浙江大學
Zhejiang University

Summer 2022: Big Data Security and Privacy Protection

Homework Assignment, **Due Time, Date** *11:59 PM, xxx-xxx, 2023*

Student Name: ______________________________ Matriculation Number: _________________

Instructor: Rongxing Lu
The marking scheme is shown in the left margin and [100] constitutes full marks.
Q.1-Q.3 belong to Set 1, **Due Time, Date** 11:59 PM, August 28, 2023.
Q.4-Q.7 belong to Set 2, **Due Time, Date** 11:59 PM, August 31, 2023.
Q.8-Q.11 belong to Set 3, **Due Time, Date** 11:59 PM, September 3, 2023.
For Set 1, please send your solutions in pdf with name "S23-HW-YourName-StudentID-Set-1.pdf" to Xiaoyu_z@zju.edu.cn and rxlu.cn@gmail.com before the due date, time (GMT+8). The title of your email is "S23-HW-YourName-StudentID-Set-1.pdf". Similar cases are "S23-HW-YourName-StudentID-Set-2.pdf" for Set 2 and "S23-HW-YourName-StudentID-Set-3.pdf" for Set 3.

[**5**] 1. Consider an automated cash deposit machine in which users provide a card or an account number to deposit cash. Give examples of confidentiality, integrity, and availability requirements associated with the system, and, in each case, indicate the degree of importance of the requirement.)

[**5**] 2. One way to solve the key distribution problem is to use a line from a book that both the sender and the receiver possess. Typically, at least in spy novels, the first sentence of a book serves as the key. The

particular scheme discussed in this problem is from one of the best suspense novels involving secret codes, *Talking to Strange Men*, by Ruth Rendell. Work this problem without consulting that book! Consider the following message:

```
SIDKHKDM AF HCRKIABIE SHIMC KD LFEAILA
```

This ciphertext was produced using the first sentence of *The Other Side of Silence* (a book about the spy Kim Philby):

The snow lay thick on the steps and the snowflakes driven by the wind looked black in the headlights of the cars.

A simple substitution cipher was used.

[4] (a) What is the encryption algorithm? Please describe the algorithm, and show the plaintext.

[1] (b) How secure is the simple substitution cipher?

[**10**] 3. We describe a special case of a **Permutation Cipher**. Let m, n be positive integers. Write out the plaintext, by rows, in $m \times n$ rectangles. Then form the ciphertext by taking the columns of these rectangles. For example, if $m = 4, n = 3$, then we would encrypt the plaintext "CRYPTOGRAPHY"

```
CRYP
TOGR
APHY
```

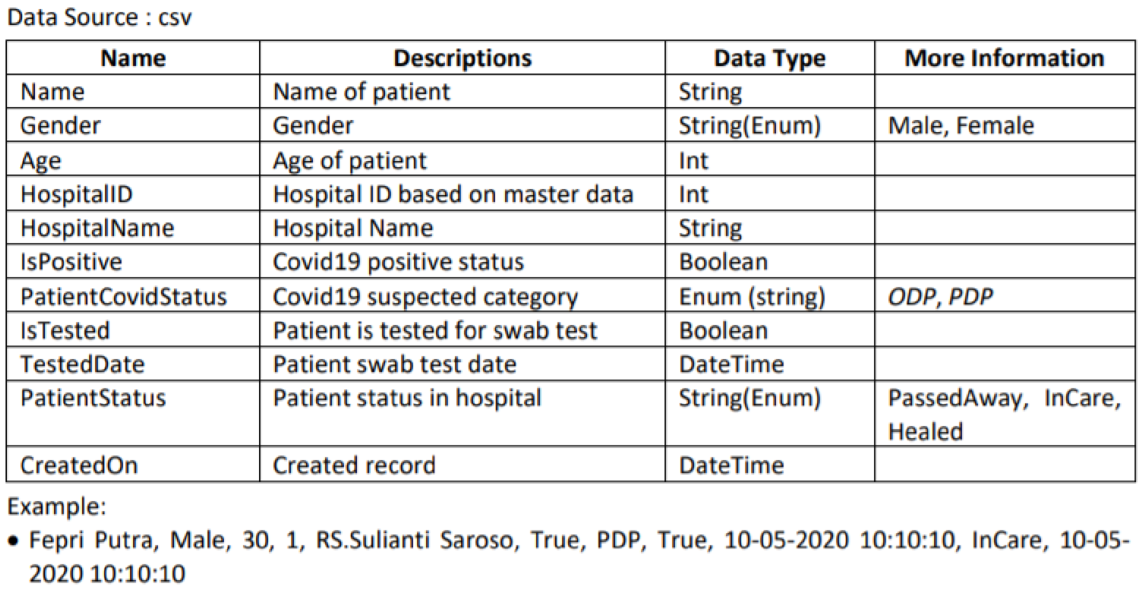
The ciphertext would be "CTAROPYGHPRY".

(a) Given a ciphertext encrypted with the above method, describe how you would decrypt the ciphertext (given values for m and n).

(b) Decrypt the following ciphertext, which was obtained by using this method of encryption:

```
    MYAMRARUYIQTENCTORAHROYWDSOYEOUARRGDERNOGW
```

[**10**] 4. Consider a dataset about Hospital Covid19 Case, which mainly describes the status of covid19 patient in care at hospital with the following attributes. Now, the hospital wants to release the dataset for some potential scientific analytics, please prepare a design guideline for the hospital so that a released version of the dataset can balance the patients' privacy and the whole dataset's utility.



Data Source : csv

Name	Descriptions	Data Type	More Information
Name	Name of patient	String	
Gender	Gender	String(Enum)	Male, Female
Age	Age of patient	Int	
HospitalID	Hospital ID based on master data	Int	
HospitalName	Hospital Name	String	
IsPositive	Covid19 positive status	Boolean	
PatientCovidStatus	Covid19 suspected category	Enum (string)	*ODP, PDP*
IsTested	Patient is tested for swab test	Boolean	
TestedDate	Patient swab test date	DateTime	
PatientStatus	Patient status in hospital	String(Enum)	PassedAway, InCare, Healed
CreatedOn	Created record	DateTime	

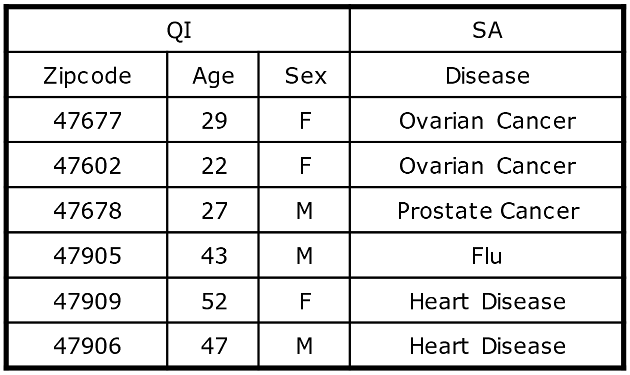
Example:

- Fepri Putra, Male, 30, 1, RS.Sulianti Saroso, True, PDP, True, 10-05-2020 10:10:10, InCare, 10-05-2020 10:10:10

[**10**] 5. Incognito is one of approaches to implement the k-anonymity. Given a table below, the full-domain generalizations described by "domain vectors" are represented as follows.

- $Z_0 = \{47677, 47602, 47678, 47905, 47909, 47906\} \rightarrow Z_1 = \{476**, 4790*\}$
- $A_0 = \{29, 22, 27, 43, 52, 47\} \rightarrow A_1 = \{2*, [43, 52]\}$
- $S_0 = \{M, F\} \rightarrow S_1 = \{*\}$

Please apply (Z_1, A_1, S_1) to generalize the original table, and discuss what is the value of k in your generalized table? In your generalized table, if we apply the definition of distinct l-diversity (a table is l-diverse if each of its QI groups contains at least l "well-represented" values for the SA), what is the value of l? If we apply the definition of entropy l-diversity (Entropy l-diversity: for each QI group g, $entropy(g) \geq \log(l)$), what is the value of l?

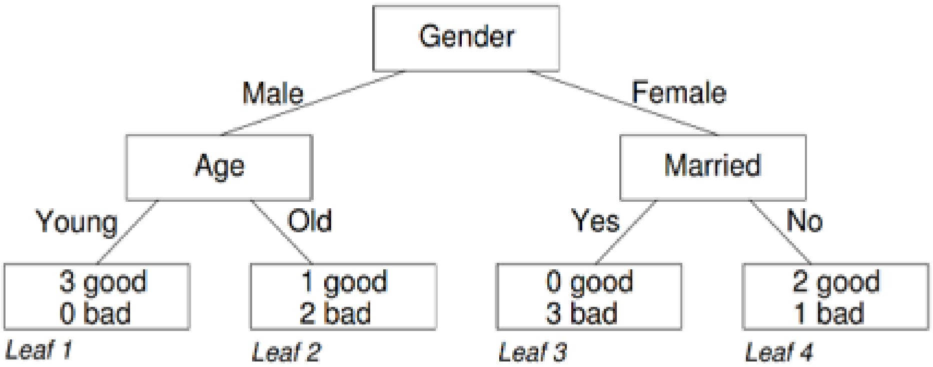


QI			SA
Zipcode	Age	Sex	Disease
47677	29	F	Ovarian Cancer
47602	22	F	Ovarian Cancer
47678	27	M	Prostate Cancer
47905	43	M	Flu
47909	52	F	Heart Disease
47906	47	M	Heart Disease

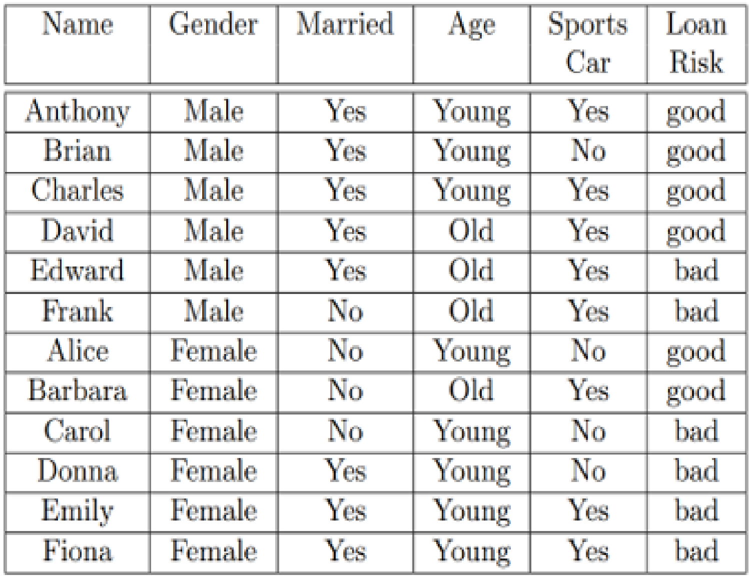
[**10**] 6. Decision tree learning is one of the predictive modelling approaches used in statistics, data mining and

machine learning. It uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). Consider the following dataset, which describes the loan risk information of a mortgage company with the following attributes: Gender, Married, Age, Sports Car, and Loan Risk. Now, the mortgage company wants to release the table. Taking the privacy into account (balancing the privacy and utility), can you help the company to generate a k-anonymous version of the table to fit a better utility of the mortgage company decision tree, i.e., the decision tree can be correctly generated from the released k-anonymous table. Further, what is the value of k? If considering l-diversity, what is the value of l?

Mortgage company decision tree



Mortgage company data



Name	Gender	Married	Age	Sports Car	Loan Risk
Anthony	Male	Yes	Young	Yes	good
Brian	Male	Yes	Young	No	good
Charles	Male	Yes	Young	Yes	good
David	Male	Yes	Old	Yes	good
Edward	Male	Yes	Old	Yes	bad
Frank	Male	No	Old	Yes	bad
Alice	Female	No	Young	No	good
Barbara	Female	No	Old	Yes	good
Carol	Female	No	Young	No	bad
Donna	Female	Yes	Young	No	bad
Emily	Female	Yes	Young	Yes	bad
Fiona	Female	Yes	Young	Yes	bad

[**10**] 7. Assume a data owner $\mathcal{A}$ has a table D below, showing the one-day electricity uses of all users in one residential area. From the table, we can observe that the electricity use of each user ranges from 0 to 5. Now, in the interactive database query (IDQ) model, a client wants to launch a statistical function query

"$Sum(D)$", i.e., "What is the total electricity consumption in the residential area at that day?" Please try your best to answer the following questions.

- [5] What is the sensitivity of the function $S(F)$ of $Sum()$ in this table D?
- [5] Consider we set the privacy level as ϵ, how to randomly choose a random noise from a Lapacian distribution so that we can achieve ϵ-Differential Privacy in the statistical function query "$Sum(D)$". Please follow the lecture note to prove your result, i.e.,

$$\Pr[A(D+I) \in T] \leq e^{\epsilon} \Pr[A(D-I) \in T]$$

where $A(*) = Sum(*) + Lapacian\ noise.$

ID#	Name	Electricity Use (0-5/day)
1	Alice	4
2	Bob	2
3	Carlo	1
4	David	4
5	Elvas	5
6	Ford	0
7	Geoge	2
8	Hilton	4
9	Ives	2
10	Jack	5

[**10**] 8. Please answer the following RSA related sub-questions.

[5] (a) In a public-key system using RSA, you intercept the ciphertext $C = 8$ sent to a user whose public key is $e = 5, n = 35$. What is the plaintext M?

[5] (b) The reason that you can recover the plaintext M in Question 1(a) is that $n = 35$ is too small, you can factor n and obtain the private key d. However, when we set the length of n is 1024 bits, i.e., $|n| = 1024$, the large integer factoring problem becomes hard. Now, when you intercept a ciphertext $C \equiv M^e \bmod n$, where $M \in \{0,1\}^{160}$, $e = 5$, and $|n| = 1024$, can you recover the message M from C without using the brute force? Why or why not?

[**10**] 9. Please answer the following ElGamal encryption related sub-questions.

[5] (a) Consider an ElGamal encryption scheme with a common prime $q = 11$ and a primitive root $\alpha = 2$. If B has public key $Y_B = 3$ and A chooses the random integer $k = 2$, what is the ciphertext of $M = 9$?

[5] (b) As we discussed in class, the message M cannot be 0 in the ElGamal encryption. Then, what strategy can you use to encrypt a message 0 in the ElGamal encryption? Please describe your strategy as detail as possible.

[**10**] 10. Use the Chinese Remainder Theorem (CRT) to solve x, where

$$\begin{cases} x & \equiv & 1 \bmod 3 \\ x & \equiv & 3 \bmod 5 \\ x & \equiv & 5 \bmod 7 \end{cases}$$

[**10**] 11. Please prove the following two results.

[5] (a) Let $q \geq 7$ be a prime number, prove the number $\underbrace{11 \cdots 1}_{q-1 \ \ 1's}$ can be divisible by q.

[5] (b) Let $x \geq 1$ be a positive integer, prove $Y = x + \sum_{i=1}^{x} 2^{2i-1}$ can be divisible by 3.

Solutions.

Q1. The system must keep personal identification numbers confidential, both in the host system and during transmission for a transaction. It must protect the integrity of account records and of individual transactions. Availability of the host system is important to the economic well being of the bank, but not to its fiduciary responsibility. The availability of individual teller machines is of less concern.

Confidentiality requirements:

a) The communication channel between the ATM and the bank must be encrypted

b) The Pin must be encrypted (whereever it is stored)

Integrity requirements:

a) the actions performed via the ATM must be associated to the account associated with the card

Availability requirements:

a) the system must be able to serve concurrent users at any given time

b)the system must be available 99.9% of the time.

Q2.a The first letter t corresponds to A, the second letter h corresponds to B, e is C, s is D, and so on. Second and subsequent occurrences of a letter in the key sentence are ignored. The result

```
ciphertext: SIDKHKDM AF HCRKIABIE SHIMC KD LFEAILA
plaintext:  basilisk to leviathan blake is contact
```

Q2.b It is easily breakable.

Q3.a

Given values for m and n, to decrypt the ciphertext we should write ciphertext vertically for n rows and get m columns in the end. After that read it horizontally from the top to the bottom.

Q3.b

MYAMRARUYIQTENCTORAHROYWDSOYEOUARRGDERNOGW

(1) The length of the ciphertext is 42.

(2) Analyze the segment length. 42 = 21*2 = 14*3 = 7*6 Possible segment lengths are 21,14,6

(3) Factor each segment lengths and get (m,n) and see if plaintext is meaningful.

6 = 2*3 = 3*2 14 = 2*7 = 7*2 21 = 3*7 = 7*3

So the answer is

MARY MARY QUITE CONTRARY HOW DOES YOUR GARDEN GROW

Q4:

1. to classify the types of Identifer, QI, and SA.

2. Remove Identifer.

4. According to the background knowledge of attack, apply tuple suppression, attribute generalization. To prevent the possible linking attack, homogeneity attack, we can consider the k-anonymity technique, l-diversity, even t-closeness.

5. Sometimes, permutation technique can be also applied.

Q5.

QI			SA
Zipcode	Age	Sex	Disease
47677	29	F	Ovarian Cancer
47602	22	F	Ovarian Cancer
47678	27	M	Prostate Cancer
47905	43	M	Flu
47909	52	F	Heart Disease
47906	47	M	Heart Disease

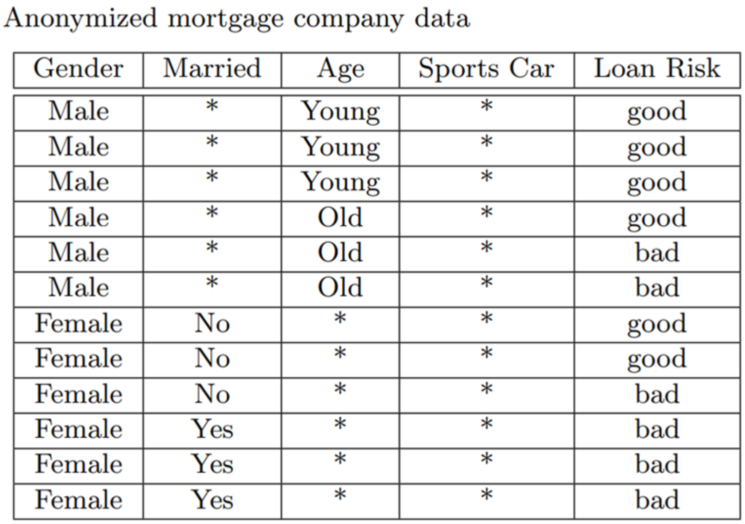
Solution: Obviously, from the generalized table, $k = 3$. When considering the distinct l-diversity, $l = 2$. While for the entropy l-diversity, in the first group, $P_1 = \Pr(OvarianCancer) = 2/3$, $P_2 = \Pr(ProstateCancer) = 1/3$. Therefore, from $entropy(thefirstgroup) \geq \log l$.

$$Entropy(thefirstgroup) = -(P_1 \log P_1 + P_2 \log P_2) = -(2/3 \log(2/3) + 1/3 \log(1/3)) = 0.276 \geq \log l$$

We have $l = 1.88$. Similarly, for the second group, we will also have $l = 1.88$. Therefore, the entropy l-diversity is $l = 1.88$.

QID			SA
Zipcode	Age	Sex	Disease
476**	2*	*	Ovarian Cancer
476**	2*	*	Ovarian Cancer
476**	2*	*	Prostate Cancer
4790*	[43,52]	*	Flu
4790*	[43,52]	*	Heart Disease
4790*	[43,52]	*	Heart Disease

Q6.



Anonymized mortgage company data

Gender	Married	Age	Sports Car	Loan Risk
Male	*	Young	*	good
Male	*	Young	*	good
Male	*	Young	*	good
Male	*	Old	*	good
Male	*	Old	*	bad
Male	*	Old	*	bad
Female	No	*	*	good
Female	No	*	*	good
Female	No	*	*	bad
Female	Yes	*	*	bad
Female	Yes	*	*	bad
Female	Yes	*	*	bad

K=3, L=1

Q7.

a. Solution: Because the range is $[0,5]$, When one record is different, $S(F) = |Sum(D_1) - Sum(D_2)| = |$ max value of the different record $| = 5$.

b. Proof: Suppose that $A = F(x) + Lap(\lambda) = Sum(*) + Lapacian\ noise$, and $D_1 = D + I$ and $D_2 = D - I$ are any two adjacent DBs. Thus, $A(D_1) = F(D_1) + x_1$ and $A(D_2) = F(D_2) + x_2$, where x_1 and x_2 are $Lap(\lambda)$ distributed. Since $\lambda = S(F)/\epsilon$, the probability density for x_1 is proportional to $e^{-|x_1|_1(\frac{\epsilon}{S(F)})}$. Similarly, the probability density for x_1 is proportional to $e^{-|x_2|_1(\frac{\epsilon}{S(F)})}$. Therefore, for any $T \in range(A)$

$$\frac{Pr[A(D_1) = T]}{Pr[A(D_2) = T]} = \frac{Pr[F(D_1) + x_1 = T]}{Pr[F(D_2) + x_2 = T]} = \frac{Pr[x_1 = T - F(D_1)]}{Pr[x_2 = T - F(D_2)]} = \frac{e^{-||T-F(D_1)||_1(\frac{\epsilon}{S(F)})}}{e^{-||T-F(D_2)||_1(\frac{\epsilon}{S(F)})}}$$

$$= e^{(||T-F(D_2)||_1 - ||T-F(D_1)||_1)(\frac{\epsilon}{S(F)})} \leq e^{||F(D_2)-F(D_1)||_1(\frac{\epsilon}{S(F)})}$$

where the inequality follows form the triangle inequality. By the definition of sensitivity,

$$S(F) = Max_{D_1,D_2:|D_1-D_2|=1}|F(D_1) - F(D_2)|$$

.

Thus, $e^{||F(D_2)-F(D_1)||_1(\frac{\epsilon}{S(F)})} \leq e^{\epsilon}$. The ration is bounded by e^{ϵ}, yields ϵ-Differential Privacy.

Q8.

1. Since $n = 35$, we have $p = 5$ and $q = 7$.

   Then,

$$\phi(n) = (p-1)(q-1) = 24$$

As we know $ed \equiv 1 \mod \phi(n)$, so

$$d \equiv e^{-1} \mod \phi(n)$$
$$d \equiv 5^{-1} \equiv 5 \mod 24$$

According to the RSA decryption algorithm, $M \equiv C^d \equiv \mod n$.

Therefore,

$$M \equiv 8^5 \equiv 8 \mod 35$$

2. Because the message space is $\{0,1\}^{160}$, the length of M^5 is around 800 bits, which is less than 1024 bits. Therefore, from $C \equiv M^5 \bmod n$, we have $C = M^5$ without the module. Then, given $C = M^5$, it is easy to recover $M = C^{1/5}$. Therefore, we can recover the message M from C without using the brute force.

Q9.

1. (4,4) Since,

$$C_1 = \alpha^k = 2^2 = 4 \mod 11 = 4$$
$$C_2 = M \cdot Y_B^k = 9 \cdot 3^2 = 4 \mod 11 = 4$$

2. We can use coding method, or padding technique, or the modified ElGaml encryption discussed in our class, so that a message 0 can be converted one in the message space of ElGamal encryption.

Q10.

Let $m_1 = 3, m_2 = 5, m_3 = 7, a_1 = 1, a_2 = 3, a_3 = 5$. We have $M = m_1 \cdot m_2 \cdot m_3 = 3 \times 5 \times 7 = 105$. $M_1 = M/m_1 = 35, M_2 = M/m_2 = 21, M_3 = M/m_3 = 15$.

$\alpha_1 = M_1^{-1} \mod m_1 = 35^{-1} \mod 3 = 2, \alpha_2 = M_2^{-1} \mod m_2 = 21^{-1} \mod 5 = 1, \alpha_3 = M_3^{-1} \mod m_3 = 15^{-1} \mod 7 = 1$.

Therefore,

$$x = a_1 \cdot \alpha_1 \cdot M_1 + a_2 \cdot \alpha_2 \cdot M_2 + a_3 \cdot \alpha_3 \cdot M_3 = 1 \times 2 \times 35 + 3 \times 1 \times 21 + 5 \times 1 \times 15 \mod M = 103$$

Q11.

1. Proof. Because $gcd(9, q) = 1$, the inverse $9^{-1} \bmod q$ exists, and $gcd(10, q) = 1$, from the Fermat's Little Theorem, we have

$$\underbrace{11 \cdots 1}_{q-1 \;\; 1's} = \frac{10^{q-1} - 1}{9} \Rightarrow 9^{-1}(10^{q-1} - 1) \equiv 0 \bmod q$$

As a result, we have

$$q | \underbrace{11 \cdots 1}_{q-1 \;\; 1's}$$

2. Proof. Since $2 + 1 \equiv 0 \mod 3$, we have $2 \equiv -1 \mod 3$. Then,

$$2^{2i-1} \equiv (-1)^{2i-1} \mod 3$$

Because, $2i - 1$ is odd, we have

$$2^{2i-1} - (-1)^{2i-1} \equiv 0 \mod 3 \Longrightarrow 2^{2i-1} + 1 \equiv 0 \mod 3 \Longrightarrow 3|2^{2i-1} + 1$$

.

For $i = 1, 2, \cdots, x$, we have

$$\sum_{i=1}^{x}(2^{2i-1} + 1) \equiv 0 \mod 3 \Leftrightarrow x + \sum_{i=1}^{x} 2^{2i-1} \equiv 0 \mod 3$$

Because $Y = x + \sum_{i=1}^{x} 2^{2i-1}$, we have

$$Y \equiv 0 \bmod 3 \Rightarrow 3|Y.$$