KFUPM

College of Computer Science and Engineering
Computer Engineering Department
COE 526: Data Privacy
Fall 2020 (201)

Assignment 1: Due date Tuesday 6/10/2020

Objectives

The objectives of this assignment is the following

1. Conduct data linkage attack
2. Implement a k-anonymization algorithm and analyze the tradeoff between privacy and utility, and
3. Understand the difference between k-anonymization and l-diversity

Dataset description

The dataset used in this assignment is the IPUMS data extracted from the 2001 US Census. The dataset has 8 attributes as described in Table 1. The size of the dataset is 20,000 tuples (rows). All attributes include numerical values only. For example, Gender attributes can be either 1 or 2, which represents Male and Female, respectively. The Income attribute is the annual income in thousand USD, for example, an income of 20 means 20,000 (20K) annually.

Age	Gender	Marital	Race status	Birth place	Language	Occupation	Income (K)

Table 1: Scheme of Census dataset

Tasks

Task1: Linkage attack (20 pts)

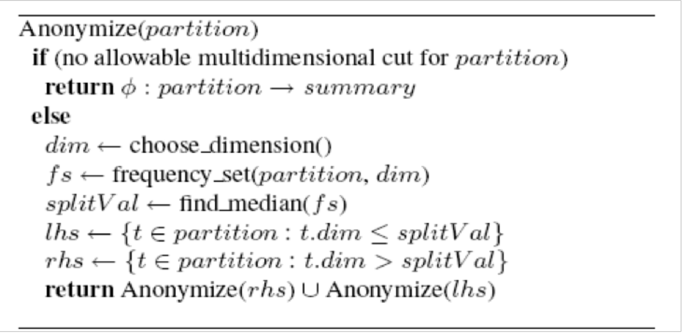
Download the file named ”ipums.txt” from blackboard and unzip it. Using Table 2 as external background information, perform a data linkage attack to find the annual salary of each person in the table. You are free to use any tool/programming language to complete this task, e.g., Excel, Python, Java, etc.

Name	Age	Birth place
Ahmed	28	110
Fatma	44	4
Ali	17	199
Abeer	34	260
Muhamad	40	15

Table 2: Background table

Task2: K-anonymization Implementation (40 pts)

Implement the greedy partitioning algorithm that was discussed in the class using your preferred programming language. The sensitive attribute is Income. The remaining attributes are Quasi-Identifiers. The steps of the algorithm is shown in Figure 1. Please read below for instructions on how to find and select the mean value.



```
Anonymize($partition$)
  if (no allowable multidimensional cut for $partition$)
    return $\phi : partition \rightarrow summary$
  else
    $dim \leftarrow$ choose_dimension()
    $fs \leftarrow$ frequency_set($partition, dim$)
    $splitVal \leftarrow$ find_median($fs$)
    $lhs \leftarrow \{t \in partition : t.dim \leq splitVal\}$
    $rhs \leftarrow \{t \in partition : t.dim > splitVal\}$
    return Anonymize($rhs$) $\cup$ Anonymize($lhs$)
```

Figure 1: Greedy Partitioning algorithm

How to find the median value from frequency set?[1]

(a) Number of records (n) is odd: the median is the value at the position $\frac{n+1}{2}$ of the sorted list of values.

(b) Number of records (n) is even:

i. Find the value at position $\frac{n}{2}$

ii. Find the value at position $\frac{n}{2}+1$

iii. The median is either the value at position $\frac{n}{2}$ or $\frac{n}{2}+1$

the median is the value at the position $\frac{n}{2}$ of the sorted list of values.

(Note: if you couldn't implement the algorithm in Task2, you may use available k-anonymization tool (e.g., [1] and [2]) to conduct Task3 and 4.)

Task3: Utility-privacy trade off (20 pts)

Using your implementation of the anonymization algorithm in Task2, find the anonymized table with k=3,5,7, and 9.

For each anonymized table compute the Discernibility metric C_{DM} and generalized information loss $ILOSS$ given by the following equations.

$$C_{DM} = \sum_{E \in EC} |E|^2 \tag{1}$$

$$ILOSS = \frac{1}{|T| \cdot n} \sum_{i=1}^{n} \sum_{j=1}^{|T|} \frac{U_{ij} - L_{ij}}{U_i - L_i} \tag{2}$$

where E is an equivelance class of the set of all equivalence classes EC, $|E|$ is the size of the equivlance class E, $|T|$ is the size of the table, n is the number of attributes, U_ij and L_ij are the upper and lower values of the i^{th} attribute in the j^{th} record, respectively, and U_i and L_i is the upper and lower values of the i^{th} attribute, respectively.

Then, draw a figure for each metric against the value of k to depict the privacy trade off. The x-axis should be the value of

[1] https://www.youtube.com/watch?v=t2BSuUXfftA

k, while the y-axis should be the value of the respective utility metric.

Task4: **ℓ-diversity** (20 pts)
Using the anonymized table with $k = 3$ from Task 9, check if the 9-anonymized table is distinct ℓ-diverse for each $\ell = 2$ and 5. In the case when the 9-anonymized table violates the ℓ-diversity requirements, print at least one equivalence class that violates the diversity requirement.

Submission

The due date of this assignment is 11:59PM 6/10/2020. Please upload all files on the assignment page on BlackBoard. You need to submit the following:

1. A report containing your response to tasks 1,3, and 4.

2. The source code of the implementation of the greedy partitioning algorithm.

References

[1] ARX - Data anonymization tool. https://arx.deidentifier.org/.

[2] Python implemntation for Mondrian. https://github.com/qiyuangong/Mondrian.