ITUPublications
Recommendations

International Telecommunication Union
Standardization Sector

Supplement

ITU-T X Suppl.39 (09/2023)

SERIES X: Data networks, open system communications and security

Supplements to ITU-T X-series Recommendations

ITU-T X.1148 – Supplement on requirements for data de-identification assurance

ITU-T X-SERIES RECOMMENDATIONS

Data networks, open system communications and security

PUBLIC DATA NETWORKS	X.1-X.199
OPEN SYSTEMS INTERCONNECTION	X.200-X.299
INTERWORKING BETWEEN NETWORKS	X.300-X.399
MESSAGE HANDLING SYSTEMS	X.400-X.499
DIRECTORY	X.500-X.599
OSI NETWORKING AND SYSTEM ASPECTS	X.600-X.699
OSI MANAGEMENT	X.700-X.799
SECURITY	X.800-X.849
OSI APPLICATIONS	X.850-X.899
OPEN DISTRIBUTED PROCESSING	X.900-X.999
INFORMATION AND NETWORK SECURITY	X.1000-X.1099
SECURE APPLICATIONS AND SERVICES (1)	X.1100-X.1199
CYBERSPACE SECURITY	X.1200-X.1299
SECURE APPLICATIONS AND SERVICES (2)	X.1300-X.1499
CYBERSECURITY INFORMATION EXCHANGE	X.1500-X.1599
CLOUD COMPUTING SECURITY	X.1600-X.1699
QUANTUM COMMUNICATION	X.1700-X.1729
DATA SECURITY	X.1750-X.1799
IMT-2020 SECURITY	X.1800-X.1819

For further details, please refer to the list of ITU-T Recommendations.

Supplement 39 to ITU-T X-series Recommendations

ITU-T X.1148 – Supplement on requirements for data de-identification assurance

Summary

De-identified data incurs the risk of re-identifying individuals. It is therefore important to assess the threat of de-identified data being used to identify individuals through re-identification methods. De-identification methods, which can be used for re-identification risk assessment, may be selected according to the following considerations:

– Data risk assessment: Data composition, data distribution, possession of other data;

– Data use environment risk assessment: Confidence level of data recipient, impact during re-identification, inadvertent re-identification;

– Using and managing de-identification data: Security measures for de-identification data, monitoring of re-identification possibilities, compliance with de-identification data provision or consignment contracts.

This Supplement defines data de-identification assurance. It also provides a set of requirements for managing data de-identification assurance, including data risk assessment, risk assessment of the data use environment, and using and managing de-identified data.

History [*]

Edition	Recommendation	Approval	Study Group	Unique ID
1.0	ITU-T X Suppl. 39	2023-09-08	17	11.1002/1000/15528

Keywords

De-identification assurance, requirement.

[*] To access the Recommendation, type the URL https://handle.itu.int/ in the address field of your web browser, followed by the Recommendation's unique ID.

FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications, information and communication technologies (ICTs). The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of ITU. ITU-T is responsible for studying technical, operating and tariff questions and issuing Recommendations on them with a view to standardizing telecommunications on a worldwide basis.

The World Telecommunication Standardization Assembly (WTSA), which meets every four years, establishes the topics for study by the ITU-T study groups which, in turn, produce Recommendations on these topics.

The approval of ITU-T Recommendations is covered by the procedure laid down in WTSA Resolution 1.

In some areas of information technology which fall within ITU-T's purview, the necessary standards are prepared on a collaborative basis with ISO and IEC.

NOTE

This is an informative ITU-T publication. Mandatory provisions, such as those found in ITU-T Recommendations, are outside the scope of this publication. This publication should only be referenced bibliographically in ITU-T Recommendations.

INTELLECTUAL PROPERTY RIGHTS

ITU draws attention to the possibility that the practice or implementation of this publication may involve the use of a claimed Intellectual Property Right. ITU takes no position concerning the evidence, validity or applicability of claimed Intellectual Property Rights, whether asserted by ITU members or others outside of the publication development process.

As of the date of approval of this publication, ITU had not received notice of intellectual property, protected by patents/software copyrights, which may be required to implement this publication. However, implementers are cautioned that this may not represent the latest information and are therefore strongly urged to consult the appropriate ITU-T databases available via the ITU-T website at http://www.itu.int/ITU-T/ipr/.

Table of Contents

Supplement 39 to ITU-T X-series Recommendations

ITU-T X.1148 – Supplement on requirements for data de-identification assurance

1 Scope

This Supplement defines data de-identification assurance. It also provides a set of requirements for managing data de-identification assurance, including data risk assessment, risk assessment of data use environment, and using and managing de-identified data.

2 References

[ITU-T X.1148] Recommendation ITU-T X.1148 (2020), *Framework of de-identification process for telecommunication service providers.*

3 Definitions

3.1 Terms defined elsewhere

This Supplement uses the following terms defined elsewhere:

3.1.1 adversary [b-ISO/IEC 27559]: An individual or unit that can, whether intentionally or not, exploit potential vulnerabilities.

NOTE – Adversary, attacker, intruder, snooper, and other similar terms are often used interchangeably in the de-identification literature.

3.1.2 anonymization [b-ISO/IEC 29100]: Process by which personally identifiable information (PII) is irreversibly altered in such a way that a PII principal can no longer be identified directly or indirectly, either by the PII controller alone or in collaboration with any other party.

3.1.3 assurance [b-ISO/IEC/IEEE 15026-1]: Grounds for justified confidence that a claim has been or will be achieved.

3.1.4 data recipient [b-ISO/IEC 27559]: The person or organization with whom data is accessed, shared or released.

3.1.5 de-identified data [b-ISO TR 18307]: Data resulting from personally identifiable information (PII) after the process of removing or altering one or more attributes so that the (direct or indirect) identification of the relevant person without knowledge of the initial information is either impossible or requires an unreasonable amount of time and manpower.

3.1.6 de-identified dataset [b-MEDSEC]: Dataset resulting from the application of a de-identification process.

3.1.7 de-identification technique [b-ISO/IEC 20889]: Method for transforming a dataset with the objective of reducing the extent to which information is able to be associated with individual data principals.

3.1.8 de-identification process [b-ITU-T X.1058]: Process of removing the association between a set of identifying data and the data principal, using de-identification techniques.

3.1.9 equivalence class [b-ISO/IEC 20889]: Set of records in a dataset that have the same values for a specified subset of attributes.

3.1.10 personally identifiable information [b-ISO/IEC 29100]: Any information that (a) can be used to identify the personally identifiable information (PII) principal to whom such information relates, or (b) is or might be directly or indirectly linked to a PII principal.

NOTE – To determine whether a PII principal is identifiable, account should be taken of all the means which can reasonably be used by the privacy stakeholder holding the data, or by any other party, to identify that natural person.

3.1.11 PII controller [b-ISO/IEC 29100]: Privacy stakeholder (or privacy stakeholders) that determines the purposes and means for processing personally identifiable information (PII) other than natural persons who use data for personal purposes.

NOTE – A PII controller sometimes instructs others (e.g., PII processors) to process PII on its behalf while the responsibility for the processing remains with the PII controller.

3.1.12 PII processor [b-ISO/IEC 29100:2011]: privacy stakeholder that processes personally identifiable information (PII) on behalf of and in accordance with the instructions of a PII controller.

3.1.13 pseudonymization [b-ISO/IEC 20889]: De-identification technique that replaces an identifier (or identifiers) for a data principal with a pseudonym in order to hide the identity of that data principal.

NOTE 1 – Pseudonymized data can be restored to its original state with the additional information which then allows individuals to be re-identified, while anonymized data can never be restored to its original.

NOTE 2 – Additional information is auxiliary information created during the pseudonymization process which is required to be managed separately from the original information. Additional information is required to be kept separately from the pseudonymized data.

3.1.14 re-identification [b-ISO/IEC 20889]: Process of associating data in a de-identified dataset with the original data principal.

3.2 Terms defined in this Supplement

This Supplement defines the following terms:

3.2.1 data de-identification assurance: The confidence that a data de-identification process meets a set of security requirements.

3.2.2 dataset: Collection of data.

NOTE – This definition is based on [b-ISO 19115-1] but the word 'identifiable' has been deleted from the definition.

4 Abbreviations and acronyms

This Supplement uses the following abbreviation:

PII Personally Identifiable Information

5 Conventions

This Supplement uses the term 'data subject' with the same meaning as 'data principal' as used in clause 3.

6 Overview

6.1 Concept of de-identified data

"De-identification" is the process of removing the association between a set of identifying data and the data subject. The fundamental objective of de-identification is to protect the privacy of individuals because once de-identified, a dataset is considered to no longer contain personally identifiable information (PII). If a dataset does not contain PII, its use or disclosure will not violate the privacy of individuals.

De-identified data are created from PII by the process of removing or altering one or more attributes so that the direct or indirect identification of the relevant person without knowledge of the initial information is either impossible or requires an unreasonable amount of time and human resources [b-Kor-GuidelineROK].

De-identification is an important tool which protects the privacy and identity of the individuals associated with the data for all types of businesses.

When using data including personal information for research, marketing, testing applications, statistical trending or other legitimate purposes, the involvement of specific individuals has to be clarified in order to meet the data usage goals. In such cases, de-identification of PII is highly recommended.

6.2 When to use de-identified data

The degree of use varies depending on the de-identification method such as, for example, pseudonymous information or anonymous information. Since there is risk of identifying an individual through additional information, pseudonymous information should be utilized with technical and managerial security measures for limited purposes such as for preserving records of public interest, scientific research and statistical purposes.

Anonymous information is intended to make it impossible to connect, recover the original data and identify an individual considering all reasonably expected means and their cost, time and technological development. Due to these characteristics, such data are then freely available including for disclosure.

6.3 Data re-identification attack type

Data re-identification attacks can be divided into three types:

1) Where the adversary knows that targeted individual data are in the dataset;
2) Where the adversary does not, or cannot, know whether targeted individual data are in the dataset;
3) Attacks on all individuals rather than on a targeted individual whose data may be in the dataset.

6.4 Re-identification risk

Data de-identification assurance requirements depend on the re-identification risk for that de-identified data. Organizations may need to specify how the data re-identified risk can be measured. The overall risk of re-identification is equal to the data re-identification risk multiplied by the context re-identification risk. The details for data re-identification risk measurement and context re-identification risk measurements can be found in Appendix I.

6.5 Adequacy assessment of de-identified data

De-identified data generally refers to data from which PII has been removed relating to an individual's identity in a given dataset. Some datasets may indirectly reveal the identities of a specific person or individuals even when the data seemingly contains no PII. In addition, some de-identified datasets may contain "re-identification codes". Re-identification codes might, for example, allow researchers to match two anonymous datasets when conducting a study.

Consequently, before using de-identified data, the adequacy of the de-identified data should be assessed in order to assure that de-identification has been processed to the level sufficient to protect privacy[b-Kor-Guideline].

7 Security requirements for data de-identification assurance

De-identification assurance refers to determining whether or not to disclose de-identified data through evaluation in order to ensure that de-identified data has been adequately processed to a level that can sufficiently reduce the risk of re-identification. Organizations that want to (re)use de-identified data shall consider first, the data itself, second, the data-use environment, and third, the requirements for safe use and management of the data to determine whether to disclose them during de-identification processing or de-identified data adequacy assessment. Analysis of the data itself can consider the existence of PII data (data composition), data distribution, and on whether there is possession of other data (see clause 8). The data use environment refers to the confidence level of the de-identified data recipient, the impact of re-identification and inadvertent re-identification. Data use and management refers to the security measures for the de-identified data, the monitoring of re-identification possibilities and compliance with de-identified data provision or consignment contracts. Figure 7-1 shows security requirements for data de-identification assurance.

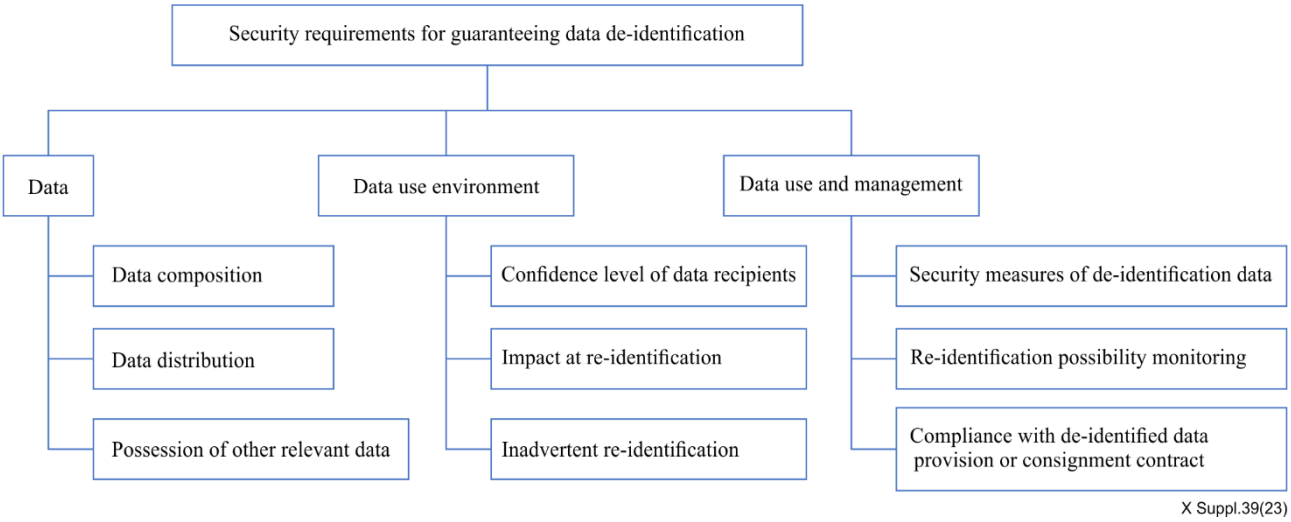

Figure 7-1 – Classification of security requirements for data de-identification assurance

8 Requirements for data risk assessment

This clause discusses requirements for data itself that are subject to de-identification, and concerns identifying whether there are PII elements in data subject to de-identification. Reviews should be carried out on, first, whether there are components of the data itself that are subject to de-identification; second, whether there are items that are identifiable by connecting with other items; and third, whether there are other data outliers such as any information that could be linked to the data in question. Clauses 8.1 to 8.3 should be selectively reviewed as risk factors in more details.

8.1 Data composition

SR 1: Organizations should identify the composition of data related to the data subject.

Data composition refers to the components of the data itself that are subject to de-identification. The following should be reviewed:

– Number of data columns: Are there many data columns that can be used, thus having high possibility for linking with other data columns? (Riskier when higher.)

– Statistical features of data: Is the data subject to de-identification research data for just one attribute or is it a study for comprehensively analysing multiple and various attributes? (The latter is riskier.)

– How data are provided: Is the data subject to de-identification provided a single time to another department in the same organization or to an outside third party, or is it provided two

or more times, or is it provided regularly during a certain period of time? (The more frequent provision is riskier.)

– Size of population: Is the population of the data subject to de-identification for all citizens, or is the population a sample of all citizens? (The former is riskier.) If it is a sample, what is the ratio?

– Hierarchical features: Does the data subject to de-identification have hierarchical relations that can have a higher possibility of personal identification such as family relations or positional relations in specific departments? (Riskier when there are hierarchical relations.)

– Time features: Do the values of items in each data column have time-based connectivity in terms of behaviour or location? (Riskier when there is time-based connectivity.)

– Up-to-date: How recent a period (i.e., less than 6 months ago, 1-2 years ago, 2 years or longer ago) does the data subject to de-identification comprise? (Riskier when data are more recent.)

– Rigidity: Do the values of each column in the data include columns that do not change values along with the passage of time? (Riskier when such columns are included.)

8.2 Data distribution

SR 2: Organizations should identify the distribution of data related to a data subject.

Data distribution refers to the components of the distribution of values within data columns that are subject to de-identification, and the following should be reviewed:

– Distribution: For example, in continuous numerical data, do the data (including data on one side along the distribution curve) on both ends of the distribution of data values have several records for data with *n* persons or fewer? (Riskier when the number of records is higher.)

  Whether or not there is a singular value: Are the values within data columns unique, i.e., do they have values that are unique from other values? (Riskier when they exist and more risky, the more there are.)

  - Occupational description from which personal identification is possible: For example, politicians in specific regions, the coach of a disabled women's fencing team, a professional baseball player in a certain area when age is included.
  - Data with high possibility of identification compared with disclosed data: For example, when the population in a specific area is too high compared with the population census data of the National Statistics Office.
  - Data that can easily specify individuals: For example, location data of vehicles departing from specific coordinates (terminals, stations, etc.) at specific times and arriving at specific coordinates, and vehicle data.
  - Data that is identifiable compared with the data of all citizens: For example, a person who is much older than the average person or far beyond the average life expectancy.
  - Data with increased identifiability of individuals by combining two or more columns: For example, street name and number with height.

8.3 Possession of other relevant data

SR 3: Organizations should identify other data that can be used to de-identify the data related to the data subject.

In the event that de-identification occurs within the organization in charge, other relevant data refers to personal knowledge or data that a PII controller or PII processor possesses excluding additional

information[1] that could be known during de-identification of original data. Furthermore, it refers to personal knowledge or data possessed by a third party when the de-identified data are provided to a third party outside the corresponding organization. In this regard, the UK Anonymisation Network (UKAN) defines other data as: "Other data are any information that could be linked to the data in question, thereby enabling re-identification. There are four key types of other data: personal knowledge, publicly available sources, restricted access data sources, and other similar data releases. A vital question is whether those other data are themselves identified or identifiable" [b-UKAN]. Controllers or processors in charge of de-identification must always review whether personal identification is possible by connecting data subjects to handling data and other data during processing.

9 Requirements for risk assessment of de-identified data use environment

According to the UKAN, "You must look at both the data and their context to ascertain realistic measures of risk. This is called the data situation approach. The basic intuition is that all data are held in some sort of context, which we call the data use environment. Formally, a data situation is the aggregate set of relationships between some data and the set of their environments." [b-UKAN]. The data use environment discussed in this clause also falls within the same context, and accordingly, the confidence level of the data recipient, the impact at re-identification and the risk for inadvertent re-identification shall be reviewed as risk factors.

9.1 Confidence level of data recipients

SR 4: Organizations should determine the confidence level of the data recipient when transferring the de-identified data.

The confidence level of the data recipient is to be considered when consigning de-identified data from one organization to an outside third party or when consigning personal data processing work including de-identified data to an outside third party. The higher the confidence level of the recipient provided with and using the de-identified data, the lower the risk for re-identification. The following sample questions should be reviewed:

- Is the organization using de-identified data from the government or from a commercial organization? (Riskier when the data are from a commercial organization.)
- Are there financial or commercial profits to be made from using de-identified data? (Riskier when the answer is yes.)
- Is there a possibility of gaining non-economic benefits (political advantage, manipulation of public opinion, self-interest, reputational benefits, etc.) or causing social issues by using the data? (Riskier when the answer is yes.)

[1] EU GDPR Preface Paragraph 26: "Personal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person".

9.2 Impact of re-identification

SR 5: Organizations should determine the impact of the re-identification of de-identified data.

The impact of re-identification refers to the possibility of violating the privacy of the subject of the data due to the re-identification of de-identified data, and the following sample questions should be reviewed:

- Is there a possibility of violating the privacy of the subject of the data due to the re-identification of de-identified data? (Riskier when the answer is yes.)
- Can re-identification of de-identified data result in discrimination against the subject of the data or the group that they are part of? (Riskier when the answer is yes.)
- Is there a possibility of misuse for political or religious reasons through the re-identification of de-identified data? (Riskier when answer is yes.)
- Is there a possibility of causing economic or non-economic losses for the data provider due to the re-identification of de-identified data? (Riskier when the answer is yes.)

9.3 Inadvertent re-identification

SR 6: Organizations should identify possible risks of re-identification that can occur through inadvertent re-identification.

Inadvertent re-identification refers to an analyst working with data inadvertently recognizing an acquaintance from the de-identified data. This is called an 'inadvertent attack' defined in [b-HITRUST] as an attack that "…transpires when a data analyst working with the data recipient (or the data recipient herself) inadvertently re-identifies someone in the dataset. For instance, this could occur when the recipient is already aware of the identity of someone in the dataset, such as a friend, relative, or more generally, an acquaintance." [b-HITRUST] presents the following as examples:

- Birth registry: Does the analyst know a person with a baby?
- Breast cancer medical records database: Does the analyst know a person with breast cancer?
- Special pharmacy database: Does the analyst know a person needing special medicines?
- Healthcare system billing database: Does the analyst know a person who joined Medicare?

Therefore, special attention should be given to the possible risk of inadvertent re-identification.

10 Requirements for using and managing de-identified data

Despite the implementation of the same de-identification methods, data risk levels may differ depending on the security infrastructure of the organization. To ensure clear de-identification, various security factors should be checked, including physical and technical connections across the infrastructure, factors that can affect the data environment and systems inside the enterprise.

The safe use of more highly utilized data is assured through management and protection measures for the technical, managerial and physical perspectives of various environments wherein de-identified data are utilized.

In addition, by observing management and protection measures in the follow-up stages of the de-identified data utilization process, it is possible to utilize de-identified data more usefully.

10.1 Security measures for de-identified data

SR 7: Organizations should implement necessary managerial and technical security measures for de-identified data.

Leaked de-identified data can become identifiable by being combined with other data, and therefore, the following mandatory protection measures should be carried out:

– **Managerial measures**: Designation of person in charge of managing de-identified files, prevention of sharing the data related to de-identification measures, disposal when purpose of use is achieved, etc., and the following measures are necessary:
  - Designation of person in charge of managing de-identified data;
  - Management of de-identified data registry;
  - Prevention of sharing the data related to de-identification measures between the original data management department/Institute and de-identification data management department/Institute;
  - Immediate destruction of data upon achieving the purpose of use;
  - Establishment of an incident response plan in case of de-identified data being leaked;
  - Separate storage of additional information during pseudonymization.
– **Technical measures**: Restricted access to de-identified data, management of connection records, installation and operation of security programs, etc., and the following measures are necessary:
  - Management of access authority and access control for de-identified data;
  - Accessing record management for de-identified data storage system;
  - Installation and operation of security program for preventing malicious codes, etc.
– **Protective measures when de-identified data are leaked**:
  - Management and technical protection measures for leak cause analysis and additional leak prevention;
  - Collection and destruction of leaked de-identified data.

10.2 Monitoring of re-identification possibilities

SR 8: Organizations should continuously monitor re-identification possibilities of the de-identified data.

Organizations, institutes, etc. that use de-identified data or that intend to provide de-identified data to third parties should regularly monitor the re-identification possibilities of the corresponding de-identified data when there is any change of internal factors or any change to the external environment. If monitoring results show that any of the following inspections are applicable, additional de-identification measures should be taken:

– Change of internal factors:
  - When collecting or being provided with additional information with concerns of re-identification by connecting with de-identified data;
  - When new information is generated by connecting additional information generated during the course of using de-identified data with an additional set of de-identified data;
  - When there is a request from the department using the de-identified data to lower the de-identification level for the de-identified data to below the original levels;
  - When new or additionally constructed systems incur major changes on the security system that manages and controls access to the de-identified data;
  - When it is known that de-identification cases have been re-identified using methods similar to de-identification methods applied to a set of de-identified data.
– Changes to the external environment
  - When new technologies appear or are disclosed that can incapacitate de-identification methods and technologies applied for de-identified data;

- When it is known that data that can be newly connected to de-identified data has appeared or been disclosed.

When other possibilities for re-identification by a person providing or consigning de-identified data are discovered, the person who processes the de-identified data should immediately notify and request suspension of processing and take necessary measures such as collection and destruction of the corresponding data.

10.3 Compliance with de-identified data provision or consignment contract

SR 9: Organizations should comply with de-identified data provision when transferring de-identified data to a third party or put contract provisions in place when de-identified data processing is outsourced [b-Kor-Guideline].

When providing de-identified data to a third party or when outsourcing its processing, the following measures on re-identification risk management should be included in the contract:

- **Prohibition of re-identification** – Clearly indicate that organizations, etc. that receive or process de-identified data are prohibited from making re-identification attempts through combination with other data;
- **Limitations of re-provision and re-consignment** – Clearly indicate in the contract by prescribing the scope for re-provision or re-consignment to the party consigned for providing or processing the de-identified data;
- **Notification in case of risk of re-identification** – Clearly indicate that data processing is stopped in the case of re-identification or if there is a high possibility of re-identification, and that there is a duty of notification to the de-identified data provider or consignor.

Appendix I

Data re-identification risk measurement and context re-identification measurement [b-IPCO]

I.1 Data re-identification risk measurement

A two-step process can be used to measure the amount of re-identification risk to a dataset:

1) a process to calculate the probability of re-identification of each row, and

2) a process to apply the appropriate risk measurement method based on the release model used.

Each row in a dataset containing the data of individuals contains PII about the data subject. Accordingly, each row has a probability of re-identification. For a given row, the probability of re-identification can be calculated from how many other rows in the dataset have the same values for variables that are indirect identifiers. The indirect identifier is an attribute that, together with other attributes that in the dataset or external to it, enables the unique identification of a data principal within a specific operational context.

All rows in a dataset with the same values for variables that are indirect identifiers form an "equivalence class". The equivalence class is a set of records in a dataset that have the same values for a specified subset of attributes. For example, in a dataset with variables for gender, age and highest level of education, all the rows corresponding to 35-year-old men with university degrees would form an equivalence class. The size of an equivalence class is equal to the number of rows with the same values for indirect identifiers.

For each row, the probability of re-identification is equal to 1 divided by the size of its equivalence class. For example, each row in an equivalence class of size 10 has a probability of re-identification of 0.1.

There are three data release models:

1) Public data releases: maximum risk;

2) Non-public data releases: strict average risk; and

3) Semi-public data releases (conditional download available to anyone): maximum risk.

For public data releases, it is assumed that someone will attempt an attack to garner public attention. These kinds of attacks will target the most vulnerable rows in the dataset, which are those with the smallest equivalence classes and highest probability of re-identification. In this model, the maximum probability of re-identification is used across all rows to measure data re-identification risk.

For non-public data releases, because access to the dataset is limited to a select number of identified recipients, it is assumed that no row is more vulnerable than others to a data re-identification attack. Here, the average probability of re-identification across all rows to measure the amount of data re-identification risk in the dataset should be used. However, to protect against unique rows or equivalence classes with a high risk of data re-identification, the average should be a "strict" average where no row may have a probability of re-identification that is greater than a specific value. A threshold of 0.33 is often proposed, that is, the smallest size of equivalence class in the dataset should be 3. In practice, however, a maximum probability of re-identification of 0.5 may also be used, which in the case of strict average ensures that there are no unique rows and that the average risk is acceptably small.

Because semi-public data releases are available to anyone for download, it is assumed that the most vulnerable rows will be more at risk of attack than others. Because of this, as with public data releases, the maximum probability of re-identification across all rows should be used to measure the amount of re-identification risk.

I.2 Measurement of context re-identification risk

The context re-identification risk is defined as the probability of one or more re-identification attacks being launched against a dataset and depends on the release model used.

I.2.1 Public data releases

Because the dataset is made available to anyone for download or use without any conditions, any adversary can attempt a demonstration attack to garner publicity. The probability of a re-identification attack being launched against the dataset, context re-identification risk, is "1".

I.2.2 Non-public data releases

For non-public data releases, the probabilities of three different re-identification risks should be determined depending on attack types: deliberate insider attack, inadvertent recognition of an individual in the dataset by an acquaintance, and data breach.

Attack 1: Deliberate insider attack

The probability of a recipient of a non-public data release attempting to re-identify one or more individuals in the dataset is based on two factors:

1) The strength of the controls regarding the privacy and security of the data;
2) The motives and capacity of the recipient regarding performing a re-identification attack.

Depending on the privacy and security controls for a non-public data release, the probability of a recipient attempting to launch a re-identification attack may vary. The higher the level of privacy and security controls, the lower the probability of a re-identification attack being launched.

Additional factors to consider when determining the probability of a recipient attempting to launch a re-identification attack are their motives and capacity. The more motivated and more capable the recipient is with respect to re-identifying one or more individuals in the dataset, the higher the probability of a re-identification attack being launched.

Table I.1 provides a guideline in determining the probability of a re-identification attack being launched against non-public datasets.

Table I.1 – Probabilities of context re-identification attacks [b-IPCO]

Strength of controls	Adversary motives and capability	Probability of context re-identification attack
High	Low	0.05
	Medium	0.1
	High	0.2
Medium	Low	0.2
	Medium	0.3
	High	0.4
Low	Low	0.4
	Medium	0.5
	High	0.6

Attack 2: Inadvertent recognition of an individual by an acquaintance

The recipient of a non-public de-identified data may also inadvertently re-identify one or more individuals. This could happen if, while analysing the data, recipients recognize a friend, colleague, family member or acquaintance. The probability of such an "attack" occurring is equal to the probability of a random recipient knowing someone in the dataset.

To calculate this, the following equation that assumes a geometric distribution may be used:

$$\boldsymbol{P[X \leq x] = 1 - (1 - p)^{m}} \quad \text{(I.1)}$$

In this equation, *p* is the probability of success (percentage of individuals in the population who have the condition or characteristic discussed in the dataset) and *m* is the number of people, on average, that an individual knows. Take, for example, a dataset about individuals who carpool to work. Based on the values of *p* and *m*, the equation would give the probability that a random individual knows someone who carpools to work.

The value of *p* should be determined by recent population statistics. On the other hand, the value for *m* may vary depending on the kind of relationship with an individual required to have knowledge about them regarding the condition or characteristic discussed in the dataset. For friends, you should in general use a value of *m* between an average of 150 and 190, where 150 represents Dunbar's number, a suggested limit to the number of connections that someone can maintain at once) and 190 [b-El Eman].

Attack 3: Data breach

If a data breach occurs at the recipient's facilities, an external adversary can attempt a re-identification attack. Therefore, the probability of such an attack occurring is equal to the probability of a breach occurring at the recipient's facilities. To calculate this value, publicly available data on the prevalence of data breaches in the recipient's respective industry can be used.

I.2.3 Semi-public data release

The possible re-identification attacks for semi-public data releases can be considered the same as those for non-public data releases. Accordingly, to measure the context risk for semi-public data releases, the same method and equations are used as for non-public data releases, with one adjustment. With respect to "Attack 1: Deliberate insider attack", it should be assumed that the recipient has high motives and capacity and that privacy and security controls are low.

Bibliography

[b-ITU-T X.1058] Recommendation ITU-T X.1058 (2017), *Information technology – Security techniques – Code of practice for personally identifiable information protection*.

[b-ISO 19115-1] ISO 19115-1:2014, *Geographic information – Metadata – Part 1: Fundamentals*.

[b-ISO/IEC 20889] ISO/IEC 20889:2018, *Privacy enhancing data de-identification terminology and classification of techniques.*

[b-ISO/IEC 27559] ISO/IEC 27559:2022, *Information security, cybersecurity and privacy protection – Privacy enhancing data de-identification framework.*

[b-ISO/IEC 29100] ISO/IEC 29100:2011, *Information technology – Security techniques – Privacy framework.*

[b-ISO/IEC/IEEE 15026-1] ISO/IEC/IEEE 15026-1:2019, *Systems and software engineering – Systems and software assurance – Part 1: Concepts and vocabulary*.

[b-ISO TR 18307] ISO TR 18307:2001, *Health informatics – Interoperability and compatibility in messaging and communication standards – Key characteristics*.

[b-Kor-GuidelineROK] Republic of Korea (2016), *Guideline for De-identification of Personal Data*. https://www.privacy.go.kr/cmm/fms/FileDown.do?atchFileId=FILE_000000000830764&fileSn=0

[b-El Eman] El Eman, K. (2013), *Guide to the De-identification of Personal Health Information*, Boca Raton, FL, Routledge, p. 213.

[b-HITRUST] Health Information Trust Alliance (2015), *HITRUST De-Identification Framework*. https://hitrustalliance.net/documents/de_id/HiTrustDeIdentificationPresentation.pdf

[b-IPCO] Information and Privacy Commissioner of Ontario (2016), *De-identification Guidelines for Structured Data*. https://www.ipc.on.ca/resource/de-identification-guidelines-for-structured-data

[b-MEDSEC] MEDSEC (1998), *Draft standard – High level security policies for health care establishments, European Commission, Directorate General III, Health care security and privacy in the information society*.

[b-UKAN] UK Anonymisation Network (2020), *The Anonymisation Decision-Making Framework: European Practitioners' Guide, 2nd Edition* (). https://ukanon.net/framework/

SERIES OF ITU-T RECOMMENDATIONS

Series A	Organization of the work of ITU-T
Series D	Tariff and accounting principles and international telecommunication/ICT economic and policy issues
Series E	Overall network operation, telephone service, service operation and human factors
Series F	Non-telephone telecommunication services
Series G	Transmission systems and media, digital systems and networks
Series H	Audiovisual and multimedia systems
Series I	Integrated services digital network
Series J	Cable networks and transmission of television, sound programme and other multimedia signals
Series K	Protection against interference
Series L	Environment and ICTs, climate change, e-waste, energy efficiency; construction, installation and protection of cables and other elements of outside plant
Series M	Telecommunication management, including TMN and network maintenance
Series N	Maintenance: international sound programme and television transmission circuits
Series O	Specifications of measuring equipment
Series P	Telephone transmission quality, telephone installations, local line networks
Series Q	Switching and signalling, and associated measurements and tests
Series R	Telegraph transmission
Series S	Telegraph services terminal equipment
Series T	Terminals for telematic services
Series U	Telegraph switching
Series V	Data communication over the telephone network
Series X	**Data networks, open system communications and security**
Series Y	Global information infrastructure, Internet protocol aspects, next-generation networks, Internet of Things and smart cities
Series Z	Languages and general software aspects for telecommunication systems

Printed in Switzerland
Geneva, 2024