

MATHÉMATIQUES ET INFORMATIQUE

Les travaux de Madhu Sudan sur les codes correcteurs d'erreurs

Daniel Augot

Nous présentons les travaux de Madhu Sudan en théorie des codes correcteurs, qui, parmi d'autres¹, lui valurent de recevoir le prix Nevanlinna en août 2002², qui est le pendant de la médaille Fields, dans le domaine des aspects mathématiques de l'informatique. La percée majeure de M. Sudan est d'avoir produit un algorithme de décodage des codes de Reed-Solomon bien au-delà de la capacité de correction de ces codes, en autorisant de pouvoir décoder en retournant comme résultat une liste de solutions plutôt qu'une unique solution, ce que l'on appelle le « décodage en liste ».

Il s'agit d'une avancée théorique fondamentale dont les conséquences pratiques et théoriques ne sont pas encore mesurées. Dans cette présentation, nous introduirons l'arrière-plan pratique de la théorie des codes, puis les codes de Reed-Solomon et les codes géométriques. Enfin, nous présenterons l'algorithme de M. Sudan, comme une généralisation de l'algorithme de Berlekamp-Welsh.

Cet exposé repose sur la présentation que M. Sudan a lui-même faite dans [6].

Introduction et définitions

La théorie des codes est la discipline des mathématiques appliquées dont le sujet est la transmission fiable d'informations sur un canal de transmission bruité, en utilisant des objets combinatoires et algorithmiques appelés *codes correcteurs d'erreurs*. Pour introduire le sujet, il est d'abord nécessaire de préciser les notions de base du codage.

Suivons pour cela un message sur le chemin depuis la source jusqu'au récepteur, et observons les notions intéressantes qui apparaissent. Il y a trois entités impliquées dans le processus : l'émetteur, le récepteur et le canal de transmission. L'objectif de l'émetteur est de communiquer au récepteur un *message*, m , appartenant à \mathcal{M} où \mathcal{M} est un ensemble fini, *l'espace des messages*. Le canal de transmission bruité est capable de communiquer des suites arbitrairement longues de symboles d'un alphabet Σ , qui est « petit » (un des cas les plus intéressants étant $\Sigma = \{0, 1\}$). Alors l'espace des messages à coder est Σ^k , l'ensemble des suites de symboles de longueur k .

Émetteur et récepteur se mettent d'accord sur la longueur n des suites codées transmises, appelée la *longueur du code*, les messages échangés appartenant

¹ Ses autres travaux portent sur la théorie de la complexité, et il est l'un des auteurs du célèbre théorème PCP [8], en rapport avec la conjecture $P \neq NP$.

² <http://www.maa.org/news/fields02.html/>

donc à Σ^n , que l'on appellera l'*espace ambiant*. L'émetteur et le récepteur se mettent aussi d'accord sur une *fonction de codage*, E , injective.

$E : \mathcal{M} \rightarrow \Sigma^n$, utilisée pour coder les messages avant transmission. L'image $C = \{E(m), m \in \mathcal{M}\}$ est appelée le *code*. Le taux k/n , noté en général R , est appelé le *taux de transmission* ou *rendement* du code, c'est le premier paramètre fondamental d'un code, en théorie des codes.

En ce qui concerne le canal de transmission, il « bruite » les messages transmis. Ce bruit peut être vu comme une application de l'espace ambiant dans lui-même. Prescrivons maintenant une structure de corps sur l'alphabet Σ (par exemple Σ est un corps fini, de petite taille), ce qui induit une structure d'espace vectoriel sur Σ^n . Il devient alors plus commode de considérer des *codes linéaires* c'est-à-dire l'image par une application linéaire de Σ^k dans Σ^n , que l'on supposera toujours non singulière. On spécifiera dorénavant un code linéaire C par sa matrice génératrice (une base de C), ce qui est une manière compacte de décrire un ensemble a priori de taille q^k . En ce qui concerne le canal de transmission, il produit un vecteur de bruit $e \in \Sigma^n$, et le mot reçu est $y = E(m) + e$, m étant le message. Le récepteur utilise alors une fonction de décodage $D : \Sigma^n \rightarrow \mathcal{M}$. Le décodage D doit être rapide, et être tel que $D(y) = m$, avec grande probabilité. Intuitivement, le code introduit une redondance en augmentant la longueur des messages, et cette redondance sera utilisée pour décoder le message transmis, même s'il est bruité. Du point de vue de la fiabilité de la transmission, la question fondamentale de la théorie des codes est

Étant donnée une distribution de probabilité P sur le canal de transmission (*i.e.* une distribution de probabilité sur les erreurs de transmission), quelles sont les meilleures fonctions de codage et de décodage, c'est-à-dire quelle est la plus petite probabilité d'erreur

$$\min_{E,D} \left\{ \mathbf{E}_{m \in \mathcal{M}} \left(\Pr_{\eta \in P} [D(E(m) + \eta) \neq m] \right) \right\}$$

où \mathbf{E} désigne l'espérance mathématique.

Shannon a étudié les propriétés asymptotiques de cette quantité quand la distribution du bruit sur Σ^n est le produit cartésien d'une distribution sur Σ . Dans ce contexte, il existe une quantité $C_0 \in [0, 1]$, dépendant du canal, telle que pour tout $R < C_0$ et $\varepsilon > 0$, et, pour n assez grand, il existe toujours un couple codage/décodage avec un code de taux R tel que la probabilité d'erreur soit au plus ε . Dans le cadre de cet exposé, nous considérerons uniquement le cas du *canal q -aire symétrique*, défini de la manière suivante : chaque symbole de Σ transmis est préservé avec une certaine probabilité $1 - \delta$, ou bien est transformé en autre symbole parmi les $q - 1$ autres possibles avec probabilité $\delta/(q - 1)$, les événements étant indépendants d'un symbole à l'autre.

D'un autre côté, Hamming a défini les notions de *code correcteur d'erreur* et de code *détecteur d'erreur*. Définissons le *poids de Hamming* d'une séquence $x \in \Sigma^n$ comme le nombre de composantes non nulles de x , et la *distance de Hamming* entre x et y comme le poids de la différence $x - y$ (c'est-à-dire le nombre de composantes où x et y diffèrent). C'est bien une distance. On définit alors la distance minimale d'un code C comme la plus petite distance entre deux mots du code C . Le canal de transmission crée en général un vecteur η de petit

poids, par exemple de poids borné par e . On dira qu'un code correcteur corrige e erreurs si les boules de rayon e centrées sur les mots de code ne s'intersectent pas. En effet si le poids de l'erreur est inférieur à e , alors, si C est e -correcteur, il y a unicité du mot de code le plus proche. Une capacité de correction e implique que la distance minimale entre deux mots distincts du code est supérieure à $2e + 1$. La distance minimale est le deuxième paramètre fondamental d'un code. On parlera d'un code $[n, k, d]$ pour un code de longueur n , de dimension k et de distance minimale d . Du point de vue de Hamming, la question fondamentale est

Étant donné un alphabet Σ de taille q , et deux entiers n et k , $k < n$, quelle est la plus grande distance minimale d d'un code $C \subseteq \Sigma^n$ de taux de transmission k/n ?

En effet, une distance minimale élevée induit que le code est capable de corriger des erreurs de poids élevé. Signalons immédiatement que le problème de Hamming n'est pas résolu quand la taille de l'alphabet est petite. Il y a une réponse satisfaisante à la question quand $q \geq n$ (voir les codes de Reed-Solomon dans la section suivante).

Constructions de codes

Nous commencerons par la famille des codes aléatoires linéaires (en fait une *non-construction*). La borne de Varshamov-Gilbert indique qu'il existe des codes de paramètres $[n, k, d]$ si n , k et d vérifient :

$$q^k V_q(n, d) \leq q^n,$$

où $V_q(n, d)$ est le volume de la sphère de Hamming de rayon d (c'est-à-dire son cardinal). Donc il existe des codes tels que $q^k V_q(n, d) \geq q^n$. En prenant le logarithme et en approchant $V_q(n, \delta n)$ par $q^{H_q(\delta)n}$ (où $H_q(x)$ désigne la fonction d'entropie q -aire : $H_q(\delta) = -\delta \log_q(\frac{\delta}{q-1}) - (1-\delta) \log_q(1-\delta)$), on obtient l'existence de codes sur la borne suivante :

$$R \geq 1 - H_q(\delta), \quad \text{avec } R = \frac{k}{n} \text{ et } \delta = \frac{d}{n}.$$

Ce résultat s'étend « particulièrement » aux codes aléatoires : avec une probabilité tendant vers 1 quand la longueur n croît, les codes aléatoires se trouvent sur la borne de Varshamov-Gilbert. La question qui en découle est de savoir s'il existe des codes dépassant la borne de Varshamov-Gilbert.

En dehors des codes aléatoires, la théorie des codes s'est donc appliquée depuis ses fondements à produire des familles explicites de bons codes, dont la dimension et la distance puissent être déterminées à l'avance. Cela a conduit à toute une « botanique » de codes, diversement utilisés en pratique.

Citons une autre borne, celle de Singleton. Suivant cette borne, tout code C linéaire voit ses paramètres $[n, k, d]$ vérifier $k + d \leq n + 1$. Pour la démontrer, considérons une *matrice de parité* de C , il s'agit d'une matrice H de taille $(n - k) \times n$ dans laquelle sont écrites $n - k$ formes linéaires qui s'annulent sur le code C : tout mot c du code vérifie $Hc = 0$. Le rang de la matrice H est $n - k$, et comme le code ne contient pas de mot de poids strictement inférieur à d , il n'y a pas de relations linéaires entre moins de d colonnes de H : $d - 1 \leq n - k$.

Nous nous contenterons de décrire les codes de Reed-Solomon qui sont optimaux pour la borne de Singleton et les codes de Goppa géométriques.

Codes de Reed-Solomon

Un code de Reed-Solomon de dimension k et de longueur n est défini par la donnée de n éléments distincts $\alpha_1, \dots, \alpha_n$ de \mathbf{F}_q , où \mathbf{F}_q est un corps fini. Nous appellerons *points* ces éléments $\alpha_1, \dots, \alpha_n$. Soit maintenant la fonction ev définie par

$$\begin{aligned} \text{ev} : \mathbf{F}_q[x] &\rightarrow \mathbf{F}_q^n \\ f(x) &\mapsto \text{ev}(f(x)) = (f(\alpha_1), \dots, f(\alpha_n)). \end{aligned}$$

où $\mathbf{F}_q[x]$ est l'algèbre des polynômes classique avec l'indéterminée canonique x . Le code de Reed-Solomon RS_k de dimension k est défini par $\alpha_1, \dots, \alpha_n$, est

$$RS_k = \{\text{ev}(f(x)); \deg(f(x)) < k\}.$$

Il est aisé de voir que sa dimension est k , et que sa distance minimale est $n - k + 1$. En effet tout polynôme $f(x)$ de degré strictement inférieur à k a au plus $k - 1$ zéros, donc le vecteur $\text{ev}(f(x))$ a au moins $n - k + 1$ composantes non nulles. On a bien $k + d = n + 1$ pour les codes de Reed-Solomon, ce qui correspond à la borne de Singleton.

En divisant par n , le code de Reed-Solomon de taux de transmission R a donc une distance minimale approximativement de $1 - R$. Notons aussi que si ces paramètres sont bons, on ne peut faire croître la longueur n en conservant un alphabet de taille fixée q , ce qui est un inconvénient majeur.

Codes de Goppa géométriques

Les codes de Goppa géométriques ont été introduits par Goppa dans [1]. Ils sont une généralisation naturelle des codes de Reed-Solomon. Nous présentons ici une version simplifiée dite des codes « à un point ».

Soit C une courbe algébrique irréductible lisse³ définie sur \mathbf{F}_q . Soit P_1, \dots, P_n , n points distincts rationnels de C et soit P_∞ un autre point distinct de P_1, \dots, P_n . Soit encore ev la fonction d'évaluation suivante :

$$\begin{aligned} \text{ev} : L(kP_\infty) &\rightarrow \mathbf{F}_q^n \\ f &\mapsto \text{ev}(f) = (f(P_1), \dots, f(P_n)), \end{aligned}$$

où $L(kP_\infty)$ désigne l'espace de fonctions de $\mathbf{F}_q(C)$ associé au diviseur kP_∞ . Alors, de manière similaire aux codes de Reed-Solomon, on définit le code de Goppa géométrique $\Gamma(P_1, \dots, P_n, kP_\infty)$ comme suit

$$\Gamma(P_1, \dots, P_n, kP_\infty) = \{\text{ev}(f); f \in L(kP_\infty)\}.$$

Si on a choisi $k \geq 2g - 2$, où g est le genre de C , alors le théorème de Riemann-Roch nous assure que la dimension de $\Gamma(P_1, \dots, P_n, kP_\infty)$ est $k - g + 1$. De même il est facile de prouver que la distance minimale d est telle que $d \geq n - k$. Pour ces codes, on a $k + d = n - g + 1$, et on voit que le « défaut » par rapport aux codes de Reed-Solomon et à la borne de Singleton est g , le genre de la courbe. En revanche, on peut faire croître la longueur de ces codes, à *alphabet*

³ Le lecteur ne connaissant pas la théorie des courbes algébriques peut sauter ce paragraphe.

fixé, en augmentant le nombre de points des courbes sur lesquelles est fondée la construction.

En construisant une famille de courbes dont le genre ne croît pas trop vite, Tsfasman, Vladut et Zink [7] ont montré l'existence de codes de distance relative δ et de taux de transmission R supérieur ou égal à $1 - \delta - \frac{1}{\sqrt{q-1}}$. Ces codes, lorsque l'alphabet est de taille supérieure à 49, sont meilleurs que les codes aléatoires, c'est-à-dire qu'ils dépassent la borne de Varshamov-Gilbert, ce qui constitua une surprise de taille lors de leur découverte, les chercheurs en théorie des codes pensant que cette borne était optimale. Toutefois, dans le cas binaire ($q = 2$), qui est le plus intéressant en pratique, on ne sait toujours pas si la borne de Varshamov-Gilbert est optimale ou s'il existe des codes dépassant cette borne.

Algorithmes de décodage

La problématique

Le problème du décodage est une tâche difficile, pour laquelle les algorithmes naïfs présentent de très mauvaises complexités⁴. Par exemple la *recherche exhaustive* qui consiste à passer en revue tous les mots du code pour trouver le plus proche a une complexité en temps exponentielle en la longueur du code (pour peu que la dimension croisse linéairement avec la longueur du code).

Notons aussi que le problème du décodage n'est pas évident à formuler, et présente de nombreuses variantes dans la littérature. La communauté des chercheurs considère, hélas sans preuve, que le décodage des codes aléatoires est difficile. Cela signifie que si les codes aléatoires ont de bons paramètres, il est impossible de les décoder de manière efficace.

Supposons la famille de codes à décoder fixée (Reed-Solomon, codes géométriques), il reste à définir proprement le problème. Suivant l'article de Madhu Sudan, nous en resterons aux définitions suivantes :

NCP (Nearest Codeword Problem : problème du mot le plus proche) Il s'agit de trouver le mot de code le plus proche du mot reçu au sens de la métrique de Hamming.

LD (List Decoding : décodage en liste) Une borne e est donnée. Le problème est de trouver *tous* (éventuellement aucun) les mots de code à distance e du mot reçu.

BDD (Bounded Distance Decoding : décodage borné) Une borne e est donnée. Le problème est de trouver *un* mot parmi les mots de code à distance e du mot reçu (s'il en existe).

UD (Unambiguous Decoding : décodage non ambigu) Ici on se donne $e = (d-1)/2$, où d est la distance minimale du code, et on cherche le mot de code à distance e du mot reçu (s'il existe)⁵

⁴ La *complexité (en temps)* d'un algorithme est ici le nombre d'opérations élémentaires nécessaires à son exécution. Pour les algorithmes de décodage, nous distinguerons la *performance* d'un algorithme, qui est une mesure de sa capacité à corriger des erreurs de poids plus ou moins élevé, de sa *complexité*, qui mesure son temps d'exécution.

⁵ Signalons, pour être complet, une grande classe de codes et d'algorithmes de décodage, très performants en pratique, qui est celle des turbo-codes et des codes LDPC, conjointement

Classiquement, le problème étudié en théorie des codes est ce dernier problème (décodage non ambigu).

Décodage non ambigu (UD)

Ce problème a été résolu pour toutes les classes de codes introduites ici, d'une manière efficace. Il est à noter que chacun des algorithmes est non trivial, et que le progrès le plus spectaculaire a été de réussir à décoder les codes de Goppa géométriques (voir le résumé [3]). Nous ne détaillerons pas ces algorithmes, mais notons que ces algorithmes ont en général une complexité quadratique ou presque quadratique⁶ en la longueur, et chacun de ces algorithmes est construit « ad hoc » pour tel ou tel code, un algorithme « générique » de bonne complexité n'existant pas.

À titre d'exemple, nous présentons l'algorithme de Berlekamp-Welsh pour le décodage des codes de Reed-Solomon tels que nous les avons présentés.

Le problème est le suivant : étant donnés des points $\alpha_1, \dots, \alpha_n$ dans \mathbf{F}_q , des valeurs y_1, \dots, y_n dans \mathbf{F}_q^n , et un entier $e < \frac{n-k}{2}$, trouver tous les polynômes $f(x)$ dans $\mathbf{F}_q[x]$ de degré strictement inférieur à k tels que $f(\alpha_i) = y_i$ pour au moins $n - e$ valeurs de i .

Pour cette valeur de e , on sait qu'il y a au plus une solution $f(x)$.

Soit donc $f(x)$ le polynôme solution du problème posé, et considérons le polynôme unitaire $E(x)$ qui est de degré e (e étant le poids de l'erreur), tel que $E(\alpha_i) = 0$ s'il y a une erreur à la position i . Alors, à coup sûr, on a, pour tout i , $y_i = f(\alpha_i)$ ou $E(\alpha_i) = 0$, ce qui se traduit algébriquement en $E(\alpha_i)y_i = E(\alpha_i)f(\alpha_i)$, pour tout i .

L'algorithme de Berlekamp-Welsh consiste en les deux étapes suivantes : trouver deux polynômes $E(x)$ et $N(x)$ de degrés respectifs au plus e et $k+e-1$, tels que $E(\alpha_i)y_i = N(\alpha_i)$, pour tout i . On voit qu'il s'agit d'un problème d'algèbre linéaire, dont les inconnues sont les coefficients de $E(x)$ et de $N(x)$. En particulier, on est sûr que ce système a une solution quand le nombre d'inconnues est inférieur au nombre d'équations, c'est-à-dire quand $e+k+e < n$, ce qui donne la condition $e < \frac{n-k}{2}$, ce qui est bien la capacité de correction des codes de Reed-Solomon. La deuxième étape est de retourner le résultat $N(x)/E(x)$ (qui est bien un polynôme).

Décodage en liste (LD)

Le problème du décodage en liste se pose dès que l'on veut décoder e erreurs avec $e > d/2$, où d est la distance minimale du code à décoder. En effet, il n'y a plus unicité du mot de code à distance e , et il devient alors nécessaire de retourner une liste de candidats.

Le principal progrès dû à Madhu Sudan dans son article [5], concerne le décodage des codes de Reed-Solomon, pour lesquels il a produit un algorithme capable de décoder bien au-delà de la capacité de correction $\lfloor \frac{d-1}{2} \rfloor$ du code.

Le problème est toujours le même : étant donnés des points $\alpha_1, \dots, \alpha_n$ dans \mathbf{F}_q , des valeurs y_1, \dots, y_n dans \mathbf{F}_q , et un entier e , trouver tous les polynômes $f(x)$ dans $\mathbf{F}_q[x]$ de degré inférieur à k tels que $f(\alpha_i) = y_i$ pour au moins $n - e$

avec les algorithmes de décodage itératifs. Ces algorithmes de décodages ne font apparaître que marginalement la notion de distance minimale.

⁶ $O(n^{7/3})$

valeurs de i . Mais dans ce cas précis, on autorise e à être plus grand que $\frac{n-k}{2}$, donc il y a plusieurs solutions possibles. L'algorithme est une généralisation de l'algorithme de Berlekamp-Welsh, et l'idée est la suivante. L'algorithme de Berlekamp-Welsh décrit ci-dessus cherche en fait un polynôme $Q(x, y) = N(x) - yE(x)$ de degré 1 en Y tel que $Q(x_i, y_i) = 0$ pour tout i , avec les conditions $\deg N(x) \leq k + e$ et $\deg E(x) \leq e$.

L'idée de M. Sudan est de chercher un polynôme $Q(x, y) = \sum_i Q_i(x)y^i$ de degré supérieur à 1 en y tel que $\deg Q_i(x) < n - e - i(k - 1)$, et tel que $Q(x_i, y_i) = 0$ pour tout i . Alors, tout polynôme $f(x)$ solution du problème de décodage vérifie $Q(x, f(x)) = 0$.

En effet, le polynôme $Q(x, f(x))$ est de degré strictement inférieur à $n - e$ par construction de $Q(x, y)$. De plus, comme $f(x_i) = y_i$ pour au moins $n - e$ valeurs de i et que $Q(x_i, y_i) = 0$ pour tout i , on a que le polynôme $Q(x, f(x))$ a plus de racines que son degré, donc il est identiquement nul.

En conséquence, l'algorithme de décodage de M. Sudan se déroule en deux étapes :

1. trouver un polynôme $Q(x, y)$ satisfaisant les conditions ci-dessus ;
2. trouver les facteurs $y - f(x)$ de $Q(x, y)$.

La première étape est encore une étape d'algèbre linéaire, où l'on cherche $Q(x, y) = \sum_i Q_i(x)y^i$ tel que $Q(x_i, y_i) = 0$ pour tout i , avec les conditions précédentes sur les degrés des $Q_i(x)$, et les inconnues sont les coefficients de $Q(x, y)$. Pour être assurés d'avoir une solution, on doit avoir $N_Q > n$ où N_Q est le nombre de monômes apparaissant dans $Q(x, y)$. Notons que le degré du polynôme $Q(x, y)$ est au plus $\lfloor \frac{n-e}{k-1} \rfloor$: c'est un majorant du nombre de solutions à l'équation $Q(x, f(x)) = 0$, donc du nombre de solutions à notre problème. Un calcul du nombre de termes de $Q(x, y)$ et la condition $N_Q > n$ donnent (grossièrement) la relation suivante en e :

$$(1) \quad e < n - \sqrt{2kn}.$$

Cette borne est à comparer avec la borne du décodage non ambigu $e < \frac{n-k}{2}$. Elle s'exprime mieux en divisant par n , où nous avons $\varepsilon < 1 - \sqrt{2R}$, avec $\varepsilon = e/n$ et $R = k/n$, et nous comparons les performances sur la figure 1.

La première remarque est que la capacité de correction de l'algorithme de Sudan n'est pas toujours plus élevée que celle de l'algorithme de Berlekamp-Welsh, notamment pour les codes de grand rendement k/n . La deuxième remarque est que lorsque le rendement du code est proche de zéro, alors le taux d'erreurs toléré approche 1.

La deuxième étape de l'algorithme de M. Sudan consiste, après avoir trouvé le polynôme $Q(x, y)$ à trouver les solutions $f(x)$ de l'équation $Q(x, f(x)) = 0$, c'est-à-dire de trouver les facteurs $y - f(x)$ de $Q(x, y)$. Notons rapidement qu'il existe un algorithme de complexité polynomiale pour factoriser les polynômes à plusieurs variables sur un corps fini. M. Sudan en conclut donc que son algorithme est de complexité polynomiale. Cela n'est pas suffisant en pratique quand on souhaite obtenir la plus grande efficacité, et beaucoup de chercheurs s'attachent à trouver les meilleurs algorithmes pour accomplir les deux étapes de l'algorithme de Sudan, afin de le rendre complètement effectif de telle sorte qu'on puisse le programmer dans des circuits électroniques.

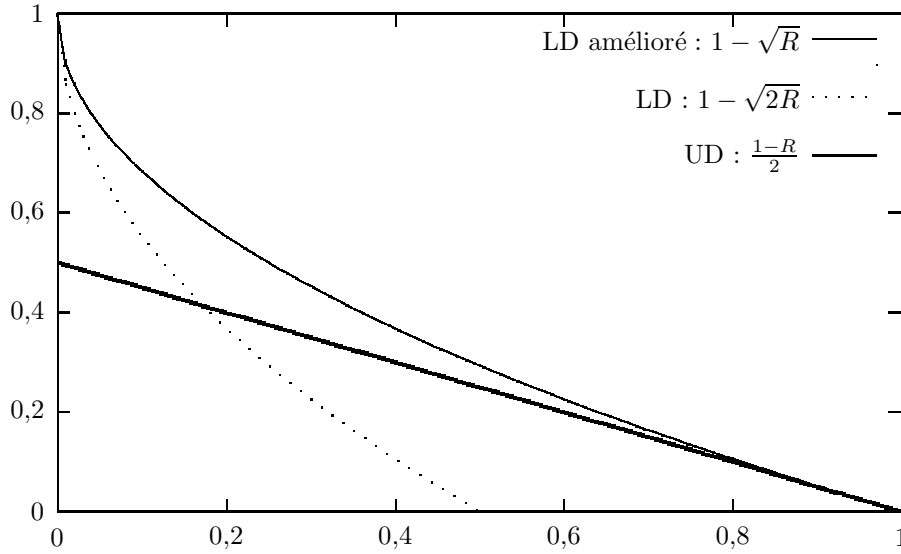


FIG. 1. Performances des algorithmes de Berlekamp-Welsh, Sudan et Guruswami-Sudan. En abscisse : le taux R de transmission ; en ordonnée : la capacité de correction.

Décodage en liste amélioré

V. Guruswami et M. Sudan ont rapidement proposé un algorithme étendant la capacité de correction à la borne $1 - \sqrt{R}$, qui est toujours meilleure que la borne classique $\frac{1-R}{2}$ (cf. figure 1). Sans entrer dans les détails de l'article [2], nous faisons la remarque suivante. Si $f_1(x)$ et $f_2(x)$ sont deux solutions au problème de décodage, alors il peut y avoir des indices i tels que l'on ait $y_i = f_1(x_i) = f_2(x_i)$. Sachant que $y - f_1(x)$ et $y - f_2(x)$ divisent tous deux $Q(x, y)$, alors ce dernier polynôme présentera une multiplicité d'ordre au moins deux au point (x_i, y_i) . C'est cette remarque qui est à la base de l'algorithme amélioré : on cherche maintenant un polynôme $Q(x, y)$ tel que $Q(x_i, y_i) = 0$ avec une certaine multiplicité r . Le degré de $Q(x, y)$ étant bien choisi ainsi que la multiplicité, alors on aura, pour tout polynôme $f(x)$ solution du problème de décodage, $Q(x, f(x)) = 0$. Une optimisation des paramètres auxiliaires (degré de $Q(x, y)$, ordre r de multiplicité), conduit à la borne $1 - \sqrt{R}$.

Pour conclure cette section, nous citons [2, 4] pour indiquer que ces algorithmes (Sudan, Guruswami-Sudan) se généralisent facilement aux codes géométriques. De plus, d'une manière surprenante, ces généralisations conduisent à des algorithmes conceptuellement plus simples que ceux déjà inventés pour le décodage classique des codes géométriques.

Conclusion

La découverte d'un algorithme de décodage des codes de Reed-Solomon avec un tel pouvoir de correction a relancé l'intérêt pratique de ces codes, ainsi

que des codes géométriques, puisque leur *performance* s'en trouve grandement améliorée. Si la question se pose de savoir en pratique comment choisir une solution dans la liste des solutions proposées par l'algorithme, force est de constater que l'algorithme retourne une unique solution avec une probabilité quasi égale à un, ce qui est fort intéressant. Reste maintenant à améliorer l'implémentation de cet algorithme pour être performant dans les applications.

Les algorithmes de Sudan et Guruswami-Sudan ont déjà eu des applications en cryptologie (pour cryptanalyser un algorithme de chiffrement) et aussi dans le domaine de la protection des droits d'auteurs. Dans les deux cas, c'est le haut pouvoir de correction, proche de 1 à taux de transmission proche de zéro, qui est utilisé. Nul doute que, du point de vue théorique, cet algorithme aura encore de nombreuses applications dans diverses branches des mathématiques appliquées et deviendra un classique parmi les grands algorithmes de l'informatique.

Références

- [1] V.D. Goppa. Codes associated with divisors. *Problems of Information Transmission*, 12(1) :22–27, 1977.
- [2] V. Guruswami and M. Sudan. Improved decoding of Reed-Solomon and Algebraic-Geometric codes. *IEEE Transactions on Information Theory*, 45 :1757–1767, 1999.
- [3] Tom Hoholdt and Ruud Pellikaan. On the decoding of algebraic geometry codes. *IEEE Transactions on Information Theory*, 41(6), 1995.
- [4] M. Amin Shokrollahi and Hal Wasserman. Decoding algebraic geometric codes beyond the error-correction bound. *IEEE Transactions on Information Theory*, 45 :432–437, 1999.
- [5] Madhu Sudan. Decoding of Reed-Solomon codes beyond the error-correction bound. *Journal of Complexity*, 13, 1997.
- [6] Madhu Sudan. Coding theory : Tutorial and survey. In *Proceedings of the 42nd Annual Symposium on Foundations of Computer Science*, pages 36–53, 2001.
- [7] M. A. Tsfasmann, S. G. Vlăduț, and T. Zink. Modular curves, shimura curves, and goppa codes better than varshamov-gilbert bound. *Math. Nachr.*, "109" :21–28, "1982".
- [8] S. Arora, C. Lund, R. Motwani, M. Sudan and M. Szegedy. Proof verification and hardness of approximation problems. *Journal of the ACM*, 45(3) :501-555, 1998.

PRIMES is in P, une avancée accessible à « l'homme ordinaire » *

Folkmar Bornemann

En août 2002, trois chercheurs indiens annonçaient qu'ils avaient trouvé un test de primalité en temps polynomial, résolvant ainsi une conjecture déjà ancienne en théorie de la complexité. Le résultat était d'importance, il fut rapidement introduit dans nombre de programmes de cours avancés en cryptographie ou théorie analytique des nombres. François Morain fit un exposé sur ce sujet au séminaire Bourbaki de mars 2003. Le texte que nous vous proposons, déjà publié en allemand dans le DMV-Mitteilungen de fin 2002 et en anglais dans les Notices de mai 2003, nous relate cette histoire extraordinaire mais aussi ses avatars dans la presse généraliste.

« New Method Said to Solve Key Problem in Math » titrait le New York Times du 8 août 2002, ce qui signifiait la preuve de $\text{PRIMES} \in \mathcal{P}$, un gros problème ouvert en théorie algorithmique des nombres et en informatique théorique. Manindra Agrawal, Neeraj Kayal et Nitin Saxena de l' Indian Institute of Technology ont réalisé cette preuve grâce à un algorithme d'une éclatante simplicité et d'une surprenante élégance. Convaincus de sa validité après seulement quelques jours, les experts s'enthousiasmaient : « This algorithm is beautiful » (Carl Pomerance), « It's the best result I've heard in over 10 years » (Shafi Goldwasser).

Quatre jours avant le gros titre du *New York Times*, un dimanche, les trois auteurs avaient envoyé à quinze experts un preprint de neuf pages intitulé « PRIMES is in P ». Le soir du même jour Jaikumar Radhakrishnan et Vikraman Arvind leur envoyaient leurs félicitations. Très tôt le lundi, un des maîtres du sujet, Carl Pomerance, vérifiait le résultat et, dans son enthousiasme, il organisait un séminaire informel pour l'après-midi et informait Sara Robinson du *New York Times*. Le mardi le preprint était en accès libre sur Internet. Le jeudi un autre grand spécialiste, Hendrik Lenstra Jr., mettait fin à une brève polémique en diffusant sur la liste de courrier électronique NMBRTHRY le jugement :

The remarks [...] are unfounded and/or inconsequential. The proofs [...] do NOT have too many additional problems to mention. The only true mistake is [...], but that is quite easy to fix. Other mistakes [...] are too minor to mention. The paper is in substance completely correct.

* Publié dans les Notices de l'AMS (mai 2003) sous le titre « Primes is in P : a breakthrough for everyman », ce texte a été traduit par Colette Anné. L'auteur utilise un jeu de mot intraduisible avec *Jedermann*, en anglais everyman, mot à mot « chaque homme », qui désigne une forme de drame allégorique héritée du Moyen-âge et de la Renaissance, c'est aussi le titre d'un drame très populaire de Hugo von Hofmannsthal.

Et déjà le vendredi, Dan Bernstein affichait sur le web une amélioration de la preuve du résultat principal, qui tenait en une seule page.

La brièveté, inhabituelle en mathématiques, de la période de vérification reflète à la fois la concision et l'élégance de l'argument et sa simplicité technique, « suited for undergraduates ». Deux des auteurs eux-mêmes, Kayal et Saxena, venaient juste de recevoir leur diplôme de licence en informatique au printemps. Est-ce donc exceptionnel qu'une découverte sensationnelle soit accessible à l'« homme ordinaire » ?

Dans son discours au Congrès International des Mathématiciens de Berlin en 1998, Hans-Magnus Enzensberger défendit la thèse selon laquelle les mathématiques sont à la fois « au-delà de la culture » et en même temps au cœur d'un âge d'or dû à des succès d'une qualité qu'on ne connaît ni au théâtre ni en sport. Toutefois nombre de ces succès posent à bien des mathématiciens la question de l'au-delà et de l'en-deça, à l'intérieur même des mathématiques – la main sur le cœur : combien d'entre nous ne sont pas un de ces « hommes ordinaires » ? – Est-ce qu'un non-spécialiste peut vraiment comprendre, ou pleinement apprécier la preuve du dernier théorème de Fermat par Andrew Wiles, bien que des efforts de popularisation comme le livre de Simon Singh aident à avoir une petite idée des connexions. Il n'y a probablement aucun auteur qui puisse aider l'« homme ordinaire » à appréhender toutes les ramifications et significations des résultats des récipiendaires de la médaille Fields des dernières années.

Ainsi chacun ajoute des briques à son parapet dans la Tour de Babel appelée Mathématiques et juge sa construction fondamentale. Il y a rarement un tel succès comme en ce début d'août ; une pierre de fondation de la Tour que l'« homme ordinaire » peut comprendre.

Paul Leyland exprima un point de vue que beaucoup ont partagé : « Everyone is now wondering what else has been similarly overlooked. »

Est-ce que cela explique le grand étonnement d'Agrawal ? (« I never imagined that our result will be of much interest to traditional mathematicians ») ; en l'occurrence pourquoi est-ce que le site web de dédicace a eu plus de deux millions de visites les dix premiers jours, et le preprint a été téléchargé plus de trois cent mille fois ?

« When a long outstanding problem is finally solved, every mathematician would like to share in the pleasure of discovery by following for himself what has been done. But too often he is stymied by the abstruseness of so much of contemporary mathematics. The recent negative solution to [...] is a happy counterexample. In this article, a complete account of this solution is given ; the only knowledge a reader needs to follow the argument is a little number theory : specifically basic information about divisibility of positive integers and linear congruences. »

Martin Davis, *Hilbert's tenth problem is unsolvable*, Amer. Math. Monthly 80 (1973), pp. 233–269, premier paragraphe de l'introduction.

En tant que spécialiste en analyse numérique et non en théorie algorithmique des nombres, j'ai voulu tester ma fougue d'« homme ordinaire » loin de mon parapet.

Le problème

Par chance le Trio n'a pas motivé son travail par l'importance des nombres premiers pour la cryptographie ou le e-commerce, mais dans le sillage de Don Knut et de sa conscience historique, ils lui ont repris une citation du grand Carl Friedrich Gauß, article 329 des *Disquisitiones Arithmeticae* (1801) :

Problema, numeros primos a compositis dignoscendi, hosque in factores suos primos resolvendi, ad gravissima ac utilissima totius arithmeticae pertinere, et geometrarum tum veterum tum recentiorum industriam ac sagacitatem occupavisse, tam notum est, ut de hac re copiose loqui superfluum foret.[...] praetereaque scientiae dignitas requirere videtur, ut omnia subsidia ad solutionem problematis tam elegantis ac celebris sedulo excolantur¹.

À l'école, on se familiarise avec le crible d'Eratosthène : malheureusement son utilisation pour prouver que n est premier nécessite un temps de calcul à peu près proportionnel à n lui-même. D'autre part la donnée de la longueur² d'un nombre est proportionnelle au nombre de chiffres dans son écriture binaire, soit de l'ordre de $\log_2 n$, on a donc en face de soi un algorithme à temps exponentiel $O(2^{\log_2 n})$. Pour citer encore une fois Gauß, *Disquisitiones Arithmeticae*, article 329 :

Nihilominus fateri oportet, omnes methodos hucusque prolatas vel ad casus valde speciales restrictas esse, vel tam operosas et prolitax, ut [...] ad maiores autem plerumque vix applicari possint³.

Est-ce que le caractère premier de chaque grand nombre peut *en principe* être décidé efficacement ? Cette question se théorise mathématiquement à travers la théorie moderne de la complexité en demandant un temps de calcul polynomial.

¹ Le problème où l'on se propose de distinguer les nombres premiers des nombres composés, et de décomposer ceux-ci en leurs facteurs premiers, est connu comme un des plus importants et des plus utiles de toute l'Arithmétique ; tout le monde sait qu'il a été l'objet de recherches des géomètres tant anciens que modernes, et il serait inutile de donner des détails à cet égard.[...] En outre la dignité de la science semble demander que l'on recherche avec soin tous les secours nécessaires pour parvenir à la solution d'un problème si élégant et si célèbre.[trad. Pouillet-Delisle, Ed. Blanchard, 1953.]

² La différence entre la taille d'un nombre et sa longueur se voit le plus clairement avec ces géants manifestes que sont le nombre d'atomes de l'univers (environ 10^{79}), ou la totalité de toutes les opérations arithmétiques jamais menées à bien par un homme ou une machine (environ 10^{24}) : 80 (respectivement 25) chiffres décimaux peuvent être écrit assez rapidement.

³ Cependant on ne peut s'empêcher de convenir que toutes les méthodes proposées jusqu'à présent sont soit restreintes à des cas très particuliers soit si longues et pénibles [...] qu'elles ne sont pour ainsi dire pas applicables à de plus grands nombres.

Existe-t-il un algorithme déterministe⁴ qui, pour un certain exposant κ , peut décider pour chaque entier n s'il est premier ou non en un temps de l'ordre de $\log^\kappa n$? soit, en raccourci, la question ouverte depuis si longtemps : est-ce que $\text{PRIMES} \in \mathcal{P}$?

L'état des choses avant août 2002

Jamais, depuis Gauß, la question de la primalité d'un nombre n'a été séparée de celle de trouver une factorisation (partielle) dans le cas composé. Dans l'article 334 des *Disquisitiones Arithmeticae* il écrit :

...posterior autem eatenus praestat, quod plerumque calculum expeditiorem permittit, sed factores ipsos numererorum compositorum, quos quoque a primis protinus distinguit, interdum non profert,⁵

Le point de départ de beaucoup de ces méthodes est le petit théorème de Fermat. Il dit que pour tout nombre premier n et tout nombre a premier avec n on a la relation

$$a^n \equiv a \pmod{n}.$$

Malheureusement la réciproque est fautive : on ne peut pas caractériser ainsi les nombres premiers. D'un autre côté « using the Fermat congruence is so simple, that it seems a shame to give up on it just because there are a few counter examples » (Carl Pomerance). Il ne faut donc pas s'étonner que des raffinements de ce critère soient à la base d'importants algorithmes.

Un algorithme *probabiliste* élémentaire de 1976, dû à Miller et Rabin, utilise un générateur de nombres aléatoires et dit, après k opérations, soit que le nombre est de façon certaine composé, soit qu'il est premier avec une grande probabilité, la probabilité d'erreur étant inférieure à 4^{-k} . Le temps de complexité est de l'ordre de $O(k \log^2 n)$, où le grand O contient une constante relativement petite. En pratique l'algorithme est assez rapide et il trouve des applications en cryptographie et dans le e-commerce pour la production d'« industrial-grade primes » (Henri Cohen). Dans le langage de la théorie de la complexité on dit en abrégé $\text{PRIMES} \in \text{co-}\mathcal{RP}$.

Un algorithme déterministe de Adleman, Pomerance et Rumely de 1983, qui utilise beaucoup plus de théorie et une généralisation du petit théorème de Fermat aux nombres dans des corps cyclotomiques, caractérise complètement les nombres premiers. C'était le meilleur algorithme déterministe avant août 2002, avec un temps de complexité d'ordre superpolynomial en $(\log n)^{O(\log \log \log n)}$. Le triple logarithme de l'exposant croît si lentement, de toute façon, que des

⁴ C'est-à-dire un algorithme qui ne nécessite pas de nombres aléatoires, par opposition à un algorithme probabiliste qui nécessite de tels nombres, F. Morain parle dans ce cas d'algorithme randomisé [11].

⁵ La seconde [manière] a l'avantage de l'emporter le plus souvent par la brièveté des calculs, mais quelquefois elle ne donne pas les facteurs des nombres composés, [...] au reste elle distingue avec autant de facilité les nombres premiers des nombres composés.

versions concrètes de l’algorithme ont eu d’excellents résultats dans la course au record du plus grand nombre premier, avec plus de mille chiffres décimaux⁶.

Une autre catégorie d’algorithmes modernes utilise des courbes elliptiques ou des variétés abéliennes de grand genre. Ainsi Adleman et Huang, dans une monographie de 1992 très difficile et technique, pouvaient donner un algorithme probabiliste à temps d’exécution polynomial qui après k opérations donne soit une réponse définitive (sans possibilité d’erreur) soit aucune réponse, ce dernier cas ayant de toute façon une probabilité inférieure à 2^{-k} . Dans le langage de la théorie de la complexité on dit en abrégé $\text{PRIMES} \in \mathcal{ZPP}$.

Avec cet arrière plan, et étant donné le niveau de difficultés qui avait été atteint et de l’absence de nouveaux résultats pendant plus de dix ans, il était difficile d’espérer qu’il puisse exister une réponse à cette question, courte, élégante et qui puisse être compréhensible pour l’« homme ordinaire ».

Entrée de Manindra Agrawal



Manindra Agrawal

L’informaticien et théoricien de la complexité Manindra Agrawal passa son doctorat en 1991 au Département d’Informatique et d’Ingénierie de l’Indian Institute of Technology de Kanpur (IITK). Après un séjour à l’Université d’Ulm en 1995-96 comme Humboldt-Stipendiat (« I really enjoyed the stay in Ulm. It

⁶ Le héros d’une autre histoire, Preda Mihăilescu, avait, dans sa thèse à l’ETH Zürich, essentiellement développé des raffinements de cet algorithme et en l’implémentant il fut pendant longtemps un des participants au jeu du record dans les nombres premiers. Récemment, il a démontré la conjecture de Catalan, voir la *Gazette* d’octobre 2002.

helped me in my research and career in many ways. »), il retourna à Kanpur comme professeur. Il se fit connaître, il y a deux ans, en montrant une forme faible de la conjecture d'isomorphisme en théorie de la complexité⁷.

Vers 1999 il travaillait avec son directeur de thèse Somenath Biswas sur la question de la reconnaissance des polynômes par un algorithme probabiliste. Comme une simple application, un nouveau test de primalité probabiliste apparaissait dans la publication « Primality and Identity Testing via Chinese Remaindering » [1].

Le point de départ était une généralisation du petit théorème de Fermat aux *polynômes*, un exercice facile à mettre dans un cours d'introduction à la théorie des nombres ou à l'algèbre : si les entiers naturels a et n sont premiers entre eux, alors n est premier *si et seulement si*

$$(x - a)^n \equiv (x^n - a) \pmod{n}$$

dans l'anneau des polynômes $\mathbb{Z}[x]$. Bien que ce soit une caractérisation très élégante des nombres premiers, cela reste difficilement utilisable. Le seul calcul de $(x - a)^n$ nécessite un temps de calcul supérieur à celui du crible d'Eratosthène. Mais c'est précisément pour des polynômes de cette taille qu'Agrawal et Biswas avaient développé un test probabiliste d'identité, avec une probabilité d'erreur bornée, et qui évitait complètement le développement du polynôme. Malheureusement le test qui en résultait, avec un temps d'exécution polynomial, était loin d'être compétitif avec celui de Miller et Rabin. Une nouvelle idée était née, mais initialement elle n'était intéressante que comme post-scriptum dans l'histoire des tests de primalité.

Deux années plus tard, Agrawal commençait à examiner en détail, avec ses étudiants de l'IITK, les potentialités de cette nouvelle caractérisation des nombres premiers dans laquelle il avait une grande confiance.

Deux projets de licence

La procédure d'admission à l'IIT est rigoureuse et sélective. Il y a une procédure ordinaire en deux étapes appelée Joint Entrance Examination (JEE) pour l'admission à l'une des sept branches de l'IIT et deux autres institutions. L'année dernière 150 000 indiens se sont présentés, et après un examen initial de trois heures en mathématiques, physique et chimie, 15 000 étaient invités à un deuxième test consistant en un examen de deux heures dans chacune de ces matières. Finalement 2 900 étudiants reçurent une place, parmi lesquels 45 étaient en informatique au très réputé IIT de Kanpur. Il ne faut pas s'étonner que l'on gagne bien sa vie en Inde en préparant les candidats au redouté JEE et que les diplômés du IIT soient tout de suite engagés dans le monde entier.

C'était donc avec des étudiants très motivés qu'Agrawal continuait maintenant le travail sur les tests de primalité. Avec Rajat Bhattacharjee et Prashant Pandey était apparue l'idée de regarder non pas la puissance polynomiale excessivement grande $(x - a)^n$ mais plutôt son reste après division par $x^r - 1$.

⁷ La conjecture d'isomorphisme de Berman et Hartmanis implique que $\mathcal{P} \neq \mathcal{NP}$. Une preuve donnerait donc la solution du premier des sept Problèmes du Millenium du Clay-Institut et rapporterait un million de dollars.



Neeraj Kayal



Nitin Saxena

Si r est d'un ordre logarithmique en n , alors ce reste, beaucoup plus petit, peut être calculé directement en temps polynomial avec un algorithme approprié. Si n est premier, alors certainement⁸

$$(T_{r,a}) \quad (x-a)^n \equiv x^n - a \pmod{(x^r - 1, n)}$$

pour tout r et n premier avec a . Quel a et quel r permettent de conclure inversement que n est premier ? Dans leur projet de licence en commun [5], les deux étudiants ont fixé $a = 1$ et examiné les conditions sur r . En analysant les résultats pour $r \leq 100$ et $n \leq 10^{10}$, ils arrivèrent à la conjecture suivante. Si r et n sont premiers entre eux et si

$$(T_{r,1}) \quad (x-1)^n \equiv x^n - 1 \pmod{(x^r - 1, n)}$$

alors soit n est premier, soit $n^2 \equiv 1 \pmod{r}$. Pour l'un des premiers nombres premiers r d'ordre $\log_2 n$, le deuxième cas n'est pas vérifié, ainsi on aurait la preuve de la primalité de n avec un temps de calcul en $O(\log^{3+\varepsilon} n)$.

Ici interviennent les héros de notre histoire, en coulisse jusqu'à maintenant, les étudiants Neeraj Kayal et Nitin Saxena. Tous les deux étaient membres de l'équipe indienne de l'olympiade internationale de mathématiques en 1997. Ayant étudié l'informatique plutôt que les mathématiques à cause des meilleures perspectives d'emploi, ils trouvèrent avec la théorie de la complexité un moyen de continuer à travailler à un haut niveau avec les mathématiques.

Dans leur projet de licence en commun ils examinèrent la relation du test $(T_{r,1})$ avec d'autres tests de primalité connus qui donnent, comme $(T_{r,1})$, dans le cas négatif une preuve que n est composé, et dans le cas positif, aucune réponse définitive. La récolte fut riche.

Ils purent montrer que si on admet l'hypothèse de Riemann, le test $(T_{r,1})$ peut être restreint à $r = 2, \dots, 4 \log_2^2 n$ pour donner une preuve de primalité. De cette façon on devrait obtenir un algorithme déterministe avec un temps de complexité en $O(\log^{6+\varepsilon} n)$. De plus ils purent montrer que la conjecture formulée par Bhattacharjee et Pandey devrait découler d'une vieille conjecture de Carl Pomerance.

Enfin ils mirent en avant une idée liée aux investigations concernant la classe des nombres « introspectifs » qui plus tard allait devenir essentielle.

Le travail qu'ils présentèrent en commun en avril 2002 portait le titre « Vers un test de primalité déterministe à temps polynomial ». Une vision, le but est déjà clairement en vue.

Changement de point de vue

Cet été là ils ne rentrèrent pas tout de suite à la maison mais plutôt commencèrent directement leurs études doctorales. En fait Saxena aurait voulu partir à l'étranger mais – ironie du sort – il n'obtint pas la bourse à l'université qu'il avait choisie.

Un petit changement de point de vue était encore nécessaire. Les deux projets de licence étudiaient le test $(T_{r,a})$ à $a = 1$ fixé et r variable. Qu'arrive-t-il si

⁸ Je suis les notations de Agrawal & al. et note $p(x) \equiv q(x) \pmod{(x^r - 1, n)}$ si la division de $p(x)$ et $q(x)$ par $x^r - 1$ puis la division des coefficients par n donnent le même reste.

au contraire on fixe r et laisse varier a ? La percée apparut le 10 juillet : avec un choix judicieux du paramètre ils obtenaient rien moins qu'une caractérisation des puissances de nombres premiers.

Le résultat, dans la version simplifiée qu'en a donné Dan Bernstein est le suivant.

Théorème d' Agrawal-Kayal-Saxena. — Soient $n, s, q, r \in \mathbb{N}$, qui vérifient $s \leq n$, q, r premiers, $q \mid r - 1$, $n^{(r-1)/q} \not\equiv 0, 1 \pmod r$ et

$$\binom{q+s-1}{s} \geq n^{2\lfloor\sqrt{r}\rfloor}.$$

Si pour tout a , $1 \leq a < s$, on a

- (i) a est relativement premier avec n et
- (ii) $(x-a)^n \equiv x^n - a \pmod{(x^r-1, n)}$, dans l'anneau des polynômes $\mathbb{Z}[x]$

alors n est une puissance d'un nombre premier.

La preuve, simple, courte et innovante, est tellement plaisante que je ne résiste pas à l'esquisser en appendice.

Le théorème conduit maintenant directement au bien nommé **algorithme AKS**⁹

1. Décide si n est une puissance d'un entier naturel. Si c'est le cas, vas au point 5.
2. Choisis (q, r, s) satisfaisant les hypothèses du théorème.
3. Pour $a = 1, \dots, s-1$ fais ce qui suit :
 - (i) si a divise n , vas au point 5
 - (ii) si $(x-a)^n \not\equiv x^n - a \pmod{(x^r-1, n)}$, vas au point 5.
4. n est premier, c'est fini.
5. n est composé, c'est fini.

Le point 1 peut être exécuté en temps polynomial en utilisant une variante de l'itération de Newton. En utilisant une arithmétique accélérée en base FFT, le temps d'exécution du point 3, qui est le plus important, est en $\tilde{O}(sr \log^2 n)$, où le tilde sur le grand O contient des termes logarithmiques en s , r et $\log_2 n$.

En conséquence pour atteindre notre but nous devons permettre à s et r de croître au plus de façon polynomiale en $\log n$. C'est le but du point 2. On commence par montrer ce qui en principe est possible. Posons $s = \theta q$, avec un facteur θ fixé. La formule de Stirling donne la relation asymptotique

$$\log \binom{q+s-1}{s} \sim c_\theta^{-1} q.$$

En conséquence, les conditions du théorème nécessitent l'estimation asymptotique

$$q \gtrsim 2c_\theta \lfloor \sqrt{r} \rfloor \log n.$$

⁹ À <http://www.ma.tum.de/m3/ftp/Bornemann/PARI/aks.txt> se trouve une implantation exécutable pour le paquet de logiciels de théorie des nombres en accès libre PARI-GP (<http://www.parigp-home.de>).

Pour n grand, cela ne peut arriver essentiellement que s'il existe une infinité de nombres premiers r tels que $r - 1$ a un facteur premier q qui vérifie $q \geq r^{1/2+\delta}$. Ainsi on se retrouve confronté à un problème bien étudié en théorie analytique des nombres.

Sophie Germain et le dernier théorème de Fermat

Le meilleur rapport coût-bénéfice q/r est obtenu pour des nombres premiers qui portent le nom de Sophie Germain : ce sont les nombres premiers impairs q tels que $r = 2q + 1$ est aussi premier. Elle a montré en 1823 que pour de tels nombres premiers ce que l'on appelle le premier cas du dernier théorème de Fermat se réalise : $x^q + y^q = z^q$ n'a pas de solution entière avec $q \nmid xyz$. Par conséquent la question de savoir s'il existe une infinité de ces nombres premiers sympathiques devenait d'un intérêt brûlant. Malheureusement on ne connaît toujours pas la réponse, même maintenant. Cependant des considérations heuristiques amenèrent Hardy et Littlewood, en 1922, à la conjecture très précise suivante concernant la densité des nombres premiers de Germain :

$$\#\{q \leq x : q \text{ et } 2q + 1 \text{ sont premiers}\} \sim \frac{2C_2 x}{\ln^2 x},$$

où $C_2 = 0.6601618158\dots$ est la constante des nombres premiers jumeaux.

Si cette conjecture était correcte on trouverait des nombres premiers q et $r = 2q + 1$ de taille $O(\log^2 n)$ satisfaisant aux hypothèses du théorème. L'algorithme AKS aurait alors un temps d'exécution polynomial en $\tilde{O}(\log^6 n)$. Comme la conjecture a été confirmée de façon impressionnante jusqu'à $x = 10^{10}$, l'algorithme AKS se comporte comme un algorithme de complexité en $\tilde{O}(\log^6 n)$ tant que n n'a pas plus de 100 000 chiffres.

En 1985, environ dix ans avant qu'Andrew Wiles ne démontre finalement le dernier théorème de Fermat, Adleman, Fouvry et Heath-Brown démontraient que la solution ne dépendait pas des nombres premiers de Germain : plus précisément le premier cas du dernier théorème de Fermat est vérifié pour une infinité de nombres premiers [8]. En fait Adleman et Heath-Brown étudiaient, comme une généralisation des premiers de Germain, exactement ces paires (q, r) qui jouent aussi un rôle clé dans l'algorithme AKS.

Une médaille Fields

Ce dont ils avaient besoin c'est que l'estimation

$$\#\{r \leq x : q, r \text{ premiers} ; q \mid r - 1 ; q \geq x^{1/2+\delta}\} \geq c_\delta \frac{x}{\ln x}$$

soit vérifiée pour un exposant $\delta > 1/6$ convenable. La chasse au plus grand δ s'ouvrit en 1969 avec Morris Goldfeld [7] qui obtint $\delta \approx 1/12$, et se conclut, jusqu'à présent, en 1985 avec Étienne Fouvry [6] avec pour valeur de δ , $\delta = 0,1687 > 1/6$. Tous ces travaux utilisaient des méthodes profondes de la théorie analytique des nombres qui se développèrent avec *le grand crible* d'Enrico Bombieri. Il publia ce crible en 1965 à l'âge de vingt-cinq ans, et reçut la médaille Fields en 1974. Ainsi une lourde tâche incombe à l'« homme

ordinaire » qui désire comprendre la preuve de l'estimation dans le détail. En réponse à ma question pour savoir si l'un des trois entreprit cette tâche, Manindra Agrawal écrivit :

« *We tried! But Sieve theory was too dense for us – we have no background in analytical number theory. So after a while we just gave up.* »

Et aussi ils n'avaient pas besoin de le faire car « the result was stated there in precisely the form we needed », et ils pouvaient compter sur sa validité en faisant confiance au referee et un certain laps de temps écoulé depuis que le résultat de Fouvry relié au sujet brûlant du dernier théorème de Fermat était paru dans *Inventiones*.

Ou peut-être pas? Fouvry avait oublié de prendre en considération une condition supplémentaire en citant le théorème de Bombieri, Friedlander et Iwaniec. Cette condition supplémentaire ramène la valeur de δ à $\delta = 0,1683 > 1/6$. Ça aurait pu aussi être en-dessous du seuil critique. Par la suite Fouvry a parlé de cette correction à Roger Baker et celui-ci l'a publié avec Glyn Harman dans un article de survey [3] en 1996.

Incidentement, c'est dans une recherche sur Internet avec *Google* qu'Agrawal, Kayal et Saxena rencontrèrent l'article de Fouvry dans la bibliographie d'un article de Pomerance et Shparlinski. Lorsqu'ils se renseignèrent sur la meilleure valeur connue de δ Pomerance les a renvoyés à l'article de Baker et Harman.

Sans se soucier de la meilleure valeur, $\delta > 0$ suffit à garantir un triple (q, r, s) permis pour l'algorithme AKS avec la taille polynomiale nécessaire,

$$r = O(\log^{1/\delta} n), \quad q, s = O(\log^{1+1/2\delta} n).$$

Ainsi l'algorithme AKS a, de toute façon, un temps d'exécution en $\tilde{O}(\log^{3+3/2\delta} n)$. Donc le fait $\text{PRIMES} \in \mathcal{P}$ est prouvé et l'avancée est faite. Félicitations! La valeur correcte de Fouvry pour δ donne $\tilde{O}(\log^{11.913} n)$, ou de façon plus simple à retenir et aussi sans le tilde, $O(\log^{12} n)$ ¹⁰.

Le directeur de l'IIT de Kanpur, Sanjay Dhande, était tellement enthousiaste du gros titre dans le *New York Times* qu'il a déclaré qu'Agrawal devrait être élu aux plus hauts honneurs en Mathématiques¹¹. En 2006 Agrawal aura quarante ans.

¹⁰ Le 22 janvier 2003, Dan Bernstein publia sur le web une nouvelle version de son ébauche [4]. On y trouve une faible variation du théorème de Agrawal-Kayal-Saxena, qu'il a apprise de Lenstra, et qui permet de compléter la preuve de $\text{PRIMES} \in \mathcal{P}$ sans se référer à des résultats profonds de théorie analytique des nombres. Un théorème bien connu de Chebyshev, affirmant que le produit des nombres premiers $\leq 2k$ est supérieur ou égal à 2^k , suffit à garantir l'existence de nombres appropriés $r, s = O(\log^5 n)$ pour lesquels l'algorithme marche. Ainsi disparaît le dernier bloc de mathématiques difficiles qui aurait pu empêcher l'« homme ordinaire » de comprendre complètement le résultat. Paulo Ribenboim a probablement raison lorsqu'il m'écrit : « Our specialists should reflect about their convoluted reasoning. » [note du texte anglais]

¹¹ Déjà en 2002 il recevait le prix Clay pour la recherche. Les lauréats précédents étaient Andrew Wiles, les probabilistes Smirnov et Schramm et les récipiendaires de la Médaille Fields Connes, Lafforgue et Witten.

Et en pratique ?

La question des applications pratiques apparut rapidement dans les Newsgroups et les journaux, puisque jusqu'à présent les grands nombres premiers forment un ingrédient important de la cryptographie et du e-commerce. Nous croyons fermement qu'avant tout a été résolu un important problème *théorique* qui a échappé aux experts pendant plusieurs décennies. Agrawal lui-même souligne que le problème l'intéressait comme défi intellectuel et qu'actuellement l'algorithme AKS est beaucoup plus lent que ces algorithmes qui ont obtenu les records dans les preuves de primalité jusqu'à 5 020 chiffres décimaux¹². Finalement on ne devrait pas oublier que la définition d'une classe de complexité comme \mathcal{P} est une question purement théorique sur un comportement asymptotique quand $n \rightarrow \infty$. Et donc dans un cas particulier l'avantage en temps de calcul d'un algorithme polynomial par opposition à un algorithme super-polynomial pourrait très probablement ne devenir manifeste que pour n tellement grand qu'aucun des deux algorithmes ne donnerait une réponse dans la durée d'une vie avec les machines actuelles. En pratique les constantes dans le grand O de l'estimée de complexité interviennent aussi.

Des « industrial-grade primes » de moindre qualité, avec 512 chiffres binaires, peuvent être produits en une fraction de seconde en utilisant le test Miller-Rabin sur un PC de 2 GHz d'usage courant. S'il le faut, leur primalité peut même être montrée en quelques secondes grâce à la méthode ECPP de Atkin et Morain qui repose sur les courbes elliptiques¹³. Le temps de complexité de cet algorithme *probabiliste* est, très certainement « a cloudy issue » (Carl Pomerance), mais des considérations heuristiques suggèrent que la bonne valeur tourne autour de $\tilde{O}(\log^6 n)$.

D'un autre côté, à cause du coût élevé de la congruence polynomiale dans le troisième pas de l'algorithme AKS, la constante dans le $\tilde{O}(\log^6 n)$ du temps de calcul conjecturé est tellement grande que l'on estime que l'algorithme prendrait plusieurs jours pour un nombre premier à 512 bits, bien que Dan Bernstein, Hendrik Lenstra, Felipe Voloch, Bjorn Poonen et Jeff Vaaler aient déjà amélioré la constante par un facteur d'au moins 2.10^6 par rapport à la formulation originale de l'algorithme— la situation au 25 janvier 2003, cf. [4].

Il manque donc un facteur 10^5 pour atteindre un niveau compétitif. La méthode ECPP aussi a commencé par une idée nouvelle de Goldwasser et Kilian, complètement impraticable mais fondamentale. Comme la méthode qu'Agrawal, Kayal et Saxena ont produite maintenant est tellement inattendue, brillante et nouvelle, on peut prévoir avec confiance une amélioration de ses capacités après maturation de l'algorithme (voir déjà [10]).

¹² S'il vous plait ne confondez pas avec le record pour le plus grand nombre premier connu, qui est pour le moment $2^{13\,466\,917} - 1$, un nombre premier de Mersenne à 4 053 946 chiffres décimaux. Ces nombres ont beaucoup de structure ce qui permet d'utiliser des algorithmes très particuliers.

¹³ voir <http://www.znz.freesurf.fr/pages/prim0.html> pour le programme en accès libre PRIMO de Marcel Martin, qui pour le moment tient le record.

Média et téléphone arabe

À l'exception d'un compte-rendu très fouillé, techniquement correct, tout à fait lisible et détaillé dans l'hebdomadaire indien *Frontline* du 17 août 2002, les comptes-rendus des media généralistes ont été déplorables. Agrawal déjoua ma curiosité à propos de ses impressions par une politesse, « Leave aside the general public coverage ».

L'article déjà cité du *New-York times* a bien sûr célébré le résultat triomphalement, mais de façon opaque en choisissant de simplifier à l'extrême : un temps de calcul polynomial devient « quickly » ; déterministe devient « definitively ». L'article se lit donc ainsi : trois indiens ont obtenu une avancée car l'ordinateur pourra dire maintenant « quickly and definitively » si un nombre est premier. Par ailleurs l'algorithme n'a aucune application, car les méthodes qui existent déjà sont plus rapides et en pratique ne se trompent pas. « Quelle avancée ! » pourront se dire les lecteurs.

Les Associated Press (AP) transformèrent l'article du *New-York Times* en une information d'agence dans laquelle « definitively » devenait « accurately » et le côté temps de calcul disparaissait dans le fond. La triste fin de ce téléphone arabe fut le site du « Tagesschau ». Le 12 août, sous le titre « Finalement les nombres premiers peuvent être calculés exactement ! » apparut des bêtises du genre « la joie des écoles allemandes est sans borne : finalement on peut calculer des nombres premiers sans larmes ! » Ce rapport a été enlevé à la suite de protestations des participants du Newsgroup de `sci.mathematik`.

À part l'article du *New-York Times*, l'histoire a été virtuellement ignorée dans la presse américaine. Au Royaume-Uni une histoire dans le *New Scientist* du 17 août utilisait au moins les mots « polynomial time », mais ils en arrivaient à parler de « an algorithm that gives a definite answer to the problem in a reasonable time ». Un texte rétrospectif du 4 novembre dans le *Wall Street Journal* portait le titre trompeur « One beautiful mind in India is putting the Internet on alert ». Une colonne de fin d'année de Clive Thompson dans le *New-York Times* du dimanche 15 décembre affirmait « Ever since the time of the ancient Greeks, finding a simple way to prove a number is prime has been the holy grail of mathematics... This year, it finally arrived... This new algorithm could guarantee primes so massive they would afford almost perfect online security. »¹⁴

Et dans les grands journaux de langue allemande ? Le « Neue Züricher Zeitung » a publié son premier compte-rendu le 30 août 2002. L'article suggérait faussement que jusqu'à présent aucun certificat de primalité ne pouvait être calculé « dans un temps raisonnable » pour les nombres premiers utilisés en cryptographie et que les trois Indiens avaient réussi précisément cela ; le résultat, de toute façon ne fut pas tellement loué par les agences de presse parce qu'il ne pouvait pas manier le plus grand nombre premier connu.

Dans la section Arts du 9 août 2002, sous le titre « Dieux polynomiaux : des indiens prolifiques et leurs nombres premiers, » le *Frankfurter Allgemeine Zeitung* publia un texte sybilin qui commençait par un rapprochement entre

¹⁴ Ce paragraphe a été rajouté dans le texte anglais ; la presse française ne nous offre pas d'occasion d'augmenter ce florilège, l'évènement ayant été visiblement ignoré des grands media généralistes.

les mathématiques indiennes et le panthéon indou et ensuite laissaient tenir à quatre de ces divinités une courte discussion sur le nouveau résultat :

– « À qui cela est-il bon ? » protesta Agni, et Lakshmi rétorqua :

– « Aux hackers ! On a besoin des nombres premiers pour encoder les données de transmissions électroniques – il y a plusieurs algorithmes cryptographiques fameux comme RSA et le Data Encryption Standard DES ; leurs clés sont des nombres avec leur factorisation en nombres premiers, et si ça peut maintenant être fait facilement en un temps polynomial par rapport aux données... ».

– « Mais c'est déjà bien connu, par exemple par le test de Miller et Rabin, que si on itère assez, on peut trouver un test de primalité avec une probabilité aussi grande que l'on veut et correcte même pour les plus grands nombres, » contredit Rudra. « Et le codage par factorisation en nombres premiers n'a rien à voir avec un test de primalité, c'est un problème complètement différent ; pour les professionnels de la sécurité ce que ces types ont fait ne vaut rien. »

À l'aube, leur hôtesse Uschas trouva finalement les paroles de réconciliation :

– « Apprécions simplement un résultat élégant que les occidentaux aussi admirent et qui est dans la continuation d'inspiration de notre grande tradition mathématique ! »

Quel lecteur pourrait tirer de ceci les raisons de tout ce tapage ?

Plans pour le futur

Les trois ont l'intention de soumettre leur travail à *Annals of Mathematics* et ont été en contact avec Peter Sarnak à ce sujet. Ils veulent réécrire l'article « in a more 'mathematical' way as opposed to 'computer science' way as that would be more suitable in Annals ».

À propos aussi bien de l'état émotionnel que du futur des deux étudiants en thèse, Kayal et Saxena, Agrawal dit :

They are happy, but at the same time quite cool about it. I would say they are very level-headed boys. As for their PhD, yes I am sure that this work will qualify for their PhD. But I have advised them to stay back for a couple of years since this is the best time they have for learning. They still need to pick up so many things. But they are free to make the decision – they already have an offer from TIFR [Tata Institute of Fundamental Research].

Appendice

Ce qui suit est, comme promis, l'esquisse de la preuve du théorème de Agrawal-Kayal-Saxena. Je suis la présentation peaufinée par Dan Bernstein dans [4].

Esquisse de la preuve. Soit p un facteur premier de n qui vérifie déjà $p^{(r-1)/q} \not\equiv 0, 1 \pmod{r}$; nous montrons que si (i) et (ii) sont vérifiés pour tout a , $1 \leq a < s$, alors le nombre n est une puissance de p .

Pour faire cela on considère – comme Agrawal le fit en ce matin du 10 juillet quand le théorème a été trouvé – des produits de la forme $t = n^i p^j$ avec $0 \leq i, j \leq \lfloor \sqrt{r} \rfloor$. Le principe de Dirichlet donne deux paires distinctes (i_1, j_1) et (i_2, j_2) d'exposants tels que $t_1 = n^{i_1} p^{j_1} \equiv n^{i_2} p^{j_2} = t_2 \pmod{r}$. Le but est maintenant de prouver qu'en fait $t_1 = t_2$ et que donc $n = p^\ell$ pour un certain ℓ .

D'après le petit théorème de Fermat, il suit de (ii) que pour tout $1 \leq a \leq p$ et $\mu = 1, 2$

$$(*) \quad (x - a)^{t_\mu} \equiv x^{t_\mu} - a \pmod{(x^r - 1, p)}.$$

Dans leur projet de licence, Kayal et Saxena avaient appelé de tels exposants « introspectifs », et ils avaient montré que pour ceux-là la congruence $t_1 \equiv t_2 \pmod{r}$ se relève en la congruence $t_1 \equiv t_2 \pmod{\#G}$ avec $\#G \gg r$; pour un bon choix des paramètres, $\#G$ devient tellement grand que cela entraîne $t_1 = t_2$. Agrawal considère ce relèvement comme « the nicest part of the paper. »

Comment s'opère ce relèvement? Comme $t_1 \equiv t_2 \pmod{r}$, $x^r - 1$ divise la différence $x^{t_1} - x^{t_2}$, ce qui entraîne finalement, d'après (*), que

$$(x - a)^{t_1} \equiv (x - a)^{t_2} \pmod{(x^r - 1, p)}.$$

Donc $g^{t_1} = g^{t_2}$ pour tout $g \in G$, si G représente le sous-groupe multiplicatif engendré par les facteurs linéaires $(\zeta_r - a)$ dans le corps cyclotomique sur $\mathbb{Z}/p\mathbb{Z}$ obtenu par adjonction des r -ièmes racines de l'unité ζ_r . En prenant un élément primitif g , c'est-à-dire un élément d'ordre $\#G$, on montre que $\#G \mid (t_1 - t_2)$.

Par ailleurs, d'après la condition (i), et comme $p^{(r-1)/q} \not\equiv 0, 1 \pmod{n}$, le groupe G a – par un peu de combinatoire et de théorie élémentaire des polynômes cyclotomiques – au moins $\binom{q+s-1}{s}$ éléments. Donc d'après l'hypothèse sur les coefficients du binôme

$$|t_1 - t_2| < n^{\lfloor \sqrt{r} \rfloor} p^{\lfloor \sqrt{r} \rfloor} \leq n^{2\lfloor \sqrt{r} \rfloor} \leq \binom{q+s-1}{s} \leq \#G$$

d'où l'on déduit l'égalité désirée $t_1 = t_2$.

Note ajoutée à la preuve

Au début mars 2003, Agrawal, Kayal et Saxena ont publié sur le web une version révisée de leur prépublication : http://www.cse.iitk.ac.in/news/primality_v3.pdf. Elle contient l'amélioration due à Lenstra et calcule avec le nouveau temps de complexité borné par $O(\log^{7.5} n)$, voir Theorem 5.3.

Remerciements

Je remercie sincèrement Manindra Agrawal pour son empressement à répondre de façon personnelle et détaillée, malgré les milliers de courriers électroniques de félicitations, à mes questions concernant des informations de second plan.

Références

- [1] Manindra Agrawal and Somenath Biswas, *Primality and identity testing via Chinese remaindering*, in *40th Annual Symposium on Foundations of Computer Science*, 202–208, IEEE Computer Soc., Los Alamitos, CA, 1999.
- [2] Manindra Agrawal, Neeraj Kayal, Nitin Saxena, *PRIMES is in P*, IIT Kanpur, Preprint of August 8, 2002, <http://www.cse.iitk.ac.in/news/primality.html>.
- [3] Roger C. Baker, Glyn Harman, *The Brun-Titchmarsh Theorem on Average*, in *Proceedings of a conference in honor of Heini Halberstam, Vol. 1*, pp. 39–103, 1996.
- [4] Daniel Bernstein, *Proving Primality after Agrawal-Kayal-Saxena*, version of January 25, 2003, <http://cr.yp.to/papers.html#aks>.
- [5] Rajat Bhattacharjee, Prashant Pandey, *Primality Testing*, Bachelor of Technology Project Report, IIT Kanpur, April 2001, <http://www.cse.iitk.ac.in/research/btp2001/primality.html>.
- [6] Étienne Fouvry, *Théorème de Brun-Titchmarsh; application au théorème de Fermat*, Invent. Math. 79, 383–407, 1985.
- [7] Morris Goldfeld, *On the number of primes p for which $p + a$ has a large prime factor*, Mathematika 16, pp. 23–27, 1969.
- [8] D. Roger Heath-Brown, *The First Case of Fermat's Last Theorem*, Math. Intelligencer 7, no. 4 (1985), 40–47, 55.
- [9] Neeraj Kayal, Nitin Saxena, *Towards a Deterministic Polynomial-Time Primality Test*, Bachelor of Technology Project Report, IIT Kanpur, April 2002, <http://www.cse.iitk.ac.in/research/btp2002/primality.html>.
- [10] P. Mihăilescu, R. Avanzi, *Efficient « Quasi »-deterministic Primality Test Improving AKS. Draft*, <http://www-math.uni-paderborn.de/preda/>, Avril 2003.
- [11] F. Morain, *La primalité en temps polynomial*, Séminaire Bourbaki **917**, Mars 2003.
- [12] R. Ramachandran, *A prime solution*, Frontline, India's National Magazine, Vol. 19, issue 17 of August 17, 2002, <http://www.flonnet.com/fl1917/19171290.htm>.