DT 306 Privacy in the Digital Age
Term 1, 2022-23
Assignment 3

Due: Presentations in class: Nov 24, Nov 29 and Dec 1, 2022
Grade weightage: 20% of course marks

In this assignment, you will work on sample data sets and use (or develop) software tools that help anonymize the data sets and analyze the privacy risks in the datasets..

Each team will identify an appropriate publicly available dataset, and analyze the dataset using available tools for the specified anonymization technique. The operations on the dataset can include:
:

1. Anonymize the dataset (if not anonymized)
2. Evaluate the extent of anonymization (e.g the “k” in k-anonymity). For DP, measures such as privacy budget can be used
3. Evaluate the trade-off between utility and privacy/re-identification risk.
4. Evaluate the above for different choices of quasi-identifiers, if relevant.

Datasets:
Identify a publicly available dataset for the domain assigned to you. You can modify the dataset if needed, or add synthetic data (e.g. adding a PII). The dataset need not be large, but should be sufficient to try out different scenarios of analysis. (Try to avoid very large datasets - stay well below 1GB, for instance)

Tools/Techniques:
There are a number of open source or free tools available (especially if the dataset is not too large). You can use any of them or try using more than one of them. You are also welcome to develop your own set of tools.

Teams:
The anonymization technique and data domain for each team is listed in the table below:

Anonymization Technique	**Data Domain**				
	Education	Finance	Healthcare	Transportation	Social Media
k-anonymity	1	5	14	17	20
l-diversity	9	2	6	15	18

t-closeness	12	10	3	7	16
Differential Privacy	19	13	11	4	8

Presentation:

Each team should make a brief presentation (10 mins) in class on their study. This should include:

1. Description of the data set - source, context, key attributes/identifiers
2. Tools used in the anonymization and/or analysis
3. Output/results from the analysis. Interesting and informative visualizations would be encouraged
4. A demo of the anonymization process - key steps. This could also be a pre-recorded video.

The presentation should be uploaded to LMS after the presentation.