Hashing sets

Your company has a database $S \in U$ of keys. For this database, it uses a randomly chosen hash function h from a universal family H (as seen in class); it also keeps a bit vector B_S of m entries, initialized to zeroes, which are then set $B_S[h(k)] = 1$ for every $k \in S$ (note that collisions may happen). Unfortunately, the database S has been lost, thus only B_S and h are known, and the rest is no more accessible. Now, given $k \in U$, how can you establish if k was in S or not? What is the probability of error? Under the hypothesis that $m \geq c|S|$ for some $c > 1$ (note: we do not know the actual values of c and $|S|$) can you estimate the size $|S|$, i.e. the size of S, looking at just h and B_S? What is the probability of error? Note that S is no more accessible as it disappeared.
Optional: Another database R has been found to be lost: it was using the same hash function h, and the bit vector B_R defined analogously as above. Using h, B_S, and B_R, how can you establish if k was in $S \cup R$ (union), $S \cap R$ (intersection), or $S \setminus R$ (difference)? What is the probability of error?

SOLUTION

a) To check whether $k \in U$ belong to S, we simply check $B_S[h(k)] = 1$. The probability of error is equal to $P(error) = 1 - (1 - \frac{1}{m})^{|S|}$. This problem is similar to the "Birthday paradox" (i.e. fixed a day how many people are born on the same day). In word: the probability of a collision is $\frac{1}{m}$, thus the probability to do not have a collision is $(1 - \frac{1}{m})$. If we have $|S|$ key the probability do not have any collision is $(1 - \frac{1}{m})^{|S|}$. Therefore the probability to have a collision is: $1 - (1 - \frac{1}{m})^{|S|}$. That is, the probability that there is at least one collision with of the key is S, the probability that $\exists j \in S : h(k) = h(j)$ but $j \neq k$ with $k \in U$.

b) To estimate the size of S, we first create an indicator variable $X = \sum_{i=0}^{m-1} X_i$ where

$$X_i = \begin{cases} 1 & \text{IF } B_S[h(k)] = 1 \\ 0 & \text{OTHERWISE} \end{cases}$$

Than the expectation $E[X]$ represents the expected number of 1 in the B_s table. To calculate the expectation we need to estimate the $P(B_S[h(k)] = 1)$, that, by the point a), should be $(1 - \frac{1}{m})^{|S|}$. Since we do not know $|S|$ we use the hypothesis, that is $|S| \leq \frac{m}{c}$. Therefore we have:

$$P(B_S[h(k)] = 1) = (1 - \frac{1}{m})^{|S|} \leq (1 - \frac{1}{m})^{\frac{m}{c}}$$

Hence we have $E[X] = \sum_{i=0}^{m-1}(1 - \frac{1}{m})^{\frac{m}{c}} = m(1 - \frac{1}{m})^{\frac{m}{c}}$. Now we have got a bound for $|S|$: $E[X] \leq |S| \leq \frac{m}{c}$.

c) For the optional point we have:

Union We need to check that $B_S[h(k)] = 1$ OR $B_R[h(k)] = 1$ and then the probability of error is $P[B_S[h(k)] = 1 \vee B_R[h(k)] = 1]$ that , by set theory is equal to $P[B_S[h(k)] = 1] + P[B_S[h(k)] = 1] - P[B_S[h(k)] = 1 \wedge B_R[h(k)] = 1]$

Intersection We need to check that $B_S[h(k)] = 1$ AND $B_R[h(k)] = 1$ and then the probability of error is $P[B_S[h(k)] = 1 \wedge B_R[h(k)] = 1]$

Difference We need to check that $B_S[h(k)] = 1$ AND $B_R[h(k)] = 0$ and then the probability of error is $P[B_S[h(k)] = 1 \wedge B_R[h(k)] = 0]$ that , by set theory is equal to $P[B_S[h(k)] = 1] - P[B_S[h(k)] = 1 \wedge B_R[h(k)] = 1]$