6.006 Recitation Notes 9/29/10
Jenny Barry
jlbarry@mit.edu

Notes on universal and collision-resistant hash functions. Neither of these are required course material, but they're cool!

Definition: A family of hash functions $H = \{h_0, h_1, ...\}$ is *universal* if, for a randomly chosen pair of keys $k, l \in U$ and randomly chosen hash function $h \in H$, the probability that $h(k) = h(l)$ is not more than $1/m$ where m is the size of the hash table.

This is useful because if you pick a hash function from H when your program begins in such a way that an adversary cannot know in advance which function you will pick, the adversary cannot in advance guess two keys that will map to the same value.

Example: The family of hash functions

$$h_{a,b}(x) = ((ax + b) \bmod p) \bmod m \tag{1}$$

where $0 < a < p$, $b < p$, $m < p$, and $|U| < p$ for prime p is universal.

Proof: Consider $k, l \in U$ with $k \neq l$. For a given $h_{a,b}$ let

$$r = (ak + b) \bmod p \tag{2}$$
$$s = (al + b) \bmod p \tag{3}$$

Note that $r \neq s$ since

$$r - s \equiv a(k - l) \bmod p \tag{4}$$

cannot be zero since $0 < a < p$, $k < p$, and $l < p$ so $a(k-l)$ cannot be a multiple of p.

Now consider

$$a = ((r - s)((k - l)^{-1} \bmod p)) \bmod p \tag{5}$$
$$b = (r - ak) \bmod p. \tag{6}$$

Now since $r \neq s$, there are only $p(p-1)$ possible pairs (r, s). Similarly, since we require $a \neq 0$, there are only $p(p-1)$ pairs (a, b). Equations 5 and 6 give a one-to-one map between pairs (r, s) and pairs (a, b). Therefore, each choice of (a, b) must produce a different (r, s) pair. If we pick (a, b) uniformly, at random then (r, s) is also distributed uniformly at random.

The probability that two keys k and l with $k \neq l$ have the same hash value is the probability that $r \equiv s \bmod m$. Therefore, we must have that

$$r - s \in \{m, 2m, ..., qm\} \tag{7}$$

where $qm < p$. This gives us at most $\lceil p/m \rceil - 1 \leq (p-1)/m$ possible values for s such that s can collide with r. Since the pairs are distributed at random, and $s \neq r$, we have $p - 1$ values for s that are all equally probable. Thus

$$Pr[s \equiv r \bmod m] = \frac{p - 1/m}{p - 1} = \frac{1}{m} \tag{8}$$

$$\Rightarrow \quad Pr[h(k) = h(l)] = \frac{1}{m} \tag{9}$$

This proof was taken from CLRS Section 11.3.3.

Definition: A family of hash functions $H = \{h_0, h_1, ...\}$ is *collision resistant* if there is no algorithm $p(h_i)$ running in time logarithmic in the size of the hash table m such that for all i, the probability that $p(h_i) = \{x, y\}$ where $x \neq y$ and $h_i(x) = h_i(y)$ is exponentially small in $\log m$.

Why $\log m$? We care about running times in $\log m$ because it requires $O(\log m)$ bits to specify a hash function to a table of size m. Therefore the input to p is $O(\log m)$ so p must run in time polynomial in the size of its input.

Example: The discrete logarithm hash functions $h_{g,n}(x) = g^x \bmod n$ where $n = pq$ for primes p and q is g is relatively prime to $\phi(n) = (p-1)(q-1)$ is a collision resistant hash function so long as p and q are unknown and factoring is hard.

Proof: It can be shown that if we could find x and y such that $g^x \bmod n = g^y \bmod n$ with $x \neq y$, we could factor n. I haven't been able to find a simple version of the proof yet, though. Please let me know if you do.