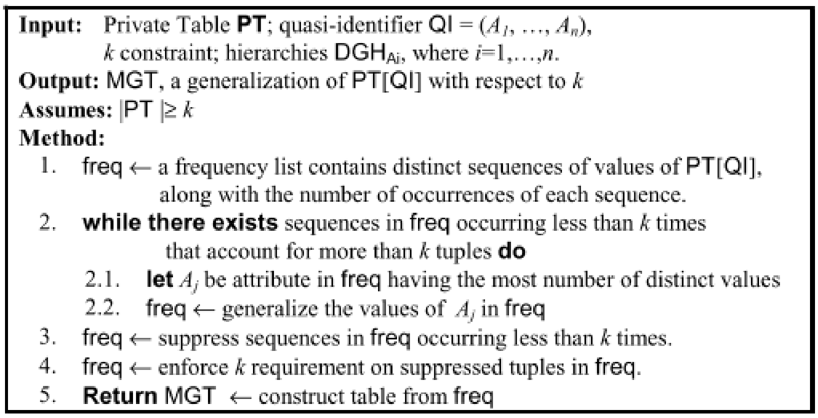
Data Protection & Privacy
First Homework: k-Anonymity

October 12, 2018

Goal

The goal of this homework is to implement the k-anonymity algorithm depicted in fig. 1 which is based on the approach presented in [1]. Feel free to implement the algorithm in your favourite programming language.



```
Input:   Private Table PT; quasi-identifier QI = ($A_1$, …, $A_n$),
         $k$ constraint; hierarchies $DGH_{A_i}$, where $i$=1,…,$n$.
Output: MGT, a generalization of PT[QI] with respect to $k$
Assumes: |PT |≥ $k$
Method:
  1.   freq ← a frequency list contains distinct sequences of values of PT[QI],
              along with the number of occurrences of each sequence.
  2.   while there exists sequences in freq occurring less than $k$ times
              that account for more than $k$ tuples do
       2.1.  let $A_j$ be attribute in freq having the most number of distinct values
       2.2.  freq ← generalize the values of $A_j$ in freq
  3.   freq ← suppress sequences in freq occurring less than $k$ times.
  4.   freq ← enforce $k$ requirement on suppressed tuples in freq.
  5.   Return MGT ← construct table from freq
```

Figure 1: Datafly algorithm

Where:

- PT : Table to anonymize.
- DGH : Domain generalization hierarchies.
- MGT: Generalization of PT that satisfies k-anonymity.

Dataset

In the homework folder there are 3 databases: `db_10000.csv`, `db_50000.csv` and `db_100000.csv` which contain 10K, 50K and 100K records, respectively. The attributes of each database are:

1) `id` that is the EI.

2) `age`, `city of birth` and `zip code` that are QI.

3) `disease` that is the SD.

All the generalizations of the QI are also provided in the folder :`age_generalization.csv`, `city_generalization.csv` and `zip_code_generalization.csv`.
Such files provide a domain generalization hierarchy in the following format: `level_0_generalization, level_1_generalization, level_2_generalization ....`
For example, for the `zip code` the following domain generalization is provided: 67002, 6700*, 670**, 67***, 6****.

Output

You are required to:

1. implement the algorithm depicted in fig. 1;
2. test your implementation by executing the algorithm for several values of k on all three databases.
3. for each test, count the number of equivalence classes/clusters which have homogeneity.
4. evaluate the relationship between the value of k and the number of homogeneous equivalence classes.

Contacts

Davide Caputo
Location (Valletta Puggia): Finsec Lab 320, 3rd floor, Via Dodecaneso 35, Genova.
Email: dave.caputo93@gmail.com

References

[1] L. SWEENEY, "ACHIEVING k-ANONYMITY PRIVACY PROTECTION USING GENERALIZATION AND SUPPRESSION," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 571–588, 2002. [Online]. Available: http://www.worldscientific.com/doi/abs/10.1142/S021848850200165X