UNIVERSITY OF CATANIA
Department of Mathematics and Computer Science
Bachelor of Computer Science Degree Program

Alessandro Resta

Implementation of a pseudonymization attack based on inference and dictionary techniques

Internet Security Project Report

Academic Year 2021 - 2022

Contents

Chapter 1

Introduction

1.1 Background

Pseudonymization is a GDPR-compliant de-identification process that allow data processors and data controllers to lower the risk of a potential data breach and safeguard personal data. More precisely, it consists in masking identifier fields with a **pseudonym**, which is a piece of information associated to an individual that cannot be reverted back to its plain form without any additional information. A re-identification process is also possible, but can only be performed by authorized personnel that can access protected information like, for example, the mapping table that associates each pseudonym to the corresponding identity. This is what distinguish pseudonymization from anonymization, which is an irreversible de-identification operation, meaning that neither data controller nor data processor can identify a particular anonymized individual.

1.2 Objectives

The main objective of this project is to demonstrate how pseudonymization can be violated through a simple, yet effective, attack on datasets about UNICT students. These last are randomly generated through a Python script, and used by the attacker to perform a **discrimination** and **re-identification** attack using **state of the art** techniques. Data contained in the datasets is described in the next chapter. At the end of the attack, it will be shown the achieved results and some best practices to take in order to avoid these drawbacks.

1.3 Hardware

The implementation and execution of the pseudonymization attack, which is the main topic of this project, was run in a desktop computer with an **Intel Core i7-7700 3.6 GHz** and **12 GB** of **Corsair VENGEANCE DDR4 RAM 3600 MHz**.

Chapter 2

Scenario Description

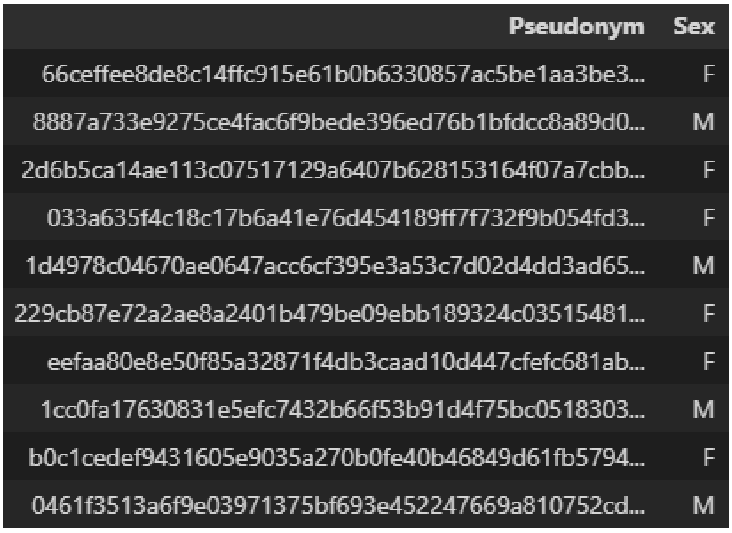
Let us consider that a malicious student named **Eve** wants to know a specific personal information about another student called **Bob**. In particular, Eve is curious to know if Bob obtained the ERSU scholarship for the academic year 2021/2022. Eve knows that ERSU published a ranking of all the participants along with their scholarship holder status, but in a pseudonymized way, so that the real identities are not shown. Instead, a pseudonym is visible, and since Eve is simply a student, he doesn't have the necessary authorizations to access the additional information used to revert back the pseudonyms. Through a simple Google search, Eve finds out that three pseudonymized datasets about UNICT students have recently been published. These datasets are the following:

- **Microsoft Outlook**: contains students pseudonyms associated to their **sex**;
- **Almalaurea**: contains students pseudonyms associated to their **off-site status**;
- **ERSU**: contains students pseudonyms associated to their **scholarship holder status**;

Each of the above datasets includes all UNICT students, so Eve is sure that Bob is present as well. Another public dataset that Eve is aware of is the one from **Internet Security** (IS) project results, that contains **50** pseudonymized students that took the IS module, along with their project result (**passed/not passed**). Let us also assume that Bob attended the IS module and is in the already mentioned IS dataset.

2.1 Pseudonymization Technique

The state of the art pseudonymization technique used in this project to pseudonymize a student identity is the **SHA256** cryptographic hash function, that takes as input the email address of the student, that must be in the format **name.surname@studium.unict.it** (to simplify the attack), and return a fixed-size hexadecimal string. This method is commonly used because is fast and doesn't require additional parameters other then the identifier itself in contrast to, for example, a symmetric-key encryption, that requires a key to be stored somewhere. Below, an extract of Microsoft Outlook dataset (the others have a similar schema):



Pseudonym	Sex
66ceffee8de8c14ffc915e61b0b6330857ac5be1aa3be3...	F
8887a733e9275ce4fac6f9bede396ed76b1bfdcc8a89d0...	M
2d6b5ca14ae113c07517129a6407b628153164f07a7cbb...	F
033a635f4c18c17b6a41e76d454189ff7f732f9b054fd3...	F
1d4978c04670ae0647acc6cf395e3a53c7d02d4dd3ad65...	M
229cb87e72a2ae8a2401b479be09ebb189324c03515481...	F
eefaa80e8e50f85a32871f4db3caad10d447cfefc681ab...	F
1cc0fa17630831e5efc7432b66f53b91d4f75bc0518303...	M
b0c1cedef9431605e9035a270b0fe40b46849d61fb5794...	F
0461f3513a6f9e03971375bf693e452247669a810752cd...	M

Figure 2.1: First 10 entry of Microsoft Outlook dataset

As one can see, sensitive data are protected by the use of pseudonyms, so it is reasonable that this dataset has been made public. Same consideration applies to the other datasets.

2.2 Eve's General Knowledge

Before moving to the next chapter about the attack, three hypothesis have to be made:

- Eve knows that Bob is a **male**;
- Eve knows that Bob **passed** the IS project;
- Eve knows that Bob is an **off-site** student;

The above assumptions are necessary to perform a discrimination attack, and can be seen as part of Eve's general knowledge. General knowledge is a key factor for this kind of attack, the more the attacker knows about the victim, the more the chances of successfully leak his personal data increase.

Chapter 3

Attacker Model and Results

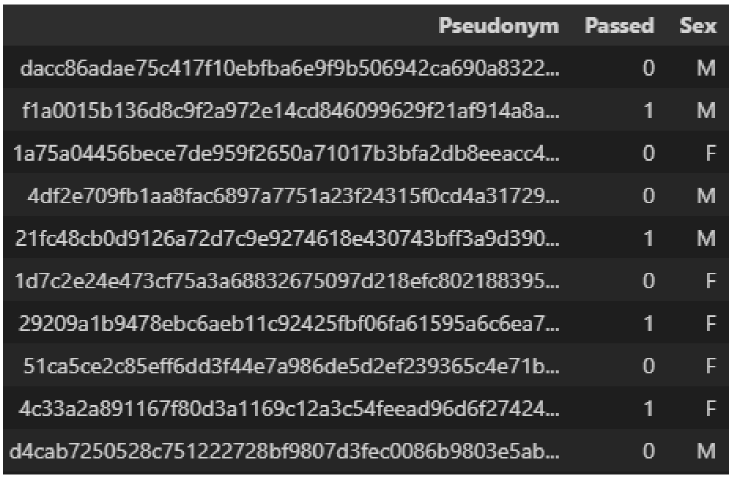
Now that the scenario is ready, Eve can put all the pieces together to access Bob's private information. First, a discrimination attack will reduce the number of candidate students that can be Bob, next a re-identification attack will tell the real identity of each of them (Bob's included). In this case Eve is an external attacker, meaning that he does not have direct access to the pseudonymization secret or other relevant information. However he has access to the four pseudonymized dataset, described in chapter 2, and will use them for the purpose.

3.1 Discrimination Attack

The goal of the discrimination attack is to identify properties of a pseudonym holder (at least one). These properties may not directly lead to uncovering the identity of the pseudonym holder, but may suffice to discriminate him or her in some way. In the proposed scenario, it has been used the **inference** method, which is a state of the art technique, performed by analyzing data in order to illegitimately gain knowledge about a subject or database. A subject's sensitive information can be considered as leaked if an attacker can infer its real value with a **high confidence**.

3.1.1 Implementation of the Discrimination Attack

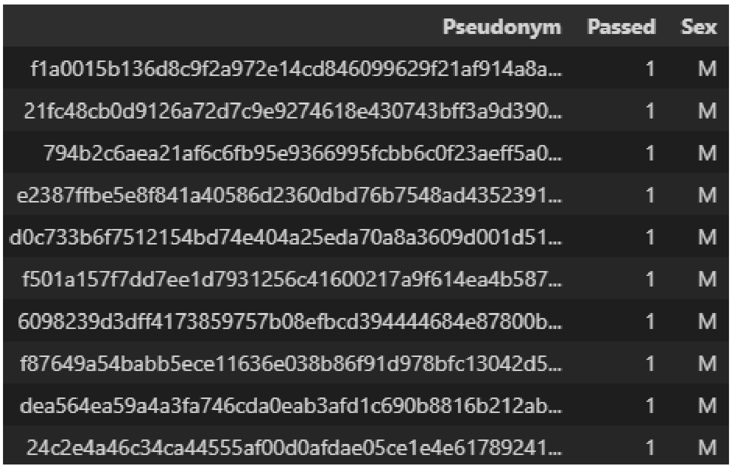
Starting from the dataset containing IS project results, a merge on the pseudonym with the dataset provided by Microsoft Outlook is performed as shown in figure 3.1.



Pseudonym	Passed	Sex
dacc86adae75c417f10ebfba6e9f9b506942ca690a8322...	0	M
f1a0015b136d8c9f2a972e14cd846099629f21af914a8a...	1	M
1a75a04456bece7de959f2650a71017b3bfa2db8eeacc4...	0	F
4df2e709fb1aa8fac6897a7751a23f24315f0cd4a31729...	0	M
21fc48cb0d9126a72d7c9e9274618e430743bff3a9d390...	1	M
1d7c2e24e473cf75a3a68832675097d218efc802188395...	0	F
29209a1b9478ebc6aeb11c92425fbf06fa61595a6c6ea7...	1	F
51ca5ce2c85eff6dd3f44e7a986de5d2ef239365c4e71b...	0	F
4c33a2a891167f80d3a1169c12a3c54feead96d6f27424...	1	F
d4cab7250528c751222728bf9807d3fec0086b9803e5ab...	0	M

Figure 3.1: First 10 row of the merge between IS and Microsoft Outlook datasets

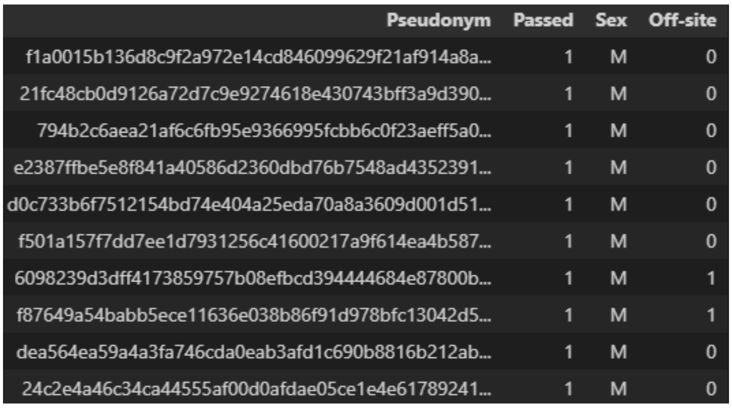
From the first and second hypothesis, Bob is a male and passed the IS project, so female students and students that failed the project have to be filtered out (see figure 3.2).



Pseudonym	Passed	Sex
f1a0015b136d8c9f2a972e14cd846099629f21af914a8a...	1	M
21fc48cb0d9126a72d7c9e9274618e430743bff3a9d390...	1	M
794b2c6aea21af6c6fb95e9366995fcbb6c0f23aeff5a0...	1	M
e2387ffbe5e8f841a40586d2360dbd76b7548ad4352391...	1	M
d0c733b6f7512154bd74e404a25eda70a8a3609d001d51...	1	M
f501a157f7dd7ee1d7931256c41600217a9f614ea4b587...	1	M
6098239d3dff4173859757b08efbcd394444684e87800b...	1	M
f87649a54babb5ece11636e038b86f91d978bfc13042d5...	1	M
dea564ea59a4a3fa746cda0eab3afd1c690b8816b212ab...	1	M
24c2e4a46c34ca44555af00d0afdae05ce1e4e61789241...	1	M

Figure 3.2: Previous dataset showing only the first 10 male students that passed the project

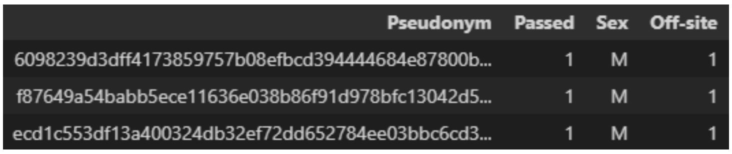
The second merge between the previous dataset and the one from Almalaurea leads to the result shown in figure 3.3



Pseudonym	Passed	Sex	Off-site
f1a0015b136d8c9f2a972e14cd846099629f21af914a8a...	1	M	0
21fc48cb0d9126a72d7c9e9274618e430743bff3a9d390...	1	M	0
794b2c6aea21af6c6fb95e9366995fcbb6c0f23aeff5a0...	1	M	0
e2387ffbe5e8f841a40586d2360dbd76b7548ad4352391...	1	M	0
d0c733b6f7512154bd74e404a25eda70a8a3609d001d51...	1	M	0
f501a157f7dd7ee1d7931256c41600217a9f614ea4b587...	1	M	0
6098239d3dff4173859757b08efbcd394444684e87800b...	1	M	1
f87649a54babb5ece11636e038b86f91d978bfc13042d5...	1	M	1
dea564ea59a4a3fa746cda0eab3afd1c690b8816b212ab...	1	M	0
24c2e4a46c34ca44555af00d0afdae05ce1e4e61789241...	1	M	0

Figure 3.3: First 10 row of the second merge between the previous and Almalaurea datasets

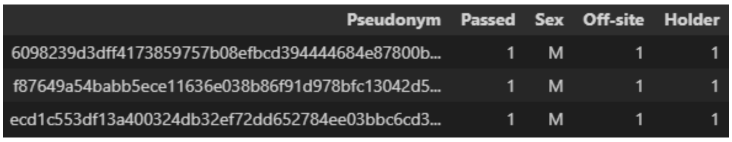
From the third hypothesis, Bob is an off-site student, therefore all not off-site students can be filtered out (figure 3.4).



Pseudonym	Passed	Sex	Off-site
6098239d3dff4173859757b08efbcd394444684e87800b...	1	M	1
f87649a54babb5ece11636e038b86f91d978bfc13042d5...	1	M	1
ecd1c553df13a400324db32ef72dd652784ee03bbc6cd3...	1	M	1

Figure 3.4: Previous dataset showing all male and off-site students that passed the project

It is clear that the number of rows of the starting dataset has been reduced, hence discriminated. Last merge is with the ERSU dataset, shown in figure 3.5.



Pseudonym	Passed	Sex	Off-site	Holder
6098239d3dff4173859757b08efbcd394444684e87800b...	1	M	1	1
f87649a54babb5ece11636e038b86f91d978bfc13042d5...	1	M	1	1
ecd1c553df13a400324db32ef72dd652784ee03bbc6cd3...	1	M	1	1

Figure 3.5: Last merge between the previous and ERSU datasets

Only three students that match Bob's profile left and all of them had the scholarship granted. That being said, it is possible to inference that **Bob obtained the ERSU scholarship**, so his personal information has been leaked by this attack.

3.2 Re-identification Attack

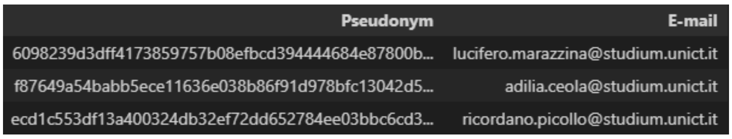
Now that the domain of possible students that match Bob's profile is restricted to only three element, it will be performed a dictionary attack, which is another state of the art technique. A dictionary is composed by a list of couples (pseudonym, identity) that is used to "quickly" **look-up** the real identity for a given pseudonym. Depending on the size of this list, the search can take a long time until it finishes, but it is still an optimization of a brute force attack, where the pseudonyms have to be computed at each look-up. In a dictionary, the pseudonyms value is pre-computed, hence a search can run faster.

3.2.1 Implementation of the Re-identification Attack

In chapter 2, it was pointed out that students email address has to be in the format **name.surname@studium.unict.it**. Having different formats other than this one increase the size of the dictionary, and a following look-up may take too long with the hardware used in this project, so this is the reason why it was specified earlier. Moreover, this is a commonly used convention in universities so, in a real scenario, an attack that is based on this format could easily allow some pseudonyms to be reverted back. The dictionary is built up from two lists: Italian name and Italian surname. These lists are available on Github and contain about **9000** Italian names and **21000** surnames respectively. For each name and surname an email address is generated, for example:

- name = **Alice**
- surname = **Pennisi**
- generated email = **alice.pennisi@studium.unict.it**

Before storing the generated email in a file, the corresponding pseudonym is computed by the application of the SHA256 cryptographic hash function. The pair **(key,value)** is then stored, where key is the hashed value (pseudonym) and value the generated email address (identifier). Once the above operation is done for all the possible combinations of names and surnames, the hash of the three students can be used to look-up their real identifier. Figure 3.6 shows the results.



Pseudonym	E-mail
6098239d3dff4173859757b08efbcd394444684e87800b...	lucifero.marazzina@studium.unict.it
f87649a54babb5ece11636e038b86f91d978bfc13042d5...	adilia.ceola@studium.unict.it
ecd1c553df13a400324db32ef72dd652784ee03bbc6cd3...	ricordano.picollo@studium.unict.it

Figure 3.6: Partial mapping table computed on candidates students

The attacker has now a partial mapping table, that contains the associations between pseudonyms and identifiers of candidates students that can be Bob, so the attack on pseudonymization was a success. Note that one cannot say which of the email address of the above table belongs to Bob, it is just there for sure. The algorithm used to crack the hashes takes **4m 6s** (1m 22s for each look-up), so it was quite fast considering that it looked-up three times a dictionary containing nearly **200 Mln** entries.

Chapter 4

Conclusions and Recommendations

The attacks illustrated in previous chapter may be feasible with high success rates, and this is what organizations should be aware of before sharing pseudonymized personal data about their customers or employees. Best practices to avoid the unpleasant consequences of these kind of attacks are:

- sharing pseudonymized data only to trusted parties and make them private when possible;
- don't use cryptographic hash functions as are subject to brute force and dictionary attacks (as it was seen). Symmetric encryption could be a better choice;
- don't use the same pseudonymization function (with the same key in case of symmetric encryption) on multiple dataset, especially if the intersection between them is not empty;
- make the mapping table available only to few authorized employees and monitor the access logs.

With the above precautions, personal data can be stored and analyzed reducing, significantly, the risk of having them leaked by a malicious attacker.

Bibliography

[1] Enisa: pseudonymization techniques and best practices. https://www.enisa.europa.eu/publications/pseudonymisation-techniques-and-best-practices/@@download/fullReport.

[2] Pseudonymization. https://dataprivacymanager.net/pseudonymization-according-to-the-gdpr/.

[3] Inference attack. https://en.wikipedia.org/wiki/Inference_attack.

[4] Italian names. https://gist.github.com/pdesterlich/2562329.

[5] Italian surnames. https://github.com/PaoloSarti/lista_cognomi_italiani/blob/master/cognomi.txt.