Mathématiques et sciences humaines
Mathematics and social sciences

194 | Eté 2011
Varia

Le fossé de Sloane

Sloane's Gap

Nicolas Gauvrit, Jean-Paul Delahaye et Hector Zenil

Édition électronique
URL : http://journals.openedition.org/msh/12014
DOI : 10.4000/msh.12014
ISSN : 1950-6821

Éditeur
Centre d'analyse et de mathématique sociales de l'EHESS

Édition imprimée
Date de publication : 1 septembre 2011
Pagination : 5-17
ISSN : 0987-6936

Référence électronique
Nicolas Gauvrit, Jean-Paul Delahaye et Hector Zenil, « Le fossé de Sloane », *Mathématiques et sciences humaines* [En ligne], 194 | Eté 2011, mis en ligne le 02 septembre 2011, consulté le 19 avril 2019. URL : http://journals.openedition.org/msh/12014 ; DOI : 10.4000/msh.12014

Math. & Sci. hum. / Mathematics and Social Sciences (49e année, n° 194, *2011*(2), p. 5–17)

LE FOSSÉ DE SLOANE

Nicolas GAUVRIT[1], Jean-Paul DELAHAYE[2], Hector ZENIL[3]

RÉSUMÉ – *La base de Sloane (Online Encyclopedia of Integer Sequences) réunit plusieurs dizaines de milliers de suites mathématiques considérées comme « intéressantes » par certains mathématiciens. La représentation graphique de la fréquence d'occurrence de n en fonction de n montre une fonction rapidement décroissante, et un nuage qui semble séparé en deux par une zone claire qu'on nomme ici le fossé de Sloane.*

La décroissance et la forme générale s'expliquent assez facilement mathématiquement, mais l'explication du fossé nécessite d'autres considérations.

MOTS CLÉS – Base de Sloane, Complexité de Kolmogorov

SUMMARY – Sloane's Gap

Sloane's (Online Encyclopedia of Integer Sequences) contains thousands of numerical sequences considered to be particularly interesting by some mathematicians. The graphic representation of the frequency with which a number n appears in that database (as a function of n) shows that the underlying function is rapidly decreasing, and that the points are split into two sub-clouds, separated by a clear zone that we shall call Sloane's gap.

The decrease and global shape are easily explained mathematically. On the other hand, explaining the presence of the Sloane's gap requires further investigation.

KEYWORDS – Kolmogorov complexity, Online Encyclopedia of Integer Sequences

L'encyclopédie des suites numériques de Neil Sloane[4] est un objet remarquable car elle est le résultat d'un travail de sélection et d'exploration du monde des suites de nombres entiers poursuivi avec méthode et obstination depuis plus de 40 ans [Cipra, 1994]. Elle recense aujourd'hui plus de 150 000 suites numériques (158 000 en mai 2009). Son élaboration a impliqué des centaines de mathématiciens, ce qui lui confère une sorte d'homogénéité et d'objectivité mathématique générale (cette objectivité délicate à définir sera discutée plus loin).

[1]Laboratoire de Didactique André Revuz (LDAR), EA 1547, Centre Chevaleret, Université Paris VII, 175 rue du Chevaleret 75013 Paris, adems@free.fr

[2]Laboratoire d'Informatique Fondamentale de Lille (LIFL), UMR CNRS 8022, Université des sciences et technologie de Lille, bâtiment M3 59655 Villeneuve d'Ascq Cedex, jean-paul.delahaye@lifl.fr

[3]Laboratoire d'Informatique Fondamentale de Lille (LIFL), UMR CNRS 8022, Université des sciences et technologie de Lille, bâtiment M3 59655 Villeneuve d'Ascq Cedex, et IHPST, Paris I/ENS, 13 rue du Four 75006 Paris, hector.zenil@lifl.fr

[4]Cette encyclopédie est consultable à l'adresse : http://www.research.att.com/∼njas/sequences/. Consultée le 3 août 2009.

Nous nous intéresserons plus particulièrement au nombre d'occurrences $N(n)$ d'un entier n dans la base. Ce nombre $N(n)$ marque l'importance de n et il varie très sensiblement d'un nombre à l'autre, même pour des nombres voisins. Cette importance peut être mathématiquement objective (2^{10} est un exemple de nombre « important » en ce sens) ou seulement en fonction d'une culture mathématique partagée (10^9 est plus important pour nous que 9^{10} parce que nous utilisons un système de numération décimale). La complexité de Kolmogorov est un outil qui – au moins en principe – nous donne des indications sur l'allure que devrait présenter la courbe représentative de N fondée sur une mesure d'importance « objective ». L'écart observé vis-à-vis d'une courbe qui se baserait sur la complexité de Kolmogorov est cependant très net. Le *fossé de Sloane* est une zone creuse de points $(n, N(n))$ récemment repérée par Philippe Guglielmetti[5]. Ce fossé est inattendu et nécessite une explication.

1. PRÉSENTATION DE LA BASE

L'encyclopédie se présente comme un catalogue de suites de nombres entiers et non pas comme une liste de nombres. Cependant la façon dont elle est conçue en fait aussi un dictionnaire de nombres permettant de rechercher quelles sont les propriétés particulières d'un entier donné et combien de propriétés remarquées possède un nombre entier donné.

Un usage courant de l'encyclopédie de Sloane consiste à rechercher quelle pourrait bien être la logique d'une suite d'entiers. Si vous lui soumettez : $3, 4, 6, 8, 12, 14, 18, 20, ...$ vous obtenez instantanément qu'il s'agit (sans doute) de la suite des nombres premiers augmentés de 1 : 2+1, 3+1, 5+1, 7+1, 11+1, 13+1, 17+1, 19+1, ...

Plus intéressant peut-être, le programme permet d'interroger la base sur un nombre isolé. Considérons à titre d'exemple le nombre de Hardy-Ramanujan 1729 (le plus petit entier somme de deux cubes de deux façons différentes). Le programme indique qu'il connaît plus de 350 suites auxquelles appartient 1729. Chacune identifie une propriété de 1729 qu'il est possible d'examiner. Les réponses sont classées par ordre d'importance, notion basée sur les citations des séquences dans les commentaires mathématiques et les références croisées que contient l'encyclopédie. Apparaît en premier la propriété que 1729 est le troisième nombre de Carmichael (nombre n non premier pour lequel $\forall a \in \mathbb{N}^*$, $n|a^n - a$). La seconde propriété de 1729 est qu'il s'agit du sixième pseudo-premier en base 2 (nombre n non premier tel que $n|2^{n-1} - 1$). La troisième propriété est l'appartenance de 1729 aux termes d'une série génératrice simple. La propriété repérée par Ramanujan sur son lit d'hôpital vient en quatrième. En parcourrant les réponses de l'encyclopédie, on trouvera aussi que :

- 1729 est le treizième nombre de la forme $n^3 + 1$;
- 1729 est le quatrième nombre "factoriel sextuple" c'est-à-dire un produit de termes successifs de la forme $6n + 1$: $1729 = 1 \times 7 \times 13 \times 19$;

[5]Sur son site http://drgoulu.com/2009/04/18/nombres-mineralises/. Consulté le 3 août 2009.

- 1729 est le neuvième nombre de la forme $n^3 + (n+1)^3$;
- 1729 est la somme des diviseurs d'un carré parfait (33^2) ;
- 1729 est un nombre dont la somme des chiffres est en même temps le plus grand facteur premier (car $1+7+2+9 = 19$ et $1729 = 7 \times 13 \times 19$) ;
- 1729 est le produit d'un nombre premier, 19, par son nombre inversé, 91 ;
- 1729 est le nombre de façons d'écrire 33 comme somme de 6 entiers.

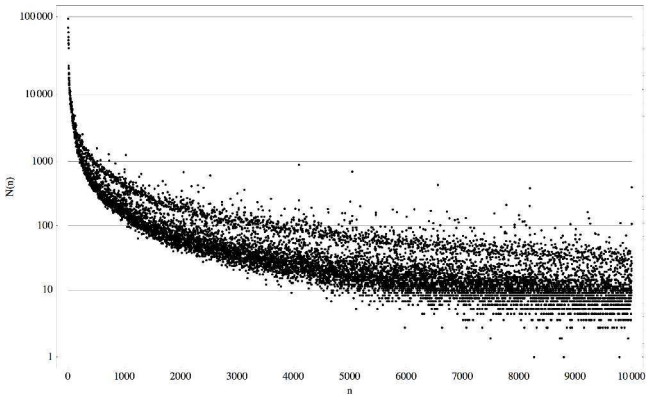

FIGURE 1. Nombre d'occurrences $N(n)$ en fonction de n, pour n variant de 1 à 10 000. Échelle logarithmique en ordonnée.

L'encyclopédie des suites de Neil Sloane comporte plus de 150 000 suites. Une version partielle ne retenant que les suites les plus importantes de la base a été publiée par Neil Sloane et Simon Plouffe en 1995. Elle retient une sélection de 5487 suites [Sloane & Plouffe, 1995] et fait écho à une publication plus ancienne de Sloane [1973].

Une quarantaine de mathématiciens participent à un « comité d'éditeurs » de la base, mais tout utilisateur peut proposer des suites. Si elles sont validées elles viennent s'ajouter à la base. La sélection se fait sur des critères d'intérêt mathématique. Neil Sloane indique qu'il adopte un comportement assez souple et ajoute donc facilement les nouvelles suites qu'on lui propose. Un certain filtre est inévitable pour que la base garde un intérêt. D'ailleurs, il existe un très grand nombre de familles infinies de suites (toutes les suites de la forme (kn), toutes les suites de la forme (k^n), etc.), dont bien sûr seules les premières sont retenues pour l'encyclopédie. Un programme annexe à utiliser en cas d'échec d'une recherche directe permet de reconnaître les suites de telles familles, non explicitement présentes dans l'encyclopédie.

Chaque suite retenue dans la base l'est sous la forme de données (ses premiers termes). La taille des données numériques associées à une suite se limite à environ 180 chiffres (les premiers termes de la suite). Cela a pour effet bien sûr que, même si la suite est facile à calculer, seuls ses premiers termes sont présents. À côté de la donnée en extension du début de la suite, l'encyclopédie propose toutes sortes

d'autres informations sur la suite, dont plusieurs définitions s'il y en a plusieurs, et des références bibliographiques.

L'encyclopédie numérique de Sloane est disponible sous la forme d'un fichier facilement exploitable ne contenant que les termes retenus pour chaque suite. On peut télécharger librement ce fichier et l'utiliser – par exemple avec un logiciel mathématique – pour étudier les nombres présents et mener un travail d'exploration statistique sur les données qu'elle contient.

On peut par exemple se poser la question : « Quels sont les nombres qui n'apparaissent pas dans l'encyclopédie de Sloane »? Lors d'un premier calcul mené en août 2008 par Philippe Guglielmetti, le plus petit nombre absent trouvé fut 8795, suivi dans l'ordre par 9935, 11147, 11446, 11612, 11630, ... En reprenant les calculs en février 2009, l'encyclopédie ayant été complétée par quelques centaines de suites nouvelles, la suite de ces nombres absents était devenue 11630, 12067, 12407, 12887, 13258, ...

L'instabilité dans le temps de cette suite est un peu ennuyeuse, et suggère d'étudier la distribution des nombres plutôt que leur seule présence ou absence. Considérons le nombre de propriétés d'un entier, $N(n)$, en le mesurant par le nombre de fois où n apparaît dans le fichier numérique de l'encyclopédie de Sloane. La valeur $N(n)$ mesure l'intérêt de n. La suite $N(n)$ est certes instable dans le temps, mais elle varie lentement et certaines notions qu'on peut tirer des valeurs de $N(n)$ sont même très stables.

Les valeurs de $N(n)$ sont représentées sur la Figure 1. On obtient un nuage à l'allure régulière dans cette représentation en échelle logarithmique.

Donnons quelques exemples : la valeur de $N(1729)$ est 380 (février 2009), ce qui est assez élevé pour un nombre de cet ordre de grandeur. Pour son prédécesseur, on a cependant $N(1728) = 622$ ce qui est encore mieux. Le nombre 1728 aurait donc été plus facile pour Ramanujan ! À l'inverse $N(1730) = 106$ et donc 1730 aurait constitué une épreuve plus difficile que 1729.

La suite $(N(n))_{n\in\mathbb{N}^*}$ est globalement d'allure décroissante, cependant certains nombres n contredisent cette règle et possèdent plus de propriétés que leur prédécesseur : $N(n) > N(n-1)$.

Nous pouvons nommer « intéressants » de tels nombres. Le premier nombre intéressant selon cette définition est 15 car $N(15) = 34183$ et $N(14) = 32487$. Viennent ensuite 16, 23, 24, 27, 28, 29, 30, 35, 36, 40, 42, 45, 47, 48, 52, 53, etc.

Insistons sur le fait que bien qu'incontestablement dépendants de certaines décisions particulières faites par ceux qui participent à l'élaboration de cette base de suites, celle-ci n'est pas arbitraire. Les contributeurs sont très nombreux, et on peut défendre l'idée que la base représente une vue objective (ou au moins intersubjective) du monde numérique, vue indépendante de chaque personne qui y contribue et reflet d'une réalité mathématique (ou culturelle) stable.

Un argument indirect de cette indépendance globale de l'encyclopédie résultant du travail cumulé d'une communauté de mathématiciens est la forme générale du nuage de points déterminé par $N(n)$ qui est étonnamment régulier, comme le serait

un nuage provenant d'une expérience de physique.

Philippe Guglielmetti a remarqué que ce nuage présente une caracéristique remarquable[6] : il est divisé en deux parties séparées par une zone claire, comme si les nombres se séparaient naturellement en deux catégories : les plus intéressants – au-dessus de la zone claire – et les moins intéressants – en dessous. Nous nommons *fossé de Sloane* cette zone claire séparant en deux le nuage représentant le graphe de la fonction $n \longmapsto N(n)$.

Notre but est de décrire la forme du nuage, puis de formuler une hypothèse explicative de cette forme.

2. DESCRIPTION DU NUAGE

Après avoir brièvement décrit la forme globale du nuage, nous nous attacherons plus particulièrement au *fossé*, et nous chercherons ce qui caractérise les points se situant au-dessus.

2.1. FORME GÉNÉRALE

Le nombre d'occurrences N est proche d'une fonction globalement décroissante concave de n, comme on peut le voir sur la Figure 1.

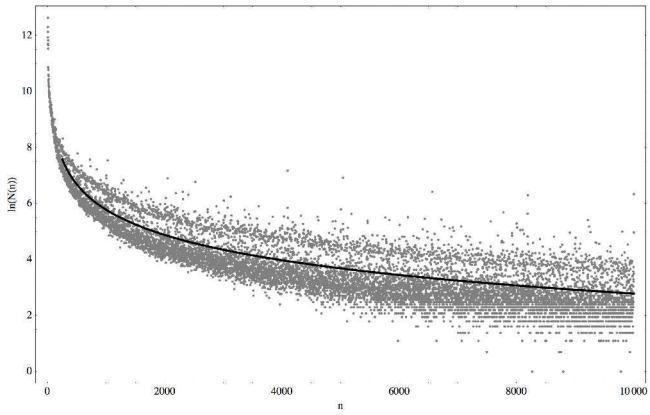

FIGURE 2. $\ln(N(n))$ en fonction de n, pour n variant de 1 à 10 000. La courbe représente la régression logarithmique de $\ln(N)$ en n.

Une régression logarithmique permet d'avoir une idée plus précise de la forme du nuage pour n variant de 1 à 10 000. Sur cet intervalle, le coefficient de détermination de la régression logarithmique de $\ln(N(n))$ en n est de $r^2 = 0,81$, et l'équation de régression donne l'estimation

$$\ln(N(n)) \simeq -1,33\ln(n) + 14,76$$

soit

[6]Communication personnelle à l'un des auteurs, 16 février 2009.

$$\hat{N}(n) = \frac{k}{n^{1,33}}$$

où k est une constante valant ici environ $2,57 \times 10^6$, et $\hat{N}$ est notre estimation de N.

Ainsi la forme de la fonction N est précisée par l'équation ci-dessus. L'existence du fossé de Sloane est-elle naturelle, ou demande-t-elle une explication spécifique ? Signalons qu'à notre connaissance, une seule publication scientifique mentionne l'existence de cette coupure [Delahaye, 2009].

2.2. LE FOSSÉ

Pour pouvoir étudier ce fossé, la première étape est de déterminer un critère de classement des points. Étant donné que le *fossé* n'est pas clairement visible pour les premières valeurs de n, nous écartons de notre étude les nombres inférieurs à 300.

Une méthode empririque de détermination de la frontière du fossé est la suivante : pour les valeurs allant de 301 à 499, nous utilisons une droite ajustée « à l'œil » à partir de la représentation de $\ln(N)$ en fonction de n. Pour les valeurs suivantes, nous prenons comme valeur limite en n le $82^{ème}$ centile de l'intervalle $[n-c, n+c]$. c est fixé à 100 jusqu'à $n = 1000$, puis à 350. Il s'agit bien sûr d'un choix purement empirique qui ne prétend en aucun cas avoir la force d'une démonstration. Le résultat correspond à peu près à ce que nous percevons comme le fossé, sachant qu'il existera toujours une zone d'incertitude, le fossé n'étant pas totalement dénué de points. La Figure 3 montre le résultat obtenu.

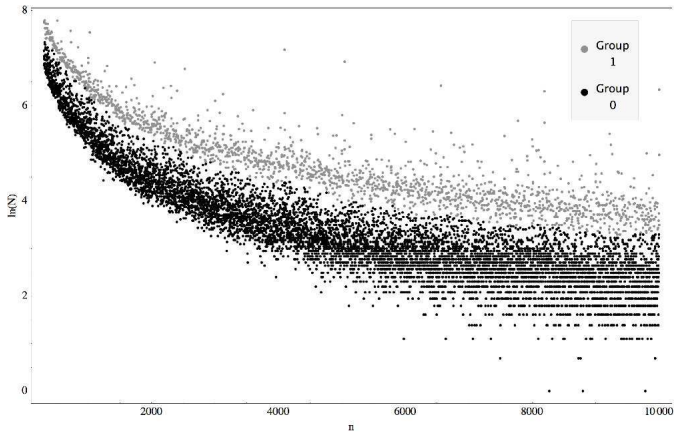

FIGURE 3. $\ln(N)$ en fonction de N. Les points grisés sont ceux classés comme étant *au-dessus* du fossé, les autres sont classés au-dessous. Ce classement automatique correspond bien à ce que nous percevons.

2.3. CARACTÉRISATION DES NOMBRES « AU-DESSUS »

Dans la suite, nous appelons A l'ensemble des abscisses des points classés « au-dessus » du fossé selon la méthode que nous avons utilisée. 18,2% des nombres

compris entre 301 et 10000 se trouvent dans A, soit 1767 valeurs. Dans cette section, nous cherchons des propriétés de ces nombres. Philippe Guglielmetti a déjà remarqué que les nombres premiers ou les puissances de deux semblent se situer plus fréquemment au-dessus du fossé. L'idée est que certaines classes de nombres particulièrement simples ou intéressants pour les mathématiques sont sur-représentées.

2.3.1. Carrés

83 carrés se trouvent entre 301 et 10 000. Parmi ceux-ci, 79 se situent au-dessus du fossé, et 4 au-dessous. Il s'agit des nombres 361, 484, 529 et 676. Bien qu'ils ne soient pas éléments de A, ces nombres sont proches de la frontière. On peut vérifier qu'ils réalisent des maxima locaux pour $\ln(N)$ dans l'ensemble des nombres classés sous le nuage. On a par exemple $N(361) = 1376$, qui est le maximum de $\{N(n), n \in [325, 10\ 000] \setminus A\}$.

La Table 1 donne, pour chacun de ces 4 nombres, le nombre d'occurrences N dans la liste de Sloane, ainsi que la valeur limite qu'ils auraient dû atteindre pour appartenir à A.

n	$N(n)$	Valeur limite
361	1376	1481
484	976	1225
529	962	1065
676	706	855

TABLE 1. Liste de tous les carrés d'entiers n'appartenant pas à A, nombre d'occurrences dans la base de Sloane, et nombre d'occurrences qu'ils auraient dû atteindre pour appartenir à A.

95,2 % des carrés se trouvent dans A, contre 17,6 % des non-carrés. La probabilité pour un nombre carré d'être dans A est donc 5,4 fois supérieure à ce qu'elle est pour les autres nombres.

2.3.2. Nombres premiers

L'intervalle considéré regroupe 1167 nombres premiers. Trois d'entre eux ne sont pas dans A. Ce sont les nombres 947, 8963, 9623. Ces trois nombres sont très proches de la frontière. 947 apparaît 583 fois, alors que la limite de A est 584. 8963 et 9623 apparaîssent 27 fois chacun, et la limite commune est de 28.

99,7 % des nombres premiers appartiennent donc à A, et 92,9 % des non premiers appartiennent au complémentaire de A. La probabilité pour un nombre premier d'appartenir à A est donc 14 fois supérieure à celle d'un non-premier.

2.3.3. Beaucoup de facteurs

Une autre classe de nombres qui semble surreprésentée dans A est celle des entiers ayant « beaucoup de facteurs ». Cette idée s'appuie sur la remarque que la proportion de nombres appartenant à A augmente avec le nombre de facteurs premiers (comptés avec leur multiplicité), comme on le voit sur la Figure 4. Pour préciser

cette idée, nous avons sélectionné les nombres n dont le nombre de facteurs premiers (avec multiplicité) dépasse le 95^e pourcentile correspondant sur l'intervalle $[n-100, n+100]$.

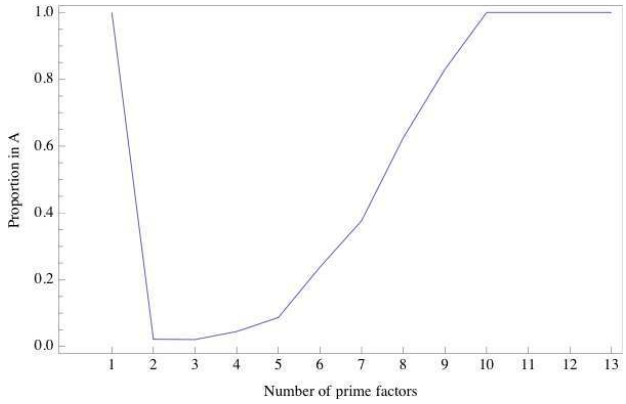

FIGURE 4. Pour chaque nombre de facteurs premiers (comptés avec leur multiplicité), on donne la proportion d'entiers appartenant à A. Dans l'intervalle qu'on s'est fixé, tous les nombres ayant au moins 10 facteurs sont dans A.

811 nombres vérifient ce critère. 39 % d'entre eux se trouvent dans A, contre 16,3 % pour les autres nombres. La probabilité pour un nombre ayant beaucoup de facteurs premiers d'appartenir à A est donc 2,4 fois supérieure à ce qu'elle est pour un nombre ayant moins de facteurs.

La Table 2 donne la composition de A en fonction des classes que nous avons considérées.

2.3.4. Autres cas

L'ensemble A regroupe donc la presque totalité des nombres premiers, 95 % des carrés, et une part importante des nombres ayant beaucoup de facteurs premiers (et tous les nombres en ayant au moins 10).

Ces différentes classes de nombres représentent à eux seuls 87,9 % de A. Parmi les nombres restants, une partie présentent des propriétés flagrantes par exemple liées à la numération décimale, comme 1111, 2222, 3333.... D'autres ont une forme simple comme 1023, 1025, 2047, 2049... qui s'écrivent 2^n+1 ou 2^n-1.

Classe	Nombre dans A	% de A	% cumulé
Nombres premiers	1164	65,9	65,9
Carrés d'entier	79	4,5	70,4
Nombres ayant beaucoup de facteurs	316	17,9	87,9

TABLE 2. Pour chacune des trois classes de nombres considérées, on indique le nombre de n de cette classe appartenant à A, le pourcentage de A que cela représente, et le pourcentage cumulé.

Lorsqu'on élimine ces cas qui, pour une raison ou une autre, présentent une simplicité « évidente », il reste une proportion de moins de 10 % de nombres dans A pour lesquels on ne voit pas immédiatement de propriété particulière.

3. EXPLICATION DE LA FORME DU NUAGE

3.1. LE REGARD DE LA THÉORIE DE LA COMPLEXITÉ DE KOLMOGOROV

Le fait de posséder beaucoup de propriétés pour un nombre implique sauf cas exceptionnels d'en posséder de simples, simple voulant dire exprimable en peu de mots. Inversement, si un nombre possède une propriété simple, alors il possédera beaucoup de propriétés. Par exemple si n est premier, alors n sera un nombre premier de rang pair, ou un nombre premier de rang impair. Être « un nombre premier de rang pair » ou « un nombre premier de rang impair » est un peu plus compliqué qu'être « un nombre premier » et on constate d'ailleurs effectivement dans la base de Sloane que de nombreuses suites présentes sont des sous-suites d'autres plus simples. En spécifiant une propriété simple on complique un peu la définition (ce qui souvent donnera une sous-suite de la première) mais, comme il y a beaucoup de façons de spécifier une propriété simple, tout nombre possédant une propriété simple possède nécessairement de nombreuses propriétés encore assez simples.

La propriété pour n de correspondre à une valeur $N(n)$ élevée semble donc liée à celle d'admettre une description *simple*. La valeur $N(n)$ apparaît ainsi comme une mesure indirecte de la simplicité de n, si on nomme simples les nombres qui ont des propriétés exprimables en un petit nombre de mots.

Il existe une théorie dans le domaine de la logique mathématique et de la théorie de la calculabilité qui fournit un indice de complexité parmi bien d'autres possibles. L'idée est qu'un objet est « simple » s'il possède une définition courte. Il s'agit de la théorie de la complexité de Kolmogorov (ou de Chaitin-Kolmogorov) née au début de la décennie 1960-1970 et dont le traité de référence est aujourd'hui l'ouvrage de Ming Li et Paul Vitànyi [1997]. En français Delahaye [1999] est une rapide introduction à la théorie. Très brièvement, cette théorie propose de mesurer la complexité d'un objet fini codé en binaire (par exemple un nombre écrit en base 2) par la taille du plus petit programme qui l'engendre. La référence à un langage de programmation universel (c'est-à-dire tel que toute fonction calculable puisse y posséder un programme) conduit à un théorème d'invariance qui garantit une certaine indépendance vis-à-vis du choix du langage de programmation.

Plus précisément, si L et L' sont deux langages universels, et si on note K_L (resp $K_{L'}$) la complexité de Kolmogorov définie par référence à L (resp. à L') alors il existe une constance c telle que $|K_L(s) - K_{L'}(s)| < c$, pour toute suite binaire finie s.

Un théorème (cf. par exemple [Li & Vitanyi, 1997, Theorem 4.3.3 p. 253]) relie la probabilité d'obtenir par hasard un objet s (c'est-à-dire en faisant fonctionner un langage universel à l'aide de programmes tirés au hasard uniformément) et $K(s)$. Le sens de ce théorème dans le cas des nombres est que, comme nous en faisions la remarque au-dessus, posséder beaucoup de propriétés est équivalent à en posséder de simples.

La traduction de ce théorème pour $N(n)$ est que si on fixait un langage universel L, qu'on fixait une borne de complexité M (on ne prend en compte que les descriptions de nombres s'exprimant en moins de M symboles) et qu'on comptait le nombre de descriptions de chaque entier, on trouverait que ce $N(n)$ est approximativement proportionnel à $\frac{1}{2^{K(n)}}$:

$$N(n) = \frac{h}{2^{K(n)+O(\log(\log(n)))}} \ (h \text{ une constante}).$$

L'impossibilité d'un calcul exact de $K(n)$ résultant des théorèmes de limitation en logique, particulièrement du théorème d'indécidabilité de l'arrêt d'un programme de Turing (même si on a pu expérimentalement déterminer un équivalent de $K\,(n)$ [Delahaye & Zenil, 2007]), et le rôle de la constante c (mentionnée au-dessus) rendent impossible un calcul précis de la valeur attendue de $N(n)$. En revanche l'analogie très forte entre la situation théorique envisagée par la complexité de Kolmogorov et la situation dans laquelle nous sommes quand nous examinons le $N(n)$ déduit de la base de Sloane conduit à penser que $N(n)$ doit être asymptotiquement lié à $\frac{1}{2^{K(n)}}$.

Certaines des propriétés de $K(n)$ sont asymptotiquement indépendantes du langage de référence choisi pour définir K. Les plus importantes sont celles-ci :

- $K(n) < \log_2(n) + 2\log_2(\log_2(n)) + c'$ (c' une constante)
- la proportion des n d'une longueur donnée (quand on les écrit en binaire) pour lesquels $K\,(n)$ s'éloigne de $\log_2(n)$ décroît exponentiellement (précisément : moins d'un entier sur 2^q de longueur binaire k, possède une complexité de Kolmogorov $K(n)$ inférieure à $k - q$).

Traduites graphiquement, ces propriétés signifient que le nuage de points qu'on obtiendrait en dessinant $\frac{1}{2^{K(n)}}$ serait situé au-dessus de la courbe définie par

$$N(n) \approx \frac{h}{2^{\log_2(n)}} = \frac{h}{n} \ (h \text{ une constante}),$$

et que tous les points se tasseraient sur la courbe, la densité des points s'écartant de la courbe décroissant très rapidement.

C'est bien cette situation que nous observons en examinant la courbe donnant $N(n)$. La théorie de Kolmogorov donne donc une bonne description de ce qu'on observe de la courbe $N(n)$. Cela justifie *a posteriori* le recours aux notions théoriques de la théorie de la complexité de Kolmogorov pour comprendre la forme de la courbe de $N(n)$. En revanche, rien dans la théorie ne laisse supposer un fossé comme celui observé. Au contraire, une forme de *continuité* vient du fait que $n + 1$ n'est jamais beaucoup plus complexe que n.

En résumé, si $N(n)$ représentait une mesure objective de complexité des nombres (plus $N(n)$ est grand, plus n est simple), ces valeurs seraient alors comparables à celles que donne $\frac{1}{2^{K(n)}}$. On devrait donc observer la décroissance rapide en moyenne, et le tassement des valeurs vers le bas contre une courbe asymptote, mais on ne devrait pas observer de fossé qui se présente donc comme une anomalie.

Pour confirmer et rendre plus probante cette conclusion que la présence du fossé résulte de facteurs spéciaux, nous avons réalisé une expérience numérique.

Nous définissons des fonctions f aléatoires de la manière suivante (grâce à un programme *Mathematica*) :

1. On choisit au hasard un nombre i entre 1 et 5 (en respectant dans ce choix les proportions de fonctions pour lesquelles $i = 1$, $i = 2$, ..., $i = 5$ parmi toutes celles définissables de cette manière).

2. Si $i = 1$, on définit f en choisissant au hasard (uniformément) une constante $k \in \{1, ..., 9\}$, un opérateur binaire φ parmi la liste suivante : $+$, $\times$, et écart, de manière uniforme, et un opérateur unaire g qui est l'identité avec probabilité 0,8, et la fonction carré avec probabilité 0,2 (pour reproduire les proportions constatées dans la base de Sloane). On pose alors $f_i(n) = \varphi(g(n), k)$.

3. Si $i \geq 2$, on définit f_i par $f_i(n) = \varphi(g(f_{i-1}(n)), k)$, où k est un entier aléatoire compris entre 1 et 9, g et φ choisis comme en (2), et f_{i-1} une fonction aléatoire choisie comme en 2.

Pour chaque fonction f ainsi produite, on calcule $f(n)$ pour $n = 1, ..., 20$. Ces termes sont regroupés et décomptés comme pour $N(n)$. Les résultats sont donnés sur la Figure 5.

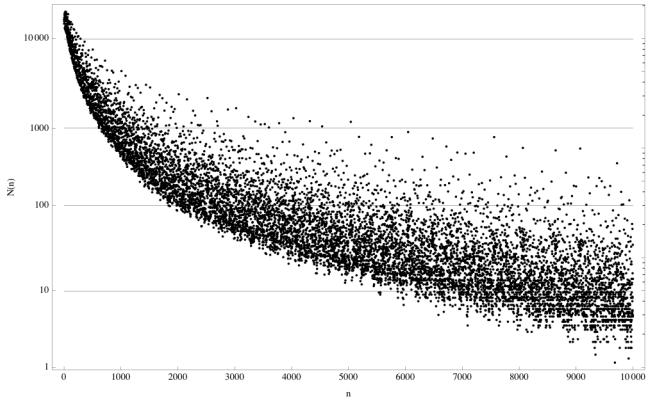

FIGURE 5. Graphe de $N(n)$ obtenu avec les fonctions aléatoires, similaire à celui de la base de Sloane (Figure 1). Huit millions de valeurs ont été produites.

Le résultat confirme ce que le rapprochement avec la complexité de Kolmogorov laissait attendre. Il y a bien décroissance asymptotique vers 0 en moyenne, tassement vers le bas des points... mais pas de fossé.

3.2. LE FOSSÉ, UN EFFET SOCIAL ?

Cette anomalie par rapport aux indications de la théorie et de la modélisation est sans doute la marque que ce que l'on trouve dans la base de Sloane n'est pas une simple mesure objective de complexité (ou d'intérêt mathématique intrinsèque) mais comporte une trace d'origine psychologique ou sociale qui en perturbe l'expression

pure. C'est l'hypothèse explicative que nous proposons ici. En tout état de cause, une vision purement mathématique fondée sur la complexité de Kolmogorov rencontre ici un achoppement, et l'hypothèse sociale est à la fois simple et naturelle, du fait que la base de Sloane, toute « objective » qu'elle soit, est également un objet social.

La Figure 6 illustre et précise notre hypothèse.

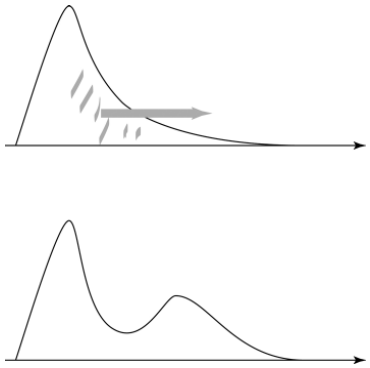

FIGURE 6. La figure du haut représente la distribution locale de N attendue sans tenir compte du facteur social. La communauté mathématique s'intéresse massivement à certains nombres de complexité moyenne ou faible (dans la zone centrale ou droite de la distribution) et, par cet intérêt, crée un décalage vers la droite d'une partie de la distribution (schématisée par une flèche grise). La nouvelle distribution qui en découle, réprésentée en bas, présente un fossé.

Nous supposons que la distribution prévue par les considérations sur la complexité est *déformée* par un effet social qui s'y ajoute : les mathématiciens s'intéressent plus particulièrement à certains nombres, liés à des propriétés sélectionnées par la communauté scientifique. Cet intérêt peut avoir des raisons culturelles de mode, ou des raisons mathématiques (de résultats déjà obtenus), mais il entraîne dans les deux cas un surinvestissement de la communauté mathématique. Les nombres sur lesquels portent cet investissement particulier ne sont pas, en général, complexes, puisque l'intérêt est porté sur des nombres parce qu'on leur a découvert certaines régularités. Ces nombres se situent donc plutôt vers le haut de la distribution théorique asymétrique. Du fait du surinvestissement de la communauté, ils se trouvent décalés vers la droite de la distribution, expliquant ainsi le fossé de Sloane.

C'est par exemple ce qui se produit avec les nombres de la forme $2^n + 1$, tous dans A, parce qu'on dispose pour ce type de nombres de résultats arithmétiques importants liées aux nombres premiers. Par suite de premières découvertes intéressantes, l'investissement des scientifiques est intense sur cette classe d'entiers, et de nombreuses suites les font apparaître. Certes, $2^n + 1$ est objectivement un nombre simple, et il donc normal qu'il se trouve au-dessus du fossé. Néamoins, la différence de complexité entre $2^n + 1$ et $2^n + 2$ est faible. Nous supposons que la différence observée est donc *aussi* le reflet d'une dynamique sociale, qui tend à augmenter $N\,(2^n + 1)$ pour des raisons que la complexité seule n'explique pas entièrement.

4. CONCLUSION

Le nuage de points représentant la fonction N présente une forme générale évoquant une fonction à décroissance rapide et « tassée vers le bas » (distribution locale asymétrique). Cette forme s'explique, au moins qualitativement, par la théorie de la complexité de Kolmogorov.

Si la forme générale du nuage était prévisible, la présence du fossé de Sloane a, en revanche, bien plus interpelé les observateurs. Ce fossé n'a pas, à notre connaissance, pu être expliqué par des considérations uniquement numériques et indépendantes de la nature humaine du travail mathématique. La complexité de Kolmogorov laisse en effet prévoir une certaine « continuité » de N, puisque la complexité de $n + 1$ est toujours proche de celle de n. La discontinuité qui prend corps dans le fossé de Sloane est donc difficilement attribuable à des propriétés purement mathématiques indépendantes des contingences sociales.

En revanche, comme nous l'avons vu, elle s'explique très bien par le fonctionnement de la recherche, qui entraîne la surreprésentation de certains nombres de faible ou moyenne complexité. Ainsi, le nuage de points représentant la fonction N présente-t-il simultanément des caractéristiques que l'on peut comprendre comme humaines *et* purement mathématiques.

BIBLIOGRAPHIE

CIPRA B. (1994), "Mathematicians get an on-line fingerprint file", *Science* 265, p. 473.

DELAHAYE J.-P. (1999), *Information, complexité et hasard*, Paris, Hermès.

DELAHAYE J.-P. (2009), « Mille collections de nombres », *Pour la science* 379, p. 88-93.

DELAHAYE J.-P., ZENIL H. (2007), "On the Kolmogorov-Chaiting complexity for short sequences", C.S. Calude (ed.), *Randomness and Complexity: from Leibniz to Chaitin*, World Scientific, p. 123-129.

LI M., VITANYI P. (1997), "An introduction to Kolmogorov complexity and its applications", Springer.

SLOANE N., (1973), "A handbook of integer sequences", Academic Press.

SLOANE N., PLOUFFE S. (1995), "The Encyclopedia of Integer Sequences", Academic Press.