Exercise 11.3-3

Consider a version of the division method in which $h(k) = k \bmod m$, where $m = 2^p - 1$ and k is a character string interpreted in radix 2^p. Show that if we can derive string x from string y by permuting its characters, then x and y hash to the same value. Give an example of an application in which this property would be undesirable in a hash function.

Solution:

First, we observe that we can generate any permutation by a sequence of interchanges of pairs of characters. One can prove this property formally, but informally, consider that both heapsort and quicksort work by interchanging pairs of elements and that they have to be able to produce any permutation of their input array. Thus, it suffices to show that if string x can be derived from string y by interchanging a single pair of characters, then x and y hash to the same value.

Let x_i be the ith character in x, and similarly for y_i. We can interpret x in radix 2^p as $\sum_{i=0}^{n-1} x_i 2^{ip}$, and interpret y as $\sum_{i=0}^{n-1} y_i 2^{ip}$. So

$$h(x) = (\sum_{i=0}^{n-1} x_i 2^{ip} \ ) \bmod (2^p - 1). \text{ Similarly, } h(y) = (\sum_{i=0}^{n-1} y_i 2^{ip} \ ) \bmod (2^p - 1).$$

Suppose that x and y are identical strings of n characters except that the characters in positions a and b are interchanged:

$$x_a = y_b \text{ and } y_a = x_b. \qquad (1)$$

Without loss of generality, let $a > b$. We have:

$$h(x) - h(y) = (\sum_{i=0}^{n-1} x_i 2^{ip} \ ) \bmod (2^p - 1) - (\sum_{i=0}^{n-1} y_i 2^{ip} \ ) \bmod (2^p - 1) \qquad (2)$$

Since $0 \le h(x), h(y) < 2^p - 1$, we have that $-(2^p - 1) < h(x) - h(y) < 2^p - 1$. If we show that $(h(x) - h(y)) \bmod (2^p - 1) = 0$, then $h(x) = h(y)$. To prove $(h(x) - h(y)) \bmod (2^p - 1) = 0$, we have:

$$\begin{aligned}
(h(x) - h(y)) \bmod (2^p - 1) &= ((\sum_{i=0}^{n-1} x_i 2^{ip} \ ) \bmod (2^p - 1) - (\sum_{i=0}^{n-1} y_i 2^{ip} \ ) \bmod (2^p - 1)) \\
&\quad \bmod (2^p - 1) && \text{by (2)} \\
&= (\sum_{i=0}^{n-1} x_i 2^{ip} \ - \sum_{i=0}^{n-1} y_i 2^{ip} \ ) \bmod (2^p - 1) && \text{by relation in footnote}^1 \\
&= ((x_a 2^{ap} + x_b 2^{bp}) - (y_a 2^{ap} + y_b 2^{bp})) \bmod (2^p - 1) && \text{as } x \text{ and } y \text{ are identical strings of } n \text{ characters except that chars. in positions } a \text{ and } b \text{ are interchanged} \\
&= ((x_a 2^{ap} + x_b 2^{bp}) - (x_b 2^{ap} + x_a 2^{bp})) \bmod (2^p - 1) && \text{as } x_a = y_b,\ x_b = y_a \text{ see (1)} \\
&= ((x_a - x_b) 2^{ap} + (x_b - x_a) 2^{bp}) \bmod (2^p - 1) && \text{by combining like terms} \\
&= ((x_a - x_b) 2^{ap} - (x_a - x_b) 2^{bp}) \bmod (2^p - 1) && \text{as } (x_b - x_a) = -(x_a - x_b) \\
&= ((x_a - x_b)(2^{ap} - 2^{bp})) \bmod (2^p - 1) && \text{by factoring out } (x_a - x_b) \\
&= ((x_a - x_b)(2^{ap}(2^{bp}/2^{bp}) - 2^{bp})) \bmod (2^p - 1) && \text{by multiplication by } 2^{bp}/2^{bp} = 1 \\
&= ((x_a - x_b) 2^{bp}(2^{(a-b)p} - 1)) \bmod (2^p - 1) && \text{by factoring out } 2^{bp} \\
&= ((x_a - x_b) 2^{bp}(\sum_{i=0}^{a-b-1} 2^{ip} \ )(2^p - 1)) \bmod (2^p - 1) && \text{by substituting } [2^{(a-b)p} - 1]^2 \\
&= 0 && \text{since one factor is } 2^p - 1
\end{aligned}$$

Because we deduced earlier that $(h(x) - h(y)) \bmod (2p - 1) = ((x_a - x_b) 2^{bp}(2^{(a-b)p} - 1)) \bmod (2^p - 1)$ and have shown here that $((x_a - x_b) 2^{bp}(2^{(a-b)p} - 1)) \bmod (2^p - 1) = 0$, we can conclude $(h(x) - h(y)) \bmod (2^p - 1) = 0$, and thus $h(x) = h(y)$. So we have proven that if we can derive string x from string y by permuting its characters, then x and y hash to the same value.

Examples of applications:

A dictionary which contains words expressed by ASCII code can be one of such example when each character of the dictionary is interpreted in radix $2^8 = 256$ and $m = 255$. The dictionary, for instance, might have words "STOP," "TOPS," "SPOT," "POTS," all of which are hashed into the same slot.

[1] Consider the congruence relation: $(m_1 \text{ o } m_2) \bmod n = ((m_1 \bmod n) \text{ o } (m_2 \bmod n)) \bmod n$, where o is $+$, $-$, or $*$

[2] Consider the equation $\sum_{i=0}^{a-b-1} 2^{ip} \ = \dfrac{2^{(a-b)p} - 1}{2^p - 1}$ (geometric series) and multiplying both sides by $2^p - 1$ to get

$$2^{(a-b)p} - 1 = (\sum_{i=0}^{a-b-1} 2^{ip} \ )(2^p - 1)]$$