# 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? [Note: you are not choosing $k$ and $S$ randomly as the they are both given... randomization here is in the choice of $h \in H$ performed when building $B_S$] Under the hypothesis that $m \geq c|S|$ for some $c > 1$, find the expected number of $1s$ in $B_S$ under a uniform choice at random of $h \in H$.

**SOLUTION**

**a)** To check whether $k \in U$ belong to $S$ we simply check $B_S[h(k)] = 1$, we return true if $B_S[h(k)] = 1$ false otherwise. For any given $k \in U$, if $k \in S$ the answer we give is correct for sure (for any possible choice of $h$), since $B_S[h(k)]$ has been set to one by hypothesis. If $k \in U$ is not in $S$, then the probability of error is the probability that the randomly cosed hash function $h$ is such that some $j \in S$ causes $B_S[h(k)]$ to be set at 1. In other words it is the probability that $h$ is such that $\exists j \in S : h(k) = h(j)$ or equivalently it is the probability that the given index $h(k)$ in $B_S$ is at 1. We see that 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: given an arbitrary index $i$ of $B_S$ the probability that one specific element of $S$ set the $B_S[i]$ at 1 is $\frac{1}{m}$, thus the probability that it does not is $(1 - \frac{1}{m})$. If we have $|S|$ key the probability that none of them sets $B_S[i]$ to 1 is $(1 - \frac{1}{m})^{|S|}$ (since we can assume they are independent events). Therefore the probability that at least one element in $S$ sets $B_S[i]$ to 1 (i.e. the probability that $B_S[i] = 1$) is: $1 - (1 - \frac{1}{m})^{|S|}$.

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

$$X_i = \begin{cases} 1 & \text{IF } B_S[i] = 1 \\ 0 & \text{OTHERWISE} \end{cases}$$

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), is $1 - (1 - \frac{1}{m})^{|S|}$. Since we do not know $|S|$ we use the hypothesis, that is $|S| \leq \frac{m}{c}$. Therefore we have:

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

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