K-ANONYMITY AS A PRIVACY MEASURE

I. SUMMARY

This technical note is intended for data processors and controllers who undertake anonymisation processes on datasets. In a reality in which independent data sources are interconnected and, by design, they may share common attributes, the possibility exists of creating an electronic trail of the individuals, even when the data which explicitly identify them have been removed, and connections may be established between those sources of information, thereby posing a risk to the privacy of those persons whose data are processed.

In application of the principle of Proactive Responsibility established in General Data Protection Regulation (EU) 2016/679, the data controller must undertake a study of the inherent risk of re-identification of the people who the data refer to and implement measures to manage it. The aim of that analysis is to achieve an adequate balance between the need to obtain results with a certain degree of fidelity and the potential cost of processing for citizens' rights and freedoms.

This note outlines one of the possible techniques to manage the risk of re-identification, known as k-anonymisation.

II. INTRODUCTION

Recital 26 of Directive 95/46 established that, in order to determine whether a person was identifiable, it was necessary to consider all of the means which could reasonably be used by the controller or anyone else to identify that person. Thus, the data protection principles were no longer applicable in those cases in which the data is rendered 'anonymous' or dissociated in such a way that it was no longer possible to identify the data subject.

Along those same lines, recital 26 of the GDPR indicates that 'pseudonymised' personal data comprise information about a natural person from which it is possible to indentify that person within reasonable probability, taking into account objective means and factors, as well as the costs, the time and the technology necessary to make that identification.

Note the difference in the terms used in the two regulations: from the limited concept of 'anonymisation', it has evolved towards a materialisation of it in the term 'pseudonymisation' in the GDPR, recognising the difficulty in achieving perfect anonymisation at the present time, or anonymisation which guarantees, in absolute terms, the masking of people's identities. However, throughout this document, we will use the term 'anonymisation' whether or not the identification of the data subject is reversible to a greater or lesser degree.

The massive processing of citizens' data by means of the use of techniques based on Big Data, Artificial Intelligence or Machine Learning makes it necessary to implement guarantees or mechanisms to preserve privacy and the right to the protection of data of a personal nature, including those based on the anonymisation of those data.

The data sources used for that processing contain personal data which are classified as '*identifiers*', because, in themselves, they are unequivocally associated with a subject, such as the National Identity Document, the full name, the passport number of the social security number. The basic anonymisation process consists of dissociating from the identifiers the rest of the more generic data associated with a subject, such as the date of birth, the place of residence, the sex, etc. The preserved data will be those necessary to fulfil the purpose of the processing, and, by means of their conservation and enrichment, to exploit them in order to extract additional information.

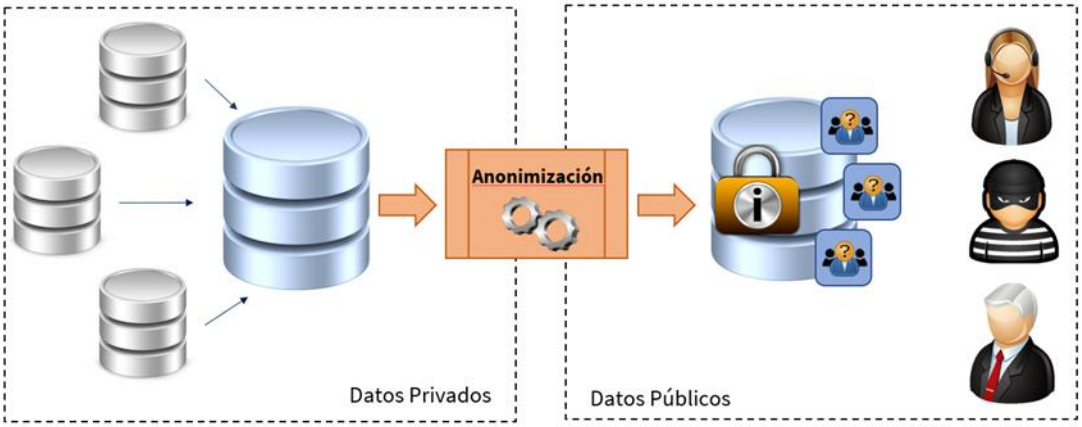

Figure 1: Anonymisation

Private Data Public Data

However, though the carrying-out of that process of anonymisation apparently makes it possible to maintain anonymity, those data, conveniently grouped together and cross-referenced with other sources of information, can identify an individual and even relate that person to special categories of data. Hence, the data which are not 'identifiers' but which could unequivocally indicate an individual are called "*pseudo-identifiers*", *"quasi-identifiers"* [1], or indirect identifiers.

There is a risk that, once a dataset has been anonymised, those data could be de-anonymised. Therefore, it is necessary to have an objective estimation of the likelihood of re-identification based on the quasi-identifiers, in order to have a measurement of that risk.

To manage that problem and avoid the de-anonymisation of data, a discipline has been developed, known as Statistical Disclosure Control (SDC)[2], whose aim is to study how to undertake additional processing on the information of data subjects in a optimal manner, maximising privacy while at the same time maintaining the objectives established in the application or service which uses those data. The techniques used in SDC can, generically, be classified as perturbative or non-perturbative, depending on whether noise is introduced into the original data source.

One of those techniques is K- anonymisation, a technique which was already cited by the working group of article 29 of Directive 45/96, in its Opinion 05/2014 [3].

III. WHAT IS *K*-ANONYMITY?

***K*-anonymity** is a property of anonymised data which makes it possible to quantify to what extent the anonymity of the subjects present in a dataset in which the identifiers have been removed is preserved. In other words, it is a measure of the risk

that external agents can obtain information of a personal nature from anonymised data.

If we classify the attributes of records according to their nature or the type of information they contain, we can distinguish the following types of data [4]:

- **Key attributes or identifiers**: they are fields which unequivocally identify the data subjects (name, ID, passport number, telephone number, etc.). Those types of data must be removed from the anonymised records.
- **Quasi-identifiers**: they are fields which, in themselves and in isolation, do not identify an individual but which, if grouped together with other *quasi-identifier*s, could unequivocally identify a subject. Anonymisation techniques work on those data, removing fields which are not necessary for processing (in application of the principle of minimisation), aggregating them or generalising them.
- **Sensitive attributes**: they are fields which contain data which could have a greater impact on the privacy of a specific individual, including the special categories of data, and which must not be linked to the data subject they belong to (illnesses, medical treatments, income level, etc.). That information may be of great interest in the object of the data processing, but unless there is some legitimation for it, it must be dissociated from a specific subject.

It is said that an individual is k-anonymous within the dataset in which he/she is included if, and only if, for any combination of the associated quasi-identifier attributes, there are at least another $K-1$ individuals who share the same values for those same attributes [5]. We must take into account that K-anonymity is not focussed on the sensitive attributes of records [4], but rather on the quasi-identifier attributes which may permit that connection.

In that way, the probability of identifying a specific individual based on that series of quasi-identifiers is at most $1/K$, and so, in order to ensure a low risk of re-identification, a minimum value of K must be guaranteed when undertaking the design of a data anonymisation or dissociation process.

For example, imagine the following dataset in which there are two attributes of quasi-identifier type, the 'postcode' and the 'age', associated with a sensitive attribute which indicates health data related to the data subjects contained in the dataset.

Postcode	Age	Cholesterol
37003	40	Y
28108	44	Y
24700	37	N
24700	37	N
37003	40	Y
28108	44	Y

Table 1: 2-anonymisation

Postcode	Age	Cholesterol
37003	40	Y
28108	44	Y
24700	37	N
24700	37	N
37003	44	Y
28108	40	Y

Table 2: 1-anonymisation

Table 1 is 2-anonymised, because each combination of values of the quasi-identifier attributes appears in at least two rows, whereas table 2 is not, because, for each record, there is not at least one other which contains identical values for those attributes.

Therefore, two conclusions are drawn in relation to the values of K in an anonymised dataset:

1. We are interested in high values of K so that, if we encounter a subject included in several sources of information and who certain attributes are associated with, it is unlikely that we will be able to know which one of them another, associated piece of data corresponds to exactly, for example, a medical treatment.
2. 1-anonymity is equivalent to saying that the individual is perfectly identifiable within the group [6]. Therefore, in the right circumstances and by properly cross-referencing the information from other sources which contain data on that individual, it could be possible to de-anonymise the identity of certain subjects of those included in the universe under study.

When designing data processing which requires the use of anonymised data, it is important to answer the following questions:

- What value of K is adequate?

  Higher values of K correspond to more stringent privacy requirements, because it would be necessary for there to be more subjects within a group with identical combinations of identifying features. In the obtainment of higher values of K, we may lose fidelity in the source data, and therefore we must determine whether, in that loss of fidelity, there is or is not a loss of information which is relevant for the purpose of the processing. If there is no loss of relevant information, that initial process must be executed. If there is a loss of relevant information, we will need to achieve a balance between the risks to the rights and freedoms of the subjects and the potential loss of fidelity in the result of the processing.

- How can we make a series of data K-anonymous?

  The next section answers that question.

IV. K-ANONYMISATION METHODS

There are two methods widely used to implement K-anonymisation and which do not introduce perturbation into the data: generalisation and suppression. Those methods are said to be non-perturbative because they achieve protection by replacing the original values of the attributes with other, more general values without introducing erroneous information into the original data source.

Generalisation

Generalisation consists of making the values of the quasi-identifier attributes less precise, transforming them or generalising them within a series or group which has the same values, either by creating ranges in the case of numerical attributes or by establishing hierarchies for nominal attributes. In that way, the number of records which have the same values for a series of quasi-identifier attributes can be increased in order to satisfy the privacy requirements, while at the same time fulfilling the purpose of the processing.

Starting from table 2, shown above, it is possible to transform it into a series of 2-anonymous data by means of a generalisation of the 'Age' attribute within a numerical range and of the 'Postcode' attribute classified in a hierarchy (figure 2). In turn, the generalisation can be global, if, given the same value for the same type of attribute, the transformation is always made in the same way (table 3), or local, if different generalisation criteria are used for each record (table 4).

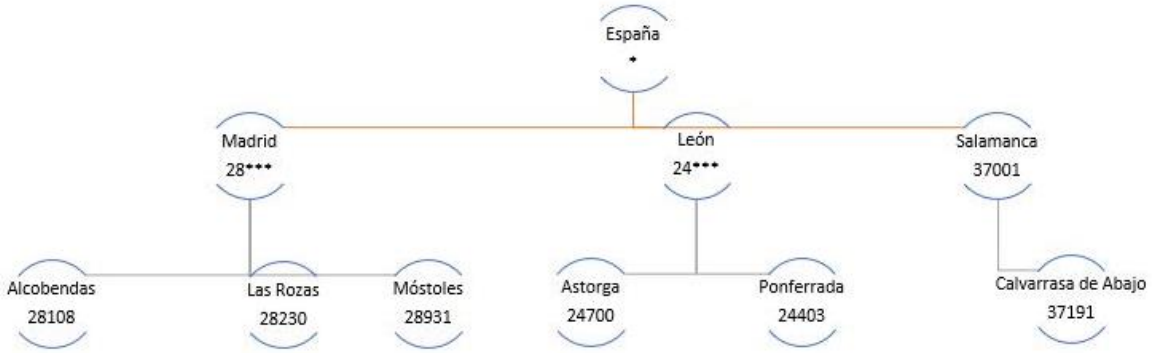

Figure 2: Hierarchy for the Postcode field

Postcode	Age	Cholesterol
37***	40 - 49	Y
28***	40 - 49	Y
24***	30 - 39	N
24***	30 - 39	N
37***	40 - 49	Y
28***	40 - 49	Y

Table 3 - Global generalisation

Postcode	Age	Cholesterol
37***	40 - 49	Y
28***	40 - 49	Y
24700	30 - 39	N
24700	30 - 39	N
37***	40 - 49	Y
28***	40 - 49	Y

Table 4 - Local generalisation

The advantage of global generalisation is that it simplifies the analysis of the data, while local generalisation, though it makes it possible to maintain more precise values, complicates the representation of the results.

Suppression

The other method to implement K-anonymity is suppression. In the above example, the values of the records were quite close to each other, which made it possible to generalise while maintaining reasonable precision. Imagine that we add the following records to table 2:

Postcode	Age	Cholesterol
37003	40	Y
28108	44	Y

24700	37	N
24700	37	N
37003	44	Y
28108	40	Y
37891	**33**	**N**
50011	**13**	**Y**

Table 5: Table 2 expanded with data outside the range

For the first six records, we can undertake a global or local generalisation as shown in tables 3 and 4, but the last one of the added records is outside the range. Trying to carry out a generalisation by defining a range which contains it could lead to a loss of precision such that the data would no longer be useful for analysis.

In those cases, the solution is to suppress or remove those records so they do not 'contaminate' the dataset and distort the results. The records with very unusual values must also be eliminated, because they significantly increase the probability of re-identification.

Applying both methods, generalisation and suppression, table 5 of the second example would become 2-anonymous, as shown in table 6:

Postcode	Age	Cholesterol
37003	40	Y
28108	44	Y
24700	37	N
24700	37	N
37003	44	Y
28108	40	Y
37891	33	N
50011	13	Y

Table 5: Original

Postcode	Age	Cholesterol
37***	40 - 49	Y
28***	40 - 49	Y
24700	30 - 39	N
24700	30 - 39	N
37***	40 - 49	Y
28***	40 - 49	Y
37***	30 - 39	N

Table 6: Generalisation + Suppression on Table 5

In trying to anonymise using the suppression method in isolation or combined with the generalisation method, we obtain datasets which contain fewer records than the original data source.

V. LIMITATIONS OF K-ANONYMISATION

Generalisation and suppression introduce different types and degrees of distortion in the anonymisation process. Anonymising based on suppression techniques may mean having to eliminate a considerable number of records from the processed dataset, introducing a bias in the original distribution of values which could distort the result of the analyses. Generalisation, for its part, means that we lose the informative potential of the atomic data, meaning that, in the dataset, we lose the ability to draw conclusions from the values of those attributes in their relation with other fields of information. Though, in the example shown, the bias which is introduced is considerable, because it is a very limited number of entries, in the case of data sources with a large number of records, the loss of a few disperse values does not excessively distort the result and it avoids the introduction of wide generalisation ranges in order to contain those extremes.

The mathematical problem which lies behind transforming a dataset into another, K-anonymous dataset, is a problem of NP-hard complexity [7]. There are different algorithms [8] , [9] to achieve a solution and on which different software solutions are constructed, both open and commercial, which make it possible to K-anonymise the dataset which is entered into them as inputs. Some examples of those kinds of tools which make it possible to implement K-anonymity are [10]:

- **ARX Data Anonymization Tool**: ARX is an open source tool which makes it possible to transform structured series of personal data using different anonymisation methods and SDC techniques. It makes it possible to remove the direct identifier attributes (for example, names) from datasets and to apply rules to the quasi-identifiers in order to minimise association attacks. The tool supports various privacy techniques, among them k-anonymity, as well as data transformation models such as random sampling or micro-aggregation. ARX is capable of handling large datasets and it has an intuitive, multi-platform graphic interface, as well as an API for integration with Java to implement data anonymisation capabilities from software developed under that programming language.

  Download link: https://arx.deidentifier.org/downloads/

- **UTD Anonymization Tool:** It is an open source tool developed in the UT Dallas Data Security and Privacy Lab which implements various anonymisation methods for public use by researchers. The algorithms can be used both directly on a dataset and through libraries of functions implemented inside other applications. It uses different anonymisation methods, including k-anonymity.

  Download link: http://cs.utdallas.edu/dspl/cgi-bin/toolbox/index.php?go=download

- **Amnesia:** Amnesia is a data anonymisation tool which makes it possible not only to remove the information associated with direct identifiers such as names or numbers of identification documents, but also to transform the quasi-identifier attributes such as the date of birth and the postcode in order to mitigate the risks of re-identification of the subjects contained in the data sources, using k-anonymity methods. It has a client version and an online version.

  Download link: https://amnesia.openaire.eu/installation.html
  Link to online version: https://amnesia.openaire.eu/amnesia/

However, though K-anonymity prevents the disclosure of a specific data subject's identity within a series of individuals who share the same values for the quasi-identifier attributes, it can still fail in protecting from the disclosure of sensitive information associated with that subject, because, if the K elements of an equivalence class share the same value for an attribute considered confidential, as occurs in the example seen in this note, the simple determination of whether an individual belongs to the K-anonymised group will mean that, without knowing his/her exact identity, he/she is associated with the protected sensitive value with complete certainty or with a very high percentage of accuracy. In our example, if we are able to determine that an individual residing in Madrid in the 40 - 49 years range belongs to the sample under study, we will know that they have cholesterol problems.

Those types of vulnerabilities have led to the appearance of additional privacy techniques which are outside the scope of this technical note, such as p-sensitive K-anonymity and l-diversity, which measure the degree of diversity or variety of the values for sensitive data within an equivalence class, and t-closeness and δ-disclosure, which measure the similarity between the distribution of the values of the sensitive attributes in each equivalence class and the global distribution of all the records. The ARX tool described above implements, in addition to the K-anonymity technique, some of these other techniques aimed at mitigating attacks to link different datasets.

VI. CONCLUSIONS

The data controller must ensure the privacy of the subjects whose data are processed. Some entities consider that suppressing or masking identifier attributes is sufficient to ensure the anonymity of those data subjects. However, it is possible that common fields present in different data sources, conveniently grouped together and cross-referenced, could become a pseudo-identifier attribute which compromises people's privacy.

Therefore, anonymisation cannot be limited to the simple, routine, passive application of certain commonly-used rules, but rather, in application of the principle of accountability, the processing controller must analyse the risks of re-identification in the anonymisation processes, adequately selecting the types of quasi-identifier attributes used with the aim of reducing the probability that the cross-referencing of those fields with others contained in external data sources could represent a risk for the rights and freedoms of the data subjects.

Therefore, during the phases of conception and design of processing of data of a personal nature, an analysis must be carried out of the degree of fidelity necessary in the result of that processing, in order to precisely determine the appropriate generalisation and suppression margins, within reasonable limits which prevent the distortion of reality.

Likewise, an analysis must be undertaken and the right balance achieved between the risks for the rights and freedoms of citizens and the legitimate benefits and benefits for society of carrying out that processing with a certain degree of precision.

Deriving from both analyses, it is necessary to achieve a balance between the benefit that would be obtained for society in carrying out processing with a given degree of fidelity and the cost that that processing implies for the rights and freedoms of the data subjects.

There are different techniques aimed at preserving the privacy of the personal data of individuals aimed at limiting the threats to privacy which could be materialised by

de-anonymising information. K-anonymity is a technique aimed at preventing the re-identification of a specific subject within a group, whether by means of the generalisation of the quasi-identifier attributes or the suppression of records which are outside the range. However, it does not provide guarantees that, if we know that a subject belongs to that group, it is not possible to infer information of a sensitive nature associated with that subject.

VII. REFERENCES

[1] Jordi Casas Rom. *Data privacy*, Universitat Oberta de Catalunya (UOC) Data Day, 2017.

[2] Agustín Solanas, Antoni Martínez-Ballesté, Josep Domingo-Ferrer, Susana Bujalande, Josep M. Mateo-Sanz. *Métodos de microagregación para k-anonimato: privacidad en bases de datos*, Dept. Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili.

[3] Opinion 05/2014 on Anonymisation Techniques, adopted on 10 April 2014, 29WP.

[4] R. Somolinos Cristóbal, A. Muñoz Carrero, M.E. Hernando Pérez, M. Pascual Carrasco, R. Sánchez de Madariaga, O. Moreno Gil, J.A. Fragua Méndez, F. López Rodríguez, C. H. Salvador. *Pseudonimización de información clínica para uso secundario. Aplicación en un caso práctico ISO/EN 13606*, Unidad de Investigación en Telemedicina y e-Salud, Instituto Carlos III, Madrid, 2014.

[5] Latanya Sweeney. *K-anonymity: A model for protecting privacy*, School of Computer Science, Carnegie Mellon University, 2002.

[6] Carlos J. Gil Bellosta. *Microdatos y K-anonimidad: un enfoque cuantitativo en el contexto español*, Dananalytics, 2011.

[7] Adam Meyerson, Ryan Williams. *On the complexity of Optimal K-Anonymity*, Computer Science Departments of University of California and Carnegie Mellon University

[8] Aris Gkoulalas-Divanis, Grigorios Loukides, Jimeng Sun. *Publishing data from electronic health records while preserving privacy: A survey of algorithms*, Journal of Biomedical Informatics - Elsevier, 2014.

[9] Zakariae El Ouazzani, Hanan El Bakkali. *A new technique ensuring privacy in big data: K-anonymity without prior value of the threshold K*, ScienceDirect – Elsevier, 2018.

[10] List of commercial and open source software related to anonymisation techniques (https://arx.deidentifier.org/overview/related-software/)