Can a fast signature scheme without secret key be secure

Paul Camion

INRIA

CENTRE DE ROCQUENCOURT

Rapports de Recherche

N° 399

CAN
A FAST SIGNATURE SCHEME WITHOUT SECRET KEY BE SECURE

Paul CAMION

Mai 1985

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P. 105
78153 Le Chesnay Cedex
France
Tél.: (3) 954 90 20

Can a Fast Signature Scheme without Secret Key be Secure

Paul CAMION

CNRS, INRIA

Mars 1985

Can a fast signature scheme without secret key be secure ?

Abstract

Another title could have been "A probabilistic factorization algorithm in GL(2,p)". However, the problem is to calculate a fast and short signature associated with a plaintext inscribed on an erasable support. The signature should be written down in a book accompanying the record in order that it could be checked anytime that the latter has not been changed. J. Bosset [1] suggest such a scheme together with an algorithm for computing a signature. The 64 characters needed for the plaintext are identified with a subset of GL(2,p), p = 997. The signature is the product of the matrices corresponding to the plaintext characters taken in the order where they appear. Such a scheme could be broken if it is possible to factorize an element of GL(2,p) into t = 16 r factors, each one in a subset U_i of GL(2,p) of size 64, i = 1, ...,t. We here assume one hypothesis only on uniform probability distributions of random variables defined on product sets $V_j = U_{jr+1} \times \ldots \times U_{(j+1)r}$, j = o , ..., 15. In consideration on which, a probabilistic factorization algorithm in GL(2,p) is introduced.

It is shown that for p = 10,007, drawing according to a uniform probability distribution a sequence of 11,952 elements in each V_j provides the whole needed material to factorizing with a probability of succes of at least 97%. The most expensive operation in the algorithm is sorting each of the sequences.

Un schéma de signature courte et rapide n'utilisant pas de clé secrète peut-il être fiable ?

Résumé

Nous étudions une solution proposée au problème d'une signature courte, calculable rapidement, associée à un texte inscrit sur un support effaçable. La (ou les) signature serait inscrite dans un livre accompagnant l'enregistrement de façon à pouvoir vérifier à tout moment que ce dernier n'a pas été modifié. J. Bosset [1] propose un tel schéma et un algorithme de signature. Il identifie l'ensemble des 64 caractères utilisés pour le texte à un sous-ensemble de GL(2,p), p = 997. La signature est le produit des matrices correspondant aux caractères du texte effectué dans l'ordre où ceux-ci apparaissent. On peut casser ce schéma s'il est possible de factoriser un élément de GL(2,p) en t = 16 r facteurs, chacun dans un sous-ensemble $\mathbf{U}_i$ de taille 64 de GL(2,p), i = 1, ...,t. Nous faisons ici une seule hypothèse de distributions de probabilités uniformes de variables aléatoires définies sur les ensembles produits $\mathbf{V}_j = \mathbf{U}_{jr+1} \times \ldots \times \mathbf{U}_{(j+1)r}$, j = o, ..., 15. Moyennant ceci, un algorithme probabiliste de factorisation dans GL(2,p) est introduit.

On montre que pour p = 10.007, un tirage selon une loi uniforme d'une suite de 11.052 éléments dans chacun des $\mathbf{V}_j$ fournit tout le matériel nécessaire à la factorisation avec une probabilité de succès supérieure à 0,97. L'opération la plus coûteuse est un tri de chacune de ces suites.

INTRODUCTION

Le problème d'une signature courte, calculable rapidement, associée à un texte inscrit sur support effaçable nous a été exposé par M. GORIN, responsable d'un groupe d'étude de la commission de l'ordinateur de compensation. Pratiquement, il s'agit de s'assurer après le transfert, par exemple d'une bande magnétique, que les données inscrites n'ont pas été modifiées entre la source et l'arrivée. La solution proposée par M. BOSSET [1] avait attiré l'attention de M. GORIN pour son intérêt pratique remarquable. En résumé, il s'agit de sectionner le texte, c'est-à-dire la suite des caractères inscrits, en blocs de, disons, 10.000 caractères et d'associer à chacun de ces blocs un nombre d'au moins 20 chiffres décimaux(*) nommé signature et calculé sur la donnée du bloc de telle sorte que tout contrôle ultérieur ferait apparaître une signature totalement différente de l'originale si une modification, même mineure, des caractères du bloc était survenue. On veut évidemment être assuré de la sécurité du test que constitue l'identité de la signature calculée à la signature originale. Le test prend d'ailleurs le nom de certification. Mais il faut aussi que ce calcul soit très rapide. Pour fixer les idées, il faut pouvoir calculer les signatures relatives à 175.000 caractères en une seconde d'unité centrale de gros ordinateur dans le cas de compensations bancaires.

M. BOSSET propose également d'être en mesure de calculer la signature d'une suite de signatures concaténées. De cette façon on pourrait "résumer" plusieurs millions de caractères en quelques centaines de chiffres décimaux qui occuperaient peu d'espace sur un support non effaçable où l'on pourrait vérifier, comme c'est l'usage pour les livres comptables, qu'il n'y a eu ni ratures ni rajouts. Si une certification devait faire défaut on pourrait demander à la source une retransmission.

(*) En vérité, M. BOSSET suggérait une signature de 12 chiffres. Compte-tenu de la possibilité d'attaques probabilistes valables pour tout schéma de signature, y compris pour un schéma utilisant des quantités secrètes, il semble raisonnable aujourd'hui de ne pas se fier à une signature de moins de 20 chiffres décimaux.

D'autres applications d'un tel schéma de signature courte et rapide sont proposées par M. BOSSET. Il s'agit essentiellement, dit-il, de prouver qu'une information n'a pas été altérée depuis sa création. Le schéma est donc applicable à tout fichier informatique dont les signatures seraient inscrites au "grand livre", et exploitées à toute édition pour certification.

M. BOSSET [1] propose dans son article un schéma particulier de signature et l'objet du présent article est de démontrer que ce schéma n'est pas fiable. Le problème reste donc posé de trouver un schéma de signature courte, rapide et qui n'imposerait pas la gestion de clés secrètes.

I. - THE GENERAL CONSIDERED SCHEME OF SIGNATURE

1.1. - The mapping

Let X be an alphabet, say, of 64 characters. Let k be an integer which will be the fixed length of a plaintext. Then k may be choosen once for all, say, from 1,000 to 10,000. A plaintext is considered to be any element from X^k. We then define a mapping σ from X^k onto the set $\{0,1,\ldots,9\}^\ell$ where ℓ is an integer. For instance $\ell = 24$. Thus for b in X^k, then $\sigma(b) = s$ is the signature of b. The signature is then an integer of 24 decimal digits.

1.2. - The aim of the considered signature scheme

The plaintext is supposed to be easily erasable or possible changed to another one. We then intend to compute the sequence of signatures corresponding to the given sequence of plaintexts. The sequence of signature is written down in a book and we assume that a fake is as unlikely to be forged as for usual writings. Hence if it was practically impossible to change a plaintext to another one with the same signature, then the only needed precaution to be taken for guaranteering data recorded, say, on a magnetic tape, would be to join with it a book containing references to all plaintexts and the corresponding signatures.

1.3. - The requirements

1.3.1. - The signature should be easily computed

1.3.2. - The probability that two plaintext have the same signature should be close to $10^{-\ell}$

1.3.3. - It should be practically impossible to change the plaintext to another one having the same signature, by any means.

1.4. - A suggested signature

1.4.1. - Computing in the group GL(2,p)

J. Bosset [1] suggest the above general scheme. He also suggest a particular function σ and the aim of the present paper is to show that σ does not fulfil requirement 1.3.3.. The alphabet X is identified with a subset of size 64 of GL(2,p) for p = 997. Hence a plaintext b is a sequence $(b_1,\ldots,b_k)$ of matrices from X and $\sigma(b)$ is nothing else but the product $b_1 \ldots b_k$.

1.4.2. - Examining the requirements

Obviously 1.3.1. is fulfilled. The author carefully chooses the subset X of GL(2,p) in view of 1.3.2. In regard to 1.3.3., the author observes that forging a false need being able to factorize any s from GL(2,p) into matrices from the small subset X. We will set the problem of forging a false by such a factorization in the next paragraph. Then in section 2 we solve the factorization problem by means of a probabilistic algorithm.

1.4.3. - Forging a fake by factorization

We denote by G the group GL(2,p) of invertible 2 by 2 matrices with entries in the finite field $\mathbb{F}_p$, p a prime. We will keep $|X| = 64$ as suggested by J. Bosset, but we make p = 10,007 to make the problem somewhat harder. An integer t will appear in the following. For the algorithm, t will be a multiple of 16. We suggest t = 48 but t = 64 could possibly provide better results for the required statistical tests.

1.4.3.1. - The considered fraud

Given a plaintext

$(b_1, b_2, \ldots, b_{i_1-1}, \ldots, b_{i_t}, \ldots, b_k) \in X^k$

of which the signature is s :

$b_1 b_2 \cdots b_k = s$,

and where $i_1, i_2, \ldots, i_t$ is any given subset of size t of $[0,k]$, the fraud is as follows. I change the whole plaintext for a new one at the exception of characters in positions $i_1, \ldots, i_t$. The new plaintext is of my own choice. Then it will be possible to adapt the values of matrices $b_{i_1}, \ldots, b_{i_t}$ in order that the new plaintext writes

$(b'_1, \ldots, b'_{i_1-1}, x_1, b'_{i_1+1}, \ldots, b'_{i_2-1}, x_2, b'_{i_2+1}, \ldots, b'_{i_t-1}, x_t, b'_{i_t+1}, \ldots; b'_k)$

having the same signature s.

1.4.3.2- The factorization problem giving the solution

Let X and t be given as before as well as t elements $u_1, \ldots, u_t$ from G.

Find a solution $(y_1, y_2, \ldots, y_t)$ *to*

$$(1) \qquad \begin{aligned} & y_1 y_2 \cdots y_t = 1 \\ & y_1 \in u_1 X,\ y_2 \in u_2 X,\ \ldots,\ y_t \in u_t X \end{aligned}$$

in feasible time.

1.4.3.3 . - How the factorization solution solves the fraud problem

The chosen new characters being

$$b'_1,\ldots,b'_{i_1-1},\ b'_{i_1+1},\ \ldots,b'_{i_2-1},\ \ldots,b'_{i_t-1},\ b'_{i_t+1},\ldots,b'_k,$$

we define

$$u_1 = b'_{i_t+1}\ b'_{i_t+2} \cdots b'_k\ s^{-1}\ b'_1 \cdots b'_{i_1-1}$$

$$u_2 = b'_{i_1+1} \cdots b'_{i_2-1},$$

$$u_t = b'_{i_{t-1}+1},\cdots b'_{i_t-1}.$$

Now by solving (1) the t unknown values $x_1, \ldots,x_t$ of the new plaintext are obtained,

$$x_1 = u_1^{-1} y_1,\ x_2 = u_2^{-1}\ y_2,\ \ldots,\ x_t = u_t^{-1}\ y_t.$$

The result is then straightforward.

1.4.4. - Toward a probabilistic algorithm

We will here introduce the ideas which preside over the settling of a clean probabilistic model used for solving (1).

Let us write t = 16r and

$$y_1 \cdots y_r = a_1,\ y_{r+1} \cdots y_{2r} = a_2,\ \ldots,\ y_{t-r+1} \cdots y_t = a_t$$

After, put moreover

$$a_1\ a_2 = b_1,\ a_3\ a_4 = b_2,\ \ldots,\ a_{15}\ a_{16} = b_8\ ;$$

$$b_1\ b_2 = c_1,\ b_3\ b_4 = c_2,\quad ,\ b_7\ b_8\ = c_4\ ;$$

$$c_1\ c_2 = d_1,\ c_3\ c_4 = d_2.$$

We now consider the chain of subgroups

$$H_o \leq H_1 \leq H_2 \leq H_3 \leq H_4 = G$$

where $H_o = \{1\}$, H_1 is the group of lower triangular matrices of the form $\begin{bmatrix} 1 & 0 \\ b & 1 \end{bmatrix}$, H_2 is the group of lower triangular matrices of the form $\begin{bmatrix} a & 0 \\ b & 1 \end{bmatrix}$ and H_3 is formed by the matrices of the form $\begin{bmatrix} a & 0 \\ b & d \end{bmatrix}$, $ad \neq 0$. Now the idea is to find a solution for (1) in which $a_1,\ldots,a_{16}$ are matrices of G such that $b_1,\ldots,b_8$ all are in H_3, and $c_1,\ldots,c_4$ are in H_2 and d_1, d_2 in H_1. This is actually a new constraint to the problem. However this permits breaking the algorithms into independent steps each step being easier than finding right away a sequence $y_1, \ldots,y_t$, $y_i \in u_i X$, $i = 1, \ldots t$ verifying $y_1 \ldots y_t = 1$.

Denote $u_i X$ by $\mathbf{U}_i$, $i = 1, \ldots,t$. The general idea is as follows. Given the sets $A_1 \subset \mathbf{U}_1 \times \ldots \times \mathbf{U}_r$ and $A_2 \subset \mathbf{U}_{r+1} \times \ldots \times \mathbf{U}_{2r}$, find a subset $B_1 \subset A_1 \times A_2$ such that the product of the components of each element in B_1 lies in H_3. Construct similar sets $B_2, \ldots,B_8$. If these sets are large enough, we will find a subset C_1 of $B_1 \times B_2$ such that the product of the components of each element in C_1 lies in H_2. We construct similar sets C_2, C_3, C_4. We then find $D_1 \subset C_1 \times C_2$ and $D_2 \subset C_3 \times C_4$ such that the product of the components of each element in D_i, $i = 1, 2$ lies in H_1. Finally we only need one element in D_1 and one in D_2 to form $(y_1,\ldots,y_t)$ such that $y_1 y_2 \ldots y_t = 1$ and moreover $y_i \in u_i X$, $i = 1, \ldots,t$. This looks feasible since apparently the probability that the product of two elements from G lies in H_3 is close to $1/(p+1) = (p-1)^2 p/(p^2-1)(p^2-p)$. Thus an average number of p+1 tries should give one success. Similar consideration lead to obtaining products from H_3 in H_2 and so on. Hence it seems that if the size of A_i, $i=1,\ldots,16$ is large enough, we may succeed within a reasonable number of computations. The problem is the settling of a probabilistic model that enable us to predict the issues of the process. Thatfor, it appeared easier to consider not subsets A_i, B_i, ... but sequences $(\varphi_1, \ldots,\varphi_n)$ and $(\psi_1, \ldots, \psi_n)$ with elements from $\mathbf{U}_1 \times \ldots \times \mathbf{U}_r$ and from $\mathbf{U}_{r+1} \times \ldots \times \mathbf{U}_{2r}$ respectively as a first step of the algorithm and similar sequences for

the other steps. This is clearly settled in section two. There, probabilistic properties of those sequences are verified. This allows the computation of the sizes of the sequences needed for a given probability of success of the algorithm. This is done in section 3. A numerical example is dealt with at the end of section 2.

II. - RANDOM VARIABLES WITH UNIFORM PROBABILITY DISTRIBUTIONS RELATED TO GL(2,p).

2.1. - The Basic lemma

2.1.1. - Lemma

The mapping $\eta : GL(2,p) \to \mathbb{F}_p \times \mathbb{F}_p^* \times \mathbb{F}_p^* \times (\mathbb{F}_p \cup \{\infty\})$ *defined by*

$\eta((x_{ij})) = (x_{21},\ x_{11},\ x_{22},\ \infty)$ *when* $x_{12} = 0$ *and*

$\eta((x_{ij})) = (x_{22}\ x_{12}^{-1},\ x_{12},\ x_{21} - x_{11}\ x_{22}\ x_{12}^{-1},\ x_{11}\ x_{12}^{-1})$ *otherwise, is one-to-one.*

Let us consider the following chain of subgroups of GL(2,p)

$$H_o \leq H_1 \leq H_2 \leq H_3 \leq G = H_4.$$

Where H_o reduces to the identify ; H_1 is the group of lower triangular matrices of the form $\begin{bmatrix} 1 & 0 \\ b & 1 \end{bmatrix}$ which is isomorphic to $(\mathbb{F}_p,+)$; H_2 is the group of triangular matrices of the form $\begin{bmatrix} a & 0 \\ b & 1 \end{bmatrix}$, $(b,a) \in \mathbb{F}_p \times \mathbb{F}_p^*$ and H_3 is formed by those matrices of the form $\begin{bmatrix} a & 0 \\ b & d \end{bmatrix}$, $(b,a,d) \in \mathbb{F}_p \times \mathbb{F}_p^* \times \mathbb{F}_p^*$.

On the other hand let us denote by E_1 the set H_1, by E_2 the set of matrices of the form $\begin{bmatrix} a & 0 \\ 0 & 1 \end{bmatrix}$, $a \in \mathbb{F}_p^*$ and by E_3 the set of matrices of the form $\begin{bmatrix} 1 & 0 \\ 0 & d \end{bmatrix}$, $d \in \mathbb{F}_p^*$. Notice that each of these sets forms a group. Finally E_4 will denote the set of matrices of the form $\begin{bmatrix} u & 1 \\ 1 & 0 \end{bmatrix}$, $u \in \mathbb{F}_p$, together with the identity matrix.

We clearly have $H_1 = E_1$, $H_2 = H_1\ E_2$, $H_3 = H_2\ E_3$. Thus $H_3 = E_1\ E_2\ E_3$ and here is defined a natural one-to-one mapping from H_3 onto $\mathbb{F}_p \times \mathbb{F}_p^* \times \mathbb{F}_p^*$:

$$\begin{bmatrix} a & 0 \\ b & d \end{bmatrix} \sim\!\!\!\!- \begin{bmatrix} 1 & 0 \\ ba^{-1} & 1 \end{bmatrix} \begin{bmatrix} a & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & d \end{bmatrix} \sim\!\!\!\!- (ba^{-1}, a, d)$$

If we now prove that the mapping of $H_3 \times E_4$ into G defined by $(x,y) \to xy$ is surjective, then the lemma will be proved since then

$$(p^2-1)\ (p^2-p) = |G| = |H_3\ E_4| \leq |H_3 \times E_4| = p(p-1)^2(p+1).$$

We thus consider any matrix $\begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix}$ of G and we show that it is obtained by the product of a matrix of H_3 by a matrix of E_4. This is obvious when $x_{12} = 0$. Otherwise we have that

$$\begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix} = \begin{bmatrix} x_{12} & 0 \\ x_{22} & x_{21}-x_{11}\,x_{22}\,x_{12}^{-1} \end{bmatrix} \begin{bmatrix} x_{11}\,x_{12}^{-1} & 1 \\ 1 & 0 \end{bmatrix} \qquad \square$$

Remarks

The inverse of the mapping given by the lemma is

$$\text{for } d = \infty : \quad (a,b,c,d) \rightsquigarrow \begin{bmatrix} b & 0 \\ ab & c \end{bmatrix} ;$$

$$\text{for } d \in \mathbb{F}_p : \quad (a,b,c,d) \rightsquigarrow \begin{bmatrix} db & b \\ da+c & a \end{bmatrix}$$

The mapping $x \to (\eta_1(x),\eta_2(x),\eta_3(x),\eta_4(x)) = \eta(x)$ clearly defines four random variables from G onto $\mathbb{F}_p$, $\mathbb{F}_p^*$, $\mathbb{F}_p^*$ and $\mathbb{F}_p \cup \{\infty\}$ respectively for any probability distribution defined on G.

2.1.2. - Right and left mappings

Let us call right mapping the mapping of the lemma. We define similarly a mapping from G onto $\{\mathbb{F}_p \cup \{\infty\}\} \times \mathbb{F}_p^* \times \mathbb{F}_p$ from the factorization of y for $y_{12} \neq 0$:

$$y = \begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & y_{22}\, y_{12}^{-1} \end{bmatrix} \begin{bmatrix} y_{21}-y_{11}\, y_{22}\, y_{12}^{-1} & 0 \\ y_{11} & y_{12} \end{bmatrix}$$

Here we explicity define the mapping by

$$y \rightsquigarrow (y_{22}\, y_{12}^{-1}, y_{12}, (y_{21}-y_{11}\, y_{22}\, y_{12}^{-1}), y_{11}(y_{21}-y_{11}\, y_{22}\, y_{12}^{-1})^{-1})$$

for $y_{12} \neq 0$ and

$$y \rightsquigarrow (\infty,\ y_{22},\ y_{11},\ y_{21}\, y_{11}^{-1})$$

for $y_{12} = 0$, with the notation

$$y \rightsquigarrow (\theta_4(y), \theta_3(y), \theta_2(y), \theta_1(y)).$$

Clearly, we have defined new random variables $\theta_1, \theta_2, \theta_3, \theta_4$ having the same properties as $\eta_1, \eta_2, \eta_3, \eta_4$.

The following corollary is straightforward.

Corollary 1

For any couple $(x,y) \in G \times G$, *we have that* $x\,y \in H_3$ *iff* $\eta_4(x) = -\theta_4(y)$. *For* $(x,y) \in H_3 \times H_3$, *we have that* $x\,y \in H_2$ *iff* $\eta_3(x) = (\theta_3(y))^{-1}$. *For* $(x,y) \in H_2 \times H_2$, $x\,y \in H_1$ *iff* $\eta_2(x) = (\theta_2(y))^{-1}$. *For* $(x,y) \in H_1 \times H_1$, $x\,y = I$ *iff* $\eta_1(x) = -\theta_1(y)$.

2.1.3. – The random variables

Let us denote by

$$x \to \eta(x) = (\eta_1(x),\eta_2(x),\eta_3(x),\eta_4(x)) = (x_1,x_2,x_3,x_4)$$
$$\in \mathbb{F}_p \times \mathbb{F}_p^* \times \mathbb{F}_p^* \times (\mathbb{F}_p \cup \{\infty\})$$

$\forall x \in G$ the mapping given in the Lemma.

Now let us be given a set E with a uniform probability distribution and a random variable ζ from E onto G. Besides, let ξ_1, ξ_2, ξ_3, ξ_4 be random variables from E onto $\mathbb{F}_p$, $\mathbb{F}_p^*$, $\mathbb{F}_p^*$, and $\mathbb{F}_p \cup \{\infty\}$, respectively. We have the

Corollary 2

Let η be such that $\xi_1 = \eta_1 o\zeta$, $\xi_2 = \eta_2 o\zeta$, $\xi_3 = \eta_3 o\zeta$ and $\xi_4 = \eta_4 o\zeta$. Then we have that ζ has a uniform probability distribution iff each of ξ_1, ξ_2, ξ_3, ξ_4, has a uniform probability distribution.

We here just sketch the proof. Let us assume that ζ has a uniform P.D. and then prove that ξ_1 as a uniform P.D.

By definition, for any $x_1 \in \mathbb{F}_p$, $P\{\xi_1 = x_1\}$ is the measure of the subset of E mapped by ζ onto G_1, where G_1 is the set of all matrices $z \in G$ such that $\eta_1(z) = x_1$.

Thus $P\{\xi_1 = x_1\} = \sum_{z \in G_1} P\{\zeta = z\}$. But $P\{\zeta = z\}$ is a constant, by hypothesis, say $P\{\zeta = z^*\}$ where z^* is any matrix of G, and $P\{\zeta = z^*\} = 1/|G|$. Now by the lemma, $|G_1| = (p-1)^2(p+1)$. Hence $P\{\xi_1 = x_1\} = |G_1|/|G| = p^{-1}$.

2.1.4. - Toward applications

Concretely the set E just introduced that we have in view will be a cartesian product $X_1 \times X_2 \ldots \times X_r$ of small subsets of G. For example if p is about 10.000, say p = 10.007, then $|G|$ is about 10^{16}, we want to make $|X_i| < 100$. The mapping ζ will be defined by $\zeta : (x_1,\ldots,x_r) \to x_1 x_2 \ldots x_r$. Then there is no hope that ζ could have a uniform P.D. unless r is larger than 8. Actually, we expect that the mapping η of E into G is a random variable with uniform P.D. when r is, say, larger than 20, since requirement 1.3.2 asks for such a uniform P.D.

However since the algorithm in view only deals separately with the random variable ξ_1, ξ_2, ξ_3, ξ_4 and if these verify the statistical tests for uniformly distributed random variables, it will be reasonable to consider smaller values of r. Our aim will then be to draw random n-sequences from E with a uniform P.D. and then consider the image under ζ of these n-sequences as random n-sequences from G with a uniform P.D. The probabilistic algorithm will rely on that technique. Observe that such sets E may be used to produce by means of ξ_4 pseudo-random binary sequences for p a Mersenne prime.

2.2. - A recursive probabilistic algorithm

2.2.1. - General considerations on random n-sequences

Given a finite set E with a uniform probability distribution, we shall call random n-sequence drawn from E or random sample of size n an element $(x_1,\ldots,x_n)$ of E^n. This is to imply that all possible samples have the same probability $|E|^{-n}$. If moreover a random variable ζ from E onto $G = H_4$ is defined with a uniform P.D., then ζ and $(x_1,\ldots,x_n)$ yield a random n-sequence $(v_1,\ldots,v_n)$ of G^n. We will now consider two sets E and F, $|E| = |F|$ and two random variables ζ and γ respectively from E and F onto G with the same properties as above. We then define $T \subset E \times F$:

$$T = \{(x,y)/\zeta(x).\gamma(y) \in H_3\}$$

and the random variable $\zeta*\gamma : (x,y) \to \zeta(x)\gamma(y)$ from T onto H_3 is well defined. It has uniform P.D. For,

$$P\{(x,y) \in T\} = |H_3|^2(p+1)/|G|^2 = 1/(p+1),$$

by the lemma, and for any $z \in G$, $P\{(x,y)|x\ y = z\} = |G|/|G|^2$.

Thus for $z \in H_3$,

$$P\{\zeta*\gamma = z\} = P\{(x,y)|x\ y = z\}/P\{(x,y) \in T\} = 1/|H_3|.$$

Thus a random n-sequence from $E \times F$ will provide a random n'-sequence from T, but n will be large compared to n' since the average number of drawings is p+1 for a single element of T.

The arguments hold here and in the following for passing from H_4 to H_3 as well as for passing from H_i to H_{i-1}, $i < 4$. But this does not provide a good algorithm.

We then need a good algorithm for transforming two random n-sequences, one from E, the other from F into a random r-sequence from T with $r < n$ but with an r not to small compared to n. Afterwards, given two other sets E' and F' defined as E dans F, we will obtain a r'-sequence from a set T' defined similarly as T. Denoting by n' the min(r,r'), we then obtain two random n'-sequences, one from T, the other from T'. The algorithm will then proceed recursively. We will show in the next section how to use it for factoring the unity of G and, from there, any element from G into factors to be found in small subsets of G.

The aim of the algorithm that we give here for finding a random r-sequence from T is double. First it need few operations and secondly, given a probability Π and an integer k, we will be able to determine which n will yield an $r \geq k$ with probability larger than Π.

2.2.2. - The basic theorem

We still denote by G the group GL(2,p) and by H_3 the subgroup of lower triangular matrices.

We will use as before a function $\eta : G \to \{\mathbb{F}_p \cup \{\infty\}\}$, $\eta(x) = x_{11}\, x_{12}^{-1}$ for $x_{12} \neq o$ and $\eta(x) = \infty$ for $x_{12} = o$; also $\theta : G \to \{\mathbb{F}_p \cup \{\infty\}\}$, $\theta(x) = x_{22}\, x_{12}^{-1}$ for $x_{12} \neq o$ and $\theta(x) = \infty$ for $x_{12} = o$.

Then our algorithm will rely on the

Theorem 1 .

Let $(w_1,\ldots,w_n)$ *be a random n-sequence from* G = GL(2,p). *Reorder it as* $(u_1,\ldots,u_n)$ *so that* $\theta(u_i) \leq \theta(u_j)$ *for* $i < j$. *Now let* $(v_1,\ldots,v_n)$ *be any given n-sequence from* G *verifying* $\eta(v_i) = -\theta(u_i), i = 1,\ldots,n$.

Then $(v_1\, u_1,\ldots,v_n\, u_n)$ *is a random n-sequence from* H_3.

We first observe that $v_i\, u_i \in H_3, i = 1,\ldots,n$. Denoting v_i by x and u_i by y, we have indeed $x_{11}y_{12} + x_{12}\, y_{22} = o$ from $\eta(v_i) = -\theta(u_i)$ (Corollary 1). Now, by the lemma, we may write, using the left mapping, $(w_1,\ldots,w_n) = ((x_1,w_1'),\ldots,(x_n,w_n'))$ where $(x_1,\ldots,x_n)$ is first drawn from $\mathbb{F}_p \cup \{\infty\}$ and then $(w_1',\ldots,w_n')$ is drawn independently from H_3. We here have $x_i = -\theta(w_i)$. Reordering $(x_1,\ldots,x_n)$ as $(y_1,\ldots,y_n)$ so that $y_i \leq y_j$ for $i < j$ defines once for all a permutation on the indices. That permutation yields a permutation of H_3^n which perserves its uniform probability distribution. Hence $(u_1,\ldots,u_n) = ((y_1,u_1'),\ldots,(y_n,u_n')$ where $y_i = -\theta(u_i)$ and where $(u_1',\ldots,u_n')$ is a random n-sequence from H_3. Now $(v_1,\ldots,v_n)$ is as well $((v_1', -y_1),\ldots,(v_n' ,-y_n))$ where $(v_1',\ldots,v_n')$ is any given fixed n-sequence from H_3. Here we used the right mapping of the lemme. Here again multiplying componentwize by $(v_1',\ldots,v_n')$ every element of H_3^n defines a permutation of H_3^n which preserves its uniform probability distribution. But from the properties of right and left mapping of the lemma, we have that $(v_1\, u_1,\ldots,v_n\, u_n) = (v_1'\, u_1',\ldots,v_n'\, u_n')$ which is as just shown, a random n-sequence from H_3. □

2.2.3. - The algorithm

The essentiel problem to be solved is the following. Given two sets E and F with $|E| = |F|$ as in 2.2.1, draw two random n-sequences, one from E, the other from F and then determine the two corresponding random n-sequence $(v_1,\ldots,v_n)$ and $(w_1,\ldots,w_n)$ from $G = H_4$. The purpose is from there to construct one r-sequence from H_3 together with the corresponding r-sequence from T with r "not too small" compared to n. The algorithm will be repeated, replacing H_4 by H_3 and H_3 by H_2, and so on. We will have to fix n large enough to finally obtain a 1-sequence in H_o which reduces to the identity matrix. Assume that the value of n is known. After determining the two random n-sequences $(v_1,\ldots,v_n)$ and $(w_1,\ldots,w_n)$ from $G = H_4$, then the following steps of the algorithm are :

2. Reorder $(w_1,\ldots,w_n)$ as $(u_1,\ldots,u_n)$ by sorting $\{w_i\}_{i \in [n]}$ according to the values $-\theta(w_i)$. (See proof of theorem 1).

3. Reorder $(v_1,\ldots,v_n)$ according to the values of $\eta(v_i)$

4. Determine the set
   $S = \{\eta(v_i) \mid i \in [n]\} \cap \{-\theta(w_i) \mid i \in [n]\}$
   (This need less than 2n comparisons help to 2 and 3).

5. Determine the r-subsequence $(u_1^*, \ldots,u_r^*)$ of $(u_1,\ldots,u_n)$ with $-\theta(u_i^*) \in S$, $i = 1,\ldots,r$.

6. Construct an r-sequence $(t_1,\ldots,t_r)$ with terms taken from $\{v_1,\ldots,v_n\}$ such that $\eta(t_i) = -\theta(u_i^*), i = 1,\ldots,r$.

7. Compute the r-sequence $(t_1\ u_1^*,\ldots,t_r\ u_r^*)$ from H_3 and keep in memory the corresponding sequence from $T \subset E \times F$.

In section 2.3 we deal with numerical values. The numbers n, r and $s = |S|$ that we are concerned here with will be denoted n_4, r_4 and s_4 in the general algorithm. We start with $r_1 = 1$ and decide of the values of

n_1, s_1. Then $r_2 = n_1$ and we decide of the values of n_2, s_2. Finally we have $n_3 = r_4$ and the value $n = n_4$ of the above algorithm will be determined.

Notice that we have to replace $-\theta(u_i')$ by $(\theta(u_i'))^{-1}$ when computing an r-sequence from H_2 or from H_1.

2.3. - Numerical values

2.3.1. - First step

The first step consist in determining the size of the random n_1-sequences from H_1 such that the product of one selected term in the first by one term in the second sequence gives the identity matrix.

We know that H_1 may be identified whith $\mathbb{F}_p$. We take $p = 10,007$. We thus first have a mapping $a : [n_1] \to m = 10,007$ and with... $|\text{Im } a| = s_1$. Then mapping $b : [n_1] \to [m]$ defining the second n_1-sequence from H_1 may be considered as a sequence of n_1 independant Bernoulli trials, each with a probability of succes s_1/m. Since we need at least one succes with probability $1-10^{-3}$, we necessarily fix s_1 and n_1 such that

$$(1-s_1/m)^{n_1} < 10^{-3} \quad .$$

We have a degree of freedom in that choice but we jointly need that the probability that $|\text{Im } a| \geq s_1$ be at least $1-10^{-3}$.

For

$$n_1 = 297, \qquad s_1 = 285,$$

we have that

$$(1-s_1/m)^{n_1} < 1.8810^{-4}.$$

On the other hand we compute an upper bound for

$$\epsilon_{s_1,n_1} = m^{-n_1} \sum_{1\le i<s_1} (m)_i \, S(n,i),$$

where $S(n,i)$ is a Stirling number of the second kind, (see the Appendix).

We have from property 1, in the Appendix that

$$\epsilon_{s_1,n_1} = (m-1)_{s_1-1} \sum_{i\ge 0} m^{-n_1-i} \, S(n_1+i,\, s_1-1)$$

We use the upper bound $S(n,s) < s^n/s!$ which is not tight but quite satisfactory for the present purpose. Notice that

$$\lim_{n\to\infty} S(n,s)/s^n = 1/s!$$

Thus using that upper bound is convenient in the application in view except in this first step where we compute exactly

$$m^{-n_1} \sum_{s_1\le i\le n_1} (m)_i \, S(n,i) = 1-\epsilon_{s_1,n_1} = 0.999572 \ldots,$$

for $n_1 = 297$ and $s_1 = 285$.

Notice that without the constraint that n_1 should be larger than s_1 from the condition $\epsilon_{s_1,n_1} < 10^{-3}$, we would have had the only condition

$$-n_1 \, s_1/m < \log 10^{-3}$$

which gives for $n_1 = s_1$,

$$n_1 = s_1 = 263 \; ; \; (1-263/10{,}007)^{263} < 9.810^{-4}.$$

From this we observe that the Bernoulli constraint prevails. Next, we use $s!\, S(n,s) < s^n$ to bound $\epsilon_{s,n}$ by a geometric series of which the sum is

$$\binom{m-1}{s-1} \left(\frac{s-1}{m}\right)^n \frac{m}{m-s+1}$$

The reader is invited to see the Appendix for further details.

2.3.2. - Second step

The scheme is the same as for the first step except that we here have to consider two n_2-sequences from $\mathbb{F}_{p^*}$. Then m = 10,006. The second n_2-sequence consist in n_2 Bernoulli trials and we must have $r_2 = 297$ success with probability larger then $1-10^{-3}$.

Thus
$$\sum_{o \le j < r_2} \binom{n_2}{j} p_{s_2}^j q_{s_2}^{n-j} < 10^{-3}$$

with $p_{s_2} = s_2/m$, where the probability that the first n_2-sequence from $\mathbb{F}_p^*$ gives at least s_2 distinct terms must be larger than $1-10^{-3}$. The binomial tail is estimated by using the normal distribution. We have that

$$\mathcal{N}(3.11) = 0.9991 \ldots ;$$

we have that

$$P\{n_2 p_{s_2} + 3.11 \sqrt{n_2 p_{s_2} q_{s_2}} \ge r_2\} = \mathcal{N}(3.11)$$

See for example W. FELLER [2].

We find out that for $n_2 = s_2$ the value $n_2 = s_2 = 1,870$ gives the expected inequality. We now have to spread the values of n_2 and s_2 in order to obtain $\epsilon_{s_2,n_2} < 10^{-3}$ together with the previous inequality.

We find out that $n_2 = 2,287$ and $s_2 = 1,534$ still satisfy the previous inequality and moreover $\epsilon_{s_2,n_2} < 10^{-3}$ for $n_2 = 2,286$ and $s_2 = 1,534$. We keep $n_2 = 2,287$.

2.3.3. - Third step

Here again m = 10,006. We put r_3 = 2,287 and we try equal values n_3, s_3 for the Bernoulli tail. We obtain $n_3 = s_3$ = 4,897. By spreading those values, we have that n_3 = 6,562 with s_3 = 3,673 maintain the required inequality. On the other hand the upper bound for ε_{s_3,n_3} makes us sure that $\varepsilon_{s_3,n_3} < 10^{-3}$ for n_3 = 6,564 and s_3 = 3,673. We thus make n_3 = 6,564.

2.3.4. - Last step

Here m = 10,008 and r_4 = 6,564. We obtain $n_4 = s_4$ = 8,172 when asking for equal values in the inequality for the binomial tail. Now spreading the values of n_4 and s_4, we get n_4 = 11,952 and s_4 = 5,638 in order to obtain 6,564 succes by n_4 trials with probability p_{s_4}. Finally the upper bound used shows that we should draw at least 11,940 elements with replacement from a set of 10,008 elements in order to obtain 5,638 distincts elements with probability $1-10^{-3}$.

2.3.5. - Final decision

The final result is that we should start with n_4 = 11,952 for a global probability of succes at least $(1-.002)^{15} > 97\%$. We start again if we failed.

2.4. - Conclusion

Coming back to the notations of 1.4.4, we need 16 n_4-sequences to start with, one from each $U_{ir+1} \times U_{ir+2} \times \ldots \times U_{(i+1)r}$, i = 0, ..., 15, with n_4 = 11,952.

Let us now consider the value of r which is to be fixed in order that the basic probabilistic hypothesis be verified. The size of each set from which the elements of the n_4-sequences are drawn is 64^r. Each of those sets is a set E as considered in 2.1.3.. Since the size of G is about 10^{16}, it is hopeless that ζ have a uniform probability distribution unless $r \geq 9$. However, all we need is that each random variable ξ_i, $i = 1,\ldots,4$ behaves as a uniformly distributed random variable. Indeed at each step of the algorithm, we deal with only one of the random variables ξ_i. Thus it is very likely that we don't actually need that ζ have a uniform probability distribution. For $r = 3$, for example, ξ_4 is a mapping from a set of $64^3 = 262,144$ elements into a set of $p+1 = 10.008$ element and it is very likely that ξ_4 will behave as a uniformly distributed random variable and that the algorithm will proceed as expected even for that small fixed value of r. Indeed, in the second step of the algorithm, we will consider an n_3-sequence with $n_3 = r_4 = 6,564$ from $\mathbf{U}_1 \times \mathbf{U}_2 \times \mathbf{U}_3 \times \mathbf{U}_4 \times \mathbf{U}_5 \times \mathbf{U}_6$. The n_3-sequence $(t_1\ u_1^*,\ t_2\ u_2^*,\ldots,t_{n_3}\ u_{n_3}^*)$ is thus drawn from $H_3 \cap \mathbf{U}_1\ \mathbf{U}_2 \ldots \mathbf{U}_6$ of which the size may be roughly estimated as $64^6/10^4 \simeq 6.810^6$. For that reason, one may expect that the further projection η_3 which maps that set into $p-1 = 10,006$ elements will behave as a uniformly distributed random variable although $t = 16r$ with $r = 3$ is theoretically to small a value for t.

III. - APPENDIX

3.1. - Probabilities related to drawing two random n-sequences from an m-set

3.1.1. - The probability P(r,n,m)

The m-set, identified with the set [m] of the first m integers, has a uniform probability distribution, briefly U.P.D.. Such n-sequences may be viewed as mappings from the n-set [n] into the m-set [m]. Let a and b two such mappings. In view of the algorithm introduced in section 2.2 we have to consider the probability P(r,n,m) that less than r elements of [n] are mapped by b into the image of [n] by a. Thus

$$P(r,n,m) = P\{|b^{-1}(Imb \cap Ima)| < r\}.$$

This is the probability of failure at each probabilistic step of the algorithm (there are 15 such steps) for various values of the parameters r, n, m. In the numerical example dealt with in section 2.3, we asked that at each step P(r,n,m) should be smaller than 2.10^{-3}.

3.1.2. - The value of P(r,n,m)

Property 1

$$(1) \qquad P(r,n,m) = m^{-n} \sum_{1\le i\le n} (m)_i \; S(n,i) \sum_{o\le j<r} \binom{n}{j} p_i^j q_i^{n-j}$$

where $p_i = i/m$, $q_i = 1-i/m$, $(m)_i = m(m-1)\ldots(m-i+1)$ *and* $S(n,i)$ *is a Stirling number of the second kind.*

Proof

S(n,i) is the number of partitions of [n] into i classes and consequently, $(m)_i$ S(n,i) is the number of mappings a : [n] → [m] such that |Ima| = i. To each such mapping corresponds a probability

$$\sum_{o \le j < r} \binom{n}{j} p_i^j q_i^{n-j}$$

that a mapping b : [n] → [m] maps less than r elements in the set Ima. □

3.1.3. - Toward a computable upper bound for P(r,n,m)

In section 2.3 we use a dynamic programming argument to decide of the size n_4 of the first n-sequence to be drawn. At each step we could use expression (1) with r and m given and compute n large enough for P(r,n,m) to be smaller than 2.10^{-3}.

However m is close to 10^4 and n may be as large. Thus computing (1) by first tabulating S(n,i) and $\binom{n}{j}$ using recurrence relations looks beyond reach. Our purpose will thus be to set up an upper bound for the probability of failure P(r,n,m). Even if the bound is not tight, so long as it is easily computed and provides reasonable values for n at each step of the computation as in section 2.3 then the scheme of J. Bosset will be shown to be breakable. The first step for obtaining such an upper bound is

Property 2

Let $1 \le s \le n$ *and* $\varepsilon_s = m^{-n} \sum_{1 \le i < s} (m)_i \, S(n,i)$.

Then we have that

$$(2) \qquad P(r,n,m) < \varepsilon_s + \sum_{o \le j < r} \binom{n}{j} p_s^j q_s^{n-j}$$

Proof

On the one hand, we have that

$$\sum_{o \le j < r} \binom{n}{j} p_i^j \; q_i^{n-j}$$

is upper bounded by 1 for any i and in particular for $i < s$. On the other hand $m^{-n} \sum_{s \le i \le n} (m)_i \, S(n,i)$ is $1-\epsilon_s$.

Consequently all is left to be proved is that if we denote by $f(x,r)$ the expression

$$\sum_{o \le j < r} \binom{n}{j} p_x^j \; q_x^{n-j},$$

we have that $f(s,r) \ge f(i,r)$ for every $i > s$.

This actually means that the probability to have few success (less than r) in running n Bernoulli trials is larger when the probability p_s of success is smaller.

Denote by $g_{s,i}(r+j)$ the difference $f(s,r+j) - f(i,r+j)$, $j=0, \ldots, n-r+1$. We then have to prove that $g_{s,i}(r)$ is nonnegative for $i>1$. We first have $g_{s,i}(1) = q_s^n - q_i^n > 0$ and $g_{s,i}(n+1) = 0$.

Now we have that

$$g_{s,i}(r+1) = f(s,r+1) - f(i,r+1) = f(s,r) - f(i,r) + \Delta_{r+1},$$

where

$$\Delta_{r+1} = \binom{n}{r} \left((p_s \, q_s^{-1})^r \, q_s^n - (p_i \, q_i^{-1})^r \, q_i^n \right),$$

$$\Delta_{r+1} = r^{-1}(n-r+1)\left((p_s \, q_s^{-1}) \, \Delta_r + \binom{n}{r-1} (p_s \, q_s^{-1} - p_i \, q_i^{-1})(p_i \, q_i^{-1})^{r-1} \, q_i^n \right).$$

From the fact that $p_s \, q_s^{-1} - p_i \, q_i^{-1} < o$, we see that $\Delta_{r+j} < o$ implies $\Delta_{r+j+1} < o$, $j \le n-r$.

This means that if $g_{s,i}(r+j)$ eventually starts decreasing, then it goes on decreasing down to $g_{s,i}(n+1)$ which is zero. Hence we have that $g_{s,i}(r)$ is nonnegative.

3.2. - A first step to numerical computation

We are not finished with setting up an upper bound to P(r,n,m) since the number ε_s should be itself again upper bounded. This will be the purpose of another section in this Appendix. What we can do right now is obtaining numerical values for n and s in order that $f(s,n) < 10^{-3}$. We actually compute, for given r, values of n and s in order that the inequality

(3) $$np_s - 3.11 \sqrt{np_s q_s} \geq r.$$

holds.

This means that the probability that the number of succes is at most $np_s - 3.11 \sqrt{np_s q_s}$ will be close to 0.009. Thus the probability that the number of success is r-1 or less will be still smaller.

Inequality (3) is given by De Moivre-Laplace limit theorem (see for example W. Feller [2]).

3.3. - An easily computable upper bound for $\varepsilon_s = m^{-n} \sum_{1 \leq i < s} (m)_i \, S(n,i)$

3.3.1. - The direct computation

First, if n is not too large, a direct computation is feasible. We first compute by recurrence

(3) $$S(n,k) = S(n-1,k-1) + k \, S(n-1,k)$$

with

$$S(n,1) = 1 \text{ and } S(n,n) = 1 \text{ for all } n.$$

Since we need an upper bound for ϵ_s, we may use for $m^{-n}(m)_i$ the upper approximation

$$(4) \qquad m^{-n}(m)_i < (m/(m-i))^{m-i+.5} \exp(-i+1/12m)\, m^{-n+i}$$

given by W. Feller [2].

But since s is expected not to be much smaller than n, the computing by recurrence of S(n,s), ..., S(n,n) need few operations and we better compute $1-\epsilon_s$ using

$$(5) \qquad m^{-n}(m)_i > (m/(m-i))^{m-i+.5} \exp(-i-1/12\,(m-i))\, m^{-n+i}$$

3.3.2. - An infinite sum for P(r,n,m)

Property 3

$$(6) \qquad m^{-n} \sum_{1\le i<s} (m)_i\, S(n,i) = (m-1)_{s-1} \sum_{i\ge o} m^{-n-i}\, S(n+i,s-1)$$

Proof

We first prove that for every $k < m$, we have that

$$(7) \qquad (m-1)_k \sum_{i\ge o} m^{-k-i}\, S(k+i,k) = 1$$

Equality (7) is obtained right away by substituting m^{-1} to u in the generating series

$$u^k(1-u)^{-1}(1-2u)^{-1} \ldots (1-ku)^{-1} = \sum_{i\ge o} S(k+i,k)u^{k+i}$$

Which is found for example in L.Comtet [3]. Now equality (6) will follow from proving that

$$(8)\qquad m^{-n} \sum_{s\le i\le n} (m)_i\, S(n,i) = (m-1)_{s-1} \sum_{o\le j\le n-s} m^{-s-j+1}\, S(s+j-1,s-1).$$

The L.H.S is the probability to obtain at least s distinct elements when drawing n elements from an m-set. That event may be considered the union of n-s+1 disjoint events. The j^{th} event, j=o, ...,n-s, is to obtain for the first time s distinct elements at the $(s+j)^{th}$ drawing. The probability of the j^{th} event is the product of the probability of having drawn exactly s-1 distinct elements at the $(s+j-1)^{th}$ drawing by the probability of draxing an element from the complementary set of size m-s+1 at the $(s+j)^{th}$ drawing. Hence the probability of the j^{th} event is

$$m^{-s-j+1}(m)_{s-1}\, S(s+j-1,s-1).(m-s+1).m^{-1}.$$

This completes the proof of equality (8) and consequently of equality (6). □

3.3.3. - Applying Property 3 to numerical computation

It is well known that

$$(9)\qquad \lim_{n\to\infty} s!\, S(n,s)s^{-n} = 1$$

and moreover $S(n,s) < s^n/s!$.

See L. Comtet [3] page 293, exercice 9.

From

$$(10)\qquad s^n = \sum_{1\le i\le s} (s)_i\, S(n,i),$$

(9) may be obtained by observing that (10) implies

$$(11)\qquad 1 > s!\, S(n,s)s^{-n} > 1-(s^{-1}(s-1))^n\binom{s}{s-1}-(s^{-1}(s-2))^n\binom{s}{s-2} - \ldots - s^{-n}\, s,$$

$$1 > s!\, S(n,s)s^{-n} > 1-((\exp(-n/s)+1)^s-1-\exp(-n)).$$

From this and Property 3, then ε_s is upper bounded by the geometric series

$$(12) \qquad \binom{m-1}{s-1} m^{-n} (s-1)^n \sum_{i \geq o} ((s-1)m^{-1})^i = \binom{m-1}{s-1} ((s-1)m^{-1})^n \, m(m-s+1)^{-1}$$

And (9) shows that when i becomes large, each term of the geometric series becomes a tight upper bound for the corresponding term of the R.H.S of (6).

For a given probability of failure e^{-t} and given s and m, we are now able to compute an n such that the L.H.S. of (6) is upper bounded by e^{-t} :

$$(13) \qquad n > (\log(m^{-1}(s-1)))^{-1}(-t+\log(1-m^{-1}(s-1)) - \sum_{o \leq i < s-1} (\log(m-i-1)-\log(i+1))).$$

Another upper bound for S(n,s) permitting to bound (6) by a geometric series appears efficient when n-s is expected to be small and when m is large. We actually need for applying that bound that $c < m$, with $c = s(s+1)/2$. If n-s is small, the upper bound c^{n-s} to S(n,s) which results from Theorem D, page 207 of L. Comtet [3], may lead to satisfactory results, as we here show on a numerical example. However this does not apply to the numerical example of section 2.3 since there c is larger than m.

The upper bound here is

$$(14) \qquad c^{n-s+1} \, m^{-(n-s+1)} \, m^{-s} (m)_s \, m(m-c)^{-1},$$

So that for a given probability of failure e^{-t} and given s and m, the L.H.S. of (6) is upper bounded by e^{-t} as soon as

$$(15) \qquad n > (-t+(s-1)(\log c - \log m) - c(m-c)^{-1}(1-c(2m)^{-1})$$

$$-(m-s+0.5)s(m-s)^{-1}(1-s(2m)^{-1}) + s-(12m)^{-1})/\log(c\, m^{-1})$$

We use in that formula the inequalities

$$(m(m-s)^{-1})^{m-s+0.5} \exp(-s+(12m)^{-1}) > (m)_s\, m^{-s} \text{ as well as}$$

$$-\log(1-cm^{-1}) < c(m-c)^{-1}(1-c(2m)^{-1}).$$

Example

For $m = 10^8$, $s = 2{,}300$, $e^{-t} = 10^{-6}$ the R.H.S of (15) is worth 2,302.803. We thus fix $n = 2{,}303$. On the other hand, the R.H.S of (13) gives 2,515.096.

3.4. - Computing the average size of Ima for a random mapping $a : [n] \to [m]$

Taking the n^{th} derivative in t of the generating series

$$(e^t + z-1)^m = \sum_{n\geq o} \sum_{o\leq s\leq m} (m)_s\, S(n,s)\, z^{m-s}\, t^n/n!,$$

We obtain

$$\sum_{o\leq j\leq m} \binom{m}{j} (z-1)^{m-j}\, j^n\, e^{jt},$$

Which gives for $t = o$

$$f(z) = \sum_{o\leq j\leq m} \binom{m}{j} (z-1)^{m-j}\, j^n = \sum_{o\leq s\leq m} (m)_s\, S(n,s)\, z^{m-s}$$

The average size $\bar{s}$ of Ima for a random mapping $a : [n] \to [m]$ is thus

$$\bar{s} = m - m^{-n} f'(z)\big|_{z=1} = m-(m-1)^n\, m^{-n+1}$$

Thus conversely, if we want to know to which n for a given value of m will correspond an average size $\bar{s}$ we have that

(16) $\quad n = \log(1-\bar{s}\, m^{-1})/\log\,(1-m^{-1})$.

Example

In 2.3.3. we found out that for $m = 10,006$ we have a probability larger than $1-10^{-3}$ that $|\mathrm{Im}\ a| \geq s_3 = 3,673$ when the size of n is 6,564. Here, for an average size of 3,673, we find by (16) $n = 4,576.624$.

IV. - SOME COMBINATORIAL IDENTITIES RAISED BY INVESTIGATING UPPER BOUNDS

Relation (8) of section 3.3.2 may write

$$\sum_{s<i\leq n} (m)_i \, S(n,i) = (m)_{s+1} \sum_{o\leq j<n-s} m^{n-s-j-1} \, S(s+j,s)$$

This is a polynomial relation of degree n in m and since it hold for infinitely many values of m, we may replace m by the indeterminate u for obtaining the polynomial identity of degree n-s-1.

$$(1) \qquad \sum_{o\leq i<n-s} (u-s-1)_i \, S(n,i+s+1) = \sum_{o\leq i<n-s} u^i \, S(n-i-1,s),$$

where $(u-s-1)_o = 1$.

We observe that the R.H.S is a polynomial of $\mathbb{Z}[u]$ expressed in the basis $(u^i)_{i\in\mathbb{N}}$ while the L.H.S expresses the same polynomial in the basis $((u-s-1)_i)_{i\in\mathbb{N}}$. Since the derivative operator D maps the basis $(u^i/i!)_{i\in\mathbb{N}}$ of $\mathbb{Q}[u]$ onto itself and the difference operator Δ maps onto itself the basis $((u-s-1)_i/i!)_{i\in\mathbb{N}}$ of $\mathbb{Q}[u]$, the use of those operators permit deriving relations among the Stirling numbers of the second kind. We will first observe how (1) entails easily well known recurrence relations.

Substituting s+1 to u and then changing s+1 to k yields

$$(2) \qquad S(n,k) = \sum_{o\leq i\leq n-k} k^i \, S(n-i-1,k-1)$$

Which is quoted as vertical recurrence relation in L. Comtet [3], chap. 5, Th. B [3d]. The same author gives in Th C. [3e] an horizontal relation.

$$(3) \qquad S(n,s) = \sum_{o\leq i\leq n-s} (-1)^i \, \langle s+1\rangle_i \, S(n+1,s+i+1),$$

where $\langle x\rangle_i$ writes $x(x+1)\ldots(x+i-1)$.

This follows from putting $u \neq 0$ in (1). We now show what can be derived by applying t times the Δ operator to both sides of (1) and then substituting s+1 to u. We recall that $\Delta f(x) = (E-I)f(x) = f(x+1)-f(x)$.

$$t!\ S(n,s+t+1) = \sum_{t\le i<n-s} \Delta^t\, u^i\big|_{u=s+1}\ S(n-i-1,s)$$

Now

$$\Delta^t\, u^i = (E-I)^t\, u^i.$$

$$\Delta^t\, u^i\big|_{u=s+1} = \sum_{o\le j\le t} \binom{t}{j}(s+1+j)^i(-1)^{t-j}$$

$$= \sum_{o\le k\le i-t} (s+1)^k \binom{i}{k} \sum_{o\le j\le t} j^{i-k} \binom{t}{j} (-1)^{t-j}$$

The last factor in that sum appears to be t! S(i-k,t). We thus have that

$$(4) \qquad \Delta^t\, u^i\big|_{u=s+1} = t! \sum_{t\le j\le i} S(j,t)(s+1)^{i-j} \binom{i}{j}$$

Actually, the range of j may remain undefined in (4).

We then obtain the

Theorem

$$(5) \qquad S(n,s+t) = \sum_{t-1\le i<n-s} \sum_{t-1\le j\le i} \binom{i}{j} S(j,t-1)(s+1)^{i-j}\, S(n-i-1,s).$$

The theorem may be considered a generalization of (2) since from the recurrence

$$(6) \qquad S(n,k) = k\, S(n-1,k) + S(n-1,k-1)$$

We may write $S(j,o) = o$ for $j>o$ and $S(o,o) = 1$. On the other hand the ranges of i and j may be left undefined in (5), then the R.H.S being a convolution product and since

$$\sum_{n\geq o} S(n,s+t)u^n = u^t(1-(s+1)u)^{-1} \ldots (1-(s+t)u)^{-1}\, u^s\,(1-u)^{-1} \ldots (1-s\,u)^{-1},$$

We have the

Corollary 1

$$(7)\qquad u^t(1-(s+1)u)^{-1} \ldots (1-(s+t)u)^{-1} = \sum_{n\geq o} \sum_{t-1\leq j\leq n} \binom{n}{j} S(j,t-1)(s+1)^{n-j}u^{n+1}$$

We observe that since

$$(8)\qquad (1-(s+2)u)^{-1} - (1-(s+1)u)^{-1} = u(1-(s+2)u)^{-1}(1-(s+1)u)^{-1},$$

then generalizing (2) by (5) with t=2 remains easy. It gives

$$(9)\qquad S(n,s+2) = \sum_{1\leq i<n-s} ((s+2)^i - (s+1)^i)\, S(n-i-1,s).$$

Relations (6) and (8) show how the sequences $(S(n,i) \bmod 2)$, $i=1, \ldots,n$ behave.

Also corollary 1 gives for s=o

$$(10)\qquad S(n+1,t) = \sum_{t-1\leq j\leq n} \binom{n}{j} S(j,t-1)$$

which is relation [3c] in Theorem B already quoted from L. Comtet [3]. From Corollary 1 we also obtain

Corollary 2

$$(11) \qquad S(n+1,t) = \sum_{t-1 \le j \le n} (-1)^{t-j+1}\, t^{n-j} \binom{n}{j} S(j,t-1)$$

Proof

From the generating series for $S(n,t)$ we know that

$$S(n,t) = \sum_{i_1 + \ldots + i_t = n-t} 1^{i_1} \ldots t^{i_t}$$

Consequently, putting $s = -t-1$ in (7), we have that

$$u^t(1+t\,u)^{-1} \ldots (1+u)^{-1} = \sum_{n \ge o} (-1)^{n-t} S(n,t) u^n \qquad \square$$

REFERENCES

[1] J. BOSSET
"Contre les risques d'altération, un système de certification des informations"
01 Informatique n° 107, Février 1977.

[2] W. FELLER
"An introduction to probability theory and its applications"
Wiley, 1968.

[3] L. COMTET
"Advanced combinatorics"
D. Reidel, 1974.

Imprimé en France
par
l'Institut National de Recherche en Informatique et en Automatique