고려대학교
KOREA UNIVERSITY

Qualifying Exam

Spring, 2012

1. Bob wants to access Alice account illegally using a hash table for password available at the black market. Bob knows that the hash function is given as $y = x^2 + 2\, mod\, 999$ without reduction function. Bob picks her hash value of her password which 447. What is Alice' password? He also wants to use a rainbow table for this hash function. What is the role of salt in the hash computation?

Start of the Hash Value	End of the Hash value
189	420
2	11
367	524
409	11

2. A database is about to be anonymized using 4-anonymity method. The goal is to minimize the objective function that is the sum of the all square of range span values. When the range of age is [29-30], the range span is 2 (i.e., 30-29+1), and the rage of zip is [1234*], its range span is 10 (i.e., 12349-12340+1). Rearrange the database given below so that its objective function is minimized. What is l-diversity? Explain it using your rearranged database.

id	zip	age	condition
1	12345	29	HIV
2	12346	30	HIV
3	12344	27	anemia
4	12347	26	TB
5	12344	26	HIV
6	12345	28	HIV
7	12346	19	HIV
8	12346	28	HIV
9	12347	26	cancer
10	12348	26	HIV

3. Let $C(X)$be a clock of an event $X.$ Is the following statement "If $C(X) < C(Y)$, then X happened before Y" true? State why. What is the role of vector clock? Is the vector clock (1, 3, 4) and (2, 4, 5) are concurrent? What is the definition of concurrent processes?

4. What is a block cipher and what is a stream cipher? What are pros and cons of these methods in terms of computation time, security, and error propagation? Why the stream ciphers are preferred in mobile and wireless environment?

5. Suppose that a Bayesian spam filter is trained on a set of 11,000 spam messages and 4,000 messages that are ham. The word "improved" appears in 2,500 spam messages and 2000 messages that are ham, while the word "aloe" appears in 8000 spam messages and 200 messages that are ham. Estimate the probability that a received message containing both the words "improved" and "aloe" is spam. Will the message be rejected as spam if the threshold for rejected as spam if the threshold for rejecting spam is 0.9?

6. RSA algorithm is summarized as follows:

    A. Pick two random numbers p and q. (Here, $p = 11$ and $q = 17$.

    B. Compute $z = (p-1) \cdot (q-1)$.

    C. Find an integer d that is relative prime to z. (Here, $d = 19$).

    D. Find e which satisfies $e \cdot d \equiv 1 \bmod z$. Find e.

Solutions

1. $\mathrm{mod}\ (447^2 + 2, 999) = 11$. It means that starting point of the rainbow table is either 2 or 409.

   $\mathrm{mod}\ (2^2 + 2, 999) = 6$. $\mathrm{mod}\ (6^2 + 2, 999) = 38$. $\mathrm{mod}\ (38^2 + 2, 999) = 447$.

   Thus, her password is highly likely to be 38.

   The starting value 409 does not hit 447 before hitting 11. Thus, fourth row misleads attackers.

   Describe your own idea about the role of salt in hashing.

2. Since there are 10 records, possible clusters have either (four records, six records), (five records, five records), or (six records, four records). For each clustering, there are many solutions at your disposal. However, metric value should be as small as possible. One possible solution is given as follows:

id	zip	age	condition
1	[12344-12347]	[26-27] [29-30]	HIV
2			HIV
3			anemia
4			TB
5			HIV
6	[12345-12348]	[19, 26, 28]	HIV
7			HIV
8			HIV
9			cancer
10			HIV

Metric value of this case is given computed as follows:

$$5 \cdot (5^2 + 4^2 + 4^2 + 3^2) = 330$$

Describe your own idea about the role of l-diversity. In this example, each cluster has at least two different conditions such as HIV, anemia, and TB or HIV and cancer. Thus, this clustering satisfies 2-diversity.

Better solution than above one is given as follows:

id	zip	age	condition
1	**[12344-12346]**	**[26-29]**	**HIV**
3			**anemia**
5			**HIV**
6			**HIV**
8			**HIV**
2	[12346-12348]	[19, 26, 30]	HIV
4			TB
7			HIV
9			cancer
10			HIV

Metric value of this case is computed as follows:

$$5 \cdot (3^2 + 3^2 + 3^2 + 4^2) = 215$$

Since this value 215 is smaller than that value 330, this case is said to be better than that case.

3. False since concurrent events do not meet the inequality condition. See what the definition of the concurrent events is. Since (1, 3, 4) is smaller than (2, 4, 5), these two vector clocks are not concurrent.

4. (intended blank)

5. Assume that I be "improved", A be "aloe", and S be spam, and H be ham, where two events I and A are independent.

$$P(S|I\&A) = \frac{P(I|S)P(A|S)P(S)}{P(I|S)P(A|S)P(S) + P(I|H)P(A|H)P(H)}$$

where $P(I|S) = \frac{2500}{11000}$, $P(A|S) = \frac{8000}{11000}$, $P(I|H) = \frac{2000}{4000}$, $P(A|H) = \frac{200}{4000}$

1) If we assume that $P(S) = P(H) = 0.5$,

$$P(S|I\&A) = \frac{P(I|S)P(A|S)}{P(I|S)P(A|S) + P(I|H)P(A|H)} = \frac{\frac{2500}{11000} \cdot \frac{8000}{11000}}{\frac{2500}{11000} \cdot \frac{8000}{11000} + \frac{2000}{4000} \cdot \frac{200}{4000}}$$

$$= 0.8686$$

2) If we assume that $P(S) = \frac{11}{15}, P(H) = \frac{4}{15}$,

$$P(S|I\&B) = \frac{P(I|S)P(A|S)P(S)}{P(I|S)P(A|S)P(S) + P(I|H)P(A|H)P(H)}$$

$$= \frac{\frac{2500}{11000} \cdot \frac{8000}{11000} \cdot \frac{11}{15}}{\frac{2500}{11000} \cdot \frac{8000}{11000} \cdot \frac{11}{15} + \frac{2000}{4000} \cdot \frac{200}{4000} \cdot \frac{4}{15}} = 0.9479$$

Thus, acceptance depends on the marginal probabilities.

6. Extended Euclid algorithm shows:

$8 = 160 \cdot 1 - 19 \cdot 8$ $\quad$ $3 = 19 \cdot 17 - 160 \cdot 2$ $\quad$ $2 = 160 \cdot 5 - 19 \cdot 42$

$1 = 19 \cdot 59 - 160 \cdot 7$

Therefore, 59 is the solution.