The Two Dimensions of Data Privacy Measures

Orit Levin *Javier Salido*

Corporate, External and Legal Affairs, Microsoft

Abstract

This paper describes a practical framework for the first phase of the design of privacy measures for big data that is independent from the content of data and applicable across different industries. The framework recognizes that the first two factors (i.e., "axis") to be considered in the design of data privacy measures are (1) the desired data usefulness, implying suitable de-identification techniques, and (2) the anticipated data sharing scenarios with the data protection measures feasible for use between the data source and the data user. Then, the framework shows the different levels of potential risk to data subjects for possible combinations of de-identification techniques and data sharing scenarios.

The conclusions provided by this initial phase will help to narrow the range of de-identification techniques a practitioner should consider and thus help to guide the detailed de-identification design. The paper also suggests using the framework to clarify guidance for the General Data Protection Regulation (GDPR) by specifying a set of high level guidelines that practitioners should consider in an initial phase of de-identification design. GDPR guidance related to "pseudonymization" and "organizational and technical measures" could be clarified by specifying the level of potential risk within the two axis in the form of the framework introduced in this paper

1 Introduction

It is commonly agreed among regulators, practitioners in different industries, and academics that data de-identification (a.k.a., anonymization) improves the privacy of data subjects. It is also well understood that the de-identification of data comes at the expense of its usefulness. As a result, the determination of suitable de-identification techniques with their parameters for each use case remains a complex job reserved for a small community of technically expert practitioners.

It is generally accepted that identifying appropriate de-identification techniques and mitigating a risk of re-identification is based on consideration of various factors. Several existing publications in the area (see Bibliography) propose a data-centric, risk-based approach to determine the right approach to de-identification for a specific use case. According to these approaches, data "sensitivity", dependent on the content of data, is one of the first factors to be considered in the risk analysis. As a result, it has been a challenge to produce a set of generic practical guidelines that would scale across different industries and use cases.

In this paper, we introduce a framework that allows a practitioner to narrow the set of suitable de-identification techniques before taking into consideration data specifics, and thus not requiring deep technical knowledge of de-identification techniques and statistics in a first phase. Inputs from stakeholders, perhaps without expert level knowledge of the specific de-identification techniques, about intended purpose and sharing of de-identified data can be included in this first phase of design or assessment of data privacy measures. Later on, these first phase results can be either further evaluated and refined by subsequently applying existing data-centric risk-based analysis for the particular use case. Performing calculations for the recommended approach, the specific dataset, and the particular use case on a narrowed range of applicable de-identification techniques should be constrained and thus more practical.

Terminology we use throughout the document:

- Data source: an entity in charge of the original data; can be the creator of the original dataset or the owner of individual data.
- Data user: a recipient or a user of the data (e.g. data analyst).
- Legal entity: an organization or an individual.

2 The Analysis

Known guidelines for the selection of appropriate de-identification techniques are not trivial to implement or evaluate. This is caused by the need to consider a long list of seemingly independent factors before choosing the de-identification techniques and the parameters for each use case.

In order to simplify this multi-dimensional problem, we aim to identify a small number of factors that fit the following condition: are necessary to consider, can be qualified in a way comprehendible to all stakeholders, and are representative of a broad range of cases. We start with examining the factors that are regularly included in a risk assessment process, and grouping them based on logical correlations between them:

(1) the *value* of data correlates to the level of incentives that an entity (i.e., a recipient or a user of the de-identified data) would have to use the data for purposes other than approved (thus becoming an attacker), and consequently correlates to the amount of resources that the entity will be willing to invest to re-identify the data;
(2) the *sensitivity* of data correlates to the amount of harm in case of data re-identification, which includes two aspects: (1) harm to data subjects which is based on the content and the level of detail in the data, and (2) harm to the data source, which may depend on the number of data subjects in the dataset;
(3) the *types of attacks* that need to be taken into consideration (such as prosecutor or journalist) depend on all of the factors in this list, but mainly they depend on the content of de-identified data;
(4) the *fitness to purpose* of data (a.k.a., data usefulness or utility) after it has been de-identified correlates to the de-identification techniques used;
(5) the *data sharing scenario* defines the extent to which the data user is bound to preserve the privacy of subjects in the dataset and stipulates the controls feasible for the particular sharing model.

Now we can observe two distinct categories of factors: those dependent on the data content and those that are independent. We will use this distinction to specify a set of high level guidelines which are independent from the content of specific data and applicable across different industries.

Factors in (1) through (3) above are data content dependent. As a result, they need to be examined in the industry context or the application context, and are specific to each use case.

Factors in (4) and (5) are independent from the data content and are applicable across different industries, applications, and use cases. We also observe that it is possible to choose a single factor from (1) and (2) that represents the rest of the factors in the group and fits the condition set above.

Data usefulness depends on properties of the de-identified data, which in turn are a function of the de-identification techniques being applied to the data. ISO/IEC JTC1 CD 20889 “Privacy enhancing data de-identification techniques” classifies known techniques for de-identification of tabular data and describes their characteristics. For the purpose of the framework, we will order the techniques by the extent to which the data retains its structure and content after being de-identified (See Figure 1).

Data sharing scenarios are characterized by the data protection measures used between the data source and the data user, and can include technical, organizational, and legislative measures. We discuss the sharing scenarios in Section 4 below. In order to keep the framework independent from the data content, risk assessments against data breaches by illegitimate parties, or deliberate violation of agreement between the sharing parties are considered complementary to our discussion. These are incremental components that can be performed in the second stage separately from this framework using conventional data-centric risk-based calculations.

3 The Proposed Model

The discussion above leads to a model that arranges the data privacy enhancing measures in two dimensions: de-identification techniques (horizontal axis) and data sharing scenarios (vertical axis) as shown in the chart below. For each use case, the point of intersection between these two axes characterizes to what extent the privacy of data subjects is protected from disclosure by the data user.

Figure 1: The Model

9	Raw data is collected from individuals	FFS	FFS	FFS	FFS	FFS	FFS	FFS	FFS	FFS	FFS	FFS	NA
8	De-identified data is collected from individuals	NA	NA	FFS	FFS	FFS	FFS	FFS	NA	NA	NA	NA	O
7	Data is published to the general public	I	I	I	I	I	I	I	R	R	R	O	O
6	Public access is provided to data	I	I	I	I	I	I	I	R	R	R	O	O
5	Data is published under SLA or contract	R	R	R	R	R	R	R	O	O	O	O	O
4	Access to data is provided under SLA or contract	R	R	R	R	R	R	R	O	O	O	O	O
3	Data is published within a legal entity	R	O	O	O	O	O	O	O	C	C	C	C
2	Access to data is provided within a legal entity	R	O	O	O	O	O	O	O	C	C	C	C
1	Data is restricted to an atomic legal entity	O	C	C	C	C	C	C	C	C	C	C	C
	Sharing scenario / De identification technique	None	Pseudonymization with controlled re-identification	Pseudonymization	Masking of identifiers	Masking of outliers and selective quasi-identifiers	Generalization of selective quasi-identifiers	Randomization of selective quasi-identifiers	Implementing K-anonymity model for quasi-identifiers	Creating synthetic data	Generating aggregated data / statistics	Implementing DP server model	Implementing DP local model
		1	2	3	4	5	6	7	8	9	10	11	12

Key:

C	O	R	I	FFS	NA
Conservative	Optimal	Risky	Inappropriate	For future study	Not applicable

NOTE: The levels of potential risk shown in the table are only examples based on the authors' knowledge and experience to illustrate the ways in which the framework can be used by different stakeholders as discussed in the following sections. As such, the details of data collection scenarios are left for future study at this stage of our research.

The focus of the proposed framework is on structured data that can be presented in a tabular form. Such data contains records related to data subjects and is called microdata. We order the techniques by the degree to which the data retains its structure and content after being de-identified. Techniques 2 to 9 retain their records format with increasing data loss; techniques 10 and 11 provide statistical results; technique 12 can be used to generate both effectively anonymous microdata and accurate statistics. For description of the de-identification techniques, we use the terminology defined in ISO/IEC JTC1 CD 20889 "Privacy enhancing data de-identification techniques" (see Annex for more information).

The data sharing scenarios are described in the next clause.

4 Data Sharing Scenarios

The data sharing dimension indicates the extent to which a legitimate data user is bound to preserve the privacy of subjects in the dataset. We assume that the protection of data privacy from illegitimate data users is achieved through conventional data protection measures, including technical security controls such as data encryption or access control mechanisms.

In order to keep the framework independent from the data content, we assume full reliability of technical security measures implemented by both parties: the data source and the data user.

Specifically, the framework doesn't address the possibility of data breaches beyond the boundaries protected by the data source (a.k.a., the "trust boundary"), which might include the leak of the original data, de-identified data, or the meta-data generated for the purposes of de-identification or controlled re-identification. The risk from such a data breach can be calculated independently from this framework as it would have been calculated for the original data.

Furthermore, the framework doesn't address the possibility of de-identified data breaches as a result of a leak of the de-identified data beyond the boundaries agreed or declared to be protected by the data user. The risk of such a data breach would depend on the content of the de-identified data and, if needed, could become a subject of a detailed risk assessment analysis. As a result, it could be covered by a separate compensation clause based on the risk calculations.

Based on all the assumptions above, the scenarios in which data is being collected from individuals by illegitimate parties or for illegitimate purposes are not covered by our framework. These scenarios would typically be covered by national laws and regulations for specific industries and use cases.

We classify the common data sharing scenarios in the next sections.

1. Data is restricted to an atomic legal entity

A single legal entity or an organization plays the role of both the data source and the data user. Typically, the data protection measures against improper use or exposure of data by employees is covered by the contract between the organization and an employee. In this case, data privacy protection measures would be implemented by adding clauses specific to different aspects of data privacy to the internal contract with the employees.

2. Access to data is provided within a legal entity

The data source and data user are two entities within a single organization. The data source allows authorized internal parties outside of the data source entity to use the data by providing access to a protected dataset. In this case, the data source implements protection controls in the form of access control lists or authentication protocols. The organization has the ability to request the authorized internal parties sign a special internal contract with the organization as appropriate for the case.

3. Data is published within a legal entity

The data source and data user are two entities within a single organization. The data source allows authorized internal parties outside of the data source entity to use the data by creating separate instances (a.k.a., publishing) of the dataset. In this case, the data source implements protection controls in the form of access control lists or authentication protocols. The organization has the ability to request the authorized internal parties sign a special internal contract with the organization as appropriate for the case.

4. Access to data is provided under SLA or contract

An organization playing the role of the data source provides access to another organization, or an individual, under a signed SLA or a contract. The signed agreement contains conditions regarding the data use, and limitations regarding data re-identification by the data user. The agreement might also include a compensation clause covering the case of data breach on the data user side. Note that this case is different from the next case in that the data source retains control over the dataset and can monitor its use on an ongoing basis.

5. Data is published under SLA or contract

An organization playing the role of the data source provides a dataset to another organization or an individual under a signed SLA or a contract. The signed agreement contains conditions regarding the data use, and limitations regarding data re-identification by the data user. The agreement might also include a compensation clause covering a case of data breach on the data user side.

6. Public access is provided to data

An organization that is playing the role of a data source selectively provides information from or about a protected dataset to the public with or without specifying limitations in addition to those in existing regulations or law.

7. Data is published to the general public

An organization that is playing the role of a data source makes a dataset available to the public with or without specifying limitations in addition to those in existing regulations or law.

8. De-identified data is collected from individuals

In this case, the data is being de-identified at the point of collection in a way transparent to the data source. Note that such “data collecting” is sometimes being referred to as “data reporting”.

This row can be further broken down into the following more specific cases:

- An individual playing the role of the data source provides the data to an organization under a signed written agreement. The agreement covers the usage of data including the re-identification limitations as requested by the data source. The agreement might also include a compensation clause covering a case of data breach on the data user side.

- Data from an individual who becomes a data source is collected with a written public or implicit clause detailing the usage of data and its re-identification limitations.
- Data from an individual who becomes a data source is collected without the knowledge of the individual for legitimate purposes.

We include this case for completeness, but leave the discussion about it for future study.

9. Raw data is collected from individuals

In this case, the collecting entity, which plays a role of data user, receives the data and then performs its de-identification. This row can be further broken down into more specific cases as in the previous case.

We include this scenario for completeness, but leave the discussion about it for future study.

4.1 Use by Regulators and Policy Makers

Regulators may define a "sufficiently protected area" within the two axes such that the level of data protection inside this area would be considered sufficient in the eyes of data subjects and regulators and applicable to a wide range of industries and use cases.

It is also possible to create different instances of this model tailored to more specific needs, such as per industry, per data sensitivity, or per geopolitical area.

For example, we think that the GDPR guidance related to "pseudonymization" and "organizational and technical measures" could be clarified by specifying a region with an acceptable level of potential risk within the two axis in the form of the framework introduced in this paper.

4.2 Use by Practitioners

Given a specific use case, in the first phase of a de-identification system design, a practitioner would need to identify the desired usefulness of data and the projected data sharing scenarios that correspond to the two axes of the chart. The "guidelines" retrieved by identifying the applicable areas on the chart could be shared with different internal stakeholders, who would then be able to provide their feedback on the intended usefulness of the data, the acceptable level of risk, and therefore the appropriate de-identification techniques before more resource-consuming analysis and calculations are performed.

Afterwards, a detailed risk-based analysis mainly related to the content of data (and based on the factors (1) to (3) from Page 2 of this document), can be performed to ensure that neither party is taking a risk exceeding its accepted limit. The results of the risk-based analysis can help to tune the exact placement of the use case on the chart as long as it remains within the "protected area".

5 Conclusion

In this paper, we introduced a practical framework for designing or assessing privacy measures for big data by identifying the factors considered when choosing de-identification techniques, and separating those factors into two distinct sets.

We used ISO/IEC JTC1 CD 20889 "Privacy enhancing data de-identification techniques" as the source for terminology, classification, and understanding of the characteristics of known techniques for de-identification of tabular data. We introduced a list of common data sharing scenarios characterized by the availability of measures constraining a legitimate data user to preserve the privacy of subjects in the dataset.

By isolating and abstracting the data-independent factors, we were able to provide a two-dimensional "check list" that practitioners can consider in an initial phase of data privacy measures design, independent from the content of a specific dataset, and applicable across different industries and use cases.

We also suggested that such a framework can help to write high level guidelines by policy makers to clarify existing and new data protection regulations such as GDPR.

6 Bibliography

[1] Khaled El Emam, Luk Arbuckle, *Anonymizing Health Data: Case Studies and Methods to Get You Started*, December 2013.

[2] National Institute of Standards and Technology, NIST IR 8053, *De-identification of Personal Information*, October 2015.

[3] Information and Privacy Commissioner of Ontario, *De-identification Guidelines for Structured Data*, June 2016.

[4] ISO/IEC JTC1 CD 20889, work in progress.

Appendix: De-identification techniques and data usefulness

The ISO/IEC JTC1 CD 20889 "Privacy enhancing data de-identification techniques" classifies known techniques for de-identification of tabular data and describes their properties. In this Appendix, we provide a brief overview of the techniques with an emphasize on the resultant data usefulness. We order the techniques by the degree to which the data retains its structure and content after being de-identified.

1. None

Data that has not be de-identified retains its original form and usefulness.

2. Pseudonymization with controlled re-identification

Identifier is defined as set of attributes in a dataset that enables unique identification of a data subject within a specific context. Pseudonymization of a data record in a dataset means replacing the identifier (or identifiers) of the data subject with a pseudonym in order to obscure his or her identity. While pseudonyms don't reveal the identities of the data subjects directly, they still allow linking of records belonging to a particular data subject across datasets. In a sense, pseudonymization fully retains the original data format and usefulness.

Pseudonymization with controlled re-identification means implementing specific measures so that the de-identified data can be fully re-identified (i.e., linked to the original data subjects) in a deterministic way. In this case, not only the data retains its original usefulness, but re-association with the original data subjects, when performed, is guaranteed to be correct.

3. Pseudonymization

Pseudonymization without controlled re-identification, means that the mapping between the identifiers and the pseudonyms is intentionally destroyed as a part of the local policy. Pseudonymized data retains its original usefulness, but linking the pseudonyms to the original data subjects cannot be guaranteed in all cases and with full certainty.

4. Masking of identifiers

Removing identifiers from a dataset is the most basic form of de-identification. Typically, it would be the first step before additional de-identification techniques are applied to a data set. The resultant de-identified data retains its original format and usefulness with the exception of two properties: (1) the ability to link between records relating to a particular data subject across multiple datasets and (2) the ability to re-identify a data subject in a deterministic way.

5. Masking of outliers and selective quasi-identifiers

Quasi-identifier is an attribute in a dataset that, together with other attributes that may be in the dataset or external to it, enables unique identification of a data subject within a specific context. In addition to removing identifiers, removing outliers and selective quasi-identifiers, makes re-identification by inference more difficult. The loss of data usefulness is relatively small and depends on the use case. The removed values can be marked as such, allowing data analysts to adjust their statistical calculations accordingly.

6. Generalization of selective quasi-identifiers

Generalization of a value of an attribute means reducing its granularity. Generalization can be performed on attribute values containing numbers, addresses, etc. While generalization might greatly decrease data usefulness (e.g. quality), it retains "data truthfulness" on record level meaning that the data can be used in legal procedures, etc.

7. Randomization of selective quasi-identifiers

Randomization of a value of an attribute means altering the value such that it differs from the original value in a random way. Randomizing can be achieved in many different ways such as noise addition, permutation, or microaggregation. Randomization doesn't preserve data truthfulness.

8. Implementation of K-anonymity for quasi-identifiers

The objective of K-anonymity is to reduce the number of unique combinations of values for subsets of attributes, so that each subset of values is shared by at least K multiple records in the resultant dataset, making it difficult to single out a specific data subject. Masking, generalization, and microaggregation can be applied to different types of attributes in a data set to achieve the desired results. K-anonymity is a systematic approach that allows to quantify the risk of re-identification in mathematical terms.

9. Creation of synthetic data

A synthetic data set is one that has been generated artificially, using a number of techniques, and contains no "real" personal data. Nevertheless, the synthetic data set is representative in a statistically relevant way of some "real" personal data set. This approach has been used successfully for years by entities like the Census bureau and the Department of Education in the U.S.A. to create row-level data sets, representative of the relevant populations, that can then be made available to the public at large. A number of techniques exist that have successfully been used to generate synthetic datasets from "real" ones, including permutation (a.k.a., row-level data swapping) and differential privacy.

10. Computation of statistics

Refers to the computation of certain quantities, such as mean, mode, median, variance, percentiles, etc. that summarize the characteristics and/or the behavior of the population, or segments of the population, in a data set. The result of this process are specific measures, graphs and equations that, unlike other

techniques described in this document, aim primarily at describing the aforementioned characteristics and behaviors, rather than affording privacy to the individuals whose information is part of that same data set.

11. Implementation of Differential Privacy server model

Differential Privacy is a mathematical model that quantifies the privacy loss in a dataset when aggregate data about it is released. Privacy loss refers to the cumulative knowledge that a theoretical attacker (that aims to re-identify the dataset's records) acquires over time from the aggregate data that has been released. Differential privacy provides strong guarantees that the presence or absence of any particular data principal in the dataset cannot be inferred from the de-identified dataset or from system's responses. These guarantees are maintained even if the attacker has access to other, related, datasets, so long as the privacy loss is limited to a certain level.

In the server model of Differential Privacy, personal data has been collected and is stored in a central database applying minimal or not de-identification protective measures to the raw data. Analysts or applications then query the database by interacting with a software agent, known as the "curator", that receives the query, fetches the data from the database, computes the correct statistical answer, and then responds to the analyst or application with a randomized version of the correct answer.

12. Implementation of Differential Privacy local model

In the client model, de-identification takes place at the end-user device by applying a selected randomization function to the individual data, before the data is reported back to a central database. The randomized data is being made available to the data users (e.g., analysts or applications) either as a synthetic microdata or as a set of statistics.

The local model is essential in scenarios where the entity that is receiving the differentially private reports is not trusted by the individuals providing the data, or as a measure taken by the entity receiving the reports to minimize risk and costs associated with having to protect the personal data.