Deliverable 1:
Literature Review

ODI Tender RDP8-001

Authors: Gemma Galdon Clavell and Carmela Troncoso

Research assistants: Victoria Peuvrelle and Miguel Vallbuena

Eticas Research and Consulting
C/ Ferlandina, 49
08001, Barcelona
(+34) 93 600 54 00
www.eticasconsulting.com

Table of contents

Glossary

Term	Meaning
Adversary	An entity with access to an anonymise dataset that seeks to re-identification of an individual or learn more information about them.
Anonymisation	Process that transforms a dataset to ensure that an adversary cannot recover information about individuals.
Anonymised dataset	A dataset: - where no individual can be identified - where no information can be linked to an individual - that cannot be used to infer information about an individual
Anonymity set	Set of identities that could be linked to a record in a dataset.
Article 29 Data Protection Working Party	An advisory body of the European Union providing expert advice and make recommendations to States regarding data protection.
Attack	An attack is a process that takes as input an anonymised dataset and outputs information related to an individual.
Attribute	Characteristic associated to a record in a dataset (e.g.: age, sex, name, phone number, diagnostic, salary, etc.). For each record the attribute takes a *value* (e.g.: 42 years old, female, Jane.).
Confidential attribute	An attribute considered to be sensitive information. Example: Salary, sexual orientation, political views, health conditions.
Dataset	A set of records with associated attributes. It can be published as a spreadsheet format with records (rows) and attributes

	(columns); or in other forms, e.g. as a relational (e.g., MySQL) or non-relational (e.g., NoSQL) database.
Equivalent class	Records that share identical values within certain attributes. Example: records sharing the same ZIP code.
Encryption	The process of converting readable data into unreadable code in order to protect it. Only authorised parties with a decryption key have access to the data.
Identifier	An attribute that identifies the individual to which it refers directly. Examples: Passport number, fingerprint, name.
Inference	The fact of being able to link a concrete attribute value to an individual. If this attribute is an identifier, the inference becomes re-identification.
Linkability	The ability to be link attributes and/or records within a dataset, or across datasets.
Open Data	Data that can be freely accessed, used and shared by anyone.
Pseudonymisation	The replacement of a value, normally an identifier, by another value to render it more difficult to re-identify.
Quasi-identifier	An attribute that in itself does not lead to re-identification but may do so if combined with other attributes. Example: ZIP code, birthdate, gender.
Non-identifier	Attributes that are neither identifiers, nor quasi identifiers and do not enable the re-identification of an individual.
Original dataset	A dataset before going through the process of anonymisation.
Record	A set of attributes that are linked together in a dataset. Examples: all attributes relating to one individual, a tuple (latitude, longitude,

	value) associated to a measurement.
Re-identification	The action of inferring the identity of the individual to which an anonymised record (or set of records) relates to.
Singling-out	The action of reducing the anonymity set of a record to only one person. Sometimes known as “uniquely identify”.

Introduction

Data is everywhere. In the digital age, our data footprints include information on what we do, where we go, who we know, what we have, what we like or how we feel. We generate this information while we work, walk, interact, speak, protest or search online. The activities we engage in generate data in their turn, and all this information is useful to shape services, products and cities, for instance, and to promote transparency and accountability. We have seen data improve care, public transport routing, policing and advertising. But we have also seen data be stolen or manipulated, and data processes infringe upon fundamental rights and privacy, discriminate or go wrong.

Realising the potential of increased data gathering, sharing and publishing, thus, requires an understanding of where and how data can improve outcomes, but also of the risks involved in the process. The same piece of health data that can be crucial to save someone's life when made available to their doctor, may also mean they are not considered for a job or see their insurance cost increase. The implications of data sharing change depending on who is at the receiving end and how they access the data.

This poses an important challenge for data sharing and open data in general. Choosing to never share data and keep it only in the sphere where it is gathered (medical data in a medical setting, to follow on the previous example) is an option that may mean we lose the possibility to explore health risks more generally -data on the health condition of people in one area made available to environmental services may mean we identify the unknown presence of toxic substances. Or linking search data with medical data may reveal undivulged side effects. In different domains, linking crime data with social service data may also help identify trends previously unknown. Employment data can also help improve public transport by better predicting flows and demand. In less crucial settings, data sharing can help improve commercial services by coupling supply and demand more efficiently, or generating demand. Profiling and targeting also mean that messages can be tailored to a specific target audience - something which can be great in case of emergencies (specific messages to older people, or to those already rescued, for instance), but problematic in politics.

One of the approaches to responsible data sharing involves the use of *anonymisation techniques*, “sanitising” databases to remove personal traits from the data before it is shared. Anonymisation involves masking or removing information that could directly or indirectly identify individuals in such a way that information in the database does not enable *re-identification* and cannot be used to learn new information about these individuals other than the information one has *a priori*.

Defining anonymisation

Anonymization is not a new concept. The history of science of data confidentiality (also known as statistical disclosure limitation, statistical disclosure control or statistical database privacy) began in 1977 with a seminal paper published by Swedish statistician Tore Dalenius: *Towards a methodology for statistical disclosure control*. Later in the 1980s, the theory of tabular disclosure limitation of Lawrence H. Cox and others helped the field progress. In 1998, the work of Samarati and Sweeney (2011) on the concept of *k-anonymity* became a reference framework, though the techniques quickly presented shortcomings. *Differential privacy*, introduced by Dwork in 2006 was another stepping-stone attempt at improving existing anonymization techniques in a context of increased data sharing, but it also showed its limits (Dwork 2008).

Removing personal data from a dataset is an easy process. What is difficult is to eliminate personal data *and* to keep the utility of the dataset. Therefore, all anonymisation techniques strive to find an optimal balance between protecting privacy and enhancing security, on the one hand, and keeping the utility of the information in meaningful ways. The utility of an anonymized dataset and the level of anonymization, i.e., the degree to which it leaks information about individuals, are opposed to one another. If a dataset is anonymized to a high level, its utility drastically decreases. Similarly, the more useful the dataset, the less anonymized it is (Ohm, 2009; Lubarsky, 2017). Depending on the anonymization technique, the trade-off is different. That is, each technique has its advantages, shortcomings and associated issues.

An anonymisation effort may work in one setting and be a recipe for disaster in another, depending on the threat model of a particular case. A dataset containing information that is very valuable to some actors (health data for insurance companies, or political opinions for non-democratic regimes) will need to take additional precautions and use more robust anonymisation techniques that a dataset with information on public transport use, for instance. Therefore, anonymity is not only a technical issue, but one that depends on a more complex and contextual understanding on who or what poses a “threat” to the integrity of a particular database.

It is also worth noting the difference between anonymity and *pseudonymity*. As we will see, robust anonymisation in a context of Big Data is almost impossible to guarantee. As we do not control what inferences or datasets others will have available, the risk of re-identification is always present and difficult to predict. That is why there are more and more voices that highlight the need to speak of pseudonymised data, and to differentiate it from anonymised data.

Pseudonymisation is a way to ensure the continued identification and linkage to one or more datasets without directly identifying the person. It normally involves the replacement of a value, normally an identifier, by another value. The individual whose record has been pseudonymised will still be identifiable due to the attribution of this new value. For instance, John Smith becomes User 3849562. With this system, a person having taken an exam can look for their test result in the database with the unique ID they were given, without others being able to identify them. This is an alternative method to anonymisation that sometimes proves sufficient, depending on the data and the uses.

However, if quasi-identifiers remain within the dataset, the individual is still re-identifiable. The scientific community widely agrees that a pseudonym is not useful to protect privacy if the same unique pseudonym is continually used throughout one or multiple datasets (Garfinkel, 2015;

Lubarsky, 2017; Article 29 Working Party, 2014), especially as the amount of attributes linked to a record grows (Barocas and Nissenbaum, 2014). Possibilities for linkage, singling out and inference (techniques we will explore below) remain the same in a pseudonymised dataset and the original dataset. This is especially the case if a predetermined algorithm to pseudonymise a dataset is used (Lubarsky 2017). This is why in its Opinion 05/2014 on Anonymisation Techniques[1] the Working Party Article 29 strongly emphasised in its 2014 opinion paper that a pseudonymised dataset is not anonymised and does not meet anonymisation standards.

Nevertheless, as we will see, pseudonymisation can be used in combination with other anonymisation techniques to robustly protect a dataset and prevent re-anonymisation. In the rest of the document we will first introduce the fallibility of anonymisation through some high-profile examples. Secondly, we summarise the relevant regulatory framework in the EU. We then propose a reading guide of the relevant academic literature, putting forward the most relevant articles covering the current discussion on this topic. Following this overview of the literature, we delve into different anonymisation techniques currently used.

Anonymisation gone wrong

In 2006, AOL released a list of all the search queries of 650,000 users made over a period of three months. This amounted to 20 million search queries. The dataset was published to provide researchers with a large and detailed source of information that was normally not accessible, and AOL was aware of the privacy risks that could arise from the release of such data. Therefore, the dataset was pseudonymised and each user name was replaced with a unique random number.

But soon people started trying to re-identify users based on their queries. These attempts were often revealed for entertainment purposes, revealing funny situations, but some search queries also identified people struggling with depression or abusive partners (Barbaro and Zeller, 2006). It took only six days for Michael Barbaro and Tom Zeller of *The New York Times* to re-identify user No. 4417749, Thelma Arnold, a 60-year-old woman from Lilburn, Georgia. She had searched for harmless information such as “landscapes in Lilburn Ga”, which were used to locate her. Other searches, however, were more private, and Ms. Arnold had to witness how the world learned that she searched for “60 single men” and “dog that urinates on everything”. Other user searches that were made public were also extremely intimate, outing some individuals’ mental health issues, or showing evidence of illegal activity. Many people also searched for information that directly identified them, such as their address or even their names (Arrington, 2006). The day after the release, uproar had already begun, and AOL soon took down the database. But a mirror site had already been created and the data had already been

[1] Opinion 05/2014 on Anonymization Techniques adopted 10 April 2014 by the Working Party Article 29. Available at: https://www.pdpjournals.com/docs/88197.pdf (accessed 28/08/2018).

downloaded hundreds of times, meaning the possibility of its dissemination was virtually endless.

AOL's open data move was an immense failure. It demonstrated how easily weak pseudonymisation could lead to re-identification. Simply hiding the name did not prevent users from being identified because personal information was plainly available in the searches, even if their names were coded.

The same year as the AOL scandal, Netflix published a sample of their movie ratings consisting of 10 million ratings from half a million users. The publication took place in the context of a contest where the research team that could provide the best improvement for Netflix's suggestion algorithm would win a million dollars (Schneider, 2007). But two researchers from the University of Texas analysed the data with a different purpose: to re-identify specific users by correlating their Netflix ratings with ratings from the Internet Movie Database (IMDB). When the attacker knew the approximate date (with a 3 day-error) of the rating of 2 movies, they were able to de-anonymise a user 68% of the times (Narayanan and Shmarikov, 2008).

The issue with re-identifying a user and having access to their movie ratings is the information it provides about the user: religious beliefs, political opinions and sexual orientations can be inferred from their views on movies. This example shows how information that was not considered sensitive and was anonymised, could in fact reveal sensitive information about individuals and was not immune to re-identification.

Anonymisation in Data Protection legislation

- GDPR and WP29 Opinion on Anonymisation Techniques

Anonymisation and pseudonymisation are two concepts covered by the General Data Protection Regulation (GDPR) in order to make compatible the processing of data for legitimate purposes and the protection of the data subject's privacy. The former European 95/46/EC Directive defined anonymous data as non-identifiable data and did not mention pseudonymised data, so data was either identifiable (personal) or non-identifiable data. The GDPR introduced a middle ground by formalising the notion of pseudonymised data.

Anonymised data is introduced in recital 26 of the GDPR in the following terms:

> *The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes.*

In the same paragraph, pseudonymised data is defined as follows:

> *Personal data which have undergone pseudonymisation, which could be attributed to a natural person using 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.*

The recognition of this middle ground has been a notable innovation of the GDPR. Kotschy (2016) argues that legislators have thus incentivised data controllers to implement pseudonymisation measures. While some scholars (Ohm, 2014) decry the definition of anonymised data because it is extremely difficult to reach complete anonymisation, other scholars such as Brasher (2018) support the European Union’s decision to introduce the concept of pseudonymised data, arguing that it constitutes a serious attempt to make compatible the right to privacy of data subjects with the value that data processing can bring to society.

The 95/46/EC Directive set up a working group to go over issues related to the protection of personal data, called the Article 29 Working Party (WP29). This body, which has now become the European Data Protection Board, published an important opinion paper in 2014 that clarified the differences between anonymisation and pseudonymisation:

> *Pseudonymisation is not a method of anonymization. It merely reduces the linkability of a dataset with the original identity of a data subject, and is accordingly a useful security measure.*

The GDPR does not directly provide a clear set of criteria to assess whether personal data is no longer identifiable. However, the WP29 did address the issue:

> *[T]hus it is critical to understand that when a data controller does not delete the original (identifiable) data at event-level, and the data controller hands over part of this dataset (for example after removal or masking of identifiable data), the resulting dataset is still personal data. Only if the data controller would aggregate the data to a level where the individual events are no longer identifiable, the resulting dataset can be qualified as anonymous. [...] once again depending on the context and purposes of the processing for which the anonymised data are intended.*

The mention of context and purpose is important. It means that the appropriate threshold of anonymisation is to be assessed on a case by case basis. This contextual approach is especially interesting since it accounts for the limitations that both pseudonymisation and even anonymisation techniques present in terms of allowing for re-identification to occur.

To summarise, the GDPR only applies to personal identifiable data, which encompasses the newly defined concept of pseudonymisation. Anonymised data is data which cannot be linked back to an individual and as such falls outside of the scope of the Regulation. However, the criteria with which an anonymisation process can be qualified as robust is less clear and context dependant, which could be qualified as a ‘’risk-based approach’’ (Stalla-Bourdillon and Knight, 2016, p.5).

- E-Privacy Directive

The EU directive on Privacy and Electronic Communications (also known as the E-Privacy Directive or ePD) came into force in 2002 and was meant to cover the areas not specifically covered by the former EU Data Protection Directive in the electronic communications sector.

The directive thus addressed subject matters such as security and confidentiality of the information, data retention, cookies, traffic data, location data and emails.

In terms of anonymisation, the directive requires in article 6(1) that data no longer needed for transmission should be destroyed or anonymised. Article 9(1) allows for the processing of location data without the consent of the users if it is anonymised. Recital 9 encourages member states to use anonymous or pseudonymous data where possible. However, very little is provided as to how this should be achieved.

The directive is to be repealed and replaced by the Regulation on Privacy and Electronic Communications in 2019, aimed to adapt the changes the GDPR brought about in the electronic communications sector. The approach to anonymisation remains essentially the same in the current proposal, and so the difference between anonymised and pseudonymised data is expected to play a crucial role, with pseudonymised data requiring further security and privacy efforts by the data controllers, as it does in GDPR.

Anonymisation in the academic literature

Below we reference some of the key academic references in the field of anonymisation and pseudonymisation. This is a very active field, with theories and approaches being published and challenged all the time. Therefore, this list is provided as a reference, and we encourage the reader to follow on the debates that will surely arise in the future.

- General, non-technical papers

Article 29 Data Protection Working Party. (2014). Opinion 05/2014 on Anonymisation Techniques. Retrieved from https://www.pdpjournals.com/docs/88197.pdf

The WP29 was a group set up by the European Union to tackle data protection issues in the EU. It is composed of the heads of national data protection agencies. As such, papers published by this group are highly regarded. As such, this opinion paper is also an important piece of work in anonymisation literature. As Ohm's article, it is non-technical and aimed at the general public and policy-makers. It goes over all the main techniques, analysing the common mistakes and issues each pose.

Ohm, P. (2009). Broken promises of privacy: Responding to the surprising failure of anonymization. Ucla L. Rev., 57, 1701.

Ohm's article is important as it is a comprehensive overview of anonymisation techniques and shortcomings, aimed at lawyers and policy-makers. As such, the article is not technical and very easy to follow. While the anonymisation debate was still very much limited to technical issues, he red-flagged the situation and created a discussion in areas of law and policy, adding context and a more multi-disciplinary perspective to the debates around anonymisation.

- On re-identification and k-anonymity

Sweeney, L. (2002a). k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05), 557-570.

Latanya Sweeney is a pillar in anonymisation research. She famously re-identified a governor's medical record by matching it with a publicly available voter list. She also created the Datafly algorithm, widely used in the United-States to anonymise datasets, and along with McSherry introduced the concept of k-anonymity to ensure adequate anonymisation of a dataset. Though k-anonymity has since then been shown to have serious shortcomings, it did become a standard. While this is not her first article on k-anonymity, it is a comprehensive guide of the technique as she envisioned it. She re-explains how she re-identified the governor, gives the mathematical definitions of certain terms such as quasi-identifiers and attributes and explains k-anonymity and its shortcomings. It is the most cited article concerning anonymity, with over 5,000 citations.

- On re-identification

Narayanan, A., & Shmatikov, V. (2008, May). Robust de-anonymization of large sparse datasets. In Security and Privacy, 2008. SP 2008. IEEE Symposium on Security and Privacy (pp. 111-125). IEEE.

This article by Narayanan and Shmatikov's article, in which they demonstrate that the Netflix anonymisation was not effective, is a turning point in how anonymisation was approached by both academia and industry. In this article, they provide an algorithm to de-anonymise a large sparse dataset, clearly demonstrating the shortcomings of anonymity techniques. By using the Netflix dataset, they show that information which would not be considered personal, or which would not be considered as quasi-identifiers, i.e. movie ratings, can in fact become so. They thus demonstrate that almost any dataset can identify individuals, and disclose private information, in this case political and religious views, and sexual preferences.

- On differential privacy

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.

Having been cited almost four thousand times on google scholar, this article is a reference in the anonymisation literature. Two years after introducing differential privacy, Dwork takes a step back and goes over three differentially private algorithms and theories found in the literature. She analyses the results and compares them. Though technical, the article has a clear introductory section defining what differential privacy roughly consists in.

- On l-diversity and t-closeness

Machanavajjhala, A., Gehrke, J., Kifer, D., & Venkitasubramaniam, M. (2006, April). l-Diversity: Privacy Beyond k-Anonymity. In null (p. 24). IEEE.

In this article, the authors challenge k-anonymity and introduce the concept of l-diversity. They analyse two attacks on k-anonymous datasets to demonstrate the technique's privacy issues. They then go over their own concept -capable of countering such attacks- in detail. This new anonymisation technique was another breakthrough in anonymisation science, and the article was cited over 4000 times.

Li, N., Li, T., & Venkatasubramanian, S. (2007, April). t-closeness: Privacy beyond k-anonymity and l-diversity. In Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on (pp. 106-115). IEEE.

This article addresses the shortcomings of k-anonymity and l-diversity and introduces the concept of t-closeness, another prominent anonymisation technique.

Anonymisation methods and techniques

The literature on the topic does not provide a unified categorisation of anonymisation methods, so the one proposed here is of our own device based on the literature and practices.

For the purpose of this text, an anonymisation *method* refers to the different possible changes made to a dataset, whereas *techniques* use the different methods to reach a certain level of anonymisation.

- Methods

To anonymise a dataset, the set needs to go through a process of alteration. The literature generally agrees that there are two ways to alter data: by *disrupting* it, *aggregating* it or by *suppressing* it. The following section delves into these methods.

○ Disruption

▪ Noise addition

Noise addition consists in making the attribute values in the dataset less accurate. To this end, one adds "noise" to the value prior to publication. This noise is another value drawn from a random distribution. As an example, a person's height will only be accurate to more or less 10 centimetres. The choice of the distribution from which the noise is drawn determines the level of privacy achieved. The WP29 (2014) stresses that it is not a sufficient method in and on itself and must be paired with others. Furthermore, some authors have argued that noise addition modifies the correlation coefficients of a dataset, leading to skewed statistics. A solution could be to mask the data with correlated noise addition. However, it must be noted that adding correlated noise may not provide more protection in terms of individual privacy.

- Permutation

Also known as swapping, “the basic idea behind the method is to transform a database by exchanging values of confidential attributes among individual records.” (Domingo-Ferrer, 2016). Experts agree that this method is not very used in practice. However, the WP29 (2014) argues that it is useful when the exact distribution of each attribute within the dataset is to be retained. The opinion mentions the example of a medical dataset containing the attributes “reasons for hospitalisation/symptoms/department in charge”. Due to the strong logical relationship between these three fields, simply swapping one would not be sufficient and could be reversed.

○ Generalisation

Bayardo and Agrawal (20015) oppose generalisation to the previously mentioned methods because the values making up a record remain truthful. The WP29 gives the following definition of the generalisation method in their Opinion paper (2014): “Generalisation consists in generalising the attributes of data subject by modifying the respective scale or order of magnitude.” By rendering the data less precise, generalisation strikes a better balance between utility and privacy than suppressing identifiers (Sweeney, 2002b). Example: Instead of writing in the precise salary amount for each individual, generalised interval values are given. 28,450$ thus becomes 20,000-30,000$. Instead of writing a precise birthdate, only the birth year is given. As the range is made larger, the utility of the data decreases, but conversely ensures a higher level of privacy. The first and last range may be made larger than the other intermediary range to satisfy privacy requirements, in a method of generalisation known as top and bottom coding.

○ Suppression

Another way to anonymise data is to directly suppress parts of data in order to prevent attacks that can link it back to individuals or enable inferences about them. Suppression, also known as *masking*, can mean the deletion of values, be it entire records, or entire attributes. It is acknowledged that simply suppressing/masking an attribute is not sufficient in itself, and it becomes less effective as the amount of quasi-identifiers in the dataset augments (Lubarsky, 2017; Narayanan, 2008). It must also be noted that the deletion must be permanent, as simply making the attribute not visible does not reduce the risk as the information still exists. This permanent change can be achieved either through deleting the data value itself, or by masking it with a constant symbol. The Personal Data Protection Commission of Singapore (2018) explains that suppression, though having the advantage of being irreversible, changes the dataset’s statistics, significantly lowering its utility.

Techniques

K-anonymity

K-anonymity is based on the generalisation of attributes to ensure that one individual cannot be singled out by grouping them with k-individual, thus making every record indistinguishable from at least k-1 other records with respect to quasi-identifiers (Machanavajjhala, 2006). The concept was introduced by Latanya Sweeney and Pierangela Samarati in 1998 (Li, 2007; Sweeney, 2002a). According to them, K-anonymity was not the first attempt at anonymising data. Indeed, as early as the mid-80s, the census bureau already had a threshold rule of minimum 15 people and 5 households sharing the same attribute to disclose certain records (Lawrence H. Cox 1985). However, there was no formal foundation for anonymity against linkage, and K-anonymity soon became the standard for creating anonymity algorithms (Meyerson and Williams, 2004; Bayardo and Agrawal 2005; LeFevre et al., 2005). K-anonymity has since then been criticised for not been a sufficiently robust technique to protect privacy. Li (2007) argues that though it protects against identity disclosure, it does not protect against attribute disclosure, Machanavajjhala *et al.* (2006) argue that k-anonymity is susceptible to homogeneity and background knowledge attacks. A background knowledge attack occurs when the attacker has background information which can lead to the re-identification of an individual. A homogeneity attack occurs when there is a lack of diversity for a given sensitive attribute. Example: If an attacker is looking for the medical record of Roger, a 53-year-old man living in a neighbourhood with the ZIP-Code 03019, and all the males over 50 living in that neighbourhood in the attacked database have breathing issues, it could infer that Roger has breathing issues.

ZIP Code	Age	Disease
0301*	≥50	breathing issues
0301*	≥50	breathing issues
0301*	≥50	breathing issues
030**	4*	cancer
030**	4*	pneumonia
030**	4*	flu

Table 1. Example of table that could lead to an inference attack.

l-diversity and t-closeness

To tackle the limitations of k-anonymity, Machanavajjhala *et al.* (2006) introduced the concept of l-diversity in 2006, though they acknowledge that a specific instance of their concept had been previously discussed by Ohrn and Ohno-Machado in 1999. l-diversity demands at least l amount of values to represent a sensitive attribute in each equivalent class. Li *et al.* (2007), argued against l-diversity, by purporting that it may not raise the level of privacy, due to its inability to take into account semantics: a table may be l-diverse but still allow for the inference of a value, or at least a range of that value. For instance: a health record table may successfully hinder an attacker from finding out the disease of an individual but know that this individual has a salary within a certain range, as visible in the table below.

ZIP Code	Age	Salary	Disease
080** 080** 080**	2* 2* 2*	6K 9K 8K	pneumonia flu bronchitis
0801* 0801* 0801*	≥30 ≥30 ≥30	12K 14K 16K	flu gastritis gastric ulcer
080** 080** 080**	4* 4* 4*	15K 10K 13K	cancer pneumonia flu

Table 2. A 3-diverse table

t-closeness was introduced by Ninghui Li, Tiancheng Li and Suresh Ventakasubramanian in 2007 in response to k-anonymity and l-diversity. In a t-diverse dataset, “the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table”.

○ Differential Privacy

Differential privacy is a privacy model first put forward by Cynthia Dwork, Frank McSHerry, Kobbi Nissim and Adam Smith in 2006. The term is broadly used to refer both to the technique proposed by Dwork *et al.* (2006), as well as to the privacy definition associated to this technique. Differential privacy reduces privacy risks when releasing aggregated statistics. Though the utility of the data may be lowered, and the risk for re-identification still exist, this is a good way to minimise it. Dwork (2008) explains that there are two ways to release statistical data containing personal information, one interactive and one non-interactive. In the non-interactive context, the data is used to create statistics, the statistics are published unilaterally and the raw, source data is destroyed, i.e., it will not be used anymore. In this case privacy concerns may stem from the statistics themselves though, and as such must be correctly anonymised to prevent re-identification. In the interactive setting on the other hand, the holder of the database sits between the end user and the database: the user makes a query into the database, the output of which are statistics. This allows for more targeted statistics which better fit the needs of end users, but means that data cannot be destroyed, thus incurring heightened privacy risks.

In this latter setting, differential privacy would work in the following way (Jain, 2016):

- An analyst wishing to access the data makes a query
- The privacy guard (the intermediary software) analyses the query and evaluates its privacy risks
- The privacy guard accordingly distorts the original dataset choosing noise to obfuscate the result of the query according to a differentially privacy function and provides it to the analyst.

Differential privacy, however, is not infallible. Pyrgelis *et al.* (2017) argue that differential privacy is not immune to temporal as well as spatial correlation attacks. The advocacy organisation *Access Now* did a series of articles on differential privacy in 2017, with one specifically focused on this technique's shortcomings, including the "privacy budget", which they define as "the limits on the privacy loss that any individual or group is allowed to accrue to protect the privacy of the data." (Cyphers, 2017) If the privacy *budget* is not well set, an attacker, by querying over and over the same dataset, can accumulate knowledge, thus exposing the data. *Collusion* is another issue: if two or more parties make similar queries and then share them, the anonymised value can be better inferred. The more parties make the same query, the more accurate the data becomes. *Correlation* is the last issue discussed, and a controversial one within the research community. There can be a correlation between various data points, which could give up information which was not meant to be disclosed. For instance, the value of a monthly rent may help infer the value of a salary. A way of approaching this issue is by treating the two queries (value of the salary and value of the rent) as one and the same. This solution, however, makes analysis more difficult.

○ Synthetic data traces

Donald B. Rubin framed a proposal for synthetic datasets to address privacy risks when publishing anonymised datasets in *Statistical disclosure limitation* in 1993. His proposal was related to, though more radical, than that of Roderick J. Little, published the same year and entitled "Statistical analysis of masked data". A synthetic dataset is a dataset which randomly generates data that follows certain statistics or internal relationships present in the original dataset (Domingo-Ferrer, Sánchez and Soria-Comas, 2016). According to these authors, there are three types of synthetic datasets:

- Fully synthetic datasets, where every data item has been synthesised,
- Partially synthetic data sets, where only some variables of some records are synthesised (usually the ones that present a greater risk of disclosure),
- Hybrid datasets, where the original data is mixed with the synthesised data.

The utility of the synthetic data depends on the accuracy of the adjusted model. Thus, if the synthesised data is close enough to the original dataset in terms of characteristics, the utility of the changed dataset is not lowered. This would mean, according to Rubin (1993), complete privacy with maximum utility. However, Bindschaedler and Shokri (2016) point out an issue arising when synthetic data is used to query location-based services. In their opinion, the system drowns out the real query with fake ones, but these often fail to mimic the behaviour of humans, thus exposing the real query. Synthetic traces thus need to resemble real traces. Even if the initial proposals in this space are not new, research on synthetic data has recently sparked more interest and its applications are beginning to be explored. Therefore, use cases are still scarce. This is, however, one of the most promising approaches to anonymisation at the moment.

Assessing anonymisation

Applying the techniques above does not guarantee that anonymisation will be achieved. Depending on the nature of the dataset and the potential attacks, the parameters of the algorithms may provide different levels of protection. Therefore, it is important to assess anonymisation risks both *before* and *after* implementing the suitable anonymisation techniques.

One of the most crucial steps to take in order to define which are the best methods and techniques to be used with a dataset, is to assess the *threat model* and consider the potential attacks that a dataset can be subject to. Attacks on anonymity can take various forms, as described by the Working Party 29:

- The singling-out of an individual by isolating records identifying that individual. For instance, Sweeney calculated that with only gender, ZIP-Code and date of birth, an individual is re-identifiable 87% of the time, i.e., the anonymity set of a record with a given (gender, ZIP, DoB) is one (Sweeney, 2000).
- The linkability of two datasets, leading to the re-identification of an individual. If an attacker can use a public dataset or another available dataset to re-identify an individual through the correlation of both datasets. For instance, if both sets include the attributes sex, ZIP Code and date of birth, re-identification would be quite straight-forward.
- Inference as being the ability to infer the value of an attribute for a certain record.

The success of one, or a combination, of these attacks may lead to two types of re-identification according to a general consensus in the literature: identity disclosure and attribute disclosure (Lambert, 1993; Duncan and Lambert, 1986; Li, 2007). Li (2007) explains that identity disclosure often leads to attribute disclosure and both can be equally harmful. If a sensitive attribute is revealed, even if false, it can lead to a differential treatment of the individual to whom this attribute belongs.

Therefore, whenever considering releasing a database, the general threat model and the potential types of attack need to be described and the measures needed to minimise risks need to be designed. This is important not only to avoid reputational harm, but also for legal compliance, especially since the passing of EU’s Data Protection Regulation (GDPR), which defines and creates specific obligations in this regard, as we saw above.

After anonymising and releasing a dataset, it is also important to run an assessment of how much anonymity has been achieved. In other words: how difficult it is to re-identify people?

How much information can be inferred using the anonymised dataset? To this end one must i) define the adversary, and ii) define metrics to quantify this risk. Then, compute the likelihood of the adversary's success. There is very little in the literature on guidelines to assess the risk, but there are tools that can help build a robust risk assessment, provided that the threat model is well defined. Some literature exists on specific applications, and frameworks are available for location privacy (Shokri et al., p.11) or for text in health applications (Scaiano et al., 2016), and academics and hackers regularly expose vulnerabilities that help improve current anonymisation and data security approaches.

Bibliography

Arrington, M. (2006). *AOL Proudly Releases Massive Amounts of Private Data. [*online*]* Tech Crunch*.* Retrieved from https://techcrunch.com/2006/08/06/aol-proudly-releases-massive-amounts-of-user-search-data/?guccounter=1

Barbaro, M. and Zeller, T. (2006). *A Face Is Exposed for AOL Searcher No. 4417749*. [online] Nytimes.com. Available at: https://www.nytimes.com/2006/08/09/technology/09aol.html [Accessed 10 Aug. 2018].

Bayardo, R. J., & Agrawal, R. (2005, April). Data privacy through optimal k-anonymization. In *Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on* (pp. 217-228). IEEE.

Bindschaedler, V., & Shokri, R. (2016, May). Synthesizing plausible privacy-preserving location traces. In *Security and Privacy (SP), 2016 IEEE Symposium on* (pp. 546-563). IEEE.

Barocas, S., & Nissenbaum, H. (2014). Big data's end run around anonymity and consent. *Privacy, big data, and the public good: Frameworks for engagement*, *1*, 44-75.

Brasher, E. A. (2018). Addressing the Failure of Anonymization: Guidance from the European Union's General Data Protection Regulation. Colum. Bus. L. Rev., 209. Available at: https://cblr.columbia.edu/wp-content/uploads/2018/06/6_2018.1_Brasher_Final.pdf

Byun, J. W., Kamra, A., Bertino, E., & Li, N. (2007, April). Efficient k-anonymization using clustering techniques. In *International Conference on Database Systems for Advanced Applications* (pp. 188-200). Springer, Berlin, Heidelberg.

Cox, L. H. (1980). Suppression methodology and statistical disclosure control. *Journal of the American Statistical Association*, *75*(370), 377-385.

Cox, L. H., Johnson, B., McDonald, S. K., Nelson, D., & Vazquez, V. (1985, March). Confidentiality issues at the Census Bureau. In *Proceedings of the First Annaul Census Bureau Research Conference, Washington, DC: US Government Printing Office* (pp. 199-218).

Cox, L., Karr, A., Kinney, S., Domingo-Ferrer, J., Duncan, G., O'Keefe, C., & Shlomo, N. (2011). Risk-Utility Paradigms for Statistical Disclosure Limitation: How to Think, But Not How to Act [with Discussions]. *International Statistical Review / Revue Internationale De Statistique, 79*(2), 160-199. Retrieved from http://www.jstor.org/stable/41305021

Cyphers, B. (2017). Differential privacy, part 2: It's complicated - Access Now. [online] Access Now. Available at: https://www.accessnow.org/differential-privacy-part-2-complicated/ [Accessed 24 Aug. 2018].

Dalenius, T. (1977). *Towards a methodology for statistical disclosure control*. Statistisk Tidskrift 5:429-444.

Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications). Available at: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex%3A32002L0058

Domingo-Ferrer, J., Sánchez, D., & Soria-Comas, J. (2016). Database anonymization: privacy models, data utility, and microaggregation-based inter-model connections. *Synthesis Lectures on Information Security, Privacy, & Trust*, *8*(1), 1-136.

Duncan, G. T., & Lambert, D. (1986). Disclosure-limited data dissemination. *Journal of the American statistical association*, *81*(393), 10-18.

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006, March). Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference* (pp. 265-284). Springer, Berlin, Heidelberg.

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.

Garfinkel, S. L. (2015). De-identification of personal information. *NISTIR*, *8053*, 1-46.

Gedik, B., & Liu, L. (2005, June). Location privacy in mobile systems: A personalized anonymization model. In *Distributed Computing Systems, 2005. ICDCS 2005. Proceedings. 25th IEEE International Conference on* (pp. 620-629). IEEE.

Jain, P., Gyanchandani, M., & Khare, N. (2016). Big data privacy: a technological perspective and review. *Journal of Big Data*, *3*(1), 25. Available at: https://link.springer.com/content/pdf/10.1186%2Fs40537-016-0059-y.pdf

Kotschy, W. (2016). The new General Data Protection Regulation-Is there sufficient pay-off for taking the trouble to anonymize or pseudonymize data. Available at: https://fpf.org/wp-content/uploads/2016/11/Kotschy-paper-on-pseudonymisation.pdf

Lambert, D. (1993). Measures of disclosure risk and harm. Journal of Official Statistics-Stockholm-, *9*, 313-313.

LeFevre, K., DeWitt, D. J., & Ramakrishnan, R. (2005, June). Incognito: Efficient full-domain k-anonymity. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data* (pp. 49-60). ACM.

Li, N., Li, T., & Venkatasubramanian, S. (2007, April). t-closeness: Privacy beyond k-anonymity and l-diversity. In *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on* (pp. 106-115). IEEE.

Little, R. J. (1993). Statistical analysis of masked data. *Journal of Official Statistics-Stockholm-*, *9*, (pp. 407-407).

Lubarsky, B. (2017). *Re-identification of "Anonymized Data."* Georgetown Law Technology Review, 202.

Machanavajjhala, A., Gehrke, J., Kifer, D., & Venkitasubramaniam, M. (2006, April). l-Diversity: Privacy Beyond k-Anonymity. In *null* (p. 24). IEEE.

Meyerson, A., & Williams, R. (2004, June). On the complexity of optimal k-anonymity. In *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* (pp. 223-228). ACM.

Narayanan, A., & Shmatikov, V. (2008, May). Robust de-anonymization of large sparse datasets. In *Security and Privacy, 2008. SP 2008. IEEE Symposium on* (pp. 111-125). IEEE.

Ohm, P. (2009). Broken promises of privacy: Responding to the surprising failure of anonymization. Ucla L. Rev., 57, 1701. Available at: https://pages.uoregon.edu/koopman/courses_readings/phil407-net/ohm_broken_promises_privacy.pdf

Personal Data Protection Commission (2018). *Guide to Basic Data Anonymization Techniques*. Singapore.

Proposal for a Regulation of the European Parliament and of the Council concerning the respect for private life and the protection of personal data in electronic communications and repealing Directive 2002/58/EC (Regulation on Privacy and Electronic Communications). Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2017:0010:FIN

Pyrgelis, A., Troncoso, C., & De Cristofaro, E. (2017). What does the crowd say about you? Evaluating aggregation-based location privacy. *Proceedings on Privacy Enhancing Technologies*, *2017*(4), 156-176.

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679

Rubin, D. B. (1993). Statistical disclosure limitation. *Journal of official Statistics*, *9*(2), 461-468.

Samarati, P., & Sweeney, L. (1998). *Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression* (pp. 101-132). technical report, SRI International.

Scaiano, M., Middleton, G., Arbuckle, L., Kolhatkar, V., Peyton, L., Dowling, M., ... & El Emam, K. (2016). A unified framework for evaluating the risk of re-identification of text de-identification tools. *Journal of biomedical informatics*, *63*, 174-183.

Schneider, B. (2007). *Why 'Anonymous' Data Sometimes Isn't*. [online] Wired. Retrieved from https://www.wired.com/2007/12/why-anonymous-data-sometimes-isnt/

Shokri, R., Freudiger, J., Jadliwala, M., & Hubaux, J. P. (2009, November). A distortion-based metric for location privacy. In *Proceedings of the 8th ACM workshop on Privacy in the electronic society* (pp. 21-30). ACM.

Stalla-Bourdillon, S., & Knight, A. (2016). Anonymous Data v. Personal Data-False Debate: An EU Perspective on Anonymization, Pseudonymization and Personal Data. Wis. Int'l

LJ, 34, 284. Available at: https://fpf.org/wp-content/uploads/2016/11/16.10.29-A-false-debate-SSB_AK.pdf

Sweeney, L. (2000). Simple demographics often identify people uniquely. *Health (San Francisco)*, *671*, 1-34.

Sweeney, L. (2002a). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(05), 557-570.

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