Un identifiant universel unique est-il vraiment unique ?

Probabilités, dénombrement, logarithme népérien, limites de fonctions

Dans le monde de l'informatique, il arrive souvent d'avoir 2 éléments ou plus du même nom, ou ayant les mêmes caractéristiques. Il faut alors recourir à un identifiant propre à chacun pour pouvoir les différencier. Par exemple, 2 utilisateurs peuvent posséder le même nom, donc pour que le programme en sélectionne un seul ; il le sélectionne par son identifiant supposé unique.
Ainsi, depuis 1988, un format spécifique est utilisé pour créer des identifiants uniques : le UUID, ou identifiant unique universel. Une longue chaîne de caractères générés chacun aléatoirement : il devrait normalement ne jamais y avoir de générations identiques. Mais, la probabilité que par pur hasard, 2 UUID identiques soient générés n'est pas nulle, donc un identifiant unique universel est-il vraiment unique ?

Tout d'abord, une chaîne UUID comporte 5 sections :
xxxxxxxx-xxxx-4xxx-xxxx-xxxxxxxxxxxx

Chaque caractère (x) est un nombre hexadécimal (base 16), allant de 0-9 puis a-f en informatique.
Un chiffre hexadécimal correspond à 4 bits, c'est-à-dire une séquence de 4 chiffres binaires (base 2).

Ainsi, avec 32 chiffres car on ne compte pas les tirets, il y a au total $32 \times 4 = 128$ bits dans un UUID.
Or, 6 bits sont utilisés pour désigner la version de l'UUID.

Il reste donc $128 - 6 = 122$ bits aléatoires.

En base 2, on sait donc qu'on a un total de 122 chiffres aléatoires et chaque chiffre pouvant avoir une des 2 valeurs : 0 ou 1.
En d'autres termes, un bit est un élément d'un ensemble fini $E = \{0, 1\}$.
On cherche donc un produit cartésien de notre ensemble par lui-même, 122 fois. D'après le principe multiplicatif, le nombre d'UUID possibles vaut :
$Card^{122}(E) = 2^{122} \approx 5,3 \cdot 10^{36}$
Ce qui signifie qu'il existe 5,3 sextillions d'UUID différents.

Pour évaluer les limites de nos UUID, il est plus utile de calculer une probabilité.
Pour ceci, une analyse du problème des anniversaires est généralement utilisée pour déterminer la rareté d'une collision.
En théorie des probabilités, le problème de l'anniversaire demande la probabilité que, dans un ensemble de n personnes choisies au hasard, au moins 2 partageront un anniversaire. Il s'agit d'un paradoxe dans le sens où c'est une vérité mathématique qui contredit l'intuition, car en effet, ce nombre n vaut 23, un nombre bien plus petit que ce à quoi on s'attend.

Notre but est donc de trouver le nombre minimum d'UUID nécessaires pour que la probabilité que 2 d'entre eux soient identiques soit supérieure ou égale à 50% / ½.

Considérons l'ensemble F contenant $Card(F) = 2^{122}$ éléments, c'est-à-dire le nombre d'UUID possibles. On dira n pour parler du cardinal de F.

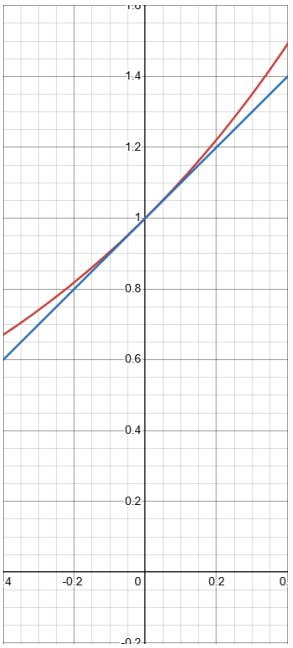
Ainsi $p(k)$: la probabilité qu'au moins 2 UUID soient identiques. L'événement contraire est donc "tous les UUID sont différents". On remarque que ceci est plus facile à déterminer par un calcul.
En effet, pour calculer $\overline{p}(k)$, c'est le rapport entre un k-uplet d'éléments distincts de F et un k-uplet de F.
Dans le dénombrement, on parle d'arrangement pour désigner un k-uplet d'éléments distincts. Il se calcule de cette manière :
$n \times (n-1) \times (n-2) \times \ldots \times (n-k+1)$
Un k-uplet de F est simplement : n^k

Donc la probabilité est :

$$p(k) = \frac{n\times(n-1)\times(n-2)\times \dots \times(n-k+1)}{n^k}$$

$$= \frac{n\times(n-1)\times(n-2)\times \dots \times(n-k+1)}{n\times \ n \ \times \ n \ \times \dots \times \ n} = \frac{n}{n} \times \frac{n-1}{n} \times \frac{n-2}{n} \times \ \dots \ \times \frac{n-k+1}{n}$$ ceci k fois.

$$= \prod_{i=0}^{k-1} \frac{n-i}{n} = \prod_{i=0}^{k-1} 1 - \frac{i}{n}$$



Désormais, faisons une petite parenthèse qui nous aidera par la suite, à propos de la fonction exponentielle.

Pour x très proche de 0, on remarque visuellement que les valeurs de e^x sont très proches de celles de la tangente de cette même fonction en abscisse 0.

Avec $f(x) := e^x$ on peut déduire l'équation de la tangente en $x = 0$:

$y = f'(0)(x - 0) + f(0) = e^0 x + e^0 = x + 1$ en sachant que $(e^x)' = e^x$

On se retrouve alors avec l'approximation : $e^x \approx x + 1$ si x tend vers 0.

En revenant à notre probabilité, si on calcule la limite du produit :

$\lim_{n \to +\infty} \frac{i}{n} = 0$ donc par produit avec -1, $\lim_{n \to +\infty} -\frac{i}{n} = 0$

Or pour nous $n = 2^{122}$ ce qui est un nombre très grand, donc on peut utiliser les valeurs de limite.

Ainsi, si l'on considère $x = -\frac{p}{n}$ très proche de 0 (car limite égale à 0) on en déduit :

$1 - \frac{i}{n} \approx e^{-\frac{i}{n}}$ pour une écriture finale :

$\prod_{i=0}^{k-1} exp(-\frac{i}{n}) = exp(-\frac{\sum_{i=0}^{k-1} i}{n})$ * et sachant que la somme des entiers de 0 à $k - 1$ vaut $\frac{k(k-1)}{2}$, on a alors :

$exp(-\frac{k(k-1)}{2n})$

Enfin, revenons à la probabilité qu'on a énoncé au début. On a :

$p(k) = 1 - \bar{p}(k) \approx 1 - exp(-\frac{k(k-1)}{2n})$

Maintenant qu'on a une expression assez simplifiée pour pouvoir la manipuler, résolvons l'équation pour $p(k) = 50\% = \frac{1}{2}$:

$\frac{1}{2} \approx 1 - exp(-\frac{k(k-1)}{2n})$ **

$\Leftrightarrow k^2 - k - n\,ln(4) \approx 0$

Pour résoudre l'équation du second degré, on utilise :

$\Delta = (-1)^2 - 4 \times 1 \times (-2^{122} ln(4)) = 1 + 1952\, ln(2) > 0$ il y a 2 solutions :

$k = -\frac{-1 \pm \sqrt{\Delta}}{2\times 1}$ on omet la solution négative car $k \in \mathbb{N}$ donc :

$k \approx 2,714922669 \cdot 10^{18}$

c'est-à-dire $p(2,714922669 \cdot 10^{18}) = \frac{1}{2}$

On tombe donc sur un nombre astronomiquement grand : ce qui est difficile pour nos cerveaux d'interpréter. Plutôt, si l'on génère 1 milliard d'UUID par seconde, il faudrait 86 ans pour en générer autant que le nombre k obtenu.

Pour conclure, même si une collision d'UUID peut arriver, il en faudrait au moins 2,7 milliards de milliards pour que la probabilité que 2 d'entre eux soient identiques soit supérieure à 50%. On peut donc qualifier un identifiant unique universel comme véritablement unique, étant donné la grandeur de la solution au problème des anniversaires.

*****

$\prod_{i=0}^{k-1} exp(-\frac{i}{n}) = e^{-\frac{0}{n}} \times e^{-\frac{1}{n}} \times \ldots \times e^{-\frac{k-1}{n}}$ or d'après la propriété de l'exponentielle on a $e^a \times e^b = e^{a+b}$:

$= e^{-\frac{0}{n}-\frac{1}{n}-\ldots-\frac{k-1}{n}} = e^{-\frac{0+1+\ldots+(k-1)}{n}} = exp(-\frac{\sum_{i=0}^{k-1} i}{n})$

******

$\Leftrightarrow exp(-\frac{k(k-1)}{2n}) \approx 1 - \frac{1}{2}$

$\Leftrightarrow ln(exp(-\frac{k(k-1)}{2n})) \approx ln(\frac{1}{2})$ on applique la fonction ln strictement croissante sur $]0 ; + \infty[$

$\Leftrightarrow -\frac{k(k-1)}{2n} \approx ln(1) - ln(2)$

$\Leftrightarrow k(k-1) \approx n \cdot ln(4)$