E \subset \mathbb{N}$ we associate the language $L_q(E) = \{rep_q(n), n \in E\} \subset \{0, \ldots, q-1\}^*$.

Many questions concerning arithmetic sequences can then be expressed in the framework of the theory of formal languages, thus establishing a link between number theory, language theory and combinatorics on words.

Finite automata

Definition 1. A finite q-automaton is a quadruplet $\mathcal{A}_q = (S, S_f, s_0, \varphi)$ with

i) S is a finite alphabet, (states) ;

ii) $S_f \subset S$, (final states) ;

iii) $s_0 \in S$, (initial state) ;

iv) φ a map from $S \times \{0, ..., q-1\}$ to S.

For any $(s, d) \in S \times \{0, ..., q-1\}$, we put $\varphi(s, d) = s.d$ and we extend φ into a map $\tilde{\varphi}$ from $E \times \mathbb{N}$ to E in the following way : if $rep_q(n) = n_\ell \ldots n_0$, then we put for any $s \in S$,

$$\tilde{\varphi}(s, n) = s.n = (\ldots((s.n_\ell).n_{\ell-1})\ldots).n_0.$$

Definition 2. We say that $E \subset \mathbb{N}$ is recognizable by the finite q-automaton $\mathcal{A}_q = (S, S_f, s_0, \varphi)$ if

$$E = \{n \in \mathbb{N} | \tilde{\varphi}(s_0, n) \in S_f\}.$$

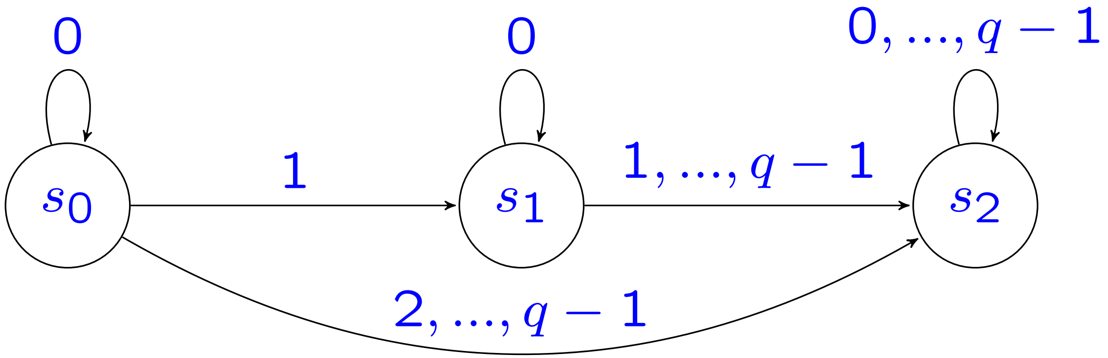
Graph associated to a finite automaton

The labeled graph $\mathcal{G}(\mathcal{A}_q)$ associated to the finite automaton $\mathcal{A}_q$ is the graph $\mathcal{G}(\mathcal{A}_q) = (S, U)$ for which
i) S is the set of vertices of $\mathcal{G}(\mathcal{A}_q)$;
ii) $U = \{(s, s', d) \in E \times E \times \{0, ..., q-1\} | \varphi(s, d) = s'\}$ is the set of labeled edges of $\mathcal{G}(\mathcal{A}_q)$.

Example 2. The set $\{q^n, n \in \mathbb{N}\}$ is recognizable by the finite q-automaton

$$\mathcal{A}_q = \{\{s_0, s_1, s_2\}, \{s_1\}, s_0, \varphi\}$$

whose graph $\mathcal{G}(\mathcal{A}_q)$ is



Christol, Kamae, Mendes-France and Rauzy theorem (1980)

G. Christol, Kamae, Mendes-France and Rauzy, *Suites algébriques, automates et substitutions,* Bull. Soc. Math. France 108 (1980), 401-419.

Theorem 1. *The formal power series* $\sum_{n\in E} X^{-n} \in \mathbb{F}_q[[X^{-1}]]$ *is algebraic over* $\mathbb{F}_q(X)$

if and only if

E *is recognizable by a finite* q*-automaton.*

Example 3. If $E = \{q^n, n \in \mathbb{N}\}$, then

$$f(X) = \sum_{n\in E} X^{-n} = \sum_{n\geq 0} X^{-q^n} = X^{-1} + f(X^q) = X^{-1} + f(X)^q.$$

Prime numbers

The prime numbers constitute a fascinating sequence which poses many difficult questions. Let us mention some known results and some open problems :

Known prime numbers

- primes of the form $an + b$ (Dirichlet theorem) ;
- primes such that $\alpha p \pmod 1$ belongs to some prescribed interval $I \in [0, 1]$, for $\alpha \in \mathbb{R} - \mathbb{Q}$ (Vinogradov-Davenport theorem) ;
- primes of the form $[n^c]$ where $1 < c < c_0 \approx 1.1$ (Piatetski-Shapiro theorem) ;
- primes of the form $a^2 + b^4$ (Friedlander-Iwaniec theorem) ;
- primes of the form $a^3 + 2b^3$ (Heath-Brown theorem) ;
- arbitrarily long arithmetic progressions of primes (Green-Tao theorem).

Prime numbers

Open problems

- are there infinitely many primes of the form $p+2$ (prime twins) ?
- are there infinitely many primes of the form n^2+1 ?
- are there infinitely many primes of the form 2^n-1 (Mersenne primes, i.e. primes with no digit 0 in their representation in base 2) ?
- are there infinitely many primes of the form $2^{2^n}+1$ (Fermat primes, i.e. primes with exactly two digits 1 in their representation in base 2) ?

Prime numbers and finite automata

If $E = \mathbb{P}$ is the set of primes, it is natural to ask whether there is a simple algorithm for deciding whether a given integer n does belong to E or not.

Minsky and Papert had shown in 1966 that $L_q(\mathbb{P})$ is never a rational language, i. e. the set of prime numbers is not recognizable by a q-finite automaton. This fundamental result has been generalized by Hartmanis and Shank and Schützenberger in 1968, showing that no infinite subset of primes is recognizable by a finite automaton (or even by a pushdown automaton).

q-finite automaton can be identified to special cases of morphisms (or substitutions) on a finite alphabet : the morphisms (or substitutions) of constant length equal to q. Mauduit showed in 1992 that the set of prime numbers can not be generated by a morphism (or substitution) on a finite alphabet and introduced in 2006 a notion of q-infinite automaton for which Cassaigne and Le Gonidec showed the non-recognizability of the set of primes.

On the other hand, this question has received a new light with the development by Agrawal, Kayal and Saxena in 2004 of a polynomial time algorithm to solve it.

Prime numbers in q-automatic sets

The search of prime numbers in a set recognizable by a q-finite automaton is a problem in general extremely difficult. Thus, the sets $\{2^n + 1, n \in \mathbb{N}\}$ and $\{2^n - 1, n \in \mathbb{N}\}$ are both recognizable by a 2-finite automaton and the associate problems correspond respectively to the research of Fermat and Mersenne primes.

When E is a set recognizable by an irreducible q-finite automaton (i. e. that the graph of the automaton is strongly connected), it follows from a remark due to Fouvry-Mauduit (1996) that the set E contains infinitely many almost prime numbers But the following problem remains open :

Problem 1. For any $E \subset \mathbb{N}$ recognizable by an irreducible q-finite automaton, find an asymptotic estimate of the number of primes less than x belonging to the set E.

When the finite automaton is not irreducible, the situation becomes very difficult. The first problem to be studied in this direction concerns probably the search of prime numbers with missing digits.

The Cantor sequence

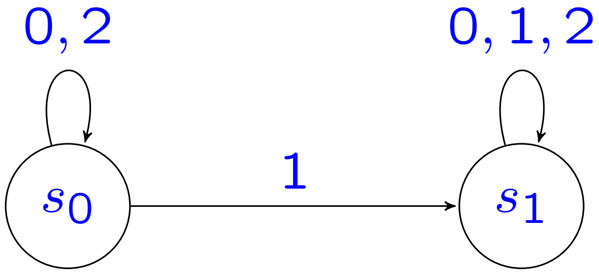
$c \in \{0,1\}^{\mathbb{N}}$ is the fixed point of the substitution σ defined on the alphabet $\{0,1\}$ by

$$\sigma(0) = 010 \quad \text{and} \quad \sigma(1) = 111.$$

We have :

$$c = (c_n)_{n\in\mathbb{N}} = 0101110101111111110101110101111111111111111111.......$$

It is an easy exercise to check that for any integer $n \in \mathbb{N}$ we have $c_n = 0$ if and only if $rep_3(n) \in \{0,1\}^*$, so that he set $\{n \in \mathbb{N}, c_n = 0\}$ is recognizable by the finite 3-automaton $\mathcal{A}_3 = \{\{s_0, s_1\}, \{s_0\}, s_0, \varphi\}$ whose graph $\mathcal{G}(\mathcal{A}_3)$ is



Integers with missing digits

Erdős, Mauduit and Sárközy studied the repartition in residues classes of integers with missing digits (1998) and showed an Erdős-Kac inequality for the function ω (counting the number of distinct prime factors) resticted to integers with missing digits (1999).

But we could not solve the following problem that remains still open :

Problem 2. For any given $D \subset \{0, \ldots, q-1\}$, find an asymptotic estimate for $\text{card}\{p \leq x,\ rep_q(p) \in D^*\}$.

The Thue-Morse sequence

$t \in \{0,1\}^{\mathbb{N}}$ is the fixed point of the substitution σ defined on the alphabet $\{0,1\}$ by

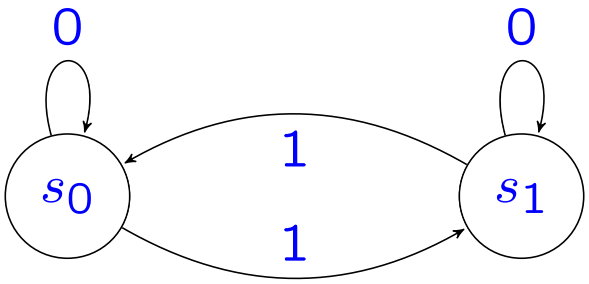
$$\sigma(0) = 01 \quad \text{and} \quad \sigma(1) = 10.$$

We have :

$$t = (t_n)_{n\in\mathbb{N}} = 0110100110010110100101100110100 1.......$$

If $n = \sum_{j\geq 0} n_j q^j$ with $n_j \in \{0, ..., q-1\}$ is the representation of the integer n in base q, then the sum of digits in base q function is defined by $\mathsf{s}_q(n) = \sum_{j\geq 0} n_j$.

It is an easy exercise to check that for any integer $n \in \mathbb{N}$ we have $t_n \equiv s_2(n) \bmod 2$, so that he set $\{n \in \mathbb{N}, t_n = 1\}$ is recognizable by the finite 2-automaton $\mathcal{A}_2 = \{\{s_0, s_1\}, \{s_1\}, s_0, \varphi\}$ whose graph $\mathcal{G}(\mathcal{A}_2)$ is



Combinatorial properties of the Thue-Morse sequence

Theorem 2. *(Thue 1912) The sequence t doesn't contain any subword of the form WWw where w is the first letter of the word W.*

Theorem 3. *(Morse 1921) The sequence t is non periodic but any subword occuring in t occurs infinitely often with bounded gaps.*

Let us define, for any infinite sequence $w = w_0 w_1 \dots w_k \dots \in \{0,1\}^{\mathbb{N}}$ and for any non negative integer n, the complexity function p_w by :

$$p_w(n) = \text{number of distincts blocks of lenght} \quad n \text{ occuring in} \quad w$$

$$Card\{(b_1, \dots, b_n) \in \{0,1\}^n, \quad \exists k \quad \text{s.t.} \quad w_k w_{k+1} \dots w_{k+n-1} = b_1 \dots b_n\}.$$

For the Thue Morse sequence, we have $c_1 n \le p_t(n) \le c_2 n$.

Resolution of the Gelfond conjecture for prime numbers

C. Mauduit and J. Rivat, *Sur un problème de Gelfond : la somme des chiffres des nombres premiers*, Annals of Math., vol. 171(2010), 1591-1646.

The following theorem answers a question asked in 1968 by Gelfond concerning the sum of digits of prime numbers.

Theorem 4. *(Mauduit-Rivat, 2010) For any $\alpha \in \mathbb{R}$ such that $(q-1)\alpha \in \mathbb{R} \setminus \mathbb{Z}$, there exists $\sigma_q(\alpha) > 0$ such that for any $x \geq 1$,*

$$\sum_{p \leq x} \exp\big(2i\pi\alpha s_q(p)\big) \ll_{q,\alpha} x^{1-\sigma_q(\alpha)}.$$

Corollary 1. *The frequencies of 0 and 1 in the sequence*

$$t_{\mathbb{P}} = (t_p)_{p \in \mathbb{P}} = 10011101001011010.......$$

are equal to $\frac{1}{2}$.

A local theorem

M. Drmota, C. Mauduit and J. Rivat, *Prime numbers with an average sum of digits*, Compositio Mathematica 145 (2009), 271-292.

Theorem 5. *(Drmota-Mauduit-Rivat, 2009) We have, uniformly for any positive integer* $k \geq 0$ *such that* $(k, q-1) = 1$

$$\mathrm{card}\{p \leq x : s_q(p) = k\} = \frac{q-1}{\varphi(q-1)} \frac{\pi(x)}{\sqrt{2\pi\sigma_q^2 \log_q x}} \Bigg(\exp(-\frac{(k-\mu_q \log_q x)^2}{2\sigma_q^2 \log_q x}) + O((\log$$

with $\mu_q := \frac{q-1}{2}, \quad \sigma_q^2 := \frac{q^2-1}{12}$ *and any given* ε*.*

This theorem provides a local version of both Copeland-Erdős normality theorem (1946) and of Bassily-Katai central limit theorem (1995). It follows from Theorem 6 that the number of primes whose binary representation contains n digits 0 and n digits 1 is asymptotically equal to

$$\frac{4^{n-1}}{\sqrt{\pi} \log 2 n^{3/2}}.$$

But the following problem remains open :

Problem 3. Find an asymptotic estimate of the number of primes whose binary representation contains $2n$ digits 0 and n digits 1.

The Rudin-Shapiro sequence

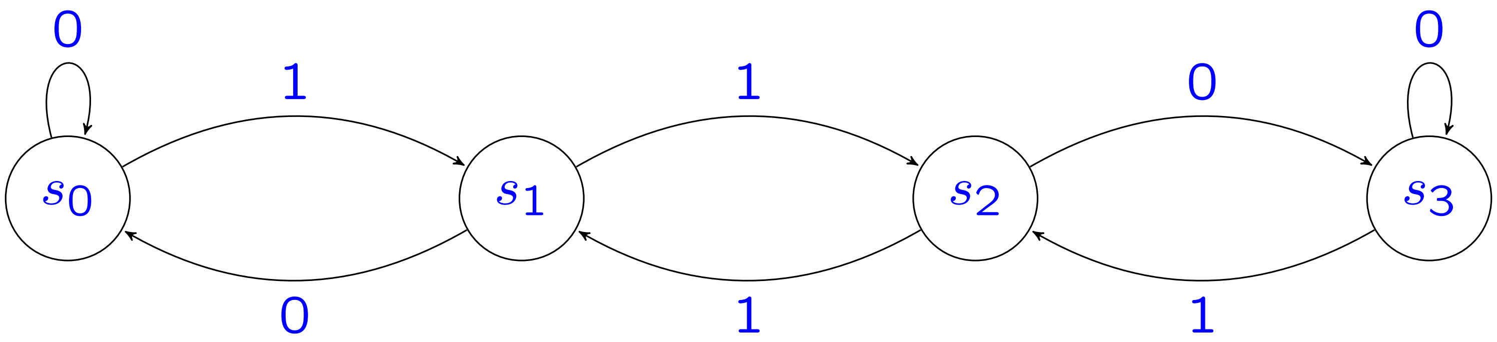
$r \in \{0,1\}^{\mathbb{N}}$ is the projection of the fixed point ρ of the substitution σ defined on the alphabet $\{a,b,c,d\}$ by $\sigma(a) = ab, \sigma(b) = ac, \sigma(c) = db$ and $\sigma(d) = dc$. We have :

$$\rho = abacabdbabacdcacabacabdb.......$$

and

$$r = (r_n)_{n\in\mathbb{N}} = 000100100001110100010010.......$$

It is an easy exercise to check that if $rep_2(n) = n_\ell \ldots n_0 \in \{0,1\}^*$ is the representation of n in base 2 , then $r_n \equiv \sum_{0\le j<\ell} n_j n_{j+1} \bmod 2$, so that he set $\{n \in \mathbb{N}, r_n = 1\}$ is recognizable by the finite 2-automaton $\mathcal{A}_2 = \{\{s_0,s_1,s_2,s_3\},\{s_2,s_3\},s_0,\varphi\}$ whose graph $\mathcal{G}(\mathcal{A}_2)$ is



Prime number theorem for Rudin-Shapiro-sequence

C. Mauduit and J. Rivat, *Prime numbers along Rudin-Shapiro sequences*, Journal of the European Mathematical Society (to appear).

Theorem 6. *(Mauduit-Rivat, 2015) There exists* $\sigma > 0$ *such that for any* $\theta \in \mathbb{R}$ *and* $x \geq 1$,

$$\sum_{p \leq x} r_p \exp(2i\pi p\theta) \ll x^{1-\sigma}.$$

Corollary 2. *The frequencies of* 0 *and* 1 *in the sequence*

$$r_{\mathbb{P}} = (r_p)_{p \in \mathbb{P}} = 010011010.......$$

are equal to $\frac{1}{2}$.

Computational complexity of the Möbius function

We denote by μ the Möbius function, defined by $\mu(1) = 1, \mu(p_1 \dots p_k) = (-1)^k$ if $p_1, \dots, p_k$ are distinct prime numbers and $\mu(n) = 0$ if n is divisible by the square of a prime number. The function μ is multiplicative, i. e. $\mu(mn) = \mu(m)\mu(n)$ for any coprime positive integers m and n and μ^2 is the characteristic function of square-free numbers.

$$\mu = (\mu(n))_{n\geq 1} = 1, -1, -1, 0, -1, 1, -1, 0, 0, 1, -1, 0, -1, 1, 1, 0, -1, 0, -1, 0, \dots$$

The results and problems presented in this talk show the independence between the multiplicative property "to be a prime number" and q-automatic properties. They are naturally connected to the Möbius randomness principle for q-automatic sequences $\mathbf{u}$. This principle often stated vaguely in the literature, says that "for any reasonable sequence $\mathbf{u} = (u(n))_{n\in\mathbb{N}}$ of complex numbers, the sum $\sum_{n\leq x} \mu(n)u(n)$ is relatively small ".

Green and Bourgain theorems

Some recent works of Green and Bourgain concern the particular case where $\mathbf{u}$ is a sequence with values in the finite alphabet $A = \{-1, 1\}$ and are motivated by the study of the computational complexity of μ. The goal is to prove the orthogonality of the Möbius function with certain classes of Boolean functions in relation with a series of questions stated by Kalai on his blog.

Theorem 7. *(Green, 2012) If $\mathbf{u} \in \{-1, 1\}^{\mathbb{N}}$ is computable by a Boolean function representable by a circuit of depth at most d and size at most n^d, then*

$$\sum_{n<2^\nu} \mu(n)u(n) = O(2^\nu \exp(d \log \nu - \nu^{1/6d})). \tag{1}$$

From a result of Linial-Mansour-Nisan, the problem turns into giving good estimates for the Fourier-Walsh transform$\sum_{n<2^\nu} \mu(n)(-1)^{s_E(n)}$, where $E \subset \mathbb{N}$ verifies $\text{card}\, E = O(\frac{\sqrt{\nu}}{\log \nu})$ and s_E is the restricted sum of binary digits function, defined by $s_E(n) = \sum_{\substack{j \le \ell \\ j \in E}} n_j$ if $rep_q(n) = n_\ell \ldots n_0$.

By generalizing the method introduced by Mauduit-Rivat to study the extreme case where $E = \mathbb{N}$ in the proof of Theorem 5, Bourgain extended in 2012 the estimate (1) to any set E by showing that

$$\max_{E \subset \{0,\ldots,\nu-1\}} \sum_{n<2^\nu} \mu(n)(-1)^{s_E(n)} = O(2^{\nu-\nu^{1/10}}). \tag{2}$$

Moreover, by studying precisely the distribution of these Fourier-Walsh coefficients, Bourgain deduced in 2013, using a result of Bshouty-Tamon concerning the localization of the Walsh-Fourier spectrum of monotone Boolean functions, a proof of the orthogonality of these functions with the Möbius function and in very recent paper a lower bounds for the number of prime numbers captured by these functions.