Privacidade de Dados - 2023

Trabalho 2 - k-Anonimato

Javam Machado

1 Objetivo:

O trabalho consiste em implementar um algoritmo que anonimize um conjunto de dados contra ataques de ligação ao registro atendendo o k-anonimato. Deverá ser implementação o modelo k-anonimato por meio da **generalização** de valores de atributos como descrito no artigo *L. Sweeney. k-anonymity: a model for protecting privacy. Int. Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570.* Voce deve ler o artigo, escolher uma técnica de construção das classes de equivalência para fazer a anonimização e seguir as instruções abaixo para aderir ao k-anonimato na geração dos datasets anonimizados.

2 Especificação:

Considere o conjunto de dados "Artists-Pseudo-02.csv". Voce deve recuperá-lo por meio do link:

`https://drive.google.com/file/d/1-RF5WDZoSL3VZQ9_UhVL73YYfZDEH4Pv/view?usp=sharing`

Este dataset contém os atributos ConstituintID, DisplayName, Region, Gender, BeginDate, Wiki QID e Income.

Carregue o conjunto de dados "Artists-Pseudo-02.csv" e faça uma limpeza no dataset para manter os seguintes atributos:

- *Semi-identificadores:* Region, BeginDate;
- *Sensíveis:* Income ($).

Você vai aplicar a técnica de generalização para gerar datasets anonimizados que atendem o k-anonimato. O valor de k deve variar no conjunto $k = \{2, 4, 8\}$. Para cada valor de k, o conjunto de dados deve ser anonimizado de forma a atender o modelo k-anonimato. Essa anonimização deve ser feita por generalização – hierarquia de três níveis no atributo *BeginDate* (Ano, Década, Século) e hierarquia de quatro níveis no atributo *Region* (Cidade, País, Sub-Região e Região). Para cada configuração, deve ser gerado um csv anonimizado com o nome "kAnonArtists.csv". Além disso, o programa deve ter uma opção no menu que recebe como entrada os níveis de generalização para cada atributo (*BeginDate* e *Region*) e mostra na tela os valores possíveis para o nível selecionado.

Procure também plotar um histograma das classes de equivalência.

Para medir a utilidade do processo de anonimização, calcule a precisão e o tamanho médio das classes de equivalência, ambas as métricas para cada um dos datasets gerados.

3 Requisitos

- Linguagens: C++ ou Python
- Duplas: as mesmas do Trabalho I
- Preparar uma Demo para explicar, mostrar o seu programa e os resultados durante a aula de entrega. Escreva um Readme.txt descrevendo o projeto.
- Zipar o seu projeto (código fonte e executável), os datasets anonimizados, os gráficos e o Readme.txt em um único pacote e submeter via **Classroom**.
- O trabalho deverá ser entregue até as 16h da segunda-feira, dia 02/10/2023 e explicado durante a aula do dia 02/10, seguindo a mesma sequência de apresentação das duplas do Trabalho I.

4 Avaliação

Na avaliação serão considerados os seguintes indicadores:

- **Corretude** do programa;
- **Precisão** pela comparação do dataset original com o dataset anonimizado;
- Clareza na **explicação** do programa durante a Demo;
- **Pontualidade** e **documentação/qualidade** do código-fonte.