Hashing—Record Distribution

- Usually hashing functions distribute the records in a better than random pattern
- Unfortunately there are no nice mathematical tools for predicting better than random distributions
- Fortunately the Poisson Distribution can be used to analyze a random distribution
- So it can be used to get a conservative estimate for out better than random distribution

We would like to know the answer to the following questions about a given address

- What is the likelihood that no key will hash to the address?
- What is the likelihood that exactly one key will hash to the address?
- What is the likelihood that exactly two keys will hash to the address?
- What is the likelihood that exactly n keys will hash to the address?
- What is the likelihood that all of the keys will hash to the address?

Suppose there are N addresses. (Pick an Address)

When a single key is hashed there are two possible outcomes

A—the key does not hash to the given address. So $p(A)$ or a is the probability that the address is not chosen

B—the key hashes to the given address. So $p(B)$ or b is the probability that the address is chosen.

So, $p(B) = b = \frac{1}{N}$ since the address has one chance in N of being chosen and

$p(A) = a = 1 - \frac{1}{N}$ since there are $N - 1$ chances of not choosing the address.

$$\text{So, for } N = 1000 \quad p(A) = a = {}^{(1000-1)}\!/_{1000} = .999$$

and

$$p(B) = b = {}^{1}\!/_{1000} = .001$$

If we want to know what the probability of choosing this address twice in 4 consecutive hashes we use

$$p(BBAA) + p(BABA) + p(BAAB) + p(AABB) + p(ABAB) + p(ABBA)$$
$$=$$
$$bbaa + baba + baab + aabb + abab + abba$$

This is just $6b^2a^2$.

Fortunately there is a nice formula that can be used for more keys.

- In general, the event "r trials results in $r-x$ As and x Bs" can happen in $r-x$ letters A can be distributed among r places.
- The probability of each such way is $a^{r-x}b^x$
- The number of such ways is $C = \dfrac{r!}{(r-x)!\,x!}$
- So $p(x) = C\left(1-\dfrac{1}{N}\right)^{r-x}\left(\dfrac{1}{N}\right)^x$

But this is awkward to compute so we use the Poisson Distribution

$$p(x) = \frac{\left(r/N\right)^x e^{-\left(r/N\right)}}{x!}$$

where

N is the number of available addresses

R is the number of records to be stored

x is the number of records assigned to a given address

We can use the Poisson Function to predict collisions in a full file

In general, if there are *N* addresses, then the expected number of addresses with *x* records assigned to them is: *Np(x)*

Can we use this to predict Collisions for a full file?

Suppose we are using a hash function that we believe will distribute records randomly and we will store $r = 10{,}000$ records in $N = 10{,}000$ addresses and we want to know the expected number of addresses that no record addresses to, exactly 1 record hashes to, exactly 2 records hash to, and exactly 3 records hash to.

x	N	r	Np(x)	N	r	Np(x)	N	r	Np(x)
	10000	10000.000		13000	10000		10000	5000	
0		0.368	3679		0.463	6024		0.6065	6065
1		0.368	3679		0.356	4634		0.3033	3033
2		0.184	1839		0.137	1782		0.0758	758
3		0.061	613		0.035	457		0.0126	126
4		0.015	153		0.007	88		0.0016	16
5		0.003	31		0.001	14		0.0002	2
6		0.001	5		0.000	2		0.0000	0

What do we do with the 1,839 + 2 * 613 = 3,065 records that hash to an already used address?

These are called overflow records.

Or what about the 3,679 addresses that no records hash to?

One technique is to allocate extra memory.

That is, allocate 13,000 addresses for 10,000 records.

Then the packing density is 10000/13000 = 0.77 = 77%

So lets use the packing density r/N in our analysis

This ratio already appears in the Poisson formula

From the earlier table we can answer the following questions:

1. How many addresses should have no records assigned to them?

$$p(0) = 13000 \times \frac{0(.77)^0 e^{-0.77}}{0!} = 13000 \times 0.463 = 6024$$

2. How many addresses should have exactly one record assigned (no synonyms)?

$$p(1) = 13000 \times \frac{0(.77)^1 e^{-0.77}}{1!} = 13000 \times 0.356 = 4634$$

3. How many addresses should have one record plus one or more synonyms?

   We can stop at x = 6 and get 902

$$13000 \times (P(2) + p(3) + p(4) + \cdots) = 2323$$

4. Assuming that only one record can be assigned to each home address how many over flow records could be expected?

$$N \times [1 \times p(2) + 2 \times p(3) + 3 \times p(4) + 4 \times p(5) + 5 \times p(6)] = 3026$$

5. What percentage of records should be overflow records?

$$\frac{3026}{10000} = 0.30 = 30\%$$

So, we can assume that with a packing density of 77% we can expect 30% of the records to be stored somewhere other than their home address.

Packing Density (percent)	Synonyms as Percent of Records
20	9.4
30	13.6
40	17.6
50	21.4
60	24.8
70	28.1
80	31.2
90	34.1
100	36.8

So what do we do with the collisions?

- Progressive Overflow?
- Buckets?
- Chaining?
- Deletion and Tombstones?