Y/D/TY

Reg. No:

VIT

Vellore Institute of Technology

Final Assessment Test – April 2025

Course: **BCSE318L - Data Privacy**

Class NBR(s): **1565/1576/1580/1604/1607/1719** Slot: **C1+TC1**

Time: **Three Hours** Max. Marks: **100**

- **KEEPING MOBILE PHONE/ANY ELECTRONIC GADGETS, EVEN IN 'OFF' POSITION IS TREATED AS EXAM MALPRACTICE**
- **DON'T WRITE ANYTHING ON THE QUESTION PAPER**

Answer ALL Questions

(10 X 10 = 100 Marks)

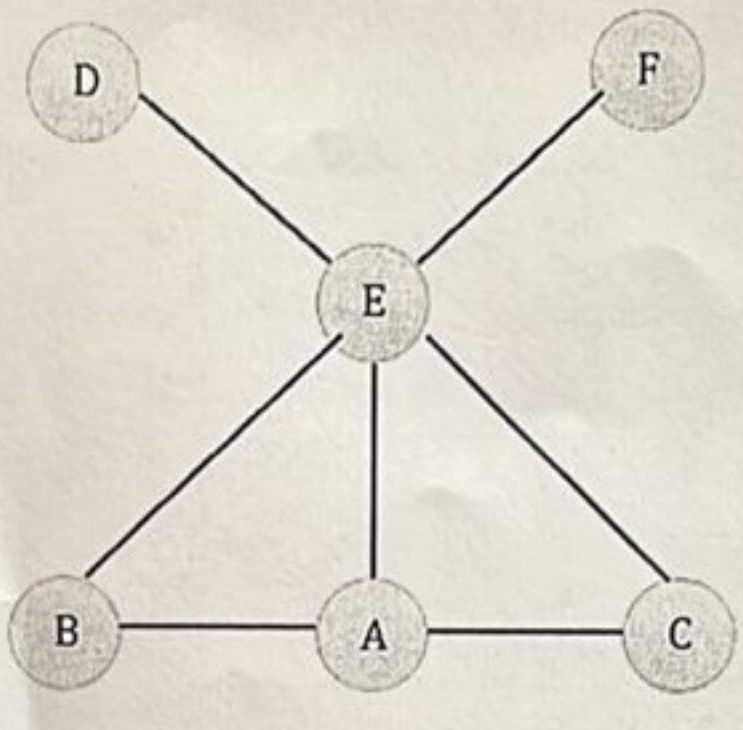
1. Modify the below dataset by removing personally identifiable information (PII), generalizing quasi-identifiers (QI), and retaining sensitive data (SD). Compare two anonymization designs: (1) generalizing zip codes and keeping income unchanged; (2) Suppressing gender and zip codes completely while keeping income and disease data intact. Analyse the impact of the two anonymization approach by evaluating privacy gain and utility loss. Discuss the trade-offs between privacy and utility in each design and determine which approach provides better privacy protection while ensuring useful data for analysis.

Name	Zip Code	Gender	Income	Disease
Alice	56001	Female	30K	Diabetes
Bob	56015	Male	45K	Hypertension
Carol	56001	Female	28K	Asthma
David	56011	Male	35K	Cancer

2. Consider your own dataset containing Age, Zipcode, Location, Illness, how would you apply non-perturbative masking to anonymize both categorical (eg. Location) and continuous (eg. Age, zipcode) attributes? Which non-perturbative techniques would be more effective for categorical data, and why might some methods fail to properly anonymize continuous variables? Provide examples to support your explanation.

3. An organization's multidimensional employee salary database contains sensitive numerical data (SD) that must be protected to prevent re-identification risks. How can organization implement a privacy framework that adjusts anonymization techniques based on the sensitivity level of salary data and potential risks of exposure?

4. In a social network, the given graph represents people and their direct friendships. Calculate the degree, closeness, and betweenness centrality of each people (node), and specify which people would be the best choice for spreading information efficiently.



5. What are the potential threats and vulnerabilities in an enterprise's anonymization design, and how do threat attacks differ between anonymizing graph data and time-series data? Which type of data poses the highest risk for re-identification, and why?

6. What are the key components involved in an independent tokenization implementation, and how do they contribute to securing sensitive data? Additionally, illustrate how does real-time tokenization protect sensitive customer data while enabling outsourced BPO operations to access and process transactions without exposing the original information?

7. Analyse a time series dataset of smart meter readings that periodically monitor household energy consumption. Explain the challenges in ensuring data privacy and explore various anonymization techniques for protecting the data.

8. What are the different test phases in software testing, and how can organizations ensure privacy-preserving test data in cases where anonymization is inadequate?

9.a) Consider the below original dataset. Assume a noise is drawn from a natural distribution N (0, 5000). Apply Additive noise masking perturbative technique using the above noise to anonymize the dataset. Finally measure the information loss using Mean Squared Error (MSE) and Mean Absolute Error (MAE).

ID	Salary ($)
1	50,228
2	65,600
3	86,075
4	56,670
5	76,063
6	87,946
7	76,651
8	76,424
9	86,297
10	71,400

OR

9.b) Apply the 3- anonymity to the below dataset. And using the Classification Metric analyse the utility of the anonymized data with respect to the original dataset to show the incorrect classification due to anonymization.

ID	Age	Zip Code	Disease
1	25	56001	Diabetes
2	27	56001	Hypertension
3	29	56001	Cancer
4	34	56015	Hypertension
5	36	56015	Hypertension
6	38	56015	Hypertension
7	45	56020	Cancer
8	48	56020	Cancer
9	50	56020	Cancer
10	53	56020	Asthma

10.a) Compare the security implications of dependent tokenization and independent tokenization when applied to SSN IDs for the given dataset. For dependent tokenization, use the hash function:

$H(SSN)=\sum(\text{Each digit} \times \text{Its position}) \bmod 1000$

Discuss how dependent tokenization, which derives tokens based on the original SSN, differs from independent tokenization, where tokens have no statistical relationship to the original data. Analyse the strengths and weaknesses of each method in the context of dynamic data protection.

CID	SSN
1	204858085
2	353848259
3	381392365
4	637172699
5	665984208

OR

10.b) A healthcare provider is preparing to share a patient dataset with a medical research institute while ensuring HIPAA compliance. The dataset includes explicit identifiers (EIs), quasi-identifiers (QIs), and sensitive data (SD) such as medical conditions. How can the healthcare provider anonymize the dataset to comply with HIPAA regulations while preserving the utility of data for medical research?

⇔⇔⇔ Y/D/TY ⇔⇔⇔