ASSIGNMENT-1

Varisha Khanna
20168067
IT

1Q) What do you understand by privacy? Explain privacy preserving.

Ans: Privacy is the ability of an individual or group to seclude themselves or information about themselves, and thereby express them selectively. When something is private to a person, it usually means that something is inherently special or sensitive to them. The domain of privacy partially overlaps with security, which can include the concepts of appropriate use, as well as protection of information.

The privacy of an individual and confidentiality of data has recieved many contributions from many field such as computer science, statistics, economics as well as social sciences.

with the current rate of growth in this area, it describes research in this area of privacy preserving, data publishing. We are mainly concerned with data such as goverment agencies, hospitals, insuarance companies and other buisnesses that help data they would like to release to analysts, researchers and anyone else who wants to use the data.

Privacy preservation in data mining is an important concept, because when the data is transferred or communicated between different parties then its compulsory to provide security to the data so that other parties do not know what data is communicated between

original parties. Preserving in data mining means hiding output knowledge of data mining by using several methods when this output data is valuable and private. Mainly two techniques are used for this, one is input privacy in which data is manipulated by using different techniques and other one is the output privacy in which data is altered in order to hide the rules.

2Q) What are the different attack models of privacy preserving?

Ans: 1) Active attacks → An active attack attempts to alters system resources or effect their operations. Active attack involve some modification of the data stream or creation of false statement. Types of active attacks are: Masquerade, modification of messages, repudiation, replay and denial of service.

2) Passive attacks → A passive attack attemps to learn or make use of information from the system but does not affect system resources. Passive attacks are in the nature of eavesdropping on or monitoring of transmission. The goal of the opponent is to obtain information being transmitted. Types of passive attacks are as following: The release of message content and traffic analysis.

3Q) Why is it essential to secure your data from un authorized users?

Ans: Data is becoming more and more valuable. Also

skills and opportunities for retrieving different types of personal data are evolving extremely fast. Unauthorized careless or ignorant processing of personal data can cause great harm to persons and companies.

Firstly, the purpose of personal data protection isn't to just protect person's data, but to protect the fundamental rights and freedoms of persons that are related to that data.

Secondly, not complying with the personal data protection regulations can lead to even harsher situations, where it's possible to extract all the money from a person's bank account or even cause a life threatening situation by manipulating health information.

Thirdly, data protection regulations are necessary for ensuring fair and consumer friendly commerce and provision of services. Personal data protection regulations causes a situation, where, for example, personal data can't be sold freely which means that people have a greater control over who make them offers and what kind of offers they make.

If personal data is leaked, it can cause companies significant damage to their reputation and also bring along penalties, which is why it's important to comply with the person data protection regulations.

4@> Explain different methods for privacy preserving.

Ans: Many privacy preserving techniques were developed, but most of them are based on anonymization of data. The following are a few techniques:

1) K anonymity:

Anonymization is the process of modifying data before it is given for data analytics, so that the identification is not possible and will lead to k indistinguishable records if an attempt is made to be identify by mapping the anonymized data with external data sources. K anonymity is prone to two attacks namely homogeneity attack and background knowledge attack. Some of the algorithms applied include, Incognito, Mondrian to ensure anonymization

2> L-diversity:

To address homogeneity attack technique called Ldiversity has been proposed. As per Ldiversity there must be L well represented values for the sensitive attribute in each equivalence class.

Implementing L diversity is not possible every time because of the variety of data. L diversity is also prone to skewness attack. When overall distribution of data is skewed into few equivalence classes attribute disclosure cannot be ensured.

3>T closeness:

Another improvement to L diversity is T closeness measure where an equivalence class is considered to have 'T Closeness' if the distance between the distributions of sensitive attribute in the class is no more than a threshold and all equivalence classes have T Closeness. T closeness can be calculated on every attribute with respect to sensitive attribute

4) Randomization technique:

Randomization is the process of adding noise to the data which is generally done by probability distribution. Randomization is applied in surveys, sentiment analysis etc. Randomization does not need knowledge of other records in the data. It can be applied during data collection and pre processing time. There is no anonymization overhead in randomization. However, applying randomization on large datasets is not possible because of time complexity and data utility which has been proved.

5) Data Distribution technique:

In this technique, the data is distributed accross many sites. Distribution of data can be done in two ways

a) horizontal distribution of data
b) vertical distribution of data

5) Multidimensional sensitivity based anonymization (MDSBA):

Multidimensional sensitivity based anonymization is an improved anonymization technique such that it can be applied on large data sets with reduced loss of information and predefined quasi identifiers.

As part of this technique Apache MAP REDUCE framework has been used to handle large data sets.

Multidimensional sensitivity based anonymization makes use of bottom generalization but on a set of attributes with certain class values where class represents sensitive attributes. Data distribution was made effectively when compared to conventional method of blocks. Data anonymization was done using four quasi identifiers using Apache Pig.

5Q)

As part of this technique, APACHE MAP REDUCE framework has been used to handle large datasets. This technique makes use of bottom generalization but on a set of attributes with certain class values where class represents sensitive attributes. Data distribution was made effectively when compared to conventional method of blocks. Data anonymization was done using four quasi-identifiers using Apache Pig.

Q5) What are the types of privacy models? Explain with examples.

Three types of privacy models all commonly considered when anonymizing data.

(1) Member Disclosure: It means that data linkage allows an attacker to determine whether or not data about an individual is contained in a dataset. While this deals with implicit sensitive attributes, other disclosure models deals with explicit sensitive attacks.

② Attribute Disclosure : This may be achieved even without linking an individual to a specific item in a dataset. It protects sensitive attributes, which are attributes from the dataset with which individuals are not willing to be linked with. As an example, linkage to set data entries allows infering information if all items share a certain sensitive attribute value.

③ Identity Disclosure : It means that an individual can be linked to a specific data entry. This is a serious type of attack, as it has legal consequences for data owners according to many laws and regulations worldwide. From the definition, it also follows that an attacker can learn all sensitive attributes in the data entry about the individual.