[redacted]

CS 4365/CS 5354 Data Processing Under Security and Privacy Summer 2016, Test 2

Name: [redacted]

1-2. In class, we learned the following algorithm for eliminating outliers:

- we compute the sample mean $\mu = (x_1 + \ldots + x_n) / n$,
- we compute sample variance $V = ((x_1 - \mu)^2 + \ldots + (x_n - \mu)^2) / (n - 1)$, and the sample standard deviation σ as the square root of V;
- then, we dismiss all the records which are not in the 2-sigma interval $[\mu - 2\sigma, \mu - 2\sigma]$ as outliers;
- after that, we repeat the same procedure for the new database, with some outliers deleted: we re-compute μ and σ and, if needed, eliminate values which are not in the new 2-sigma interval, etc.;
- this process continues until, at some stage, no more outliers are eliminated.

1. Use this algorithm to eliminate outliers from the following database:

10
10

- we have 100 records each equal to 1.9;
- we have 100 records each equal to 2.1;
- we have one record with value 10; and
- we have one record with value 1,000.

The algorithm is straightforward:

10
10

2. Explain how outlier elimination is used in computer security.

$$\mu = \frac{100 * 1.9 + 100 * 2.1 + 10 + 1000}{202} = 6.98$$

$$V = \frac{100 * (1.9 - 6.98)^2 + 100 * (2.1 - 6.98)^2 + (10 - 6.98)^2 + (1000 - 6.98)^2}{201}$$

$$= \frac{2580.64 + 2381.44 + 9.12 + 986088.72}{201}$$

$$= 4930.65$$

$$\sigma = \sqrt{V} = 70.22$$

The two-sigma interval : $[6.98 - 2 * 70.22,\ 6.98 + 2 * 70.22]$
$= [-133.46, 147.42]$.

Clearly, 1000 is the out of interval. So, we can ~~on~~ consider 1000 as an outlier and remove it from dataset.

Now we have,

100 records each with 1.9
100 records each with 2.1
1 records with 10.

$$\mu = \frac{100 * 1.9 + 100 * 2.1 + 10}{201} = 2.04$$

$$V = \frac{100 * (1.9 - 2.04)^2 + 100 * (2.1 - 2.04)^2 + (10 - 2.04)^2}{200}$$

$$= \frac{1.96 + 0.36 + 63.36}{200}$$

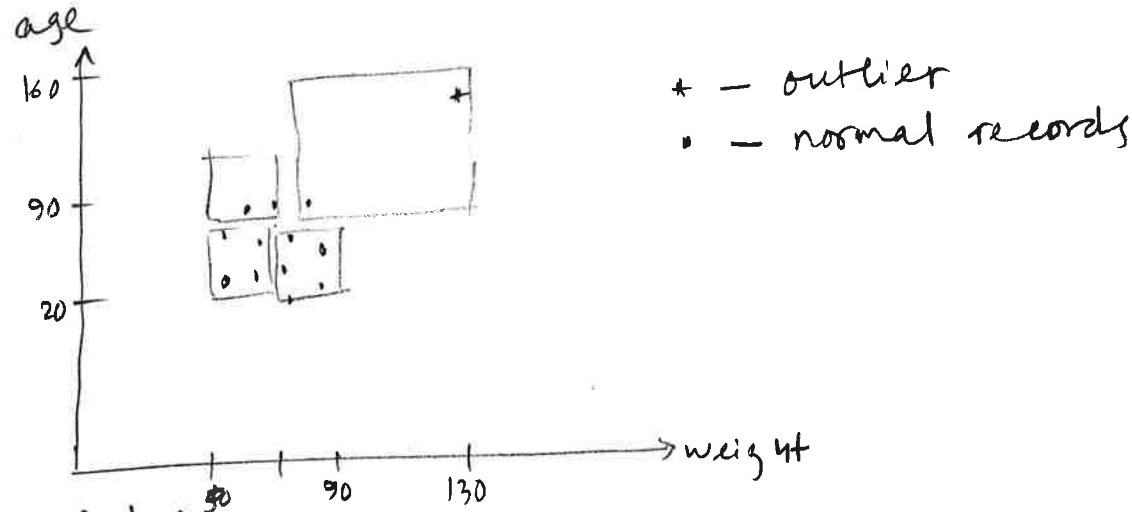
$$= 0.33$$

$$\sigma = 0.57$$

Now, the two sigma interval $[0.9, 3.18]$. Again 10 is out of the interval. We can consider 10 as another outlier and remove it from dataset.

There are no more outlier after that.

- we need to eliminate the outlier to protect the privacy. For example, age and weights are distributed



graph 1:

~~like~~ in graph 1.

if we find the interval for weight and age for K anonymity, we might find the box like the graph 1. Larger box leads to more error in computation. If we ignore the outlier, we will ~~the~~ find the nice smaller interval boxes with more accurate query result keeping the privacy.

[redacted]

CS 4365/CS 5354 Data Processing Under Security and Privacy
Summer 2016, Test 2

Name: [redacted]

1-2. In class, we learned the following algorithm for eliminating outliers:

10/10

- we compute the sample mean $\mu = (x_1 + \ldots + x_n) / n$,
- we compute sample variance $V = ((x_1 - \mu)^2 + \ldots + (x_n - \mu)^2) / (n - 1)$, and the sample standard deviation σ as the square root of V;
- then, we dismiss all the records which are not in the 2-sigma interval $[\mu - 2\sigma, \mu - 2\sigma]$ as outliers;
- after that, we repeat the same procedure for the new database, with some outliers deleted: we re-compute μ and σ and, if needed, eliminate values which are not in the new 2-sigma interval, etc.;
- this process continues until, at some stage, no more outliers are eliminated.

1. Use this algorithm to eliminate outliers from the following database:

- we have 100 records each equal to 1.9;
- we have 100 records each equal to 2.1;
- we have one record with value 10; and
- we have one record with value 1,000.

The algorithm is straightforward:

10/10

2. Explain how outlier elimination is used in computer security.

1) 100 · 1.9 , 100 · 2.1 , 10, 1000

$$\mu = \frac{(100(1.9) + 100(2.1) + 10 + 1000)}{202} = 6.98$$

$$V = \frac{(100(1.9-6.98)^2 + 100(2.1-6.98)^2 + (10-6.98) + (1000-6.98))}{201}$$

$$\sigma = \sqrt{29.64} = 5.44$$

2-sigma interval = $[(6.98) - 2(5.44), (6.98) + 2(5.44)]$

$= [-3.9, 17.86]$

Only value that is not in interval is 1000

Again

$$\mu = \frac{(100(1.9) + 100(2.1) + 10)}{201} = 2.04$$

$$V = \frac{(100(1.9-2.04)^2 + 100(2.1-2.04)^2 + (10-2.04)^2)}{200}$$

$$\sigma = \sqrt{0.3284} = .573$$

2 sigma interval = $[(2.04) - 2(5.73), 2.04 + 2(5.73)]$

$= [.894, 3.186]$

Outlier is 10

Again

$$\mu = \frac{(100(1.9) + 100(2.1))}{200} = 2$$

$$V = \frac{(100(1.9-2)^2 + 100(2.1-2)^2)}{199} = .01, \ \sigma = \sqrt{.01} = .1$$

2-Sigma $= [2 - .2, 2 + .2]$

$= [1.8, 2.2]$

No more outliers

10/10

3. In class, we derived a general formula for the optimal sizes Δ_i of a privacy-enhancing box for which the inaccuracy in the resulting estimation of a statistical characteristic C is the smallest possible under the condition of k-anonymity: $\Delta_i = c / a_i$, where a_i is the absolute value of the partial derivative of C with respect to the i-th variable x_i, and $c = (1/2) * \sqrt{(k * a_1 * a_2 * \ldots) / \rho(x)}$, where $\rho(x)$ is the data density, i.e., number of record per unit volume.

Use the general formula to find the sizes of the cell that provides the smallest possible inaccuracy in computing the covariance between weight and age for adults. Assume that:

- we are looking for k-anonymity with k = 10,
- we deal with a population of El Paso, N = 700,000 students,
- age is uniformly distributed on the interval [20, 90];
- weight is uniformly distributed on the interval [50, 90]; and
- we are interested in the cell that covers a record with age $x_1 = 65$ and weight $x_2 = 80$ kg.

For covariance, the derivative with respect to x_1 is equal to $(x_2 - \mu_2) / N$ and the derivative with respect to x_2 is equal to $(x_1 - \mu_1) / N$, where μ_i are the sample means of the corresponding values.

Given that,

$N = 700,000$

C = covariance of age and weight

age, $x_1 \in [20, 90]$

weight, $x_2 \in [50, 90]$

$K = 10$, $x_1 = 65$ and $x_2 = 80$

Since the ~~their~~ age and height is uniformely distributed,

$$\mu_1 = \frac{20+90}{2} = 55$$

$$\mu_2 = \frac{50+90}{2} = 70.$$

$$a_1 = \frac{|x_2 - \mu_2|}{N} = \frac{|80-70|}{700,000} = \frac{10}{700000} = \frac{1}{70000}$$

$$a_2 = \frac{|x_1 - \mu_1|}{N} = \frac{|65-55|}{700000} = \frac{1}{70000}.$$

$$\rho(x) = \frac{700,0\cancel{0}\cancel{0}}{7\cancel{0} * 4\cancel{0}} = \frac{1000}{4} = 250$$

$$c = \frac{1}{2}\sqrt{\frac{k * a_1 * a_2}{\rho(x)}} =$$

$$\Delta_1 = \frac{1}{2}\sqrt{\frac{k}{\rho(x)}} \cdot \frac{\sqrt{a_1 * a_2}}{a_1}$$

$$= \frac{1}{2}\sqrt{\frac{10}{250}} \cdot \frac{\sqrt{\frac{1}{70000} \cdot \frac{1}{70000}}}{\frac{1}{7000}}$$

$$= \frac{1}{2}\sqrt{\frac{1}{25}}$$

$$= \frac{1}{10} = \underline{\underline{0 \cdot 1}}.$$

$$\Delta_2 = \frac{1}{2}\sqrt{\frac{k}{\rho(x)}} \cdot \frac{\sqrt{a_1 * a_2}}{a_2}$$

$$= \frac{1}{2}\sqrt{\frac{10}{250}} \cdot 1.$$

$$= \underline{\underline{0 \cdot 1}}.$$

10/10

3. In class, we derived a general formula for the optimal sizes Δ_i of a privacy-enhancing box for which the inaccuracy in the resulting estimation of a statistical characteristic C is the smallest possible under the condition of k-anonymity: $\Delta_i = c / a_i$, where a_i is the absolute value of the partial derivative of C with respect to the i-th variable x_i, and $c = (1/2) * \sqrt{(k * a_1 * a_2 * \ldots) / \rho(x)}$, where $\rho(x)$ is the data density, i.e., number of record per unit volume.

Use the general formula to find the sizes of the cell that provides the smallest possible inaccuracy in computing the covariance between weight and age for adults. Assume that:

- we are looking for k-anonymity with k = 10,
- we deal with a population of El Paso, N = 700,000 students,
- age is uniformly distributed on the interval [20, 90];
- weight is uniformly distributed on the interval [50, 90]; and
- we are interested in the cell that covers a record with age $x_1 = 65$ and weight $x_2 = 80$ kg.

For covariance, the derivative with respect to x_1 is equal to $(x_2 - \mu_2) / N$ and the derivative with respect to x_2 is equal to $(x_1 - \mu_1) / N$, where μi are the sample means of the corresponding values.

$$\rho(x) = \frac{700{,}000}{70 \cdot 40} = 250$$

$$M_1 = \frac{20+90}{2} = 55 \quad M_2 = \frac{50+90}{2} = 70$$

$$a_1 = \frac{(80-70)}{700{,}000} = \frac{1}{70{,}000}$$

$$a_2 = \frac{(65-55)}{700{,}000} = \frac{1}{70{,}000}$$

$$\Delta_1 = \frac{1}{2} \cdot \sqrt{\frac{10}{250}} \cdot \frac{\sqrt{\frac{1}{70000} \cdot \cancel{\frac{1}{70{,}000}}}}{\cancel{\frac{1}{70{,}000}}} = .1$$

Δ_2 = Same $\Rightarrow$ .1

$$C = \frac{1}{2} \cdot \sqrt{\left(10 \cdot \frac{1}{70{,}000} \cdot \frac{1}{70000}\right) / 250}$$

$$= .00000001429$$

$$\Delta_1 = \frac{\checkmark}{1/70{,}000} = .1 \checkmark$$

10/10

4. What if in Problem 3, in addition to k-anonymity, we also require l-diversity, with l = 2, and the thresholds $\varepsilon_1 = 0.2$ and $\varepsilon_2 = 0.01$?

Reminder:

- If for the k-anonymous solution, we have $2\Delta_i \geq l * \varepsilon_i$, then l-anonymity is also satisfied.
- If for some of the variables x_i, this inequality is not satisfied, then for this variable, we select:
  - for this variable, we select $\Delta_i = (1/2) * l * \varepsilon_i$, and
  - for other variables, we select Δ_j from the condition that the cell contain k records, i.e., that $\rho(x) * 2^n * \Delta_1 * \Delta_2 * \ldots = k$.

$l = 2$, $\varepsilon_1 = 0.2$, $\varepsilon_2 = 0.01$

$\Delta_1 = \Delta_2 = .1$

Δ_1

$2(.1) \geq (2)(.2)$

$.2 \geq .4$ X

Δ_2

$2(.1) \geq 2(.01)$

$.2 \geq .02$ ✓

order

$q_1 \cdot \varepsilon_1 \geq q_2 \cdot \varepsilon_2$

$\Delta_1 = l \cdot \varepsilon_1 \cdot (1/2) = \boxed{.2 = \Delta_1}$

$\rho(x) \cdot 2^n \cdot \Delta_1 \cdot \Delta_2 = k$

$250 \cdot 4 \cdot (.2) \cdot \Delta_2 = 10$

$\Delta_2 = \dfrac{10}{250 \cdot 4 \cdot (.2)}$

$\boxed{\Delta_2 = .05}$

10/10

5. What percentage of privacy do we lose if someone detects the first digit x in the age xy? the second digit y? Assume that the age is between 20 and 90.

The age is between 20 and 90. So the width of interval = 90 − 20 = 70.

if someone hacks the first digit, x, the maximum privacy loss will be occurred if the x=9. Then, 90 is the only number.

The privacy loss = $\frac{70}{70}$ = 100%

for any other digits excepts 9, the privacy loss will be,

$\frac{70-9}{70} \approx 87\%$.

20
⋮
29

if someone knows the last digit y, the minumum privacy loss will be occurred if y=0,

privacy loss = $\frac{0}{100}$ = 0%

if y is anything other than 0, then

privacy loss = $\frac{10}{70} \approx 14\%$.

21
⋮
81

10/10

5. What percentage of privacy do we lose if someone detects the first digit x in the age xy? the second digit y? Assume that the age is between 20 and 90.

age = [20, 90]

range = 70

detects x

for x=2 => 20, 21, ..., 29

range = 9

loss of privacy = $\frac{|70-9|}{70}$ = 87 %

detects y

case 1: for y = 5 => 25, 35, ..., 85

range = 60

loss of privacy = $\frac{|70-60|}{70}$ = 14 %

Case 2: y = 0 => 20, 30, ..., 90

range = 70

loss of privacy = 0 %

Case 1 works for y=1-9, and case 2 only works for y=0, so case 1 is more prominent.

10/10

7. Use the efficient raising-to-the-power algorithm to compute 5^{21} mod 11. Where is this algorithm used in RSA coding?

$21_{10} \equiv 10101_2$
16 8 4 2 1

So,
21 can be represent as $(16+4+1)$.

Therefore,
$5^{21} = 5^{16} * 5^{4} * 5^{1}$

$5 \bmod 11 = 5$

$5^2 \bmod 11 = 3$

$5^4 \bmod 11 = (3*3) \bmod 11 = 9$

$5^8 \bmod 11 = (9*9) \bmod 11 = 4$

$5^{16} \bmod 11 = (4*4) \bmod 11 = 5$

$\therefore$ $5^{21} \bmod 11 = (5^{16} * 5^{4} * 5^{1}) \bmod 11$

$= ((5^{16} \bmod 11) * (5^4 \bmod 11) * (5^1 \bmod 11)) \bmod 11$

$= (5 * 9 * 5) \bmod 11$

$= ((5*9 \bmod 11) * 5) \bmod 11$

$= (1*5) \bmod 11$

$= 5$

This method is used in both generate encoded message and get decoded message in RSA algorithm.

20/20

8-9. Use Euclid's algorithm to compute the greatest common divisor gcd(11, 21). Then find a number d for which d * 11 mod 21 = 1. Where is this used in RSA coding?

First we use Euclid's algorithm to check that indeed $\gcd(11, 21) = 1$.

$$21 = 1 * 11 + 10 \longrightarrow (i)$$

$$11 = 1 * 10 + 1 \longrightarrow (ii)$$

The last remainder is 1, so indeed $\gcd(11, 21) = 1$.

Next, for each of the remainder 10 and 1, we represent it as a linear combination of the original numbers 11 and 21 until we get such a representation for remainder 1.

from (i),

$$10 = 21 - 11$$

from (2), $1 = 11 - 10$

$$= 11 - (21 - 11)$$

$$= 2 * 11 - 1 * 21.$$

Here, $2 * 11 - 1 * 21 = 1$.

So, $2 * 11 = 1 \bmod 21$.

The answer is $d = 2$.

This method is used in find the secret key, d, in RSA coding.