

# Anonymisation and Compliance to Protection Data: Impacts and Challenges into Big Data

Artur Potiguara Carvalho<sup>1</sup><sup>a</sup>, Edna Dias Canedo<sup>1,2</sup><sup>b</sup>, Fernanda Potiguara Carvalho<sup>3</sup><sup>c</sup>  
and Pedro Henrique Potiguara Carvalho<sup>d</sup>

<sup>1</sup>Electrical Engineering Department (ENE), University of Brasília (UnB), P.O. Box 4466, Brasília - DF, Brazil

<sup>2</sup>Department of Computer Science, University of Brasília (UnB), P.O. Box 4466, Brasília - DF, Brazil

<sup>3</sup>Law School (FD), University of Brasília (UnB), Brasília - DF, Brazil

Keywords: Anonymisation, Big Data, Privacy, Governance, Compliance.

Abstract: Nowadays, in the age of Big Data, we see a growing concern about privacy. Different countries have enacted laws and guidelines to ensure better use of data, especially personal data. Both the General Data Protection Regulation (GDPR) in the EU and the Brazilian General Data Protection Law (LGPD) outline anonymisation techniques as a tool to ensure the safe use of such data. However, the expectations placed on this tool must be reconsidered according to the risks and limits of its use. We discussed whether anonymity used exclusively can meet the demands of Big Data and, at the same time, the demands of privacy and security. We have concluded that, albeit anonymised, the massive use of data must respect good governance practices to preserve personal privacy. In this sense, we point out some guidelines for the use of anonymised data in the context of Big Data.

## 1 INTRODUCTION

News about leakage of personal information on Social Network websites is almost an every day occurrence nowadays (Mehmood et al., 2016; Joyce, 2017). In this era of Big Data, one of the most widely discussed issues is privacy and protection of personal data (Liu, 2015; Lanying et al., 2015; Dalla Favera and da Silva, 2016; Ryan and Brinkley, 2017; Casanovas et al., 2017; Popovich et al., 2017; Pomares-Quimbaya et al., 2019; Huth et al., 2019). This is a concern worldwide, which highlights the need for reflection and solutions in areas such as law, governance, and technology structure.


In this context, several countries have sought strategies to preserve privacy and guide the use of Big Data. Big Data (BD) is a massive data processing technology (De Mauro et al., 2016), which often includes all kinds of personal data, while not providing clear guidelines on how to store them. This method of aggregation impacts data protection directly (Brkan, 2019; Mustafa et al., 2019; Koutli et al.,


2019; Fothergill et al., 2019; Pomares-Quimbaya et al., 2019). Because of this, “data minimisation” has been presented as a guiding principle of the regulation of such software.


In the European context, that expression is mentioned in the General Data Protection Regulation (GDPR), Article 5(1)(c)(Regulation, 2018), which posits personal data shall be: “adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed” (‘data minimisation’).


Also, the Brazilian General Data Protection Law (LGPD), contains similar wording in its article 6 (da República, 2018), that providing the “principle of necessity”. Under this principle, data processing should be limited to the minimum necessary to achieve its purposes, using only data that is relevant, proportionate and not excessive to the purposes of the processing.

Underlying these legal structures, Data Governance (DG) has been used to foster standardization of and quality control in internal data management. To accomplish this, “data minimisation” has been proposed as a way to rationalize otherwise costly and expansive DG (Fothergill et al., 2019; Ventura and Coeli, 2018). In this scenario, compliance with these legal and governance structures has become a guar-

<sup>a</sup>  <https://orcid.org/0000-0001-7463-1487>

<sup>b</sup>  <https://orcid.org/0000-0002-2159-339X>

<sup>c</sup>  <https://orcid.org/0000-0003-4934-5176>

<sup>d</sup>  <https://orcid.org/0000-0002-0110-4069>

antee of the viability of BD tools in a society increasingly concerned about data protection (Brkan, 2019; Casanovas et al., 2017; Huth et al., 2019). The challenge is to conciliate Personal Data Regulations (PDR) and BD mechanisms, mitigating friction between companies and governments.

In this paper, we investigate an important tool for the compliance of BD mechanisms with PDR: anonymisation<sup>1</sup> techniques. These are important because once anonymised, these data are exempt from the requirements of PDR, including the principle of “data minimisation” (Regulation, 2018).

To guide this work, we present the BACKGROUND exploring the limits of expectations placed upon this tool. The question is whether anonymisation used exclusively can meet the demands of the two apparently opposing systems, in example, demands presented by both PDR and the BD. The justification, about the choice of the problem in focus, is specified by pointing the difficulties of conceptualizing the term. In this moment, an overview of the academy’s work in the area is presented. We strive to counterbalance the advantages and risks of using anonymisation as a form of compliance. We raised the hypothesis that, although anonymisation is an important tool to increase data protection, it needs to be used with assistance from other mechanisms developed by compliance-oriented governance.

The main goal is to present anonymisation risks in order to promote better use of this tool to privacy protections and BD demands. In section Related Work, we raise the main bibliographical references for the subject. We point out as a research method the literature review and the study of a hypothetical case. In section Related Work, we bring the results obtained so far, which been compared, when we bring a brief discussion about areas prominence and limitations of this work.

We conclude that it is not possible to complete BD compliance with PDR and privacy protection exclusively by anonymisation tools (Brasher, 2018; Ryan and Brinkley, 2017; Casanovas et al., 2017; Ventura and Coeli, 2018; Domingo-Ferrer, 2019). To solve this problem we aim to conduct future research about frameworks that can promote good practices that, associated with anonymisation mechanisms, can secure data protection in BD environments.

<sup>1</sup>The term is spelled with two variants: “anonymisation”, used in the European context; or “anonymization” used in the US context. We adopt in this article the European variant because the work uses the GDPR (Regulation, 2018) as reference.

## 2 BACKGROUND

Many organizations have considered anonymisation through BD to be the miraculous solution that will solve all data protection and privacy issues. This belief, which has been codified in European and Brazilian regulations, undermines an efficient review of organisations’ data protection processes and policies (Brasher, 2018; Dalla Favera and da Silva, 2016; Ryan and Brinkley, 2017; Casanovas et al., 2017; Popovich et al., 2017). For this, in this work we investigate the following research question:

RQ.1 Is anonymisation sufficient to conciliate Big Data compliance with Personal Data Regulations and data privacy at large?

In order to answer RQ.1, the concept of anonymisation, its mechanisms, and legal treatment must be highlighted. The text preceding the articles of Regulation, pertaining to the European Economic Area, guides the anonymisation is point 26 (Regulation, 2018). It states that the “principles of data protection should apply to any information concerning an identified or identifiable natural person”. Therefore, the principles are not applied to anonymous data, namely, “to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable.”

The LGPD contains a similar exclusion in its Article 12 (da República, 2018). Regulators conclude that, once anonymous, information cannot violate privacy, because data can no longer be linked to an identified or identifiable person. However, this premise implies some challenges. First, the data can be considered personal even though it is not possible to know the name of the person to whom the data refers. This is because the name is just one way to identify a person, which makes it possible to re-identify data when a personal, nameless profile is provided.

Second, in a BD context, precisely because it deals with massive data, connecting information becomes extremely easy, even when it comes to metadata or data fragments. Thus, some easy anonymisation techniques, such as masking, can be effective in closed and smaller databases, but not in BD (Pomares-Quimbaya et al., 2019).

Besides, techniques such as inference are more easily applicable in BD contexts. Inference is one of the techniques where information, although not explicit, can be assumed through the available data. Analyzing the propositional logic, we can say that there is inference when three propositions A, B and C respect the following equations:

$$A \Rightarrow B \quad (1)$$

$$B \Rightarrow C \quad (2)$$

so,

$$A \Rightarrow C \quad (3)$$

In an example (using a Shapiro's CarPool World (Shapiro, 1995)) if A is "Betty is a driver", B is "Tom is a passenger" and C is "Betty drives Tom", we can say by inference that every time that Betty is a driver and Tom is the passenger, Betty drives Tom (equation 3).

In this sense, it is possible to emphasize that, when techniques such as inference are considered, anonymisation projects in a BD context become even more vulnerable, making data normally considered auxiliary to indicate personal data.

For those reasons, PDR specifies assumptions about what is considered identified or identifiable data. The introduction to GDPR states, in point 26 (Regulation, 2018), that to determine whether an individual is identified or identifiable, all reasonable means must be taken into account, namely, "all objective factors such as costs and the amount of time required must be considered for identification, taking into account the technology available at the time of processing and technological developments". We could highlight this classification as a third point of concern, since the text assumes that, in massive contexts, data identification is possible, depending on the effort employed to re-identify it.

The fourth point of concern would be the difficulties of determining the anonymity of a certain piece of, as this identification will depend on criteria that, although specified by law, will change according to technical advances or even by the specific analysis conditions. This causes constant uncertainty regarding anonymisation.

These four concerns converge to the point that it is not possible to sustain the unexamined belief in anonymisation as a surefire way of ensuring privacy in BD contexts, which leads us to the hypothesis of the present paper. As noted, anonymisation in BD involves risks, especially to user privacy. Therefore, we argue that anonymisation should be used with the assistance of other privacy mechanisms, so as to better manage this data. Considering this, we can define the hypothesis of this research as follows:

HP.1 BD compliance with the PDR and data privacy cannot be achieved solely by anonymisation tools.

This does not mean that anonymity is a useless tool. Instead, it is an excellent ally when using BD platforms as it is one of the most powerful privacy protection techniques. The point, however, is that it is not

possible to rely solely on this technique, leaving aside the use of privacy-oriented governance. This means that to some extent BD must also adapt to privacy, either by increasing data capture criteria in the sense of minimization or by strengthening the governance of such data, even if anonymised.

The Main Goal of this paper is to present anonymisation risks and promote the better use of this tool for BD and the necessary privacy protection. We intend to expose privacy threats related to the use of anonymisation as an alternative to PDR enforcement. Therefore, we point out that anonymity tools should follow the protection guidelines to foster privacy.

As a research method, we use literature review, exploring the evolution of the concept, classification, demands, improvements on anonymity. To demonstrate the weaknesses of the tool we present a hypothetical case study. Random anonymous data were organized within a BD platform and analyzed with basic data from a specific database structure. Thus, although the data were anonymised in relation to the platform itself, they could be re-identified when exposed to external data. The result of the hypothetical case were analyzed in light of current legislation, and will be discussed in the following sections.

To guide the consultation of PDR, in particular, the GDPR and LGPD, follows a comparative table 1 of regulations, which will be used throughout the paper.

Table 1: Comparative Table of Regulations.

Concepts	GDPR	LGPD
"Data minimization" concept	Article 5(1)(c)	Article 6(III)
Anonimisation concept	text preceding the articles of GDPR, point 26; Article 4(5)	Article 5(XI)
Personal data concept	Article 4(1)	Article 5(I)
Exclusion of anonymous data from personal data classification	text preceding the articles of GDPR, point 26.	Article 12
Processing concept	Article 4(2)	Article 5(X)
Sensitive data concept	text preceding the articles of GDPR, point 51.	Article 5(II), Article 11

## 2.1 Related Work

Back in 2015, H. Liu had already announced the challenge of managing legal frameworks, privacy protection, individual autonomy, and data applications. (Liu, 2015). In 2016, Mehmood et al. detailed a group of methods and techniques that provides encryption and protection to data inside BD (Mehmood et al., 2016). In the same year, Dalla Farvera and da Silva discuss veiled threats to data privacy in the BD era (Dalla Favera and da Silva, 2016). Still in 2016, Lin et al. presents a model considering differential privacy (varying by datasets privacy loss) (Lin et al., 2016).

In 2017, Ryan and Brinkley add the vision of the organization governance model to address the new protection data regulations (PDR) issues (Ryan and Brinkley, 2017). In that year, many other authors discussed the same subject (Casanovas et al., 2017; Popovich et al., 2017; Joyce, 2017). In 2018, Ventura and Coeli introduce the concept of the right to information in the context of personal data protection and governance (Ventura and Coeli, 2018), while Brasher (Brasher, 2018) criticize the current process of anonymisation in BD.

In 2019, Domingo-Ferrer (Domingo-Ferrer, 2019) summarizes the Brashers review, mainly in BD platforms, presenting the issues of anonymisation and its specificities. Jensen et al. (Jensen et al., 2018) discuss how to realize value with BD projects and the best practices to measure and control it. Mustafa et al. (Mustafa et al., 2019) indicate a framework about privacy protection for application in the health field. They present the threats of privacy involving medical data in the light of regulation (GDPR).

According to Brasher (Brasher, 2018), “anonymisation protects data subjects privacy by reducing the link-ability of the data to its subjects”, which is similar to the concept outlined by PDR. According to this definition, it is possible to highlight two types of data: Personally Identifiable Information (“PII”), which may include the quasi-identifiers and contains security liabilities concerning personal data, and Auxiliary Data (“AD”), which can reveal the subjects referenced. These two types of data must be handled separately by anonymisation, according to the risks inherent to each.

About quasi-identifiers, Brasher (Brasher, 2018) describes: “non-facially identifiable data that can be linked to auxiliary information to reidentify data subjects”. Mehmood et al. (Mehmood et al., 2016) complements: “The attributes that cannot uniquely identify a record by themselves, but if linked with some external dataset may be able to re-identify the records.” To exemplify that description, Mehmood et

al. (Mehmood et al., 2016) show an example (Figure 1) of link quasi-identifiers from records of medical application and movie reviews application:

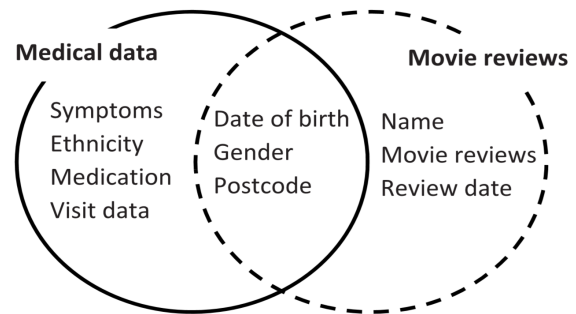


Figure 1: Quasi-Identifiers and Linking Records (Mehmood et al., 2016).

Brasher’s work (Brasher, 2018) presents the five most common anonymisation techniques: (1) **Suppression**, (2) **Generalization**, (3) **Aggregation**, (4) **Noise Addition**, and (5) **Substitution**, as shown in Figure 2.

- 1) **Suppression** is the process that excludes any PII from the base.
- 2) **Generalization** shuffles PII identifiers, without excluding any information, reducing their link-ability.
- 3) In **Aggregation**, both data types (PII and AD) go through some reducing treatment that maintains some properties of data (average, statistical distribution, or any others property) and also reduces their link-ability.
- 4) **Noise addition** adds some non-productive data to confuse the link between PII/AD and their subjects.
- 5) Finally, **Substitution** is similar to Generalization, while it differs in that: it shuffles not the identifier, but the value of the data itself, replacing the original dataset with other parameters. This process can be applied to both PII and AD (Brasher, 2018).

Mehmood et al. (Mehmood et al., 2016) also divides the privacy protection by anonymisation in five different operations: (1) **Suppression**, (2) **Generalization**, (3) **Permutation**, (4) **Perturbation**, and (5) **Anatomization**, all of which correspond to the strategies presented by (Brasher, 2018). This can be seen in Figure 2.

The **Suppression** (strategy 1 in Figure 2) strategy is criticized by Domingo-Ferrer (Domingo-Ferrer, 2019). Anonymising data in BD is not enough because re-linking the deleted identifiers becomes trivial in this massive context, especially if external data



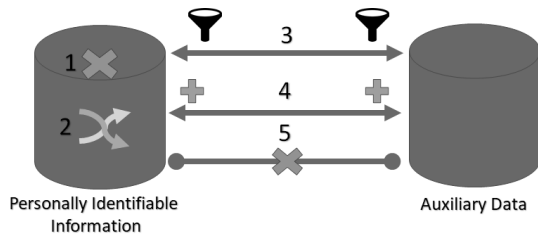


Figure 2: Anonymisation Techniques, Adapted from (Brasher, 2018; Mehmood et al., 2016).

is factored into the analysis. According to the author, concerns about the social impact of this insufficient protection are as great as to have surfaced on mainstream media (Domingo-Ferrer, 2019).

The author goes on to explain that efficient privacy protection must consider balancing these two aspects: utility loss and privacy gain of PII-based data. Supposed privacy gains occur at the expense of utility loss. When a suppressed piece of data is discarded less valuable information can be extracted (Domingo-Ferrer, 2019). BD anonymisation is still limited (Domingo-Ferrer, 2019). Domingo-Ferrer presents three main limitations to current big data anonymisation processes:

- 1) trust in data controllers, granted by PDR, is undermined by lack of actionable management criteria for the treatment of confidentiality;
- 2) the weakness of the anonymisation methods, which satisfy an insufficiently broad set of Statistic Disclosure Controls (SDC);
- 3) and the utility cost of the process of data anonymisation which may incur the difficulty of merging and exploring anonymised data.

Mehmood et al. (Mehmood et al., 2016) and Domingo-Ferrer (Domingo-Ferrer, 2019) agree about the trade-off between privacy by anonymisation and utility, and its negative relation mainly in the BD context. Applying some anonymisation strategy as the only action regarding data privacy leads to the decrease of potential insights on PII and AD.

Quoting the weakness of the anonymisation methods, Lin et al. (Lin et al., 2016) apply differential privacy to body sensor network using sensitive BD. In their work, Lin et al. (Lin et al., 2016) combine strategies 3 and 4 (figure 2) to hardening the privacy of a given dataset. But as shown, the scheme adopted by Lin only considers the information given by the internal dataset, ignoring possible attacks using other ADs available on the Internet, for example. Lin et al. (Lin et al., 2016) also discuss the risk of data loss through the anonymisation process.

### 3 PARTIAL RESULTS

To demonstrate a hypothetical example, we will use a data repository proposal on a BD platform whose inserted data represents customers of a financial institution.

Customer registration information (usually not just for financial companies) represents significant concentrations of personal data, sometimes, even sensitive. Besides, in the financial sector, it is possible to identify a customer through other non-conventional data (considered quasi-identifiers) such as identity, social register, driver's license, bank account number, among others.

The hypothetical example will use BD because, as already discussed, the large amount of data (and its intrinsic challenges) make the BD platform the infrastructure where it is easier to re-identify personal data once to treat its countless relationships (explicit or implicit) of personal data can be an arduous and expensive task in terms of processing.

Consider a certain data structure in a BD platform according to Figure 3:

This structure is implemented on a BD platform, to enable the analysis of the customer (current or potential) characteristics of a certain financial company. This analysis would contain personal data filters such as age, sex or relationship time with the company and will support several departments in this organization. Also datasets AUX\_CUSTOMER and CUSTOMER\_DETAIL were considered and classified according to Tables 2 and 3.

Now, we must consider the anonymisation applied by combining the strategies 1-5 described before, according to the showing:

- 1) Using strategy 1 (**Suppression**): The registers with CD\_CUSTOMER lower than 100500 were excluded from this table (from 100000 to 100500).
- 2) Using strategy 2 (**Generalization**): From 100500 to 100800, the identification was weakened by shuffling the CD\_CUSTOMER.
- 3) Using strategy 3 (**Aggregation**): The register with the same ID\_CPF (22464662100) was converted to a unique register (CD\_CUSTOMER = 603093, 603094, 603095 and 603096) by the sum of attribute value VL\_CURRENT\_CREDIT\_LIMIT and the max operation over attribute values DT\_EXPIRATION\_CREDIT\_LIMIT, DT\_REGISTER\_EXPIRATION, NB\_AGE and the min operation over attribute values DT\_BIRTH, DT\_CUSTOMER\_SINCE and DT\_ISSUE\_ID.

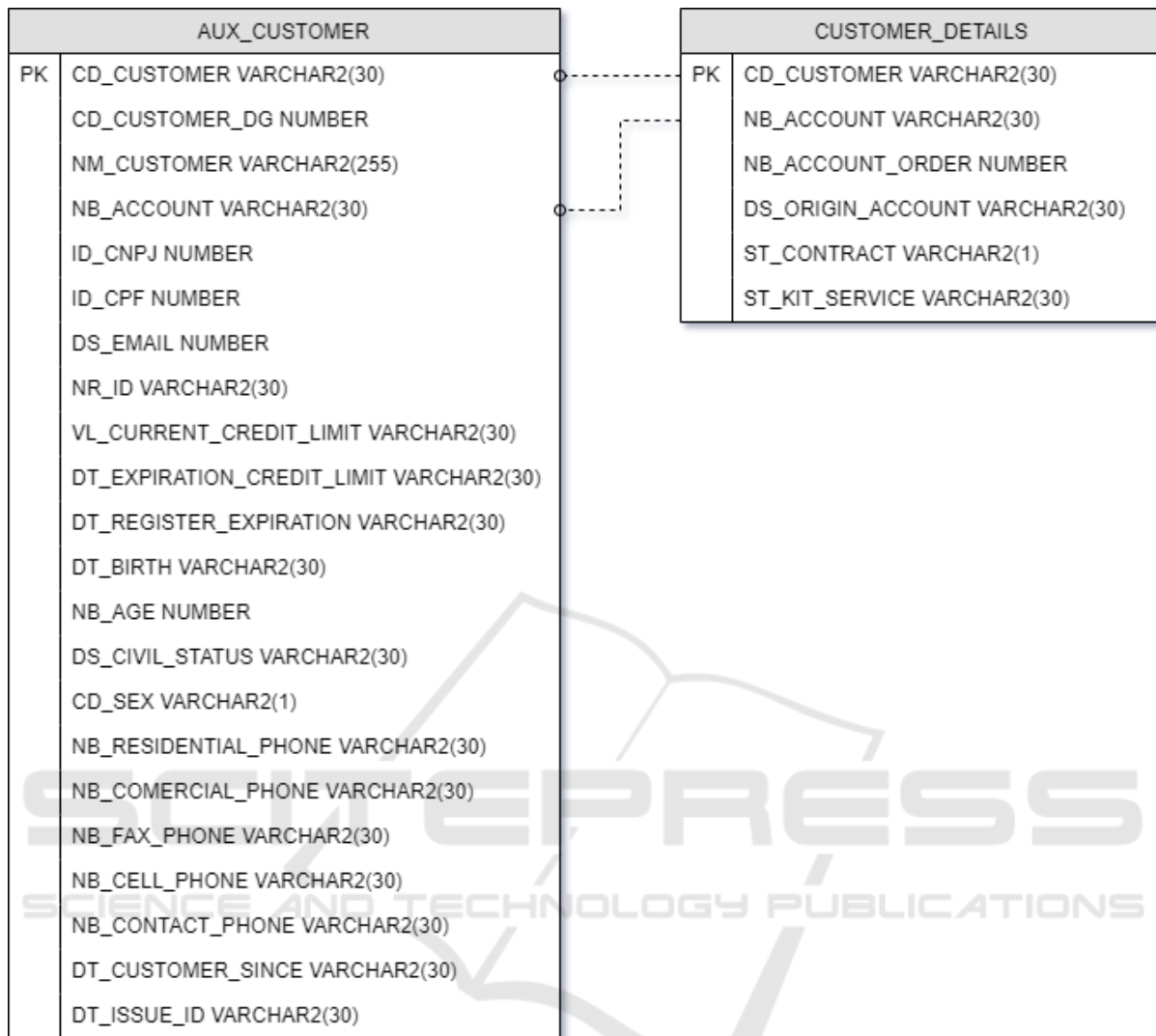


Figure 3: Hypothetical Structure Data Model.

- 4) Using strategy 4 (**Noise Addition**): Was included the register identified by the CD\_CUSTOMER 100623 with random information.
- 5) Using strategy 5 (**Substitution**): Was divided two groups of registers (G1 - from CD\_CUSTOMER 100800 to 101100 and G2 - from CD\_CUSTOMER 1013000 to 101600) and the AD attributes were shuffled between these two groups, preserving the original characteristics.

Based on the difficulty of transforming data privacy governance concepts into operational data protection actions (as described by Ventura and Coeli (Ventura and Coeli, 2018)), suppose that only part of the data in the structure shown by figure 3 has been classified as identifiable of the respective subject. Only the data contained in the dataset AX\_CUSTOMER

will be anonymised, excluding the data present in the dataset CUSTOMER\_DETAILS.

In the actual production environment, several reasons could lead to the BD information not being taken into consideration while in providing anonymisation, such as data governance process failures, misinterpretation of regulation, mistakes in internal concept of sensitive personal data, difficult to manage large amounts of data or many different datasets, among others.

Using another dataset (concerning customer details) from the same schema that was extracted from the previous customer table, it is possible to undo or disturb the anonymisation (weakening the privacy protection) according to the shown:

- 1) Concerning strategy 1 (**Suppression**): The registers excluded were identified (provide that the

Table 2: Attributes/classification of an Example Customer Table.

PII/AD	COLUMN NAME	DATA TYPE
PII	cd_customer	double
PII	cd_customer_dg	double
PII	nm_customer	string
PII	nb_account	double
PII	id_cnpj	string
PII	id_cpf	string
PII	ds_email	string
PII	nr_id	string
AD	vl_current_credit_limit	double
AD	dt_expiration_credit_limit	string
AD	dt_register_expiration	string
AD	dt_birth	string
AD	nb_age	double
AD	ds_civil_status	string
AD	cd_sex	string
AD	nb_residential_phone	string
AD	nb_comercial_phone	string
AD	nb_fax_phone	string
AD	nb_cell_phone	string
AD	nb_contact_phone	string
AD	dt_customer_since	string
AD	dt_issue_id	string

Table 3: Attributes/classification of a Example Customer Details Table.

PII/AD	COLUMN NAME	DATA TYPE
PII	cd_customer	double
PII	nb_account	double
PII	nb_account_order	double
AD	ds_origin_account	string
AD	st_contract	string
AD	st_kit_service	string

application of the anonymisation method was known) by the referential integrity (not explicit) with the table CUSTOMER\_DETAIL by the attribute CD\_CUSTOMER. Besides, exclusion is the most aggressive strategy, and produces the greatest loss of utility.

- 2) Concerning strategy 2 (**Generalization**): Using the attribute NB\_ACCOUNT (not search index, but personal data), it was possible to identify the shuffling, since this attribute can identify an individual.
- 3) Concerning strategy 3 (**Aggregation**): The presence of the register with the CD\_CUSTOMER = 603093, 603094, 603095 and 603096 in the table CUSTOMER\_DETAIL denouncement that these registers were manipulated in the original table.
- 4) Concerning strategy 4 (**Noise Addition**): The absence of the register with CD\_CUSTOMER = 100623 indicates that this register was added to the original table.

- 5) Concerning strategy 5 (**Substitution**): Combining the CD\_CUSTOMER and the NB\_ACCOUNT from these two tables its possible to identify the manipulation of these data, even if it is hard to define what exactly was modified.

Note that the data used to undo/detect the anonymisation process belonged to the same data schema as the original base. In the context of BD, it would be common that in the large universe of data there would be replications of PII or quasi-identifiers like the shown in the example.

Thus, it is possible that within the DB database there are reliable data to guide the conclusions against anonymisation. Besides, the data used to re-identify can be accessed by internet, social network, another BD or any other external data repository. Both present themselves as weaknesses in BD platforms, as they will provide insight into the anonymisation methods used.

Once the anonymisation method is detected, it is simpler to look for mechanisms to complete missing information or even rearrange and restructure information that has been merged or added noises.

For this, public databases can be an effective source for obtaining specific information.

Also, knowing which data has been anonymised greatly weakens database protection. This is because data that has not undergone the anonymisation process, for example, or data that is reorganized within the platform, will constitute a remnant base that maintains its integrity. Thus, unchanged data is known to be intact and can be used to obtain relevant information.

Finally, we clarified that all scripts used for create/populate the examples data structures are available:

```
CREATE TABLE
AUX.CUSTOMER
(
CD_CUSTOMER VARCHAR2(30)
PRIMARY KEY,
CD_CUSTOMER_DG NUMBER,
NM_CUSTOMER VARCHAR2(255),
NB_ACCOUNT VARCHAR2(30),
ID_CNPJ NUMBER,
ID_CPF NUMBER,
DS_EMAIL NUMBER,
NR_ID VARCHAR2(30),
VL_CURRENT_CREDIT_LIMIT
VARCHAR2(30),
DT_EXPIRATION_CREDIT_LIMIT
VARCHAR2(30),
DT_REGISTER_EXPIRATION
VARCHAR2(30),
```

```

DT_BIRTH VARCHAR2(30),
NB_AGE NUMBER,
DS_CIVIL_STATUS
VARCHAR2(30),
CD_SEX VARCHAR2(1),
NB_RESIDENTIAL_PHONE
VARCHAR2(30),
NB_COMERCIAL_PHONE
VARCHAR2(30),
NB_FAX_PHONE VARCHAR2(30),
NB_CELL_PHONE VARCHAR2(30),
NB_CONTACT_PHONE VARCHAR2(30),
DT_CUSTOMER_SINCE VARCHAR2(30),
DT_ISSUE_ID VARCHAR2(30) );
CREATE TABLE
CUSTOMER_DETAILS
(
CD_CUSTOMER VARCHAR2(30)
PRIMARY KEY,
NB_ACCOUNT VARCHAR2(30),
NB_ACCOUNT_ORDER NUMBER,
DS_ORIGIN_ACCOUNT VARCHAR2(30),
ST_CONTRACT VARCHAR2(1),
ST_KIT_SERVICE VARCHAR2(30) );

```

## 4 THREATS AND VALIDATION

### 4.1 The Hypothetical Case

Partial results have given us a perspective about the threats involved anonymisation. In some cases, as when attributes have been shuffled, comparative analysis with the table CUSTOMER\_DETAILS makes it possible to re-identify and rearrange the information.

But, in general, it was possible to conclude at least the existence of data processing. For example, when deleting data, comparison with the CUSTOMER\_DETAILS table reveals that information has been suppressed.

It means that the use of anonymisation is clear from a simple comparison with a table within the same database. This is true even with suppression, which is the most aggressive anonymisation technique.

This reveals which data has been modified, deleted or shuffled and provides a remnant base that maintains its integrity and can be used.

Also, it provides information to complete or organize all bases through external reinforcement, as with public base, as mentioned.

This requires that the comparison be based on information whose integrity is assured. Obtaining such secure information is not only possible but is likely,

mainly in the context of BD, that to take into account large databases, that are stored without effective governance. Also, it is likely that exist database there is unfeasible anonymisation due to the need to link information to users, as in the case of personalized services, within the same database.

Therefore, these anonymised data still present risks when they are indiscriminately shared on different bases, marketed or made available.

Lack of management increases the likelihood of leakage of this data, which could cause information to be easily obtained.

Therefore, anonymisation, taken in isolation, while providing a sense of security, contains factors that make its misuse extremely risky.

### 4.2 Validation

Considering the risks presented in this paper, and also based on the criticisms raised by (Domingo-Ferrer, 2019), we highlight some discussion points and propose, for each of them, guidelines to the use of anonymous data in the context of BD mechanisms.

- 1) Is data anonymised by comparing the entire company database or the public database?

As we discussed earlier, data is usually considered anonymous within their own platforms. However, anonymisation cannot neglect that, in our BD age, a large amount of data is available through other sources. We suggest that, as a minimum requirement for anonymisation to be considered effective, it should analyze its own database and - at least - other databases that are public, organized, and freely accessible.

- 2) Acknowledging the loss of utility caused by the anonymisation process, which data can and cannot be anonymised?

This is an important issue because, depending on the company's activity, anonymisation may be a technique that will render data unusable for certain purposes. If the company deals, for example, with personalized services, knowing which data relates to each customer becomes essential. Thus, companies need to choose which data to store, reducing the cost of maintaining large anonymised databases.

This is justified by the fact that maintaining anonymity requires continuous readjustment according to the evolution of the technique, as highlighted earlier. Besides, keeping smaller databases minimizes the risk of leakage, which increases as more data is stored. Finally, better choosing which data to anonymise forestalls



the need for an anonymous database not to be re-identified in order to meet business demands.

- 3) Is anonymisation a type of “processing of personal data”? While some researchers argue differently (Hintze and El Emam, 2018), we argue that anonymisation is a form of “processing of personal data”.

Once anonymised, the data can be used, even for purposes other than originally stated when it was collected as personal data, as we can see in article 6(4)(e) (Regulation, 2018) and the point 26 of the GDPR introductory text. But, for anonymisation to be considered a lawful processing method, it is necessary to follow the requirements outlined in the GDPR, Article 6 (1) (Regulation, 2018), such as the subject’s consent or vital interest. However, we highlight some criticisms of article 6 (1) (f) (Regulation, 2018), which will be explained in the next point.

- 4) Can anonymisation be applied by legitimate interests?

Article 6 (1)(f) (Regulation, 2018) stresses that data may be used for “legitimate interests” pursued by the controllers. On this point, the introduction of GDPR states in point 47 (Regulation, 2018) that:

“The legitimate interests of a controller, including those of a controller to which the personal data may be disclosed, or of a third party, may provide a legal basis for processing, provided that the interests or the fundamental rights and freedoms of the data subject are not overriding, taking into consideration the reasonable expectations of data subjects based on their relationship with the controller.”

Thus, legitimate interest is an abstract term that may be used to create a means of escaping regulation, rather than data protection (Zikopoulos et al., 2011). Furthermore, the same point 47 highlights that “At any rate the existence of a legitimate interest would need careful assessment including whether a data subject can reasonably expect at the time and in the context of the collection of the personal data that processing for that purpose may take place”.

As seen, the article 6(4)(e) (Regulation, 2018) and point 26 of the GDPR introductory text state that the anonymous data is not a personal data, and, therefore, processing of this data need not respect the original purpose of the data. Due to this, it is not possible to initially predict which purposes the data will serve after anonymisation.

In short, because defining legitimate interest involves a high degree of abstraction, in addition to the fact that once anonymised data can be used even for purposes other than the original, we argue that legitimate interests are insufficient to make anonymisation processing legal. On the other hand, we consider that due to risk, the best way to enact lawful processing of anonymisation is through given consent, without excluding remaining case applications described in Article 6 (1) (Regulation, 2018).

Importantly, based on the hypothetical case and the observations already exposed, we adhere to the position described by (Domingo-Ferrer, 2019) about the three main limitations to anonymisation. Also, we add the following observations:

- 1) Are data controllers trustworthy? Although granted by PDR, it is undermined by a lack of actionable management criteria for the treatment of confidentiality. Therefore, especially for anonymous data, it is likely that bad data processing will be detected only with data leakage. This is why we support stricter regulation of anonymity, as a way to increasing care with this data and promoting good governance practices for its management. This is a tool to consider objective factors for measuring trust in data controllers.
- 2) As with (Domingo-Ferrer, 2019) discourse, the many anonymisation methods proposed and its privacy models satisfies a specific SDC. A schema designed with focus in the decentralized process would fortify the data protection, approaching its issues holistically, especially because the BD platform requires the scalability property, both in terms of performance and infrastructure.
- 3) The utility cost of the data anonymisation process that can result in the difficulty of merging and exploiting anonymised data.

As indicated, anonymisation requires continuous improvement of its processing, considering the evolution of the techniques. So keep data anonymous on these platforms implies expenditure of maintenance resources.

Anonymisation also implies a reduction in the utility of data. It difficult its use in businesses using personalized services, as mentioned. These make in some cases unfeasible to use anonymisation tools. On the other hand, it is possible to overcome these hurdles in the future. One example is homomorphic encryption, which allows personalized services with anonymous data analysis without re-identifying this information. It seems

to be an alternative to this type of database analysis and to reduce risk. However, this tool needs to be refined to analyze the DB. Currently, homomorphic encryption requires an unreasonable amount of time to perform on DB platforms.

## 5 CONCLUSIONS

Both GDPR (Regulation, 2018) and LGPD (da República, 2018) describe anonymisation techniques as a tool to ensure the safe use of personal data and to exclude them from the rules governing data processing. However, as seen, the expectations placed on this tool should be reconsidered according to the risks and limitations of its use.

In the context of Big Data, even anonymous data does not ensure privacy. As highlighted, the tool has its own internal limitations. We have indicated four concerns related to the concept of anonymisation and the fact that this data is treated as distinct from personal data. They are:

- 1) subject identification when a profile is provided;
- 2) connecting information becomes extremely easy in a BD context with metadata or data fragments;
- 3) the legal concept of anonymisation accepts that, in massive contexts, data identification is possible, depending on the effort employed to re-identify it. Because of this concept, in order to mitigate risk derived from anonymisation, other mechanisms must be combined to improve the privacy protection;
- 4) difficulty in determining anonymity, as it depends on criteria that could change according to technical advances or even by the specific analysis conditions.

We also show that, compared to an internal data set, it is possible to discover the anonymisation technique used, and, in some cases, immediately re-identify subject data. Therefore, it is clear that anonymisation is not sufficient to reconcile Big Data compliance with Personal Data Regulations and data privacy. This does not mean that anonymisation is a useless tool, but it needs to be applied with the assistance of other mechanisms developed by compliance-oriented governance.

We conclude that anonymity used exclusively cannot meet the demands of Big Data and, privacy and security simultaneously. Besides, anonymisation needs to consider some other factors listed, such as interference from external data, such as public databases; the recognition of the loss of utility that this technique

involves; the need to comply with legal requirements for the processing personal data to promote anonymisation.

Some guidelines do not address anonymised data, these need to be required to manage such data through principles such as “data minimisation”. In conclusion, a Big Data-driven framework is required to recommend best practices that, coupled with anonymisation tools, ensure data protection in Big Data environments, while also addressing the compliance issue. We expect to investigate and present this proposal in future research.

## ACKNOWLEDGMENTS

This research work has the support of the Research Support Foundation of the Federal District (FAPDF) research grant 05/2018.

## REFERENCES

- Brasher, E. A. (2018). Addressing the failure of anonymization: Guidance from the European Union’s general data protection regulation. *Colum. Bus. L. Rev.*, page 209.
- Brkan, M. (2019). Do algorithms rule the world? algorithmic decision-making and data protection in the framework of the GDPR and beyond. *International journal of law and information technology*, 27(2):91–121.
- Casanovas, P., De Koker, L., Mendelson, D., and Watts, D. (2017). Regulation of big data: Perspectives on strategy, policy, law and privacy. *Health and Technology*, 7(4):335–349.
- da República, P. (2018). Lei geral de proteção de dados pessoais (lcpd). *Secretaria-Geral*, accessed in November 19, 2019. <https://www.pnm.adv.br/wp-content/uploads/2018/08/Brazilian-General-Data-Protection-Law.pdf>.
- Dalla Favera, R. B. and da Silva, R. L. (2016). Cibersegurança na união europeia e no mercosul: Big data e surveillance versus privacidade e proteção de dados na internet. *Revista de Direito, Governança e Novas Tecnologias*, 2(2):112–134.
- De Mauro, A., Greco, M., and Grimaldi, M. (2016). A formal definition of big data based on its essential features. *Library Review*.
- Domingo-Ferrer, J. (2019). Personal big data, GDPR and anonymization. In *International Conference on Flexible Query Answering Systems*, pages 7–10. Springer.
- Fothergill, D. B., Knight, W., Stahl, B. C., and Ulicane, I. (2019). Responsible data governance of neuroscience big data. *Frontiers in neuroinformatics*, 13:28.
- Hintze, M. and El Emam, K. (2018). Comparing the benefits of pseudonymisation and anonymisation under the GDPR. *Journal of Data Protection & Privacy*, 2(2):145–158.

- Huth, D., Stojko, L., and Matthes, F. (2019). A service definition for data portability. In *21st International Conference on Enterprise Information Systems*, volume 2, pages 169–176.
- Jensen, M. H., Nielsen, P. A., and Persson, J. S. (2018). Managing big data analytics projects: The challenges of realizing value. *Managing Big Data Analytics Projects: the Challenges of Realizing Value*.
- Joyce, D. (2017). Data associations and the protection of reputation online in australia. *Big Data & Society*, 4(1):2053951717709829.
- Koutli, M., Theologou, N., Tryferidis, A., Tzouvaras, D., Kagkini, A., Zandes, D., Karkaletsis, K., Kaggelides, K., Miralles, J. A., Oravec, V., et al. (2019). Secure iot e-health applications using vicinity framework and gdpr guidelines. In *2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, pages 263–270. IEEE.
- Lanying, H., Zenggang, X., Xuemin, Z., Guangwei, W., and Conghuan, Y. (2015). Research and practice of datarbac-based big data privacy protection. *Open Cybernetics & Systemics Journal*, 9:669–673.
- Lin, C., Wang, P., Song, H., Zhou, Y., Liu, Q., and Wu, G. (2016). A differential privacy protection scheme for sensitive big data in body sensor networks. *Annals of Telecommunications*, 71(9-10):465–475.
- Liu, H. (2015). Visions of big data and the risk of privacy protection: A case study from the taiwan health data-bank project. *Annals of Global Health*, 1(81):77–78.
- Mehmood, A., Natgunanathan, I., Xiang, Y., Hua, G., and Guo, S. (2016). Protection of big data privacy. *IEEE access*, 4:1821–1834.
- Mustafa, U., Pflugel, E., and Philip, N. (2019). A novel privacy framework for secure m-health applications: The case of the gdpr. In *2019 IEEE 12th International Conference on Global Security, Safety and Sustainability (ICGS3)*, pages 1–9. IEEE.
- Pomares-Quimbaya, A., Sierra-Múnera, A., Mendoza-Mendoza, J., Malaver-Moreno, J., Carvajal, H., and Moncayo, V. (2019). Anonymity: From a small data to a big data anonymization system for analytical projects. In *21st International Conference on Enterprise Information Systems*, pages 61–71.
- Popovich, C., Jeanson, F., Behan, B., Lefaire, S., and Shukla, A. (2017). Big data, big responsibility! building best-practice privacy strategies into a large-scale neuroinformatics platform. *International Journal of Population Data Science*, 1(1).
- Regulation, G. D. P. (2018). Eu data protection rules. *European Commission, Accessed in October 9, 2019*. [https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules\\_en](https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_en).
- Ryan, M. and Brinkley, M. (2017). Navigating privacy in a sea of change: new data protection regulations require thoughtful analysis and incorporation into the organization’s governance model. *Internal Auditor*, 74(3):61–63.
- Shapiro, S. (1995). Propositional, first-order and higher-order logics: Basic definitions, rules of inference, examples. *Natural Language Processing and Knowledge Representation: Language for Knowledge and Knowledge for Language*. AAAI Press/The MIT Press, Menlo Park, CA.
- Ventura, M. and Coeli, C. M. (2018). Para além da privacidade: direito à informação na saúde, proteção de dados pessoais e governança. *Cadernos de Saúde Pública*, 34:e00106818.
- Zikopoulos, P., Eaton, C., et al. (2011). *Understanding big data: Analytics for enterprise class hadoop and streaming data*. McGraw-Hill Osborne Media.