Hashing Project

The project is to explore and compare two variants of hashing. Take $M = 1000003$ (a prime), the hash table will be $0, 1, \ldots, M-1$, with calculations done modulo M. The table is initially empty.

You will inset $1, \ldots, 800000$ into the hash table in two different ways.

Let $h_1(x)$ be a hash function into $0, 1, \ldots, M-1$ and let $h_2(x)$ be a hash function into $1, 2, \ldots, M-1$. For this project you may use random number generators for h_1, h_2. (Normally that would not be permissible as then you couldn't find where the element was hashed to. But here we will only insert so that won't be a problem.

`Dumb Hashing:` Use the probe sequence $h_1(x) + i$.

`Double Hashing:` Use the probe sequence $h_1(x) + ih_2(x)$. (Warning: Once $h_2(x)$ is calculated do not recalculate it!)

In each case keep trying until the item is placed in the hash table. Keep a cumulative count of the number of collisions. (Note that if $h_1(x)$ is empty there are no collisions.) Let $COLL[i]$ be the total number of collisions after items $1, \ldots, i$ have been placed in the hash table.

Graph the data $COLL[i]$, $1 \leq i \leq 800000$ for the two methods. We said in class that "dumb hashing" is bad because you start to get long intervals in the hash table and then many collisions. Is that supported by the data? Speculate on what $COLL[i]$ looks like as a function of i.

Most of all, have fun – explore – take to heart the words of the founder of Theoretical Computer Science, Don Knuth:

> ...pleasure has probably been the main goal all along. But I hesitate to admit it, because computer scientists want to mantain their image as hard-working individuals who deserve high salaries. Sooner or later society will realise that certain kinds of hard work are in fact admirable even though they are more fun than just about anything else.