


VIT

Vellore Institute of Technology
(Declared to be University under section 3 of the UGC Act, 1956)
Final Assessment Test – April 2026

Course: BCSE318L - Data Privacy

Class NBR(s): 2063/2228/5869

Time: Three Hours

Slot: D1+TD1

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

COs	CO Statements
CO1	Characterize basic rules, principles for protecting privacy and personally identifiable information.
CO2	Formulate data that supports useful statistical inference while minimizing the disclosure of sensitive information.
CO3	Identify the list of threats on the various types of anonymized data.
CO4	Classify and analyse the methods of test data generation with Privacy and utility.

BL – Blooms Taxonomy Level (1 – Remember, 2 – Understand, 3 – Apply, 4 – Analyse, 5 – Evaluate, 6 – Create)

Answer ALL Questions

(10 X 10 = 100 Marks)

- Organizations often need to share customer data across departments or with external partners for analysis and business operations, but such sharing must not reveal an individual's identity or personal information. In this context, explain the difference between privacy and anonymity, and discuss the need for sharing sensitive data in organizations along with the major enterprise use cases for data sharing while ensuring the protection of Personally Identifiable Information. CO1 BL2
- Microdata files produced from surveys or administrative processes often contain detailed information about individuals or organizations. Before releasing such data to external users, it is necessary to follow a statistical disclosure control process to ensure confidentiality protection. Explain the concept of microdata and describe the stages involved in the statistical disclosure control process for releasing microdata, including the assessment of the need for confidentiality protection. CO1 BL2
- Consider the following **microdata dataset** containing two attributes. CO2 BL5

ID	Age	Income (₹)	Medical Expense (₹)
1	25	30000	2200
2	30	38000	2500
3	35	42000	2700
4	40	46000	3300
5	45	52000	3100

- Compute the correlation matrix for the attributes *Age* and *Income*.
- Apply perturbative methods and non-perturbative methods separately to the given dataset, modify the dataset accordingly, and explain the changes introduced by each method.
- Discuss how these methods may affect the correlation between the attributes and the overall data utility.

4. Using a suitable example dataset, explain why finding the optimal k-anonymization with minimum information loss is considered an NP-hard problem.

Age	Gender	ZIP Code	Salary (₹)
20–29	Male	6201*	30,000
20–29	Male	6201*	30,000
20–29	Male	6201*	30,000
30–39	Female	6202*	40,000
30–39	Female	6202*	45,000
30–39	Female	6202*	50,000

Using the anonymized dataset given above, identify the drawback observed in the dataset. Then modify the dataset so that it satisfies 2-diversity for the sensitive attribute Salary, and determine whether the first equivalence class satisfies t-closeness by computing the Earth Mover's Distance between the salary distribution of the equivalence class and the overall dataset.

5. Given a high-dimensional financial time series dataset, explain how background knowledge and pattern preservation can lead to privacy breaches. Evaluate at least two privacy-preserving techniques that could mitigate these risks and compare their effectiveness.

6. Enterprises release anonymized datasets for analysis, but adversaries—internal or external—can use background and external knowledge to re-identify individuals. Considering the differences in adversary location, knowledge, and access, propose a comprehensive anonymization strategy for an enterprise that balances data utility and privacy. Explain how your strategy addresses:

- Internal versus external adversaries.
- Background versus external knowledge exploitation.
- Geographic location of data processing.

7. An e-commerce company processes thousands of online transactions daily and stores customer credit card details for payment processing. To enhance data security and comply with regulations, the company decides to implement tokenization.

Based on this scenario, answer the following:

- Explain the concept of tokenization and how it works in dynamic data protection.
- Illustrate the tokenization process with a suitable example.
- Discuss different types of tokenization techniques.
- Analyse the advantages and limitations of tokenization in securing sensitive data.

CO2 BL4

CO3 BL5

CO3 BL4

CO4 BL3

8. A company uses the following anonymized dataset derived from real production data for software testing:

CO4 BL4

User_ID	Age Group	Location	Transaction Amount (₹)	Device Type	Purchase Category
U501	20–30	City X	5000	Mobile	Electronics
U502	30–40	City Y	7000	Desktop	Clothing
U503	20–30	City X	5000	Tablet	Electronics
U504	40–50	City Z	2000	Mobile	Groceries
U505	30–40	City Y	7000	Desktop	Clothing

A company uses the above anonymized production dataset for software testing, but the test results are found to be inaccurate and unreliable. Analyse why anonymized data may be insufficient for effective testing, and discuss its impact on data realism and system validation, particularly in terms of preserving real-world patterns, edge cases, and functional correctness.

- 9.a) An anonymized social network graph consists of two tightly connected communities linked by a few edges:

CO3 BL3

Nodes: {A, B, C, D, E, F, G, H}

Edges: A → B, B → C, C → A (Community 1)

D → E, E → F, F → D (Community 2)

C → G, F → G, G → H

Calculate the degree centrality for all nodes, compute the closeness centrality for nodes C, F, and G, and evaluate the betweenness centrality of node G by analysing shortest paths between the two communities. Identify the node responsible for inter-community communication and discuss how such nodes are highly vulnerable to re-identification attacks.

OR

- 9.b) A ride-sharing company releases anonymized trip data (time, location, fare). Based on the data, find:

CO3 BL5

- Analyse possible privacy threats (location tracking, re-identification).
- Explain how movement patterns can reveal user identity.
- Suggest technical and legal safeguards to prevent misuse.

- 10.a) A healthcare organization handles sensitive patient data (medical records, IDs, billing info). Propose a dynamic data protection strategy using tokenization. Also compare it with encryption, masking, and anonymization. Evaluate the overall benefits.

CO4 BL5

OR

- 10.b) "Privacy regulations are essential but difficult to implement in global systems." Discuss by explaining the importance of regulations like GDPR and HIPAA in protecting sensitive data and ensuring user trust, analysing the challenges organizations face, examining the legal penalties and financial losses resulting from non-compliance, and suggesting practical strategies to improve compliance.

CO4 BL5

Ca/D/TZ