Grant Agreement Number: 825225

Safe-DEED

www.safe-deed.eu

D5.6 Report on the application of re-identification techniques on use-case data

Deliverable number	*D5.6*
Dissemination level	*Public*
Delivery date	*25 November 2019*
Status	*Final*
Author(s)	*Alexandros Bampoulidis*

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n° 825225.

Changes Summary

Date	Author	Summary	Version
08.11.2019	Alexandros Bampoulidis	Ready for review	1.0
14.11.2019	Alexander Georg	Review	1.1
15.11.2019	Alessandro Bruni	Review	1.2
25.11.2019	Alexandros Bampoulidis	Final Version	Final

Executive summary

Task 5.4 de-anonymisation of the Safe-DEED project investigates the de-anonymisation of the use case data, in order to raise privacy red flags. In this deliverable, we describe the work carried out in the first year of the project. Specifically, we present the procedure we carried out, and which can be generalised, on the use case data, in case the data publisher decides to release, exchange, or sell their data. The procedure constists of practical (actual de-anonymisation of individuals) and theoretical (through an analysis tool) de-anonymisation, and the anonymisation measures, beyond removing personally identifying information (PII), that can be taken to reduce the risk of de-anonymisation.

Table of Contents

List of Figures

1 Introduction

Personal data contains information about individuals and is commonly used in the industry and academia for research and innovation purposes. However, the sharing of personal data increases the risk of a privacy breach, posing a threat to the privacy of the individuals whose information is contained in the dataset.

A common misconception is that removing all **personally identifying information (PII)**, such as name, address, etc., makes the data anonymous. Extensive research in re-identification (referred to as **de-anonymisation** in the rest of this deliverable) – the process of identifying individuals in a dataset - has proven this belief wrong. Sweeney [1] showed that 87% of the U.S. population is uniquely identifiable by the combination of their gender, date of birth and ZIP code. Such attributes, whose combination can serve as a unique identifier, are called **quasi-identifiers (QIs)**.

Recital 26 of the General Data Protection Regulation (GDPR) [6] refers to the de-anonymisation of individuals in datasets: "*The principles of data protection should apply to any information concerning an identified or identifiable natural person. Personal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person. To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments...".*

Safe-DEED's task 5.4-de-anonymisation aims in raising privacy red flags in the use case data, and we have designed the task in accordance to Recital 26 of the GDPR. Given the use case data provided by Forthnet (FNET) with no PIIs included, we determine whether the natural persons in the dataset are identifiable, and the costs and amount of time needed to de-anonymise an individual. Specifically, we applied a battery of de-anonymisation tests on the data (Sect. 3), developed a risk analysis tool and applied it to it (Sect. 4), and, additionally, we applied anonymisation measures, beyond removing PII, to the data (Sect. 5). In the next section, we provide a description of the use case data, in Sect. 6 we describe the published research outcomes of this task, and in Sect. 7 we conclude this deliverable by summarising the work carried out in this task and providing the future directions.

2 Use Case Data

The use case of WP6 contains customer relationship management (CRM) data having all PIIs removed, provided by the Greek telecommunications provider Forthnet (FNET). Specifically, the CRM data consists of 3 tables: *Assets*, *Invoices*, and *Support Requests (SRs)*. Tables 1, 2, and 3 provide a description of the columns of those tables.

The table *assets* contains information about the customers' contracts, with each line corresponding to a contract. The table *invoices* contains the monthly invoices sent out to customers, with each line corresponding to a revenue type per month per asset. The table *SRs* contains the support requests customers have made per month, with each line corresponding to a type of request per month per asset.

Column	Description
CUSTOMER_ID	Identifier of a customer (not a PII)
ASSET_ID	Identifier of an asset (contract)

ACTIVATION_DATE	The activation date of the contract
DEACTIVATION_DATE	The deactivation date of the contract
ASSET_STATUS_ID	Binary indicator of whether the contract is still active
INITIATION_CHANNEL	The channel from which a contract was initiated e.g. Forthnet store, call centre, retailer, etc.
INITIATION_DEALER_ID	Identifier of the contract initiator e.g. specific Forthnet store, call centre, retailer, etc.
PORTABILITY	Binary indicator of whether the customer kept his/her phone number from the previous provider
LOOP_TYPE	Binary indicator of whether the customer has another currently active contract
ASSET_STATUS_REASON	The reason why a contract was terminated e.g. no longer needed, non-payer, etc.
ASSET_STATUS_REASON_DESCR	How the contract termination was done e.g. online form, e-mail, termination of services, etc.
PROVIDER_DEST	The customer's previous telecom provider
PROVIDER_SOURCE	The customer's telecom provider after the contract termination

Table 1: Assets table

Column	Description
MONTH	Month and year of invoice
CUSTOMER_ID	Identifier of a customer (not a PII)
ASSET_ID	Identifier of the asset (contract) in the *assets* table
DATE_ISSUED	The exact date the invoice was issued to the customer
REVENUE_TYPE	Kind of revenue (service usage) e.g. monthly fee, mobile/international calls, etc.
REVENUE	Amount in € for the respective REVENUE_TYPE

Table 2: Invoices table

Column	Description
MONTH	Month and year of service requests
CUSTOMER_ID	Identifier of a customer (not a PII)
ASSET_ID	Identifier of the asset (contract) in the *assets* table
CONTACT_TYPE	Type of request e.g. technical problem, service upgrade, complaints, etc.
CONTACTS	How many times the customer made the respective CONTACT_TYPE
RESOLUTION DAYS	How many days it took to resolve the customer's issues of the respective CONTACT_TYPE

Table 3: Support Requests (SRs) table

3 De-Anonymisation Attacks

At month 6 of the project (May 2019), an employee of Research Studios Austria (RSA) worked at the premises of FNET for one week in order to get access to the data and perform de-anonymisation attacks – actually trying to identify individuals in the dataset – with support from FNET and LSTech (LST). The purpose of this procedure is to raise privacy red flags, and it is an indicator of the effort, cost and likelihood of de-anonymising any individual using publicly available information found on the web.

The procedure was carried out and can be generalised into 3 steps :

1) Gathering external information: In this step, we spent time looking for sources of information that could provide information about FNET's customers that could be matched to the information in the datasets described in Sect. 2. Those, namely, were FNET's Youtube channel, Facebook and Twitter social media accounts, and a tech forum where FNET's customers ask for support and FNET officially provides it.

2) Processing the data : At FNET's premises, FNET provided ~1.25 million lines of the table *assets*, invoices from October 2018 to March 2019 from ~570.000 customers and support requests made by ~438.000 customers from October 2018 to April 2019. The data needed cleaning and preprocessing, such as deleting corrupt lines and converting Greek characters to Latin characters. After preprocessing, the data tables were inserted into an SQLite database.

3) Constructing and executing queries: In this step, the information gathered in step 1 needs to be structured and formulated as an SQL query, in order to de-anonymise the individual in question. The more information gathered, the easier it is to de-anonymise an individual. Since this information is unstructured (free text), a considerable effort might be required to express it in an SQL language.

Due to confidentiality reasons, further details cannot be given about this procedure, however, FNET became aware of the privacy red flags in their dataset, how a de-anonymisation attack would be performed, and how much effort would be required for it, in the context of the currently publicly available information.

4 De-Anonymisation Risk Analysis

Complementary to the procedure described in Sect. 3, we developed a de-anonymisability analysis tool that raises privacy red flags, through an analysis of a dataset's QIs, and helps in identifying the QIs that are critical in de-anonymisation. Specifically, it outputs the percentage of records that are at risk for every unique combination of QIs, which can be visualised in an interactive plot.

Figure 1 depicts a snapshot of such an interactive plot produced for an example dataset containing 11 QIs. Each point represents a unique combination of QIs – 2047 unique combinations for 11 QIs – and, when moused over, indicates the percentage of individuals that are uniquely identifiable by the respective combination of QIs, i.e. the probability of de-anonymising any individual with those QIs. In this example, it can be seen that attributes 4, 5 and 8 are critical in de-anonymisation, since 80% of the individuals are uniquely identifiable by their combination, while the combination of all QIs uniquely identifies 86% of the individuals.

The tool was applied to FNET's data and revealed the extent to which the CRM dataset is de-anonymisable, assuming an attacker possesses all the QIs' information, and indicated which QIs distinguish the individuals the most.

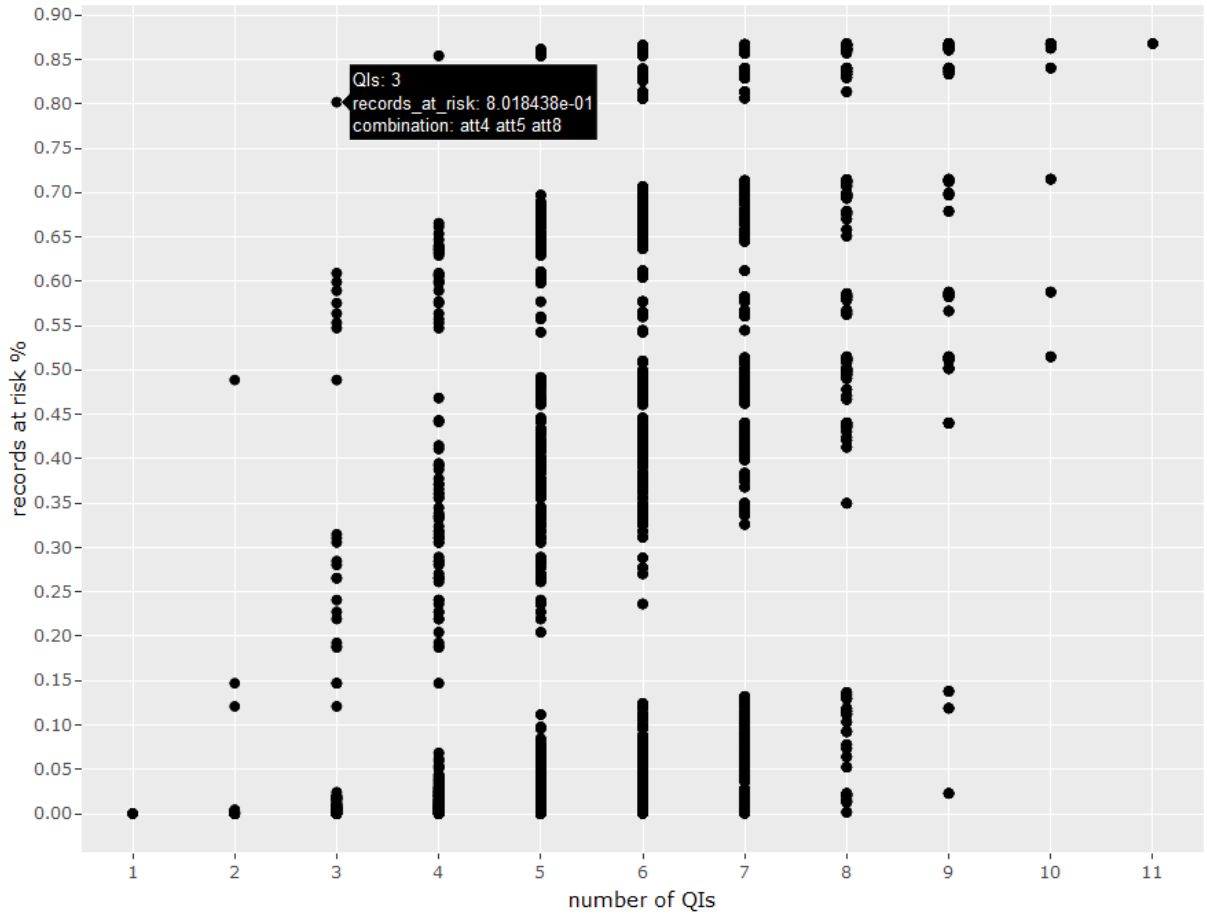

Figure 1: Example output of the de-anonymisability analysis tool

5 De-Anonymisation Risk Mitigation

In order to mitigate the de-anonymisation risks, anonymisation measures, beyond removing PII, need to be taken. Such measures rely on distorting the original values of a dataset so as to protect the privacy of its individuals, but at the cost of its utility.

The most basic anonymity principle is called *k-anonymity* and states that every individual in a dataset cannot not be distinguished by at least *k-1* other individuals, i.e. the maximum probability of identifying any individual is *1/k*. This can be achieved by defining *generalisation hierarchies* that describe the abstract values that could replace the original values of a dataset. Table 4 depicts an example of generalisation hierarchies for the attributes gender, date of birth and ZIP code.

QIs	Generalisation Levels			
	0 (Original Values)	1	2	3
gender	M or F	*	-	-
date of birth	dd/MM/yyyy	MM/yyyy	yyyy	*
ZIP code	5 digits	first 4 digits	first 3 digits	*

Table 4: Example of generalisation hierarchies

There are two ways (transformation models) of applying generalisation hierarchies: global and local recoding. In global recoding, all the values of a QI in the anonymised dataset belong to the same level of generalisation across the whole dataset, while in local recoding, different generalisation levels in different subsets of the dataset for each QI may be applied. Naturally, local recoding loses less information, since not all records and their QIs' values in a dataset may need to be transformed in order for them to conform to k-anonymity. Figure 2 depicts an example comparison of global and local recoding for 2-anonymity – each individual has at least 1 other individual with the same QIs.

gender	DOB	ZIP
M	01.01.1990	55555
M	01.01.1990	55556
F	01.01.1988	55515
M	01.01.1988	55516
F	02.01.1988	55517
M	02.01.1988	55516

Original

gender	DOB	ZIP
M	Jan 90	5555*
M	Jan 90	5555*
F	Jan 88	5551*
M	Jan 88	5551*
F	Jan 88	5551*
M	Jan 88	5551*

2-Anonymous Global Recoding

gender	DOB	ZIP
M	01.01.1990	5555*
M	01.01.1990	5555*
F	Jan 88	5551*
M	Jan 88	55516
F	Jan 88	5551*
M	Jan 88	55516

2-Anonymous Local Recoding

Figure 2: Example comparison of global and local recoding for 2-anonymity

There are other, more advanced anonymity principles than k-anonymity, that are used in case a dataset contains very sensitive information about individuals – information that they would not publicly reveal such as salary and political/religious beliefs. In the case of FNET's data, there are no such sensitive attributes and, therefore, local recoding k-anonymity is enough.

At FNET's premises in May 2019, along with the de-anonymisation attacks, we investigated how local recoding k-anonymity may be applied with the tools available at that time. In the procedure, we used the state-of-the-art tool ARX [2]. The procedure was carried out and can be generalised into 3 steps:

1) Identification of QIs : We identified the QIs for each of the 3 provided data tables: the table *assets* contains 11 QIs (all attributes except *CUSTOMER_ID* and *ASSET_ID*), the table invoices has 48 QIs (6 invoice months *x* 8 revenue types), and the table *SRs* contains 91 QIs (7 SRs months *x* 13 request types). The merging of the 3 tables results in a 150-dimensional data table. At the time of writing, there is no local recoding k-anonymity tool available that can handle such high dimensional datasets and, therefore, we only considered the table *assets* for k-anonymisation.

2) Definition of generalisation hierarchies: For each of the 11 QIs of the table *assets*, we defined generalisation hierarchies of at least 1 level, i.e. replacing the original value with "*". Some examples of the hierarchies defined are: the dates were defined as in Tab. 4, and the *INITIATION_DEALER_ID*'s hierarchies were defined as *original -> city of dealer -> county of dealer*.

3) Generating k-anonymised version(s) of the dataset: Providing as input the dataset, the list of QIs and the generalisation hierarchies to ARX [2], we generated 9 different versions of the table *assets* for $k \in [2,10$. Additionally, in order not to completely lose the information of the tables *invoices* and *SRs*, we generated aggregate information for those 2 tables : sum of revenue per type of revenue per customer, and sum of contacts per type of request per customer.

While this procedure enhances the privacy of the individuals in a dataset, it reduces its utility, and, therefore, value, since it distorts the original values of the dataset. The higher the privacy, the lower the utility. A challenge in this procedure is to find the golden mean between privacy and utility – the point of having high privacy, while still having a valuable dataset – a decision which has to be made by the data publisher.

6 Publications

During the first year of the project and within the context of task 5.4, we published two papers in peer-reviewed conferences.

A Horizontal Patent Test Collection [3] : In this paper, we introduce a novel patent research test collection, publicly available and for free that can be used on a variety of tasks beyond traditional information retrieval (IR), such as de-anonymisation. We describe how it can be used for de-anonymisation under the same solid empirical framework the IR community is used to. The paper was presented at SIGIR 2019, in Paris, July 21-25.

PrioPrivacy: A Local Recoding K-Anonymity Tool for Prioritised Quasi-Identifiers [4] : In this paper, we developed a local recoding k-anonymity tool that takes into consideration how important specific QIs are to the data publisher. The tool tries to distort these QIs as little as possible, and it is shown that our tool is capable of outperforming the state-of-the-art tool ARX [2]. The paper was presented at WI 2019, in Thessaloniki, October 14-17.

Additionally, we published a non peer-reviewed paper on arXiv.

An Abstract View on the De-Anonymization Process [5] : In this paper, we provide a taxonomy of the research in de-anonymisation from an abstract point of view, oriented towards data publishers.

7 Conclusion

In this deliverable, we presented the work carried out at task 5.4 within the first year of Safe-DEED project. Specifically, the work represents the procedure a data publisher should follow before releasing, exchanging, or selling their dataset(s), namely practical (Sect. 3) and theoretical (Sect. 4) de-anonymisation and measures to reduce the risk of de-anonymisation (Sect. 5).

The **practical de-anonymisation** consists of manual work – trying to actually identify individuals in a dataset. The procedure indicates the effort, cost and likelihood of de-anonymisation with the data that is publicly available, raises privacy red flags, and helps the data publisher in deciding the anonymisation measures that may need to be taken.

The **theoretical de-anonymisation** is carried out by using a tool that performs a risk analysis of the dataset's QIs. Similarly to practical de-anonymisation, it indicates the likelihood of de-anonymisation if the data of all QIs was available to an attacker, raises privacy red flags, and helps the data publisher in deciding the anonymisation measures that may need to be taken.

Anonymisation measures, beyond removing PII, are needed, in order to enhance the privacy of the individuals in a dataset, but at the cost of its utility. The extent to which these measures need to be taken is decided by the data publisher, being helped by the output of the practical and theoretical de-anonymisation.

Our plans for the remaining year of this task include, but not limited to, studying in more detail the procedure described above, and investigating the (de-)anonymisation of other kinds of data, such as query logs and mobility traces.

8 References

[1] Sweeney, L. (2000). Simple Demographics Often Identify People Uniquely.

[2] https://arx.deidentifier.org/

[3] Lupu, M., Bampoulidis, A. and Papariello, L., 2019, July. A Horizontal Patent Test Collection. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1213-1216). ACM.

[4] Bampoulidis, A., Markopoulos, I. and Lupu, M., 2019, October. PrioPrivacy: A Local Recoding K-Anonymity Tool for Prioritised Quasi-Identifiers. In IEEE/WIC/ACM International Conference on Web Intelligence-Companion Volume (pp. 314-317). ACM.

[5] Bampoulidis, A. and Lupu, M., 2019. An Abstract View on the De-anonymization Process. arXiv preprint arXiv:1902.09897.

[6] https://www.privacy-regulation.eu/en/recital-26-GDPR.htm