The Two Dimensions of Data Privacy Measures

Ms. Orit Levin, Principle Program Manager

Corporate, External and Legal Affairs, Microsoft

Abstract

This paper describes a practical framework that can be used in the first stage of the design of privacy measures for big data in a way that is independent from the content of data and applicable across different industries. First, the framework recognizes that the first two factors to be considered in the design of data privacy measures are (1) the desired data utility (with its corresponding de-identification techniques), and (2) the anticipated data sharing scenarios (with the corresponding feasible data security measures). Then, the framework shows the different levels of suitability 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 to clarify the General Data Protection Regulation (GDPR) statements using the described framework by specifying a set of high level guidelines that practitioners should consider in an initial phase of de-identification design.

1 Introduction

It is commonly agreed among regulators, different industries, and academics that some sort of data de-identification (a.k.a., anonymization) improves the privacy of data subjects. It has also been well understood that the de-identification of data comes at the expense of its utility. 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 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, 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 the data specifics, and thus not requiring deep technical knowledge of de-identification techniques and statistics. Inputs from stakeholders without expert level knowledge of the specific de-identification techniques can be included in this first phase of design or assessment of data privacy measures. After the initial phase of consideration, the high-level results can be either further refined or contested by subsequently applying existing data-centric risk-based analysis for the particular use case. Performing the detailed calculations for the recommended approach,

the specific dataset, and the particular use case on a narrowed range of applicable de-identification techniques should be scalable and thus more practical.

We are using the following terminology in our discussion 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.

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 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 a recipient or a user of the de-identified data would have to exploit the data for purposes other than intended, and consequently correlates to the amount of resources that the entity will be willing to invest to implement the exploit;
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 details in the data, and (2) harm to the source of data, which additionally is based on the number of data subjects in the dataset;
3) the types of attacks that need to be taken into consideration (such as incidental, 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 utility) after it has been altered 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 division to specify a set of high level guidelines which are independent from the content of specific data and applicable across different industries.

Groups 1 through 3 are data content dependent. The factors in these groups require consideration not only based on the content of the data, but they also need to be examined in the context of the industry or the application and are specific to each use case.

Groups 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 group 1 and from group 2 that represents the rest of the factors in the group and fits the condition above, i.e., a factor that is necessary to consider, can be qualified in a way comprehensible to all stakeholders, and is representative of a broad range of cases.

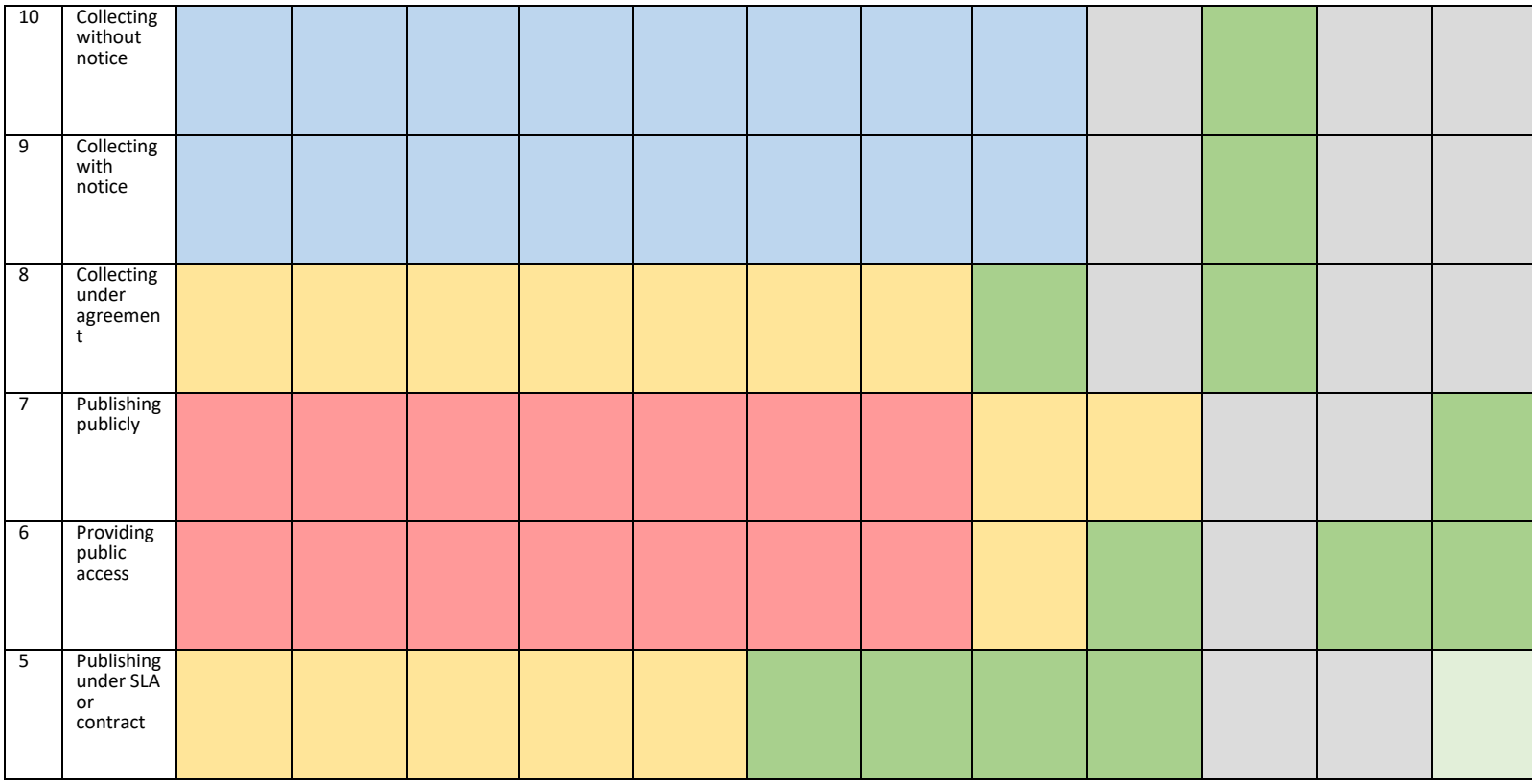
Data utility depends on data properties, which in turn are a function of the de-identification techniques being applied to the data. ISO/IEC JTC1 working draft standard 20889 “Privacy enhancing data de-identification techniques” classifies known techniques for de-identification of tabular data and describes 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 determined by the data protection measures used between the data source and the data user that can include technical, organizational, and legislative measures. We discuss the sharing scenarios in Section 4 below. Note that these measures are complementary to any existing measures to protect the data from illegitimate parties. In order to keep the framework independent from the data content, the risk assessment due to data breaches by illegitimate parties or due to deliberate violation of agreement between the sharing parties are not included in this framework. 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 and data sharing scenarios 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



4	Providing access under SLA or contract												
3	Publishing within a legal entity												
2	Providing access within a legal entity												
1	Restricting to an atomic legal entity												
	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	Implementing DP client model	Implementing DP server model	Generating aggregated data / statistics
		1	2	3	4	5	6	7	8	9	10	11	12

Note: 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. Techniques 2 to 10 retain their records format, while techniques 11 and 12 provide statistical results.

Key:

Conservative	Recommended	Inappropriate	Data content dependent	Prohibited	Not applicable

Note: The colored key is a suggestion only 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.

3.1 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 statements related to "pseudonymisation" and "organizational and technical measures" could be clarified by providing high level guidance in the form of the framework introduced in this paper.

3.2 Use by Practitioners

Give a specific use case, in the first phase of a de-identification system design, a practitioner would need to identify the desired utility 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 be able to provide their high level considerations before more resource-consuming analysis and calculations are performed.

If required, afterwards, a detailed risk-based analysis mainly related to the content of data (and based on the factors in groups 3 to 5 above), 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".

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.

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 include the possibility of data breaches beyond the boundaries protected by the data source, 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. If needed, 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 include 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. The scenarios in which data is being collected from individuals by legitimate users and for legitimate purposes without an agreement with or a knowledge of the individuals heavily depend on the data content. As such they are exceptions to the general case and cannot be addressed by our framework. These scenarios would typically be covered by laws and regulations for specific industries and use cases.

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

1. Data never leaves 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 an 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. Providing access 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 common dataset. In this case, the data source implements protection controls in the form of access 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. Publishing 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 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. Providing access under SLA or contract

An organization playing the role of the data source provides access for another organization or an individual under a signed SLA or a contract. The signed agreement contains conditions regarding the data usage 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.

5. Publishing 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 usage 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. Providing public access

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 the existent in the regulations or law.

7. Publishing publicly

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 the existent in the regulations or law.

8. Collecting under agreement

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.

Note: In cases where the data is being de-identified at the point of its collection, “data collecting” is sometimes being referred to as “data reporting”.

9. Collecting with notice

Data from an individual who becomes a data source is collected with a written public or otherwise clause detailing the usage of data and its re-identification limitations.

10. Collecting without notice

Data from an individual who becomes a data source is collected without the knowledge of the individual.

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 working draft standard 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 classified by the availability of measures bounding 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

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

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

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