LINMA2472-Homework 3

Groupe 21
Cédric Antoine, Romain Graux, Lionel Lamy

October 22, 2022

1 Introduction

> **Info:** All scripts used to create the final databases can be found on GitHub : `https://github.com/RomainGrx/LINMA2472-Homeworks/tree/main/Homework%203`

In this assignment, we will anonymize a database in order to publish it and reveal a minimum of information that could betray the anonimity of the people present in the initial database.

2 Objectives

Before starting this work, we set ourselves some objectives regarding the characteristics to be obtained for our databases. In order to observe the evolution of our usefulness we started by calculating the entropy of the database before modifying it. We obtained an entropy of 10.96. We set ourselves the goal of maintaining at least half of this entropy in order to maintain at least 50% of the usefulness compared to the initial database.

Concerning the anonymity of our database we decided to try to guarantee at least a 2-anonymity and a 2-diversity. We chose k=2 because after observing the initial database we saw that almost every line was different and that if we wanted to get k=3 the utility could drop too much. We chose l=2 because we know that l<=k and so we couldn't choose a l bigger than the chosen k.

Also in the anonymization of the database, we have decided not to use disruptive or swap methods as these methods would distort the truthfulness of the information present in this dataset. We therefore only used suppressing and generalisation methods.

3 Pre-process

First of all we started by generalizing some attributes, in order to reduce the number of unique people in the dataset. We modified the attributes *dob*, *zipcode*, *education*, *children*, *number_vehicles*, *commute_time* and *accommodation*, to make them less unique. However, after that, a large part of the people remained unique anyway.

For this reason we decided to create two different datasets for the two cases. This allowed us to take only the semi-identifiers that we thought were necessary for each case and therefore to reduce the number of attributes per dataset considerably.

4 Pseudonymisation

A first step towards anonymization is to pseudonymize people. To do this, we used the SHA-512 hashing algorithm with as argument the concatenation of the initial values of *id*, the *zipcode*, the *dob* and then we added a random salt of 256 bytes.

This process makes it possible to anonymize the identity of individuals while avoiding collisions and thus preserving distinct observations. In fact, we checked the whole dataset after our pseudonimization process and we have, as expected, a unique id for each person.

5 k-anonymity

5.1 1st database

In order to choose the semi-identifiers that we were going to maintain for the first case we started by establishing a list of attributes that we found indispensable for this case and a list of attributes that we found could be interesting for the case. We retained as indispensable attributes *employment*, *children*, *commute_ time* as these seemed to us to have the most influence on the environment of each person (1st case). As interesting attributes we chose to keep *gender*, *marital status*, *number_ vehicles*, *education* and *accommodation* as they also influence the environment of a person. We therefore chose not to take into account *dob*, *zipcode*, and *ancestry* for this case. Indeed these attributes did not seem to be related to the environment.

Then we implemented a method calculating the entropy and the number of unique rows, of the datasets of all possible combinations of attributes we considered indispensable (always keeping them all) with the attributes we considered interesting. We were thus able to compare the entropy and uniqueness of each case.

After observation of the results we opted for the combination of attributes that included all our indispensable attributes as well as all the interesting attributes except *marital_ status.* Indeed this solution is a good compromise to minimize the number of unique individuals as well as to maximize the entropy. In this case there were only 138 unique rows left (which represents 7% of the dataset) and we obtained an entropy of 6.96 after removing these 138 lines. If we consider the entropy of an optimal 2-anonymous dataset with 2000 rows (1000 classes containing always 2 lines), which is 9.96, our result is not to far away.

This way after keeping the found attributes and after removing the 138 unique rows we obtained a 2-anonymous dataset with only a loss of 36% of entropy.

After that we wondered if it was not possible to reach the 3-anonymity, however when we looked at our 205 classes, we quickly realized that 57 of them contained only 2 persons and that therefore it would be difficult or impossible to reach k=3 without losing a unreasonable amount of usefulness.

5.2 2nd database

Contrary to the first case, in the second case, we first chose the columns before generalizing to obtain a 2-anonymity.

In order to be able to achieve a 2-anonymity, we worked on the columns attribute; as we were keeping only *state* and *date of birth* (*dob*) as quasi-identifiers, we tried to generalise their attributes to achieve the 2-anonymity.

We have therefore been able to guarantee it by giving the range of years of birth over 5 years and the state corresponding to a zipcode because several zipcodes correspond to a single state.

We chose these parameters to allow the people using this dataset to have the most precise information to build the hospitals while respecting as much as possible the anonymity of the people forming this dataset. As a result, we withdrew 144 people to comply with the condition.

6 l-diversity

To come to a 2-anonymity, we have already modified the attributes of the columns; to now have a 2-diversity, we can only check if we have several diseases for the same set of quasi-identifiers. If this is not the case, we can further generalize the attributes by losing a minimum of information (this is of course subjective). Therefore, we can drop rows that do not respect this diversity.

For the first case we had to drop 10 rows (5 classes) and for the second case we had to drop 21 rows.

7 Attacks

7.1 Uniqueness attack

As we have a 2-anonymity for both databases, it will be impossible to have only one answer to a query.

7.2 Homogeneity attack

As we have a 2-diversity for both databases, we are protected against homogeneity attacks.

7.3 Semantic attack

The sensitive attributes sensible to semantics attacks in our databases are cancers. However, there was only one class of two people where all the people in the class (=2) had cancer. We have removed these people from the database so the databases are resistant to semantic attacks.

7.4 Skewness attack

Our databases are subject to skewness attacks. In order to protect them we should make them t-close. However, given the nature and size of our initial database, we don't think this is necessary. Indeed 2000 samples drawn from a large population such as the United States can not really serve as a reference. Moreover if we made such a small database t-close we would lose a lot of information. Indeed as 1/4 of our classes have only 2 people, it is difficult to have a similar distribution of several (>2) diseases in all classes.

7.5 Others attack

An attack that could work is one where the attacker knows almost all the information about a person and knows he has a disease.

He then searches in the first database and finds two people who comply with the conditions and therefore find two different diseases (2-diversity database) and he finds in the other database, two people who comply with the conditions with two different diseases too but with only one disease in common with the other query.

But here, it is unlikely that an attacker would know 100% of the information about that person without knowing their disease (but it is of course possible).

One solution could be to enlarge the range of date of birth but if the range is too large, we might as well drop the values to ensure more l-diversity and k-anonymity.

8 Finally

For the first database, here is a preview

id	education	gender	number_vehicles	accommodation	employment	children	commute_time	disease
71b16769...	High School	female	2	Own	Retired	True	0	Alzheimer's disease
848cebbf...	Bachelor	female	0	Rent	Employed	True	[0-1]	heart disease
1c245b32...	Bachelor	female	0	Rent	Retired	True	0	endometriosis
4212e9f5...	High School	female	0	Rent	Retired	True	0	gastritis
b1962816...	High School	male	1	Own	Retired	True	0	endometriosis

And here is a preview for the second database

id	state	dob	disease
962178310ebab3ca4bf3bc1f844034c5bc56f5fefb098efda013d11d...	(TX) Texas	[1940 - 1944]	multiple sclerosis
399eb8bd888120fc3c63157c05fdc1cd29ce415ae6f70493da6f93b8...	(IA) Iowa	[1965 - 1969]	heart disease
e68f162b121686cc13485585be270e1c6789fd0d0afdac5f916902e3...	(WI) Wisconsin	[1940 - 1944]	endometriosis
075209297829549515c45fa3763485eb493179bc63044e0f005dcf43...	(NE) Nebraska	[1945 - 1949]	gastritis
a3e872772562762d0ea922558f5f1e22716a887abb50387a76fc333f...	(TX) Texas	[1930 - 1934]	endometriosis