Notes on Bounding the Longest Chain

Let's consider a hash table resolving collisions using chaining, as described in Section 11.2 in the textbook. The worst-case time to search for an item in the table is the length of the longest chain, so we want to prove that this is very likely to be small. Specifically, we'll get a formula for the probability that a chain contains more than k items, as a function of k, and then we'll figure out a value of k that makes this probability so small that it is unlikely that any chain in the table contains k or more items.

We consider a hash table T with m locations, storing n items (all the items in one location are stored in the chain at that location). We assume that $n = cm$, where c is a constant, so that the expected number of items per location is c. The length of the chain at location $T[i]$ is the number of items that hash to index i. We will assume that the locations assigned to each item by the hash function $h()$ are random, with each location equally likely, and independent of each other. So the probability that $h(x) = i$ for one particular x is $1/m$.

If we choose a fixed set Y containing k items, using the assumption that the events $h(y) = i$ are all independent for $y \in Y$, we can write,

$$\Pr[h(y) = i, \forall y \in Y] = (1/m)^k$$

This is the probability that all k items in set Y hash to i.

There are $\binom{n}{k}$ possible subsets Y_j of size k, and we want to argue that the event that even one of them hashes to i is quite small. We use the *union bound* (equation C.13), which says that if A, B are events, not necessarily independent, then $\Pr[A \text{ or } B] \leq \Pr[A] + \Pr[B]$. So,

$$\Pr[\exists Y_j, \text{ such that } h(y) = i, \forall y \in Y_j] \leq \binom{n}{k}(1/m)^k$$

We want to choose k so that $\binom{n}{k}(1/m)^k$ is very small; in particular, let's find a k such that $\binom{n}{k}(1/m)^k < 1/m^2$. We observe that

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \leq \frac{n^k}{k!} \leq \frac{n^k}{(k/2)^{(k/2)}} = \frac{c^k m^k}{(k/2)^{(k/2)}}$$

So we want to choose k such that

$$1/m^2 > \frac{c^k m^k}{(k/2)^{(k/2)}} \frac{1}{m^k} = \frac{c^k}{(k/2)^{(k/2)}} = \frac{(c^2)^{(k/2)}}{(k/2)^{(k/2)}}$$

Taking both sides to the power $-2/k$ (notice we have to switch the direction of the inequality), we get

$$m^{(4/k)} < \frac{k}{2c^2}$$

Choosing $k = 4 \lg m$, we get

$$(2^{\lg m})^{1/\lg m} < \frac{2 \lg m}{c^2}$$

or

$$c^2 < \lg m$$

This is going to be true for large enough m. So we've shown is that for table size $m > 2^{c^2}$, the probability that a specific location gets $4 \lg m = k$ or more items is at most $1/m^2$, very small.

The final idea is that because the probability that any one location gets k items is so small, it is very unlikely that any of the m locations in the table will get k or more items. Again, we use the union bound:

$$\Pr[\text{any location gets } \geq 4 \lg m \text{ items}] \leq m(1/m^2) = 1/m$$

A somewhat more complicated argument that shows that the expected size of the longest chain is only $O(\lg n / \lg \lg n)$ is sketched in problem 11-2.