Grant Agreement Number: 825225

Safe-DEED

www.safe-deed.eu

D5.10 Report on the application of re-identifcation techniques on use-case data v2

Deliverable number	*D5.10*
Dissemination level	*Public*
Delivery date	*23 November 2020*
Status	*Final*
Author(s)	*Alexandros Bampoulidis*

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825225.

Changes Summary

Date	Author	Summary	Version
29.10.2020	Alexandros Bampoulidis	First draft	0.1
30.10.2020	Petr Knoth	Review and comments	0.2
02.11.2020	Alexandros Bampoulidis	Ready for review	0.3
09.11.2020	Lukas Helminger	Review and comments	0.4
11.11.2020	Hosea Ofe, Gert Breitfuss	Review and comments	0.5
21.11.2020	Alexandros Bampoulidis	Final version	1.0

Executive summary

Personal data is a necessity in many fields for research and innovation purposes, and when such data is shared, the data controller carries the responsibility of protecting the privacy of the individuals contained in their dataset. The removal of direct identifiers, such as full name and address, is not enough to secure individuals' privacy as shown by de-anonymisation methods in the scientific literature, because of the quasi-identifiers (QIs): attributes that, when combined, could create a unique identifier of individuals. Data controllers need to become aware of the risks of de-anonymisation and apply the appropriate anonymisation measures before sharing their datasets to comply with privacy regulations. Task 5.4 *De-Anonymisation* of the Safe-DEED project aims to address this need for the personal data of the project provided by the telecommunications provider Forthnet. To achieve this goal, we defined a procedure that makes data controllers aware of the de-anonymisation risks and helps them decide the anonymisation measures that need to be taken to comply with the General Data Protection Regulation (GDPR) and we showcase this procedure with use-case data. Additionally, in the duration of the task we developed two tools that are part of the execution of this procedure, and we produced three peer-reviewed publications.

Table of Contents

List of Figures

List of Tables

Abbreviations

CRM Customer Relation Management

DVC Data Valuation Component

D*x.x* Deliverable *x.x*

EDPB European Data Protection Board

EU European Union

FNET Forthnet

GDPR General Data Protection Regulation

IFX Infineon

IR Information Retrieval

MPC Multi-Party Computation

PII Personally Identifiable Information

PSI Private Set Intersection

QI Quasi-Identifier

QSD Qualified Synthetic Data

SME Small and Medium Enterprise

SR Support Request

UI User Interface

U.S. United States

WP29 Article 29 Working Party

WP*x* Work Package *x*

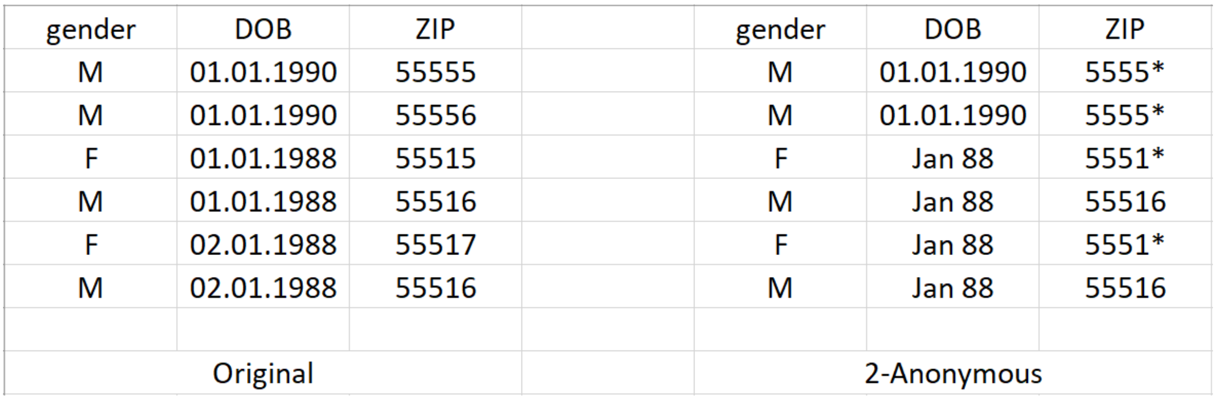
ZIP code Zone Improvement Plan code

1 Introduction

Personal data contains information about individuals and is used to advance many research fields and for foster innovation. Trace data are used to optimise public transportation, query logs for improving information search, and genomic data for cancer research are some of the example benefits of collecting and processing personal data. However, such data contains private and often sensitive information about individuals that they might not be willing to share publicly, and that could be used maliciously against them. Therefore, their privacy must be protected. Data controllers carry the responsibility of protecting the privacy of individuals in their datasets. They must be able to extract value from these datasets, while acting in compliance with regulations when collecting and processing these datasets.

Simply removing direct identifiers, such as full name and address, and releasing only a sample of a dataset [1] is not enough to protect the individuals' privacy because of the quasi-identifiers (QIs). QIs are attributes that do not directly identify individuals, but, when combined, could serve as a unique identifier of individuals. An extensive literature on de-anonymisation has proven that only a few QIs can uniquely identify the majority of individuals in a dataset. A couple of examples are: the combination of ZIP code, date of birth, and gender uniquely identifies 87% of the U.S. population [2], and four data points of location and time uniquely identify 95% of the individuals in a human mobility dataset consisting of 1,5 million individuals [3].

To counter the risks of de-anonymisation, privacy models that rely on distorting the original dataset have been introduced. The most prevalent of which are k-anonymity [4], l-diversity [5], and differential privacy [6]. Such models define a privacy principle that a dataset needs to conform to and offer various degrees of privacy represented by parameters. For example, k-anonymity defines that every individual in a dataset cannot be distinguished from at least k-1 other individuals. One way to achieve this is by substituting the original values with more abstract values (e.g., the exact date of birth substituted by month and year of birth). Figure 1 shows an example of a dataset becoming 2-anonymous in such a way. In the original table every individual is uniquely identifiable through their QIs, but in its 2-anonymised version, every individual has at least one other individual with the same QIs. If the original table is to become 3-anonymous, then more substitutions of the original values would be required.



gender	DOB	ZIP		gender	DOB	ZIP
M	01.01.1990	55555		M	01.01.1990	5555*
M	01.01.1990	55556		M	01.01.1990	5555*
F	01.01.1988	55515		F	Jan 88	5551*
M	01.01.1988	55516		M	Jan 88	55516
F	02.01.1988	55517		F	Jan 88	5551*
M	02.01.1988	55516		M	Jan 88	55516
	Original				2-Anonymous	

Figure 1: Example of k-anonymisation

A consequence of conforming to such privacy models is the decrease in the dataset's utility and, therefore, the value one can extract from it; the higher the privacy is, the more distorted the original dataset is. Before sharing their datasets, data controllers need to decide how much privacy is enough while still having a valuable dataset. While there exist methods trying to quantify the balance between privacy and utility [7], this topic remains highly subjective due to the dynamic context of, and the value one intends to extract from data sharing.

The de-anonymisation of an individual is considered a privacy breach, and it is subject to legal action against the data controller, and many regulations worldwide have been put in place addressing this issue. In this deliverable, notwithstanding the provisions that characterise the European Union (EU) privacy and data protection framework, we focus on the General Data Protection Regulation (GDPR) [8], that was put in force on 25 May 2018 in the EU, through the perspective of de-anonymisation and anonymisation. Additionally, we take into consideration the opinions published by Article 29 Working Party (WP29) [9] - replaced by the European Data Protection Board (EDPB) [10] with the introduction of the GDPR - an independent EU advisory body where representatives from all EU Member States come together to have a common interpretative approach on the provisions that touch upon personal data.

1.1 Task Motivation and Contribution

Task 5.4. *De-anonymisation* of the Safe-DEED project investigates the de-anonymisation of the data exposed in the project in a data sharing setting to raise privacy "red flags" and to ensure that the data are not *reasonably likely* to be de-anonymised.

To do so, we first needed to define a procedure that considers the GDPR, the guidelines from WP29, and the existing literature on anonymisation methods and tools. The defined procedure raises the awareness of data controllers on the de-anonymisation risks in their datasets and helps them decide the appropriate anonymisation measures.

While there exist studies on data protection under GDPR [13] and anonymisation guidelines published by WP29 and other authorities [14-17], we did not follow any specific one, but we incorporated elements from most of them in our defined procedure. In comparison to the existing guidelines, we provide more details on how a data controller becomes aware of the effort required to de-anonymise individuals from the perspective of an adversary that does not possess any prior knowledge on the individuals. We applied this procedure on the personal data of the project, namely the dataset supplied by WP6, provided by Forthnet (FNET). In addition to this procedure as a research and innovation contribution, we have developed a de-anonymisation risk analysis tool with 3 modules, a k-anonymisation tool, and three peer-reviewed research publications.

In Sect. 2, we present the (de-)anonymisation of individuals in the context of GDPR. In Sect. 3, we describe the personal data of the use-case of WP6 and in Sect. 4, we describe the aforementioned procedure and how we applied it to the use-case data of WP6. In Sect. 5, we describe the potential, minor threats to the privacy of Private Set Intersection (PSI) use-case of the project. In Sect. 6, we provide details on the developed tools. In Sect. 7, we describe the relationships of task 5.4. to the other WPs of the project. In Sect. 8, we list the scientific publications as an outcome of task 5.4. Finally, we conclude the deliverable in Sect. 9, where we also recount the current challenges, that we identified, of (de-)anonymisation in data sharing.

Note that in this deliverable, we are dealing with the personal data exposed in the project by FNET. The Qualified Synthetic Data (QSD) of the project is generated and made sure that it is not *reasonably likely* to be de-anonymised within Task 7.3 by Infineon (IFX). The work done in this task is described in detail in deliverables D7.3 and D7.9. The QSD was generated using a combination of statistical methods and widely used standard anonymisation techniques, some of which were also used in the anonymisation of FNET’s data, such as generalisation of exact dates to month and year, suppression (removal) of outliers and the addition of noise to the dataset’s values. The final QSD dataset is a dataset that mimics the patterns in the original dataset, but whose values do not correspond to actual IFX’s customers.

2 (De-)Anonymisation in the Context of GDPR

GDPR defines personal data as "*any information relating to an identified or identifiable natural person ('data subject')*" (Article 4, Paragraph 1). Any such information must be protected from the dangers of de-anonymisation. However, Recital 27 states that the regulation does not apply to personal data of deceased individuals and "*member states may provide for rules regarding the processing of personal data of deceased persons*". Therefore, the de-anonymisation of deceased individuals is not considered a privacy breach under the GDPR. The WP29 has already published a detailed opinion on the concept of personal data in 2007 (Opinion 4/2007 [11]), which is also reflected in the GDPR.

Article 4, Paragraph 5 of GDPR defines pseudoanonymisation as "*the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information...* ". This refers to the removal or masking all the direct identifiers of individuals in a dataset and not modifying the QIs of a dataset. As aforementioned, individuals may be identified through their QIs and, therefore, pseudoanonymisation is not enough as a practice to protect the individuals' privacy in data sharing. Still, the application of further anonymisation measures (e.g., k-anonymity) is not discouraged by the regulation (Recital 28). Additionally, the anonymisation process itself falls in the definition of processing (Article 4, Paragraph 2), and, therefore, its compliance with the GDPR provisions is necessary.

Recital 26 refers to the de-anonymisation of individuals through their QIs: "*Personal data which have undergone pseudoanonymisation, 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*". Here, *additional information* refers to an adversary, that has the purpose of de-anonymising individuals, knowing the values of the QIs.

Furthermore, Recital 26 calls for the data controller becoming aware of the risks of de-anonymisation: "*To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used ... to identify the natural person directly or indirectly.*". This requires a procedure with which a data controller gets to know how an adversary would de-anonymise individuals in their dataset.

It also calls for the data controller to become aware of the effort required to de-anonymise individuals in their dataset: "*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 processing and technological developments.*". Also, WP29 calls for awareness on the likelihood of de-anonymisation and the severity of its consequences (Opinion 05/2014 [12]): "*data controllers should focus on the concrete means that would be necessary to reverse the anonymisation technique, notably regarding the cost and the know-how needed to implement those means and the assessment of their likelihood and severity.*". The total effort in terms of costs, time, and know-how is only relevant for a specific perspective: an adversary that does not possess any information on the individuals prior to accessing a shared dataset and has to gather information in order to de-anonymise them. In cases where an adversary is already in possession of enough information to de-anonymise an individual (e.g., a relative or a neighbour), the only effort required would be to process the data and issue an SQL query to the dataset.

WP29 suggests that the knowledge of the effort required for de-anonymisation should be used in the anonymisation process (Opinion 05/2014): "*they*" (data controllers) "*should balance their anonymisation effort and costs (in terms of both time and resources required) against the increasing low-cost availability of technical means to identify individuals in datasets, the increasing public availability of other datasets ...*". "*The optimal solution should be decided on a case-by-case basis.*". This passage suggests that the easier to get the information of a dataset's QIs is and will be in the future, the higher the anonymisation effort should be, and that the anonymisation procedure should be approached on a case-by-case basis.

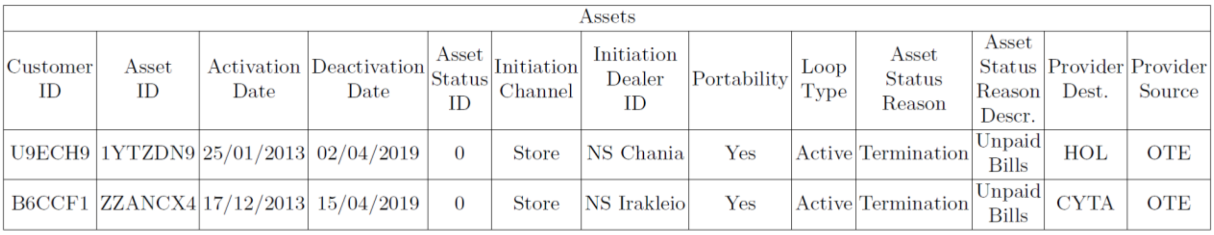
Finally, Recital 26 refers to individuals that cannot be de-anonymised after anonymisation measures have been applied: "*The principles of data protection should therefore not apply to anonymous information ... rendered anonymous in such a manner that the data subject is not or no longer identifiable.*" This particular passage suggests that anonymous data fall out of the GDPR's scope of application. Even when applying anonymity measures to a dataset, the risk of de-anonymisation is reduced, but not eliminated unless the dataset becomes completely valueless. Therefore, it is impossible to have completely anonymous data. For example, 3-anonymity implies that the maximum probability of de-anonymising any individual is 33%, and 4-diversity implies that the maximum probability of revealing any individual's sensitive attributes is 25%. The definition of anonymity is also arguable: Is it enough to consider a dataset anonymous, if "*the costs of and the amount of time required for identification*" are high (Recital 26), or should anonymity be quantified through privacy models (e.g. k-anonymity, etc.)?

3 Personal Data Description

3.1 CRM Dataset

One of the WP6 datasets is a customer relationship management (CRM) dataset having all personally identifiable information (PII) removed, provided by the Greek telecommunications provider Forthnet (FNET). Specifically, the CRM dataset consists of 3 tables: *Assets*, *Invoices*, and *Support Requests (SRs)*. Tables 1, 2, and 3, and Figures 2, 3, and 4 provide a description and depiction of those tables' columns, respectively.

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.



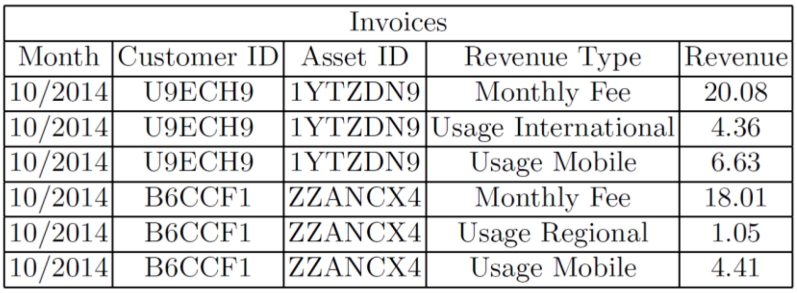
Assets												
Customer ID	Asset ID	Activation Date	Deactivation Date	Asset Status ID	Initiation Channel	Initiation Dealer ID	Portability	Loop Type	Asset Status Reason	Asset Status Reason Descr.	Provider Dest.	Provider Source
U9ECH9	1YTZDN9	25/01/2013	02/04/2019	0	Store	NS Chania	Yes	Active	Termination	Unpaid Bills	HOL	OTE
B6CCF1	ZZANCX4	17/12/2013	15/04/2019	0	Store	NS Irakleio	Yes	Active	Termination	Unpaid Bills	CYTA	OTE

Figure 2: Assets table example

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 (includes location of the dealer) 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

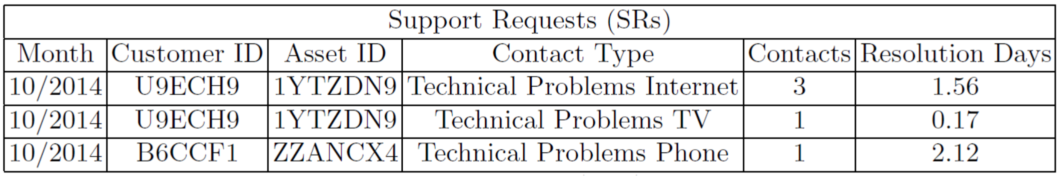


Invoices				
Month	Customer ID	Asset ID	Revenue Type	Revenue
10/2014	U9ECH9	1YTZDN9	Monthly Fee	20.08
10/2014	U9ECH9	1YTZDN9	Usage International	4.36
10/2014	U9ECH9	1YTZDN9	Usage Mobile	6.63
10/2014	B6CCF1	ZZANCX4	Monthly Fee	18.01
10/2014	B6CCF1	ZZANCX4	Usage Regional	1.05
10/2014	B6CCF1	ZZANCX4	Usage Mobile	4.41

Figure 3: Invoices table example

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



Support Requests (SRs)					
Month	Customer ID	Asset ID	Contact Type	Contacts	Resolution Days
10/2014	U9ECH9	1YTZDN9	Technical Problems Internet	3	1.56
10/2014	U9ECH9	1YTZDN9	Technical Problems TV	1	0.17
10/2014	B6CCF1	ZZANCX4	Technical Problems Phone	1	2.12

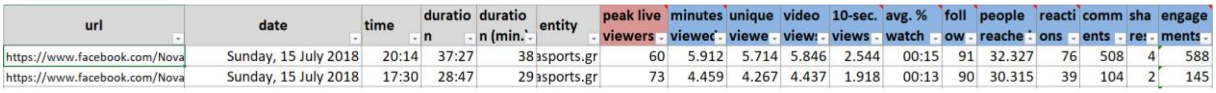
Figure 4: SRs table example

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 problems, service upgrades, 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

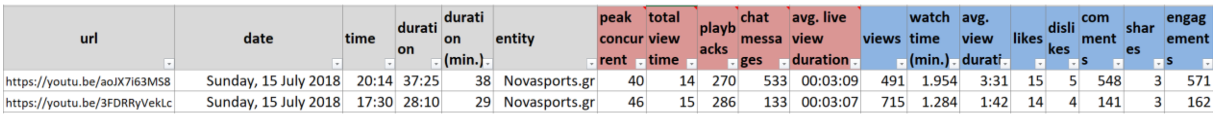
3.2 Viewership Dataset

In addition to the CRM dataset, FNET provided a viewership dataset that contains aggregate statistics of views and viewers of videos on their Facebook and Youtube channels. Figures 5 and 6 depict examples of such statistics.



url	date	time	duration	duration (min.	entity	peak live viewers	minutes viewed	unique viewe	video view	10-sec. views	avg. % watch	follow	people reache	reactions	comments	shares	engagements
https://www.facebook.com/Nova	Sunday, 15 July 2018	20:14	37:27	38	asports.gr	60	5.912	5.714	5.846	2.544	00:15	91	32.327	76	508	4	588
https://www.facebook.com/Nova	Sunday, 15 July 2018	17:30	28:47	29	asports.gr	73	4.459	4.267	4.437	1.918	00:13	90	30.315	39	104	2	145

Figure 5: Facebook viewership dataset example



url	date	time	duration	duration (min.)	entity	peak concurrent	total view time	playbacks	chat messages	avg. live view duration	views	watch time (min.)	avg. view durati	likes	dislikes	comments	shares	engagements
https://youtu.be/aoJX7i63MS8	Sunday, 15 July 2018	20:14	37:25	38	Novasports.gr	40	14	270	533	00:03:09	491	1.954	3:31	15	5	548	3	571
https://youtu.be/3FDRRyVekLc	Sunday, 15 July 2018	17:30	28:10	29	Novasports.gr	46	15	286	133	00:03:07	715	1.284	1:42	14	4	141	3	162

Figure 6: Youtube viewership dataset example

4 The (De-)Anonymisation Procedure

In this section, we present the procedure we defined to investigate the (de-)anonymisation of the datasets described in Sect. 3 and we describe how we applied it on these datasets. The procedure consists of 3 steps: data landscape analysis, threat analysis, and anonymisation measures.

4.1 Data Landscape Analysis

In this step, the data controllers become aware of the de-anonymisation risks in their datasets, and get to know how an adversary would de-anonymise individuals in their dataset. Additionally, they become aware of the effort, costs, and know-how required for de-anonymisation by an adversary that does not possess any information on the dataset's QIs prior to getting access to the dataset. This step consists of manual work and includes:

1) **Gathering external information**: The data controller needs to spend time looking for information sources that could be matched to the information contained in their dataset. Whether publicly or privately

available, the search for personal information in those sources indicates the effort an adversary would have to make to acquire enough information to de-anonymise individuals. If applicable, the costs of de-anonymising individuals are indicated by the purchasing cost of private information from companies or data brokers. The most prominent sources of publicly available information, where individuals share many potential QIs, are social media platforms, such as Facebook and Instagram, and forums, such as Quora. These sources do not only contain unstructured information (e.g. text), which requires manual extraction of QIs, but structured information as well, that could be extracted in an automatic way through APIs (e.g., date of birth and location).

2) **Processing the data**: After gathering as much information as possible, data controllers need to know how this information could be matched to their datasets. More specifically, they need to become aware of the know-how required in order to process their dataset in a way that would de-anonymise individuals with the acquired information. The processing could be as simple as importing a CSV file into a database, or it could require more complex methods, such as the application of machine learning models in order to extract further information from the dataset (e.g., sentiment analysis for product reviews).

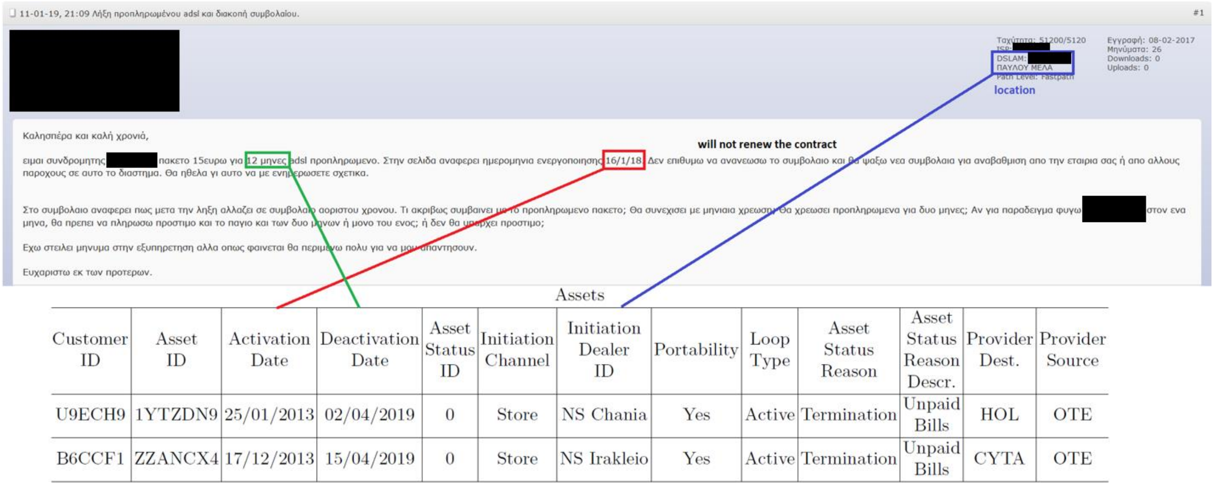
After executing this step, the data controllers should be able to answer the following questions:

1) What information contained in their dataset can be obtained by outsiders?

2) How much effort does it require to obtain this information?

3) How severe are the consequences of a de-anonymisation?

For the CRM dataset, we searched for information, that could be matched to the information in the CRM dataset, on 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. Figure 7 depicts such an example from a tech forum post: this particular customer of FNET provides the activation date of his contract (red box) and its length (green box), and states that he will not renew the contract and, additionally, his approximate location is visible in his profile (blue box). The customer’s information is matched to 3 of the columns in the *Assets* table, namely activation and deactivation date and initiation dealer id. De-anonymising this customer would lead an adversary to gain knowledge of his invoices records.



11-01-19, 21:09 Λήξη προπληρωμένου adsl και διακοπή συμβολαίου.

#1

[redacted]

Ταχύτητα: 51200/5120
ISP: [redacted]
DSLAM: [redacted]
ΠΑΥΛΟΥ ΜΕΛΑ
Path Level: Fastpath

Εγγραφή: 08-02-2017
Μηνύματα: 26
Downloads: 0
Uploads: 0

Καλησπέρα και καλή χρονιά,

ειμαι συνδρομητης [redacted] πακετο 15ευρω για 12 μηνες adsl προπληρωμενο. Στην σελιδα αναφερει ημερομηνια ενεργοποιησης 16/1/18. Δεν επιθυμω να ανανεωσω το συμβολαιο και θα ψαξω νεα συμβολαια για αναβαθμιση απο την εταιρια σας ή απο αλλους παροχους σε αυτο το διαστημα. Θα ηθελα γι αυτο να με ενημερωσετε σχετικα.

Στο συμβολαιο αναφερει πως μετα την ληξη αλλαζει σε συμβολαιο αοριστου χρονου. Τι ακριβως συμβαινει με το προπληρωμενο πακετο; Θα συνεχισει με μηνιαια χρεωση; Θα χρεωσει προπληρωμενα για δυο μηνες; Αν για παραδειγμα φυγω [redacted] στον ενα μηνα, θα πρεπει να πληρωσω προστιμο και το παγιο και των δυο μηνων ή μονο του ενος; ή δεν θα υπαρχει προστιμο;

Εχω στειλει μηνυμα στην εξυπηρετηση αλλα οπως φαινεται θα περιμενω πολυ για να μου απαντησουν.

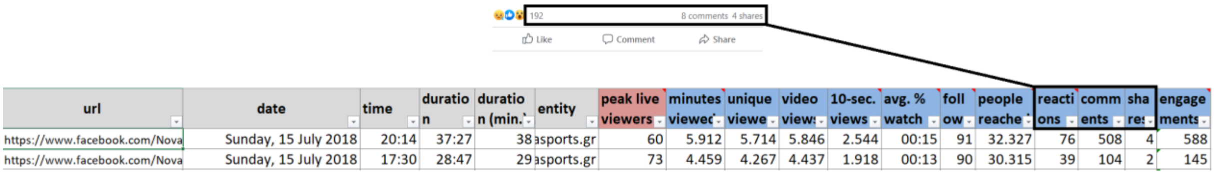
Ευχαριστω εκ των προτερων.

Assets

Customer ID	Asset ID	Activation Date	Deactivation Date	Asset Status ID	Initiation Channel	Initiation Dealer ID	Portability	Loop Type	Asset Status Reason	Asset Status Reason Descr.	Provider Dest.	Provider Source
U9ECH9	1YTZDN9	25/01/2013	02/04/2019	0	Store	NS Chania	Yes	Active	Termination	Unpaid Bills	HOL	OTE
B6CCF1	ZZANCX4	17/12/2013	15/04/2019	0	Store	NS Irakleio	Yes	Active	Termination	Unpaid Bills	CYTA	OTE

Figure 7: Example of the Data Landscape Analysis step on the CRM dataset

In the viewership dataset, this step is relatively straightforward since part of the information contained in the viewership dataset is extracted from publicly available resources; reactions, comments and shares from Facebook (Figure 8) and live chat (replay) and comments sections of Youtube.



url	date	time	duratio n	duratio n (min.	entity	peak live viewers	minutes viewed	unique viewe	video view	10-sec. views	avg. % watch	foll ow	people reache	reacti ons	comm ents	sha res	engage ments
https://www.facebook.com/Nova	Sunday, 15 July 2018	20:14	37:27	38	asports.gr	60	5.912	5.714	5.846	2.544	00:15	91	32.327	76	508	4	588
https://www.facebook.com/Nova	Sunday, 15 July 2018	17:30	28:47	29	asports.gr	73	4.459	4.267	4.437	1.918	00:13	90	30.315	39	104	2	145

Figure 8: Example of the Data Landscape Analysis step on the Facebook viewership dataset

Through this procedure, FNET became aware of the information their customers publicly reveal, and the effort required to process their dataset and match the gathered information to their dataset. Due to confidentiality reasons, further details on this step of the procedure applied on FNET's datasets cannot be provided.

4.2 Threat Analysis

In this step, data controllers become aware of the threats to the privacy of the individuals in their datasets, as well as the likelihood of a de-anonymisation by an adversary that already possesses enough information to de-anonymise individuals. This step is carried out with the help of tools that analyse the datasets and perform privacy checks. Such tools are also used for reporting purposes to justify the anonymisation measures taken by the data controller and prove a reasonable effort has been taken to do so, as required by the regulations (see Sect. 2). To address this need for the use-case of FNET, we developed three such tools corresponding to the different types of FNET's data: tabular, invoices, and aggregated. Note that, the *SRs* table does not contain any sensitive information.

Tabular: We refer to a tabular dataset as a dataset whose each line corresponds to one individual. Such is the case of the *Assets* table. The privacy threat in tabular datasets is that of an individual being de-anonymised through their QIs. This could lead to an adversary gaining knowledge about an individual's sensitive information (if such information exists in a dataset). To that end, we have developed a tool that reveals the likelihood of a de-anonymisation by an adversary that already possesses QIs of individuals, can designate the QIs that are critical for a de-anonymisation, and reveals the extent to which a dataset is de-anonymisable.

Figure 9 depicts a snapshot of the tool we have developed. Each point in the interactive plot represents a unique combination of QIs, with the x-axis referring to the number of QIs in a combination and the y-axis referring to the probability of de-anonymisation, if an adversary possesses the information of those QIs. When a user paces over a point, this information is displayed. Figure 9 depicts the probability of de-anonymisation of the individual in the example of Figure 7: there is an 80% probability of de-anonymising that individual and finding out his values of the other attributes in the *Assets* table, as well as his invoices and SRs records. Additionally, by exploring the interactive plot, it can be seen that two QIs are enough to de-anonymise ~50% of the individuals (top point in *number of QIs* = 2), while with the combination of all QIs, 88% of the individuals are de-anonymisable.

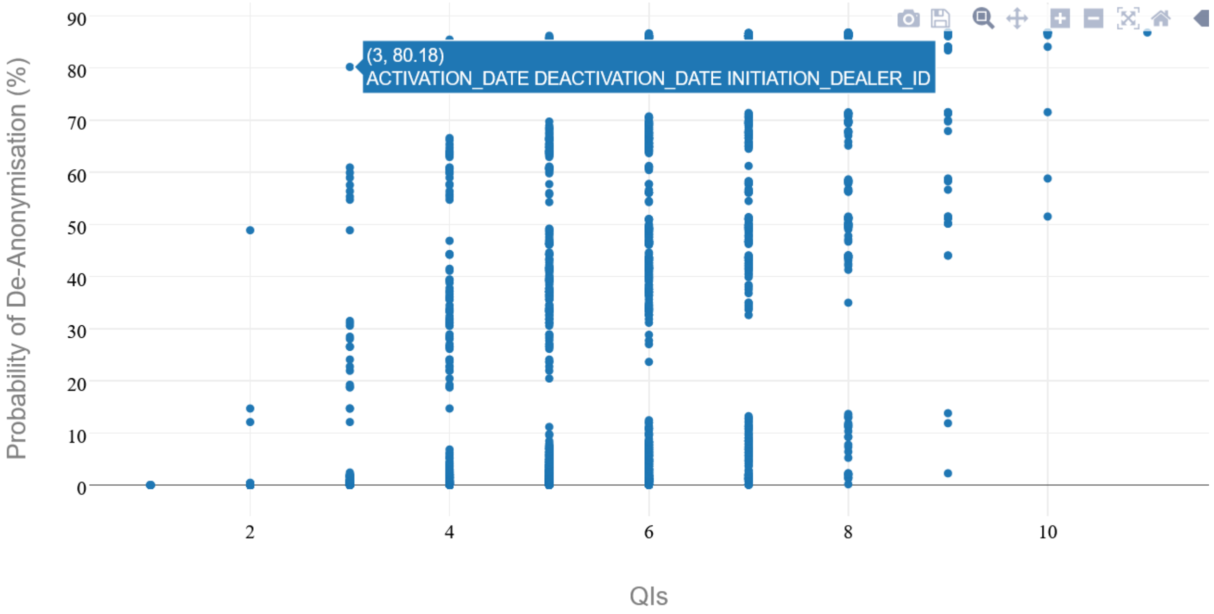


Figure 9: Tabular data de-anonymisation risk analysis

Invoices: Since every individual has more than one invoice in the table, the threat in the case of the *Invoices* table is different than *Assets*'s case. De-anonymising an individual from the exact invoice amount is highly unlikely since this information is hard to acquire by an adversary. It would rather be the sensitive information an adversary would like aim to find out. In this case, the privacy concern is that of the information contained in the *Invoices* table being sensitive. *SRs* is similar; however, it does not contain any sensitive information.

To that purpose, we have developed a tool that visualises the risks of any given invoices dataset, taking as input 3 privacy parameters: **a.** the number of individuals, **b.** an invoice amount, **c.** a timeframe. Figure 10 depicts a snapshot of this tool. Each point in the plot represents a unique invoice amount (y-axis) at a specific time (x-axis). In Figure 10, a point is coloured green if there are at least 2 distinct individuals (parameter **a**) having an invoice of ± 5000 (parameter **b**) within ± 1 month (parameter **c**); coloured red, otherwise. This output provides two insights:

1) Whether there is sensitive information. In the example, certain individuals have distinctively high invoice amounts, from which it could be inferred that they belong to a higher salary class.

2) Whether aggregating the data to the specified parameters **a, b,** and **c** conforms to the privacy notiont they specify.

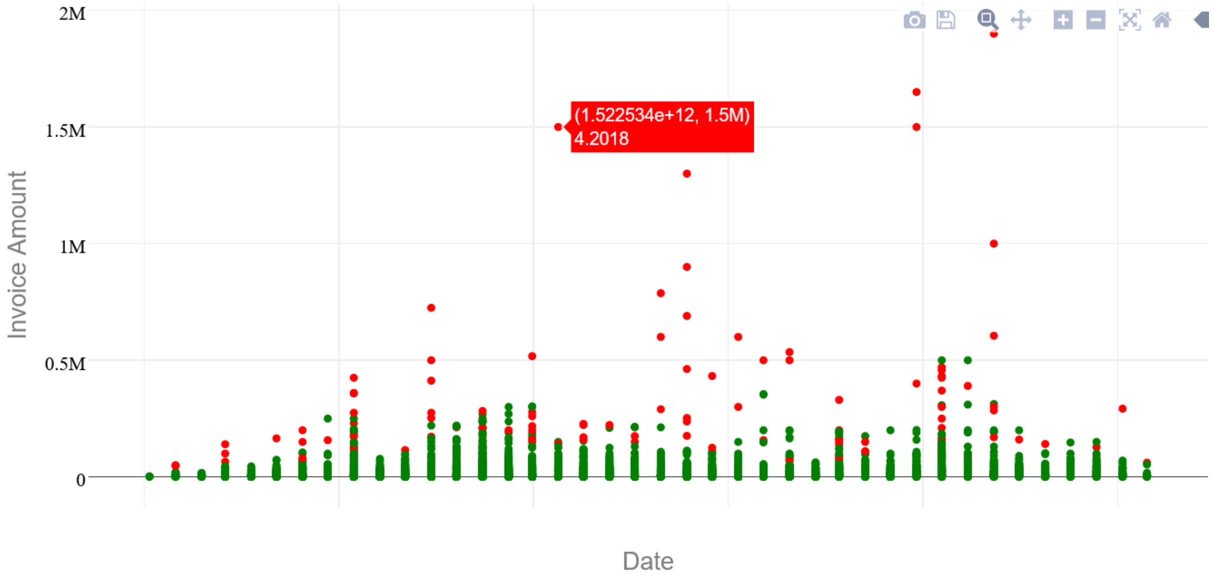


Figure 10: Invoices data risk analysis

Aggregated: An aggregated dataset is a dataset that contains aggregate values referring to individuals. Such is the case of the viewership dataset (Figures 5 and 6), where certain columns correspond to a grouping of individuals (e.g., shares on Facebook). While the individuals cannot be identified through those aggregated values, the values of potentially sensitive attributes may be inferred when the aggregate values corresponding to individuals are low. In the case of the viewership dataset, there are no sensitive attributes, but if there were such, for example, an attribute "sum of income in November of those who shared the video", then there would be a privacy breach if *shares* = 1.

To that purpose, we have developed a tool that visualises the risks of any given aggregated dataset, taking as input one privacy parameter: the number of individuals in an aggregation – *k*. Figure 11 depicts a snapshot of this tool applied to the Facebook viewership dataset, with the bars representing the *shares* column and *k* = 2. If the example described above was real, then some of the values of *shares* and of the example attribute would not be safe for release, since there are cases where *shares* = 1 (below the specified privacy parameter *k*). The tool's output provides an insight into how well the aggregate values protect their sensitive attributes, if such attributes exist.

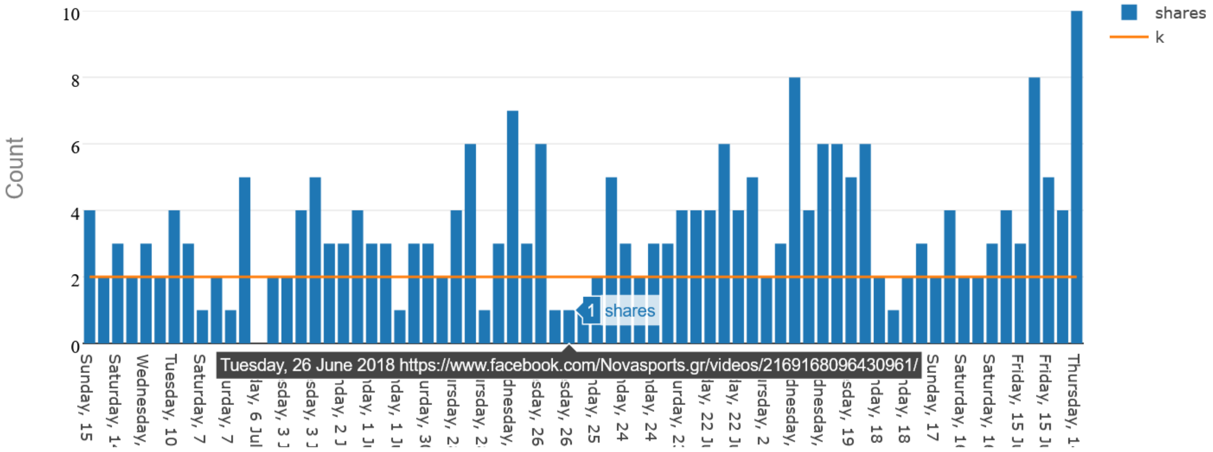


Figure 11: Aggregated data risk analysis

In this step, FNET became aware of the likelihood of de-anonymisation both from the perspective of an attacker who does not have prior knowledge of the individuals (Data Landscape Analysis) and the perspective of an attacker possessing enough information to de-anonymise individuals. FNET also became aware of the extent to which their dataset is de-anonymisable and the QIs critical in de-anonymisation. Due to confidentiality reasons, further details on this step of the procedure applied on FNET's datasets cannot be provided.

4.3 Anonymisation Measures

After becoming aware of the privacy risks in their dataset and judging the severity of a potential de-anonymisation, data controllers need to take the appropriate measures in order to mitigate them. Data controllers should follow anonymisation guidelines, such as the ones cited in Sect. 2, to decide which anonymisation measures are appropriate for their dataset to protect the individuals' privacy.

Additionally, the data controllers are helped by the output of the previous two steps in deciding the extent to which the anonymisation measures should be applied; the degree of privacy their dataset conforms to, while still being useful. For example, if, through the *Data Landscape Analysis*, the data controller finds that information about certain QIs can be easily acquired, then the data owner might consider not releasing them or distorting them more than the rest of the QIs. A data controller may reach such a decision if such QIs are shown to be critical in de-anonymising individuals (in the *Threat Analysis* step) even if currently they are not easy to acquire but may become so in the future. Similarly, a data controller may decide on the level of aggregation of invoices based on the output described in Sect. 4.2 and not release information on customers having an invoice amount above a threshold (e.g. red points in Figure 10), or partially, if at all, release an aggregate column (e.g., the example described in Sect. 4.2).

After studying existing anonymisation guidelines and anonymisation literature, we decided which anonymisation measures should be taken. Since the CRM dataset does not contain any sensitive attributes, conforming to the k-anonymity principle is a good enough measure to mitigate the de-anonymisation risks. To achieve as little loss of information as possible, we decided to use local recoding [18] as a transformation method through generalisation and suppression [19] (as in Figure 1). The viewership dataset is safe for release without any anonymisation measures, since the only information that is not already publicly available contains statistics about the videos and not individuals.

FNET, having applied all the necessary GDPR processes, provided the CRM dataset, which did not contain any direct identifiers of individuals, for a limited time (May 6-10, 2019) in M6 of the project, on their premises, allowing us to apply the procedure described in this section and anonymise the dataset. 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.

For k-anonymising FNET's CRM dataset, we used the state-of-the-art anonymisation tool ARX [20] which, among others, offers local recoding k-anonymity. However, its latest version (3.7.1), at the time of the given access to FNET's dataset, could not handle the complete dataset (the 3 tables merged) due to its high dimensionality - 150 dimensions: 11 QIs of *Assets*, 48 QIs of *Invoices* (6 months *x* 8 revenue types), and 91 QIs of *SRs* (7 months *x* 13 SRs types). Therefore, we decided to k-anonymise the *Assets* table for k∈[2,10], and generate aggregate information of the tables *Invoices* and *SRs*: sum of revenue per type of revenue per contract for the total period of 6 months, and sum of requests per type of request per contract for the total period of 7 months, respectively. In total, we generated 9 different k-anonymised versions of the CRM dataset, each having varying degrees of privacy and utility.

Inspired by FNET's needs, which consider certain QIs of their dataset more important than others and, therefore, these QIs should be distorted as less as possible in the anonymisation process, we developed *PrioPrivacy* [21], a local recoding k-anonymity tool which is capable of outperforming the state-of-the-art tool ARX [20], when the data controller specifies the importance of the QIs.

Due to confidentiality reasons, further details on this step of the procedure applied on FNET's datasets cannot be provided.

5 Minor Privacy Threats of the PSI Use-Case

PSI is a form of privacy preservation, where the actual datasets are not exposed, but rather the output of a computation. While it is a very secure way of privacy preservation, there is still a minor privacy threat in extreme cases which we describe in this section. We refer to deliverables D5.4 and D5.8 for a more detailed description of this use-case.

The PSI use-case involves two parties finding out each other's common values of a column in their datasets – in the specific use-case, postal codes. The only privacy threat is the case where one party finds out the existence of a specific individual that is uniquely identifiable by exactly one QI in the other party's dataset. Examples of such extreme cases are: exactly one individual living in a specific postal code or a unique date of birth (e.g. the oldest individual in a country). In FNET's case, the probability of this occurrence is 0%, as shown in Figure 9, i.e., no customer is uniquely identifiable by exactly one QI.

Note that in the Multi-Party Computation (MPC) use-case (also described in D5.4 and D5.8), the exposure of the output of a computation remains only with the buyer and the seller, and not competing entities, and, therefore, there is no threat to their privacy.

6 Developed Technologies

We developed a de-anonymisation risk analysis tool within the task context with three modules and a k-anonymisation tool. These tools are open-source and are available through the Safe-DEED meta-repository [22]. In this section, we describe the technical characteristics of these tools.

The **de-anonymisation risk analysis tool** is a web application with three modules, each corresponding to a type of data (see Sect. 4.2): tabular, invoices, and aggregated. It was developed in Java and uses Spring Boot as a web server. A prototypical user interface (UI) is found in its code repository, but in the WP6 demonstrator, it comes with a proper UI.

The algorithmic complexity of the *tabular* module is between O(#QIs × r), if every individual is uniquely identifiable by exactly one QI, and $O(2^{\#QIs} \times r)$, if there is no individual uniquely identifiable with any combination of QIs, with *r* being the number of rows in the dataset. The more de-anonymisable a dataset is, the lower the algorithmic complexity is. What happens, in essence, in this module is that $2^{\#QIs}$ datasets are created, each corresponding to a unique combination of QIs, and each individual in these datasets is checked for uniqueness. The algorithm is parallelised and, also, two optimisations add to the scalability of this module:

1) The first iteration is that of the dataset with the whole set of QIs, and the individuals that are not uniquely identifiable with the whole set of QIs are, also, not uniquely identifiable by any subset of QIs. Therefore, these individuals are eliminated from the checks in the next iterations. For example, in FNET's case, 12% of the dataset (see Sect. 4.2 and Figure 9) is eliminated from the next iterations.

2) If an individual is uniquely identifiable from a set of QIs, they are uniquely identifiable with any superset of these QIs, and they are uniquely identifiable with any superset of these QIs. These individuals are eliminated from the checks in the next iterations that contain these set of QIs. For example, in FNET's case, after the set of two QIs that uniquely identifies 48% of the individuals is checked (see Sect. 4.2 and Figure 9), these individuals are eliminated from checking in any superset of these two QIs.

The algorithmic complexity of the *invoices* and *aggregated* modules is O(r), linear in the number of rows.

In comparison to two existing risk analysis tools [20, 23], these tools are only able to deal with tabular data and do not offer an interactive visual exploration of the uniqueness of individuals through their QIs.

The **k-anonymisation tool** ***PrioPrivacy*** is a desktop application utilising a local recoding k-anonymity algorithm that we developed. It was developed in Java and JavaFX was used to design the UI. Its algorithmic complexity is the same as the *tabular* module of the de-anonymisation risk analysis, is parallelised and, as reported in its publication, achieves a better execution time that the state-of-the-art. The algorithm is detailed in its publication [21].

7 Connection to Other Work Packages

In this section, we describe how the output of this task connects to other WPs, namely 4, 6, and 8.

WP4 (Private and Public Data Value)**:** As mentioned in Sect. 1, anonymisation measures distort a dataset, reduce its utility and, therefore, its value. The developed tools described in Sect. 6 will be integrated into the Data Valuation Component (DVC), in which the data controller would be able to gain knowledge of the de-anonymisation risks in their dataset and k-anonymise it. The effect of anonymisation on the value of a dataset is reflected by utility metrics that are used in the anonymisation literature and which measure, each in its own way, how similar the anonymised dataset is to the original one. In the case of WP4, the Non-Uniform Entropy metric [24], for which our tool *PrioPrivacy* can outperform the state-of-the-art, will be integrated to the DVC.

WP6 (Secure enterprise data exchange (use-case personal data))**:** The personal data provided by FNET are part of WP6 and their anonymisation is a specific aim of Task 6.3 *Joint data usage between different enterprises in different domains.* Additionally, the developed tools described in Sect. 6 are integrated into the WP6 demonstrator.

WP8 (Dissemination, Communication, Exploitation, Sustainability and Market Validation)**:** The procedure described in Sect. 4 will be part of the online learning videos to be hosted at TU Delft's edX platform. The online learning video will have a length of 4 to 6 minutes and its target audience are small and medium enterprises (SMEs), serving as a guidance for them before selling their datasets to other entities.

8 Publications

A Horizontal Patent Test Collection [25]: 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, 2019.

PrioPrivacy: A Local Recoding K-Anonymity Tool for Prioritised Quasi-Identifiers [22]: 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 [20]. The paper was presented at WI 2019, in Thessaloniki, October 14-17, 2019.

Practice and Challenge of (De-)Anonymisation for Data Sharing [26]: In this paper, we present the procedure we defined which is described in Sect. 4 and how we applied it on FNET's dataset, and additionally, we recount the current challenges of (de-)anonymisation for data sharing. The paper was presented at RCIS 2020, online, September 22-25, 2020.

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

An Abstract View on the De-Anonymization Process [27]: In this paper, we provide a taxonomy of the research in de-anonymisation from an abstract point of view, oriented towards data controllers.

9 Conclusion

In this deliverable, we presented the work carried out within the context of task 5.4 *De-anonymisation* of the Safe-DEED project. After studying the literature on (de-)anonymisation and the GDPR, we defined a 3-step procedure that data controllers should follow before sharing their datasets, and we applied it to the project's datasets provided by FNET. The defined procedure makes the data controllers aware of the de-anonymisation risks in their datasets and helps them in deciding the appropriate anonymisation measures. In addition to this procedure and to make its execution possible, we developed two tools: a de-anonymisation risk analysis tool and a k-anonymisation tool.

We identified the following challenges of (de-)anonymisation for data sharing that the industrial and scientific community has to deal with during the definition of this procedure and by putting existing knowledge and tools into practice.

Lack of awareness on (de-)anonymisation: In general, laypeople that do not have a scientific or engineering background are not aware of the risks of de-anonymisation, and anonymity is viewed as simply removing the direct identifiers. Unfortunately, the same situation exists in the industry as well, as reported by popular news media [28-30]. While these news media strongly point out the dangers of de-anonymisation in personal data, they do not refer to anonymisation measures (beyond removing direct identifiers) that could be taken to protect the individuals' privacy. This challenge may be faced by authorities and news media promoting awareness on the dangers of de-anonymisation and the anonymisation measures that can be taken to mitigate those risks.

Lack of detailed guidelines: As addressed in the previous sections, privacy in data sharing is a complex issue and it should be studied in more detail and have broader coverage in the regulations and guidelines. Even though WP29 and other authorities provide guidelines that help data controllers comply with the regulations, more details should be provided on anonymising datasets case-by-case, especially, on the balance between privacy and utility, and cases where even a low privacy guarantee results in a tremendous loss of information and value [3]. This challenge may be faced by authorities providing more detailed guidelines on anonymising datasets.

Lack of open-source tools from complex data: As mentioned in Sect. 4.3, we could not k-anonymise the complete FNET dataset due to its high-dimensionality. The ARX tool is the state-of-the-art in anonymisation of datasets and provides a wide palette of anonymisation methods, but the scientific literature on anonymisation consists of many more methods that could anonymise high-dimensional datasets whose source code, however, is not open-source or not available. Here, we identify the need for reproducing the most important methods in the anonymisation literature and packaging them as open-source, easy-to-use tools. This challenge may be faced by researchers and developers reproducing existing and developing new anonymisation methods and making them open-source.

10 References

[1] Rocher, L., Hendrickx, J. M., & De Montjoye, Y. A. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. Nature communications, 10(1), 1-9.

[2] Sweeney, L. (2000). Simple demographics often identify people uniquely.

[3] De Montjoye, Y. A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the crowd: The privacy bounds of human mobility. Scientific reports, 3, 1376.

[4] Sweeney, L. (2002). Achieving k-anonymity privacy protection using generalization and suppression. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05), 571-588.

[5] Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). l-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data (TKDD), 1(1), 3-es.

[6] Dwork, C. (2008, April). Differential privacy: A survey of results. In International conference on theory and applications of models of computation (pp. 1-19). Springer, Berlin, Heidelberg.

[7] Li, T., & Li, N. (2009, June). On the tradeoff between privacy and utility in data publishing. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 517-526).

[8] https://eur-lex.europa.eu/eli/reg/2016/679/oj

[9] https://ec.europa.eu/newsroom/article29/news-overview.cfm

[10] https://edpb.europa.eu/edpb en

[11] https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2007/wp136_en.pdf

[12] https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

[13] Gruschka, N., Mavroeidis, V., Vishi, K., & Jensen, M. (2018, December). Privacy issues and data protection in big data: a case study analysis under GDPR. In 2018 IEEE International Conference on Big Data (Big Data) (pp. 5027-5033). IEEE.

[14] https://www.dataprotection.ie/en/guidance-landing/anonymisation-and-pseudonymisation

[15] https://www.fsd.tuni.fi/en/services/data-management-guidelines/anonymisation-and-identifiers/

[16] Bargh, M. S., Meijer, R., & Vink, M. (2018). On statistical disclosure control technologies. WODC.

[17] Graham, C. (2012). Anonymisation: managing data protection risk code of practice. Information Commissioner's Office.

[18] Xu, J., Wang, W., Pei, J., Wang, X., Shi, B., & Fu, A. W. C. (2006, August). Utility-based anonymization using local recoding. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 785-790).

[19] Sweeney, L. (2002). Achieving k-anonymity privacy protection using generalization and suppression. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05), 571-588.

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

[21] Bampoulidis, A., Markopoulos, I., & 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).

[22] https://github.com/Safe-DEED

[23] Hagedoorn, T. R., Kumar, R., & Bonchi, F. (2020). X2R2: a tool for explainable and explorative reidentification risk analysis. Proceedings of the VLDB Endowment, 13(12), 2929-2932.

[24] Prasser, F., Bild, R., & Kuhn, K. A. (2016, September). A Generic Method for Assessing the Quality of De-Identified Health Data. In MIE (pp. 312-316).

[25] Lupu, M., Bampoulidis, A., & 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).

[26] Bampoulidis, A., Bruni, A., Markopoulos, I., & Lupu, M. (2020, September). Practice and Challenges of (De-) Anonymisation for Data Sharing. In International Conference on Research Challenges in Information Science (pp. 515-521). Springer, Cham.

[27] Bampoulidis, A., & Lupu, M. (2019). An Abstract View on the De-anonymization Process. arXiv preprint arXiv:1902.09897.

[28] https://www.nytimes.com/2019/07/23/health/data-privacy-protection.html

[29] https://www.theguardian.com/technology/2019/jul/23/anonymised-data-never-be-anonymous-enough-study-finds

[30] https://www.cnbc.com/2019/07/23/anonymous-data-might-not-be-so-anonymous-study-shows.html