Privacy Definitions and Metrics

Exercises (Week 4), 🕒 3-4 hours

k-anonymity

The goal of this exercise is to develop and implement anonymization techniques such that (sanitized) data can be shared with privacy guarantees, while remaining useful.

We will focus on health-related data and provide an anonymization technique for sharing a medical database such that the data can be useful for answering public health questions, but the identities of the patients cannot be reliably determined using side information.

You are provided with two (real-life looking) datasets: (1) "`hospital_records.txt`" - public hospital records of 400 patients living in the canton Vaud from age 20 to 79, including the hospital identifier (e.g., pseudonym) of the patient, three quasi-identifiers (gender, age, and zip code of the patient), and diagnosis of the patient for the following diseases: HIV, flu, diabetes, and cardiovascular disease (note that some patients may not have been diagnosed with any of these diseases, and this is represented as "`---`" in the dataset). And, (2) "`voters.txt`" – a subset of public voter registration forms, including the real identifiers (e.g., names) of 420 individuals along with the same set of quasi-identifiers.

The hospital records look as below (they are taken from the dataset we provide for you, but not necessarily in the same order). Each entry represents the tuple (hospital id, gender, age, zip code, diagnosed disease) and the fields are separated with commas.

```
300336,male,65,1170,Cardiovascular Disease
333301,female,76,1170,---
946042,female,64,1000,HIV+
979053,male,72,1162,Diabetes
658395,female,36,1182,Diabetes
```

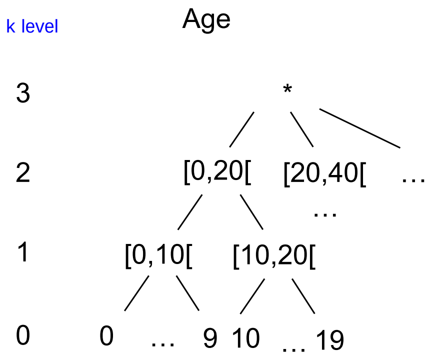
The voter dataset looks as below. Each entry represents (voter name, gender, age, zip code).

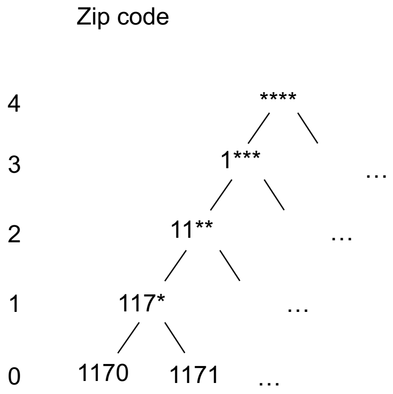
```
Raphael Kleiner,male,65,1170
Oceane Joos,female,36,1182
Audrey Lobsinger,female,36,1182
```

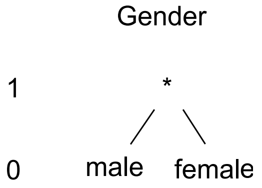
The diagnosis of the patients is privacy-sensitive information, and ideally it should not be associated with the real identity of any patient. For this reason, hospitals release their records by replacing the real identities of the patients with a pseudonym (i.e., after anonymizing the real identities of the patients). However, using these two datasets, it is possible to re-identify the anonymized patients in the hospital dataset.

In order to successfully re-identify a patient in the hospital dataset, there should be exactly one entry that has the same combination of quasi-identifier values. For instance, in the above toy examples, patient “300336" can be re-identified as “Raphael Kleiner", but patient “658395" is not successfully re-identified because there are two matching voters.

You will then produce generalized versions of the “`hospital_records.txt`” file, to be released, which prevent re-identification of the patients by ensuring k-anonymity for $k \geq 13$. To this end, you will use the following value generalization hierarchies (VGH).







The two dataset files, as well as a Python template file (named `k_anonymity_template.py`) are available on Moodle.

Question 1

Using these publicly available datasets, write a Python script to compute the identifiability of the patients in the hospital dataset (i.e., the proportion of the patients that you are able to re-identify). To this end, try to re-identify as many of the patients in the hospital dataset as you can. For each person that you have successfully re-identified in the hospital dataset, list the name of the patient and the specific diagnosis of the patient.

It has been shown that the combination of gender, zip code, and date of birth is sufficient to identify 87% of individuals in the United States. How does this compare with your findings?

Hints:

Write a function called “`reidentify_patient`”, which takes as input a hospital record and looks up a match in the voter records. If and only if *exactly one* match is found, it returns the name and diagnosis of that person. Call this function on all the entries of the `hospital_records` file. Finally, compute the identifiability rate relative to the total number of patients.

Question 2

For the given VGH diagrams, how many different generalizations are possible for the `hospital_records` file?

A generalization applies – to the entire dataset – a specified generalization level to each of the quasi-identifiers, as illustrated in the above VGH diagrams. A generalization level of 0 means that the quasi-identifier value stays the same for that record. A generalization level of i means the value of that quasi-identifier in the record is replaced by the value in the VGH tree that is i levels above it. For instance, zip code '1170' generalizes to '11**' for a level of 2 and age 8 generalizes to '*' for a level of 3. Thus, a generalization is defined by a tuple of the specified generalization levels for each of the quasi-identifiers. For example, for the generalization (1,1,1) (i.e., one level up in the VGH of each quasi identifier), the first record would become:

```
[300336, '*', '[60,70[', '117*', 'Cardiovascular disease']
```

Question 3

Write a function called "`generalize_record`" which takes as input a record (one line) from the `hospital_records` dataset file and the desired level of generations for each of the quasi-identifiers and returns a generalized record. We provide partial code for this function, you only need to complete the function "`generalize_zipcode`", which it calls.

For instance, `generalize_record([300336, male, 65, 1170, Cardiovascular disease'], 1, 1, 2)` should return

```
[300336,'*','[60,70[', '11**', 'Cardiovascular disease'].
```

Question 4

Write a function called "`compute_anonymity_level`" which takes as input a list of records and an array of quasi-identifiers and returns its k-anonymity level. For that, you will need to compute the different equivalent classes.

Question 5

Write a function called "`compute_distortion`" which takes as input the specific levels of generalization and the maximum possible levels of generalization for each of the quasi-identifiers and returns the computed distortion as:

$$\frac{1}{N_q}\sum_1^{N_q}\frac{l_q}{\max\_l_q},$$

where N_q represents the number of quasi-identifiers, l_q is the specific level of generalization for quasi-identifier q and $\max\_l_q$ is the maximum possible level of quasi-identifier q. For instance, as illustrated in the above VGH, $\max\_l_{age}$ = 3 and `compute_distortion( [1, 1, 2], [1, 3, 4]))` would return:

$$\frac{1}{3}\left(\frac{1}{1}+\frac{1}{3}+\frac{2}{4}\right) = 0.61.$$

Question 6 [🗃 Homework: deadline 31.10.19, 6pm] 🏆

Use the “generalize_record”, “compute_anonymity_level” and “compute_distortion” functions to generate all possible generalizations of the hospital_records file and return the optimal one. The optimal generalization is the one with the minimum distortion such that its k-anonymity level is ≥ 13.

🗃 Deliverable:

Upload on Moodle, for the optimal generalization that you found: the value of l_gender, l_age, l_zipcode, *distortion*, *k*, namely the levels of generalization for each of the three quasi-identifiers, the computed distortion and the computed k-anonymity for your optimal generalization.

Upload a file named "h2_GR.zip" (GR is the one or two letter code of your group name on Moodle, /e.g. "AA", "F", etc) containing exactly 2 files, as follows.

The text file is named "h2.txt". This file should contain exactly one line, as in the example file provided - the 0 values should be replaced with your obtained ones for the optimal generalization. We suggest you simply copy/paste the corresponding printed line after running the code (if you used our provided print statement, otherwise please match the format).

The Python script used to find your answer.

✓ Correction criteria:

The script does not work **and/or** no generalization satisfying the k-anonymity is found: 0 point

The script works and a solution near optimal is found: 0.5 point

The script works and your generalization is optimal: 1 point

🏆 Challenge:

The group with the best result will get a reward. In the case of a tie, the group that submitted his answer first will be the winner. The challenge is to get to the smallest distortion possible while maintaining a k-anonymity level above 13.

🎁 Optional:

This is an optional question 🎁. Implement a more realistic method to generate generalizations of the hospital_records dataset, presented by LeFevre et al. (pdf available on Moodle). This method relies on a multi-dimensional greedy partitioning technique (see Figure 6 in the paper).

K. LeFevre ; D.J. DeWitt ; R. Ramakrishnan. Mondrian Multidimensional K-Anonymity. International Conference on Data Engineering 2006[1].

[1] https://moodle.unil.ch/mod/resource/view.php?id=687433