KFUPM
College of Computer Science and Engineering
Computer Engineering Department
COE 449: Privacy Enhancing Technologies

Fall 2019 (191)

Assignment 1: Due date Thursday 3/10/2019

Objectives

The objectives of this assignment is the following

1. Conduct various type of attacks on a public dataset
2. Implement a k-anonymization algorithm, in particular, Mondrian multidimensional
3. Analyze the tradeoff between privacy and utility, and
4. Understand the difference between k-anonymization, l-diversity, and t-closeness

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 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 (30 pts)

Implement Mondrian, the k-anonymization algorithm that we discussed in the class using your preferred programming language. The sensitive attribute is Income, while the remaining attributes are QIs. The steps of the algorithm is shown in Figure 1.

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 $\frac{n}{2}$

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

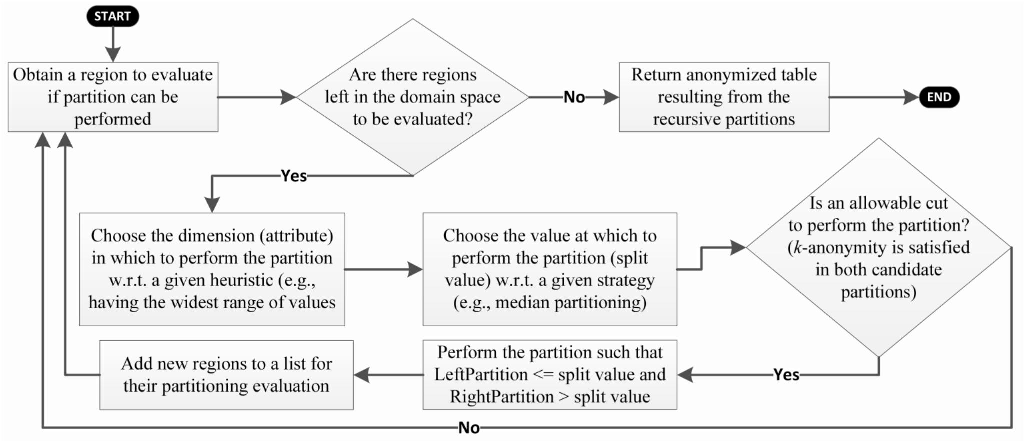
START
Obtain a region to evaluate if partition can be performed
Are there regions left in the domain space to be evaluated?
No
Return anonymized table resulting from the recursive partitions
END
Yes
Choose the dimension (attribute) in which to perform the partition w.r.t. a given heuristic (e.g., having the widest range of values
Choose the value at which to perform the partition (split value) w.r.t. a given strategy (e.g., median partitioning)
Is an allowable cut to perform the partition? (k-anonymity is satisfied in both candidate partitions)
Yes
Perform the partition such that LeftPartition <= split value and RightPartition > split value
Add new regions to a list for their partitioning evaluation
No

Figure 1: Mondrian algorithm

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.

Task3: Utility-privacy trade off (30 pts)

Using your implementation of Mondrian algorithm, or any available k-anonymization tool, e.g., [1] and [2], find the anonymized table with k=3,5,7,9, and 11.

For each anonymized table compute the Discernibility metric C_{DM}, the normalized average equivalence class C_{AVG}, and generalized information loss $ILOSS$.

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

$$C_{AVG} = (\frac{|T|}{|EC|})/k \tag{2}$$

$$ILOSS = \frac{1}{|T| \cdot n} \sum_{i=1}^{n} \sum_{j=1}^{|T|} \frac{U_{ij} - L_{ij}}{U_i - L_i} \tag{3}$$

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 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 3, check if the 3-anonymized table is distinct ℓ-diverse and entropy ℓ-diverse for each $\ell = 2$ and 5. In the case when the 3-anonymized table violates the ℓ-diversity requirements, print at least one equivalence class that violates the diversity requirement.

Task5: **(optional) Skewness and similarity attacks and t-closeness(bonus 10 pts)**
Is this dataset susciptable to the skewness and similarity attacks? Will t-closeness improve the privacy of the anonymized table?

Submission

The due date of this assignment is 11:59PM 3/10/2019. 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 Mondrian algorithm.

References

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

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