VIT
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING
CONTINUOUS ASSESSMENT TEST – II
WINTER SEMESTER 2024-2025

---

1. A healthcare organisation wants to share patient data with researchers while ensuring privacy compliance. The dataset is as follows:

Patient_ID	Name	DOB	Condition	Medication	Visit_Date	Location	Doctor	Bill_Amount
P001	John Doe	1985-06-12	Hypertension	Amlodipine	2025-01-15	New York	Dr. Smith	250.75
P002	Jane Smith	1990-09-23	Diabetes Type 2	Metformin	2025-01-16	Los Angeles	Dr. Adams	1200
P003	Michael Brown	1978-04-05	Asthma	Albuterol	2025-01-17	Bombay	Dr. Ram	89.99
P004	Sarah Johnson	2001-11-30	Migraine	Sumatriptan	2025-01-18	Delhi	Dr. Karthik	300.25
P005	David Wilson	1995-07-19	Fever	Atorvastatin	2025-01-19	Kolkata	Dr. Ram	560.4
P006	Emily Davis	1982-02-14	Diabetes Type 1	Naproxen	2025-01-20	Seattle	Dr. Garcia	47.85
P007	Robert Martinez	1971-12-10	Heart Disease	Lisinopril	2025-01-21	Dallas	Dr. Johnson	980.65

To protect patient identities, the organisation implements **k-anonymisation** before releasing the data. Show the implementation of Samarati's algorithm for k=2. Analyse the risk and then suggest a technique for improving data protection. In each case, step-by-step implementation of the anonymisation process is necessary.

ANS:
Samarati's algorithm is a **k-anonymity** technique that generalizes or suppresses data to ensure that each row is indistinguishable from at least **k-1** others. Since you specified **k=2**, we will generalize certain attributes to ensure at least two identical records in key attributes.

Step 1: Extracted Data (showing the original table)

Step 2: Apply Samarati's Algorithm (k=2)

Generalization Strategy: Students must explain the generalization process for each attributes.

1. **Suppress Patient_ID and Name** (Replacing with "PXXX" and general labels like "Patient A").

2. **Generalize DOB** (Keep only the decade, e.g., "1980s").
3. **Generalize Location** (Convert to region level).
4. **Suppress Doctor Names** (Replace with "Dr. X" to ensure anonymity).
5. **Condition Generalization** (Group similar conditions).

Patient_ID	Name	DOB	Condition	Medication	Visit_Date	Location	Doctor	Bill_Amount
PXXX	Patient A	1980s	Chronic Disease	Amlodipine	2025-01-15	East Coast	Dr. X	250.75
PXXX	Patient B	1990s	Chronic Disease	Metformin	2025-01-16	West Coast	Dr. X	1200
PXXX	Patient C	1970s	Respiratory	Albuterol	2025-01-17	India	Dr. X	89.99
PXXX	Patient D	2000s	Neurological	Sumatriptan	2025-01-18	India	Dr. X	300.25
PXXX	Patient E	1990s	General Illness	Atorvastatin	2025-01-19	India	Dr. X	560.4
PXXX	Patient F	1980s	Chronic Disease	Naproxen	2025-01-20	West Coast	Dr. X	47.85
PXXX	Patient G	1970s	Cardiovascular	Lisinopril	2025-01-21	South	Dr. X	980.65

Step 3: Privacy Analysis

k-Anonymity Achieved (k=2):

- Each DOB appears at least twice (grouped by decades).
- Location is generalized to broader regions.
- Doctor names are masked.
- Identifiable information (Patient ID, Name) is suppressed.

Apply ℓ-Diversity to Ensure Condition Diversity

ℓ-Diversity ensures that within each group, there is **enough diversity in sensitive attributes** (e.g., medical conditions), preventing adversaries from linking data to a single individual.

Refinement Strategy:

1. **Ensure Condition Diversity:**
    - Each group (based on DOB or Location) should have **at least two different conditions**.
2. **Maintain k-Anonymity (k=2)** while improving diversity.

Final Privacy Enhancements

k-Anonymity (k=2):

- Each row is indistinguishable from at least one other.

ℓ-Diversity (ℓ=2):

- **Each group has at least two distinct conditions**, reducing inference risk.

Added Bill Amount Ranges:

- Prevents exact price-based re-identification.

Masked Identifiers (Names, Patient_IDs, Doctors, Specific Locations)

2. Consider the following table and generate a graph from this.

User_ID	Age_Group	Role	Friends	Relationship_Type
U001	20-30	Student	U002, U003	Friendship
U002	20-30	Student	U001, U004	Friendship
U003	30-40	Engineer	U001, U005	Colleague
U004	40-50	Manager	U002, U006	Colleague
U005	30-40	Engineer	U003, U007	Friendship
U006	50-60	Retired	U004, U007	Family
U007	50-60	Retired	U005, U006	Family

It represents a **social network graph**, where:

- **Nodes** contain **age group & role**, removing sensitive personal details.
- **Edges** define relationships (**Friendship, Colleague, Family**).
- **Preserves Graph Structure** for network analysis while ensuring privacy.

Now, implement the k-generalization on node and edges and then represent privacy enriched graph structure.

ANS:

Step 1: Understanding k-Generalization

k-Generalization ensures privacy by **grouping similar attributes** to prevent identification of individuals while preserving the graph structure.

- **Nodes (Users) Generalization:**
    - **Age Group** → Broadened categories (e.g., merge "20-30" and "30-40" into "20-40").
    - **Role** → Generalized into broader categories ("Engineer" and "Manager" → "Working Professional").
- **Edges (Relationships) Generalization:**

- Merge relationship types into broader categories:
    - **Friendship & Colleague** → "Social"
    - **Family** → "Family" (remains unchanged)

Step 2: Implement k-Generalization

Step 3: Output and Interpretation

Privacy-Enriched Graph Adjustments:

1. **Nodes are generalized**:
    - **Age Groups are merged** (e.g., "20-30" and "30-40" → "20-40").
    - **Roles are generalized** (e.g., "Engineer" & "Manager" → "Working Professional").
2. **Edges (Relationships) are generalized**:
    - "Friendship" and "Colleague" → **"Social"**.
    - "Family" remains unchanged.

Visualization Results:

- Nodes show **generalized age groups and roles**.
- Edges display **broad relationship categories** to **protect privacy**.

<u>Student needs to represent the graphical structure of original and anonymized data.</u>

3. Consider the following healthcare time series data:

Date	Time	Patient ID	Heart Rate (bpm)	Systolic BP (mmHg)	Diastolic BP (mmHg)
2024-03-01	08:00:00	P001	72	120	80
2024-03-01	08:00:00	P002	75	122	82
2024-03-01	08:00:00	P003	78	118	78
2024-03-02	12:00:00	P001	70	119	79
2024-03-02	12:00:00	P002	74	121	81
2024-03-02	12:00:00	P003	77	117	77
2024-03-03	16:00:00	P001	73	118	78
2024-03-03	16:00:00	P002	76	120	80
2024-03-03	16:00:00	P003	79	119	79
2024-03-04	08:00:00	P001	72	121	81
2024-03-04	08:00:00	P002	75	123	83
2024-03-04	08:00:00	P003	77	118	78
2024-03-05	12:00:00	P001	71	120	80
2024-03-05	12:00:00	P002	74	122	82
2024-03-05	12:00:00	P003	78	119	79

Analyse the various risk involves in this original data and implement the relevant anonymization techniques for this time series data to protect the privacy.

ANS.

Risk Analysis of the Healthcare Time-Series Data

The dataset contains sensitive health information for multiple patients, including:

1. **Personally Identifiable Information (PII)** – The **Patient ID** could be linked back to individuals.
2. **Longitudinal Health Data** – The repeated heart rate and blood pressure readings over time could reveal **medical conditions** or trends.
3. **Re-identification Risk** – If combined with external data (e.g., timestamps and health metrics), attackers could **re-identify** patients.
4. **Regulatory Compliance** – **HIPAA, GDPR, and other data protection laws** require de-identification before sharing or storing healthcare data.

Anonymization Methods

1. Remove Direct Identifiers

- **Patient ID should be replaced** with a pseudonymized value or completely removed.
- If linking is needed, use a **random hash** instead of real patient IDs.

2. Generalization of Timestamps

- Convert exact timestamps (e.g., `"2024-03-01 08:00"`) into **broader time windows** (e.g., `"Morning of Day 1"`).
- This prevents fine-grained tracking of an individual's routine.

3. Data Perturbation (Noise Addition)

- Slightly modify **heart rate and blood pressure** values by adding **random noise** (±2-5%) to prevent exact value matching.

4. K-Anonymity for Grouping Patients

- Instead of individual records, **aggregate** data by grouping patients into categories (e.g., age groups, risk levels).

Anonymized Output :

Date	Time	Anon_ID	Heart Rate (bpm)	Systolic BP (mmHg)	Diastolic BP (mmHg)
2024-03-01	Morning	a1b2c3d4	73	121	79
2024-03-01	Afternoon	d4e5f6g7	76	124	83
2024-03-02	Morning	h7i8j9k0	71	118	78
2024-03-02	Afternoon	l1m2n3o4	75	122	80

4. Consider the following anonymized dataset. Identify the type of the dataset and also find the anonymization methods applied for this dataset. Analyse the risk of the released dataset and list all the possible threats for this data.

Date	Time	Age Group	Heart Rate (bpm)	Systolic BP (mmHg)	Diastolic BP (mmHg)	Diagnosis
2024-03-01	Morning	30-40	72	120	80	Hypertension
2024-03-01	Afternoon	30-40	75	122	82	Hypertension
2024-03-02	Morning	50-60	70	130	85	Diabetes
2024-03-02	Afternoon	50-60	74	128	83	Diabetes
2024-03-03	Morning	30-40	78	118	78	Hypertension
2024-03-03	Afternoon	30-40	77	117	77	Hypertension
2024-03-04	Morning	60-70	73	140	90	Heart Disease
2024-03-04	Afternoon	60-70	76	138	88	Heart Disease
2024-03-05	Morning	50-60	72	125	80	Diabetes
2024-03-05	Afternoon	50-60	74	124	82	Diabetes

Type of Dataset

This dataset represents **anonymized health data** as it contains **medical parameters** such as heart rate, blood pressure, and diagnosis. It belongs to the **structured healthcare dataset** category and is **time-series data**, as it records measurements at different times on different days.

Anonymization Methods Applied

The dataset has undergone **partial anonymization** using the following techniques:

1. **Generalization**:
    - The **Age** of individuals is provided in **age groups** (e.g., 30-40, 50-60) instead of exact values.
    - The **time of data recording** is generalized to "Morning" and "Afternoon" instead of exact timestamps.
2. **Removal of Identifiers**:
    - Personally Identifiable Information (PII) such as names, patient IDs, addresses, and social security numbers is removed.
3. **Data Masking (Indirect)**:
    - Although indirect identifiers like **exact dates** are still included, they are less specific due to the generalization of other attributes.

Risk Analysis & Possible Threats

Despite the anonymization efforts, **re-identification risks still exist** due to:

1. **Linkage Attack**

- If an attacker has access to an external dataset (e.g., hospital admission records, insurance claims) that includes exact dates of visits along with personal identifiers, they may **link** the diagnosis and health readings to individuals.

2. **Singling Out Attack**
    - Some **age groups** (e.g., 60-70) may have **few individuals** in a small community, making it easier to **single out** a person based on known medical conditions.
3. **Inference Attack**
    - By analyzing patterns in diagnosis and health readings over time, an attacker might infer **personal health conditions** even without knowing specific names.
4. **Attribute Disclosure**
    - If an attacker knows that a person falls within a particular **age group** and attended a medical facility on a given date, they may infer their **health condition** (e.g., Hypertension, Diabetes).

Mitigation Strategies

To enhance privacy, the following additional anonymization techniques can be applied:

- **Stronger Generalization**: Broaden age ranges further (e.g., 20-40 instead of 30-40).
- **Aggregation**: Instead of providing daily records, aggregate the data weekly or monthly.
- **Suppression**: Hide data points that might reveal specific patterns (e.g., outliers in smaller age groups).
- **Perturbation**: Slightly modify numerical values (e.g., ±5 bpm for heart rate) to prevent exact matches.

5. Consider the following dataset:

Timestamp	Age Group	Anon_ID	Transaction Type	Amount ($)	Location
2024-03-01 14:23:45	18-25	a1b2c3d4	Deposit	215.32	New York
2024-03-01 09:12:30	18-25	a1b2c3d4	Online Purchase	187.45	Los Angeles
2024-03-02 18:45:12	26-35	d4e5f6g7	ATM Withdrawal	199.85	Chicago
2024-03-03 12:32:10	36-50	h7i8j9k0	Bank Transfer	230.5	Houston
2024-03-04 16:20:55	51-65	l1m2n3o4	Withdrawal	180.75	Miami

Now, implement and analyse the tokenization for this dataset.

Output of Tokenization:

Transaction Type	Transaction_Tokens	Location	Location_Tokens
Deposit	["Deposit"]	New York	["New", "York"]
Online Purchase	["Online", "Purchase"]	Los Angeles	["Los", "Angeles"]
ATM Withdrawal	["ATM", "Withdrawal"]	Chicago	["Chicago"]

Transaction Type	Transaction_Tokens	Location	Location_Tokens
Bank Transfer	["Bank", "Transfer"]	Houston	["Houston"]
Withdrawal	["Withdrawal"]	Miami	["Miami"]

Student needs to discuss the benefit of tokenization.