

Hash function, Birthday paradox and probabilities.

Source: <http://sci.tech-archive.net/Archive/sci.math/2006-03/msg02779.html>

- *From:* fabrice.gautier@xxxxxxxxxx
 - *Date:* 15 Mar 2006 17:59:31 -0800
-

Hi,

I'm wondering about this problem, its coming from a computer science background but I think this is mainly a probability theory problem. The title is probably misleading as I dont really know how to describe the problem in one line, and I'll probably get the vocabulary all wrong, so dont hesitate to fix the way I set the problem :

Lets have:

- a hash function $H()$ that take as input a bit string of length L and return an hash of length N bits (ie an integer in the range $[0, 2^N[$)
- a iterator function F , that takes as input a bit string of length L and return another bit string of length L ,
- $S[0], S[1], \dots S[n]$, bit strings of length L , so that $S[i+1]=F(S[i])$,
- $h[0], \dots h[n]$, so that $h[i]=H(S[i])$
- M a fixed hash

Given all that lets define m , the smallest positive integer so that $h[m]=M$.

I'm wondering how A and H affect the probabilistic properties of m .

For example, the practical example is this:

- $S[0]$ is a file, that can be divided in 3 parts: $S[0]=\{T, s[0], U\}$ where $s[0]$ is a 32 bits integer. (L is the length of this file)
- The $H()$ function is the sha1 hash of the file, truncated to the last 16bits ($N=16$)
- $S[i]=\{T, s[i], U\}$, with $s[i+1]=a(s[i])$
- $a()$ is a function that transform a 32 bits integer into another one. (So $A()$ is the function that take a file, and change the 32 bit value at a fixed offset according to the function $a()$)

Now the question is what is the best function $a()$ to minize m (in average), and how to minimize the "spread" of m around its average. (There must be a better word than "spread" but I dont remember)

I have been trying two things:

Hash function, Birthday paradox and probabilities.

- $s(0)=\text{random}()$, $s(i+1)=s(i)+1$ (modulo 2^{32}),
- $s(i)=\text{random}()$

I was thinking that, whether I use random or increments, the average of m should be 2^N , and my experiences seem to agree with that.

Now, I'm not sure about the "spread", my instinct says that it might be better with random, but then again I'm not sure why...

Anybody has some thoughts about this. ?

Thanks

-- F.G.

.