



VIT
Vellore Institute of Technology
(RANKED 1st IN THE COUNTRY BY THE NIRF 2021)

Final Assessment Test – April 2025

Course: BCSE318L - Data Privacy

Class NBR(s): 1569/1573/1584/1588/1605/1609/1720/
6630

Slot: C2+TC2

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. A university publishes an anonymized dataset containing details of student exam scores, age, gender, and department. The dataset is intended for transparency and research purposes but does not include names as below.

Student ID	Age	Gender	Department	Final Exam Score
S101	19	Male	Computer Science	95
S102	21	Female	Mathematics	80
S103	22	Male	Computer Science	30
S104	20	Female	Physics	60

Explain identity disclosure, attribute disclosure and inference disclosure in the context of data privacy. Identify the possible disclosures present in the given dataset and explain their impact. Also apply appropriate privacy preserving technique to mitigate it.

2. A hospital wants to release anonymized patient records for medical research while ensuring patient privacy. The dataset includes Patient ID, Age, City, Disease, Treatment Cost, and Disease Severity is given below.

Patient ID	Age	City	Disease	Treatment Cost (₹)	Severity
P001	27	New Delhi	Diabetes	18,000	Mild
P002	35	Mumbai	Hypertension	55,000	Moderate
P003	49	Chennai	Cancer	6,00,000	Severe
P004	53	Kerala	Heart Disease	8,00,000	Critical
P005	67	Bangalore	Alzheimer's	12,00,000	Critical

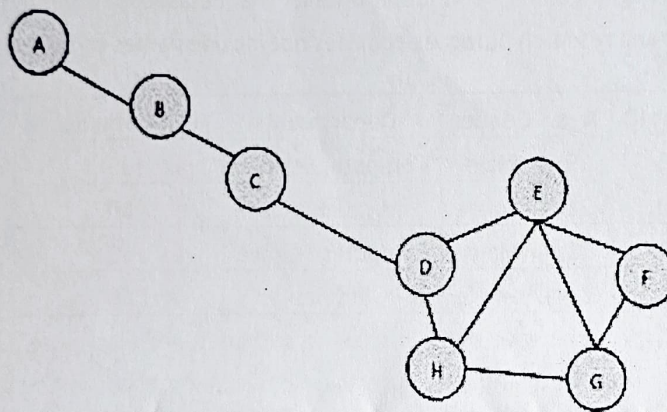
Apply Recoding, Top-Bottom Coding, and Local Suppression techniques to anonymize the above dataset. Explain each technique in detail.

3. Consider the patient dataset contains quasi-identifiers such as Age, Location and ZIP Code. To prevent identity disclosure, the hospital prefers Samarati's Algorithm for k-anonymity, ensuring at least 2-anonymity by generalizing the data (ie) each record should be indistinguishable from at least one other record based on quasi-identifiers. Show the algorithm implementation using the lattice structure.

Patient ID	Age	ZIP Code	Location	Disease
P1	29	530012	Bangalore	Diabetes
P2	31	530014	Bangalore	Flu
P3	42	540020	Bangalore	Cancer
P4	45	540022	Bangalore	Flu
P5	55	550030	Bangalore	Hypertension
P6	57	550032	Bangalore	Diabetes

4. a) Consider the following graph.

[5]



- Calculate the degree centrality of nodes in the graph.
 - Calculate the betweenness centrality of node 'D' in the given graph
- b) Outline the various identity protection privacy model in Graph data structure with suitable example.

[5]

5. A smart city wants to predict traffic congestion patterns and dynamically adjust traffic lights based on real-time vehicle density. So, the city government records vehicle flow data for congestion analysis as shown in the table below.

Timestamp	Location	Vehicle Count	Avg Speed (km/h)	Traffic Signal State	Predicted Congestion Level
2025-08-01 08:00	Junction A	200	40	Green	Low
2025-08-01 08:05	Junction A	500	20	Red	High
2025-08-01 08:10	Junction B	700	15	Yellow	Critical
2025-08-01 08:15	Junction C	250	35	Green	Moderate

Identify the various risk involved in this time series original data and implement the relevant anonymization techniques to protect the privacy.

6. A banking organization wants to enhance the security of its customers' credit card numbers while processing transactions. Instead of storing raw credit card numbers in databases, the bank decides to implement tokenization to protect sensitive data.
- Explain how tokenization can be applied in this scenario to ensure security.
 - Differentiate tokenization from encryption and explain which method is more suitable for this use case.
 - What are the potential challenges of implementing tokenization in real-time payment processing systems?
7. How does the concept of the bathtub curve apply to software reliability? What are its three primary phases, and how do they impact system failure rates? Explain the kind of data essential for each test phase with diagram.
8. Classify the various challenges in protecting Quasi-Identifiers (QIs) in data anonymization. Explain each challenge with relevant examples, highlighting how they impact privacy preservation and data utility.
- 9.a) A retail company has developed a machine learning model to predict monthly sales revenue based on past sales trends, seasonal effects, promotions, and market conditions. The company wants to evaluate the accuracy of the model. Apply Mean Absolute Error (MAE) and Mean Squared Error (MSE) for the given dataset to measure the accuracy of the model.

Month	Actual Revenue (Rs. in Lakhs)	Predicted Revenue (Rs. in Lakhs)
JAN	120	125
FEB	130	135
MAR	145	140
APR	160	155
MAY	180	172
JUN	200	190
JUL	210	215
AUG	230	220
SEP	220	225
OCT	190	195
NOV	170	165
DEC	150	155

OR

- 9.b) A hospital releases the following anonymized multidimensional dataset for research purposes, ensuring k-anonymity by generalizing quasi-identifiers such as Age and ZIP Code.

Age	ZIP Code	Disease
30-40	53000*	Diabetes
30-40	53000*	Diabetes
30-40	53000*	Diabetes
40-50	53000*	Cancer
40-50	53000*	Cancer
40-50	53000*	Cancer

Identify the possible attack in this dataset and apply appropriate technique to prevent the attack. Also outline all possible attack types in multidimensional dataset.

- 10.a) Demonstrate the key use cases for one-way and two-way tokenization with examples. In which scenarios is each method preferred? Also explain the various components of tokenization with neat diagram.

OR

- 10.b) A Swiss financial institution stores its customer data on servers located in a country that is not recognized by Switzerland as having adequate data protection laws. A data subject files a complaint, claiming that their personal information is at risk. How does the FADP regulate cross-border data transfers, and what conditions must be met for compliance?

⇔⇔⇔ BH/D/TY ⇔⇔⇔