Feedback — Week3B (Basic)

Help Center

You submitted this quiz on **Sun 15 Feb 2015 2:07 PM PST**. You got a score of **2.00** out of **2.00**.

Question 1

Suppose we hash the elements of a set S having 20 members, to a bit array of length 99. The array is initially all-0's, and we set a bit to 1 whenever a member of S hashes to it. The hash function is random and uniform in its distribution. What is the expected fraction of 0's in the array after hashing? What is the expected fraction of 1's? You may assume that 99 is large enough that asymptotic limits are reached.

Your Answer		Score	Explanation
○	The fraction of 1's is $e^{-20/99}$.		
○	The fraction of 1's is $e^{-79/99}$.		
◉ The fraction of 1's is $1-e^{-20/99}$.	✔	1.00	
○ The fraction of 0's is 20/99.			
Total		1.00 / 1.00	

Question Explanation

The probability that a given bit is set to 1 is $1-e^{-20/99}$ (assuming that $(1-1/99)^{99}$ is exactly 1/e). This formula is derived in the assoc-rules1.ppt slides. Thus, the probability that the bit remains 0 is 1 minus this expression, or $e^{-20/99}$.

Question 2

A certain Web mail service (like gmail, e.g.) has 10^8 users, and wishes to create a sample of data about these users, occupying 10^{10} bytes. Activity at the service can be viewed as a stream of elements, each of which is an email. The element contains the ID of the sender, which must

be one of the 10^8 users of the service, and other information, e.g., the recipient(s), and contents of the message. The plan is to pick a subset of the users and collect in the 10^{10} bytes records of length 100 bytes about every email sent by the users in the selected set (and nothing about other users).

The method of Section 4.2.4 will be used. User ID's will be hashed to a bucket number, from 0 to 999,999. At all times, there will be a threshold t such that the 100-byte records for all the users whose ID's hash to t or less will be retained, and other users' records will not be retained. You may assume that each user generates emails at exactly the same rate as other users. As a function of n, the number of emails in the stream so far, what should the threshold t be in order that the selected records will not exceed the 10^{10} bytes available to store records? From the list below, identify the true statement about a value of n and its value of t.

Your Answer		Score	Explanation
○ $n = 10^{10}$; t = 100,000			
◉ $n = 10^{13}$; t = 9	✔	1.00	
○ $n = 10^{11}$; t = 1000			
○ $n = 10^{14}$; t = 1			
Total		1.00 / 1.00	

Question Explanation

Suppose that the fraction of users in the sample is p. That is, 10^8p is the number of users whose records are stored. Since each user generates 10^{-8} of the emails in the stream, when n emails have been seen, the number of records stored is $10^8p10^{-8}n = pn$. Note that is number does not depend on the number of users of the service.

Since each record is 100 bytes, we can store $10^{10}/100 = 10^8$ records. That is, $pn = 10^8$, or $p = 10^8/n$. If the threshold is t, the fraction p of users that will be in the selected set is (t+1)/1,000,000. That is, $(t+1)/1{,}000{,}000 = 10^8/n$, or $t = 10^{14}/n - 1$.