INVITED PAPER Special Section on Information and Communication System Security

An Overview of De-Identification Techniques and Their Standardization Directions

Heung Youl YOUM[†], Nonmember

SUMMARY De-identification [1]–[5], [30]–[71] is the process that organizations can use to remove personal information from data that they collect, use, archive, and share with other organizations. It is recognized as an important tool for organizations to balance requirements between the use of data and privacy protection of personal information. Its objective is to remove the association between a set of identifying attributes and the data principal where identifying attribute is attribute in a dataset that is able to contribute to uniquely identifying a data principal within a specific operational context and data principal is entity to which data relates. This paper provides an overview of de-identification techniques including the data release models. It also describes the current standardization activities by the standardization development organizations in terms of de-identification. It suggests future standardization directions including potential future work items.

key words: de-identification, re-identification, pseudonym, anonymization, standardization

1. Introduction

Organizations (e.g., public or private companies) may wish to process data containing personal information, also known as PII (personally identifiable information), and commercially sensitive information. A PII is defined as any information about individual, including (1) any information that can be used to identify the PII principal (data subject) to whom such information relates, such as name, social security number, date and place of birth, mother’s maiden name, or biometric records, and (2) any information that is or might be directly or indirectly linked to a PII principal [7], such as medical, educational, financial, and employment information [25]. Each jurisdiction may restrict organizations’ capability to share these kinds of information as open data. It also requires organizations either to obtain consent from data subjects or to make data anonymous, before transferring and sharing that data to the third party, and publishing data to public for research and development [27]. Possible technical compromise to publish data is to use de-identification techniques, especially when data is processed for the purpose other than intended purpose or data is combined with many other sources of data. De-identification process is to remove the association between a set of identifying attributes in the record and the data principal [6]. A de-identification provides a practical tool that enables using data in the big data environment while balancing privacy protection and utility of data, when data is released to public or it is used for the purpose other than original purpose within the organization.

Contribution: This paper’s objective does not present theoretical results of de-identification techniques, but describes all related issues that are necessary for standardization activities for de-identification techniques. This paper differs from the previous papers mentioned above in many ways. This survey paper focuses on studies related to standardization activities and analyzing practical studies and works, including presenting future standardization directions.

The rest of the paper is organized as follows: In Sect. 2, we present the overview of de-identification such as definition, re-identification, key terminology, privacy preserving data release models, and de-identification process. In Sect. 3, various types of de-identification techniques are presented. In Sect. 4, we propose future directions of international standardization activities related to de-identification techniques. Finally, in Sect. 5 we conclude this paper.

2. De-Identification Overview

Theoretical background: There are many existing theoretical papers regarding de-identification [1]–[5], [30]–[71]. Two types of identifiers are used in the dataset: direct identifier and quasi identifier, also known as indirect identifiers. The direct identifier is data that directly identifies a single individual (or data subject) [25]. Normally, the direct identifier should be removed during de-identification process. To this end, several technical methods, such as suppression are used for de-identifying quasi-identifiers. Quasi-identifiers are identifiers that by themselves do not identify a specific individual but can be aggregated and “linked” with other information to identify data subjects [25]. Common de-identification strategies are to delete or mask identifying identifiers, such as personal name, and to suppress or generalize quasi-identifiers, such as date of birth [6], [25]. The reverse process of using de-identified data to identify individuals is known as data re-identification. The researchers [1]–[4] demonstrated successful re-identifications that caused cast negative impacts on de-identification’s effectiveness. For example, a study [2] showed that people could be identified by their “mobility traces” (a record of locations and times that the person or vehicle visited). In their study, trace data from a sample of 1.5 million individuals was processed, with time values being generalized to the hour and spatial

Manuscript received November 29, 2019.
Manuscript revised February 6, 2020.
Manuscript publicized May 14, 2020.
†The author is with the Department of Information Security Engineering, Soonchunhyang University, Korea.
a) E-mail: hyyoum@sch.ac.kr
DOI: 10.1587/transinf.2019ICI0002

Copyright © 2020 The Institute of Electronics, Information and Communication Engineers
data generalized to the resolution provided by a cell phone system. A study [5] provided a systematic review of fourteen distinct re-identification attacks that found “a high re-identification rate dominated by small-scale studies on data that was not de-identified according to existing standards.”

Privacy measurement models provide a means of assessing the effectiveness of de-identification. Typically, there are three formal privacy models: K-anonymity, L-diversity, and T-closeness. K-anonymity [6] is a privacy model that ensures that for each identifier in a dataset there is a corresponding equivalence class containing at least K records. L-diversity [6] is a privacy model that ensures that for a selected attribute each equivalence class has at least L well-represented values. T-closeness [6] is a privacy model that ensures that the distance between the distribution of a selected attribute in an equivalence class and the distribution of this attribute in the entire table is no more than a threshold T. The researchers [41], [51], [52], [54], [63] addressed the subject of K-anonymity that is a formal privacy measurement model that ensures that for each identifier there is a corresponding equivalence class containing at least K records. A dataset is a collection of data [6]. The researchers [37], [42], [50]–[52] addressed the subject of L-diversity, an enhancement to K-anonymity for datasets with poor attribute variability. The studies [7], [39] described the subject of T-closeness that is an enhancement to L-diversity for datasets with attributes that are unevenly distributed, belong to a small range of values, or are categorical. The formal privacy models such as K-anonymity, L-diversity, and T-closeness are a fundamental tool to measure the effectiveness of de-identification. Differential privacy prevents disclosure by adding non-deterministic noise (usually small random values) to the results of mathematical operations before the results are reported [25].

The researchers [7], [9], [15], [35], [44], [45], [50] addressed the subject of differential privacy that is a formal privacy measurement model that provides mathematical guarantees that the probability distribution of the output of this analysis differs by a factor no greater than a specified parameter regardless of whether any particular data principal is included in the input dataset. A study [46] presented the “server model” for differential privacy that typically preserves data in unmodified form in a secure database. The researchers [44], [51] found that in the context of differential privacy, random noise takes the form of random numbers that are generated according to a selected probability distribution.

2.1 What is the De-Identification?

A de-identification is a process of removing the association between a set of identifying attributes and the data subjects [6]. Data de-identification stages could be represented as data types that represent the degree to which an individual is directly identified by the data and how the individual is associated with characteristics (attributes) in the data. Figure 1 describes data stages from original data to de-identified data in the de-identification process [14]. Re-identification risk is the risk that de-identified records can be re-identified. Re-identification risk is typically reported as the percentage of records in a dataset that can be re-identified [25]. Each stage relates to a different possibility of re-identification risk. A data type characterizes specific stages that a dataset would go through as it is increasingly de-identified.

In the original stage of identified data, data can unambiguously be associated with a specific individual because an individual is identified with the information. In the pseudonymized data stage, data cannot be reversed by reasonable efforts of anyone other than the party that performed the alias assignment because all identifiers are substituted by aliases. This corresponds to data defined as “pseudonymization”. In the unlinked pseudonymized data stage, all identifiers are erased or substituted by aliases for which the assignment function is erased or irreversible, such that a linkage cannot be re-established by reasonable efforts by anyone including the party that performed them. In the stage of aggregated data, data forms information about enough different persons that individual-level attributes are not inferable as statistical data that does not contain individual-level entries and is combined. All aggregated data may not reach the degree of identifiability below a threshold if cell size for a given crossing of some combination of variables can
lead someone to identify a particular individual. In the de-
identified data stage, data is unlinked and attributes are al-
tered (e.g., attributes’ values are randomized or generalized)
in such a way that there is a reasonable level of confidence
that a person cannot be identified, directly or indirectly, by
the data alone or in combination with other data.

2.2 What is the Re-Identification?

A re-identification is a process of associating data in a de-
identified dataset with the original data subjects[6]. Data
subject or PII subjects is a natural person to whom the
PII relates [6]. A re-identification attack is an action per-
formed on de-identified data by an attacker with the purpose
of re-identification. Several successful re-identification at-
ttempts have raised doubts about the effectiveness of de-
identification in protecting individuals’ privacy [11–5].

2.3 Key Terminology

An identifying attribute is an attribute in a dataset that is able
to contribute to uniquely identifying a data principal within a
specific operational context. Pseudonymization [25] is a de-
identification technique that replaces an identifier (or identi-
fiers) for a data principal with a pseudonym in order to hide
the identity of that data principal. A pseudonym is a unique
identifier created for a data principal to replace the com-
monly used identifier (or identifiers) for that data principal.
An anonymization is a process by which 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. A PII con-
troller [7] is a privacy stakeholder (or privacy stakeholders)
that determines the purposes and means for processing PII
other than natural persons who use data for personal pur-
poses. Anonymized data is data that has been produced as
the output of a PII anonymization process [6]. Aggregated
data is data representing a group of data principals, such as
a collection of statistical properties of that group.

Direct identifier is an attribute that alone enables a
unique identification of a data principal within a specific op-
erational context [9]. Examples are social security number,
national resident registration number. Indirect identifier (or
quasi identifier) is an attribute that, together with other at-
tributes that can be in the dataset or external to it, enables
unique identification of a data principal within a specific op-
erational context. Examples are birth date, postal area code
(ZIP code), and sex.

Two key terms are de-identification and pseudonymiza-
tion which are critical to develop the further documents.
Table 1 compares the definition of two key terms, i.e.,
de-identification and pseudonymization, in various exist-
ing documents. Each document uses different terminol-
y with almost same definition depending on the context

| Table 1 | Comparison of terms and definition on de-identification and pseudonymization. |
|---------|--------------------------------------------------------------------------|
|         | **De-identification/anonymization**                                      | **Pseudonymization**                                          |
| ISO/IEC 20889 [6] | [De-identification] process of removing the association                  | 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 |
| ISO/IEC 29100 [7] | [Anonymization] process by which PII is irreversibly altered             | process applied to PII which replaces identifying information with an alias |
| NISTIR 8053 [25] | [De-identification] any process of removing the association              | particular type of anonymization that both removes the association with a data subject and adds an association between a particular set of characteristics relating to the data subject and one or more pseudonyms |
| IPCO [9] | [De-identification] process of removing personal information from a record or dataset. | Not applicable |
| ISO 25237 [10] | [De-identification] any process of reducing the association              | particular type of de-identification that both removes the association with a data subject and adds an association between a particular set of characteristics relating to the data subject and one or more pseudonyms |
| ICO [20] | [Anonymization] process of rendering data into a form which does not identify individuals and where identification is not likely to take place. | Not applicable |
| ISO/IEC 19944 [11] | Not applicable                                                              | process applied to PII which replaces identifying information with an alias |
of an application. For example, ISO/IEC 20889 [6] uses de-identification while ICO document [29] uses anonymization. Table 2 maps key de-identification terminology referred to by relevant documents [6]. In addition, ISO 19944 defines some terms in the cloud computing context. Taking into account ISO/IEC 19944 [6] provides all necessary de-identification techniques, the comparison of the definition of terms is of importance. Table 3 compares key terms definitions in ISO/IEC 19944 and ISO/IEC 20889.

### 2.4 Privacy-Preserving Data Release Models

There are three release models for delivering de-identified data: public, semi-public or non-public [14].

Each release model allows for different levels of availability and protection of privacy. Depending on the purposes and/or regulation requirements of the data release, the suitability of each model may vary. The release model plays an important role in the de-identification process, as the amount of de-identification required may vary depending on the release model selected [14].

In a traditional public data release, anyone can access the de-identified data without registration or conditions. Data are processed using de-identification techniques to produce a new, de-identified dataset that is distributed to users. Examples of such releases include the publicly available data from organizations and data posted to an open-access data repository such as a web portal. Organizations proactively release datasets and make them freely available to anyone for use and republishing. When releasing data publicly, it is common practice to place as few restrictions as possible on the information, including who can access it and how. As such, when individuals who download the dataset cannot be identified, these disclosures should be handled as public data releases. Semi-public data release model is more restrictive than public data release models and occurs when there exists a formal request and approval process to obtain access to data. In this case, the data recipient may agree to some terms of use or sign a “click-through” contract. Click-through contracts are online terms of use that may place restrictions on what can be done with the data and how the data are handled. Regardless, anyone can still download such data. In non-public release model, also known as Privacy Preserving Data Mining (PPDM), de-identified data are not released, but are used instead for statistical processing. The results of the processing may be released in the form of statistical tables based on summarization and aggregation. In this model, access to the dataset is limited to a number of identified recipients. No row is more vulnerable than others to a re-identification attack.

Datasets that contain PII may be shared within and among organizations only if the disclosure is permitted under country regulatory guidance. If the disclosure is not permitted and the institutions still wish to share datasets, then any PII must be removed. Non-public data releases provide the least availability, but can provide a higher amount of protection, requiring a smaller amount of de-identification.

When sharing information among organizations, because access to the dataset is limited to the organization, requirements regarding the privacy and security of the information can be set and enforced through a data-sharing agreement. For a data release to be treated as non-public,

### Table 2  Mapping of de-identification terminology among relevant documents [6].

| ISO/IEC 20889 [6] | ISO 25237 [10] | ISO/IEC 29100 [7] | EU Article 29 [28] | ICO [29] |
|-------------------|----------------|-------------------|-------------------|---------|
| De-identification | De-identification, anonymization | Anonymization | N/A | Anonymization |
| Pseudonymization with controlled re-identification | Pseudonymization reversible | Pseudonymization | Pseudonymization | Anonymization |
| Pseudonymization without controlled re-identification | Pseudonymization irreversible | Anonymization | Pseudonymization | Anonymization |
| Randomization | N/A | N/A | Anonymization | Anonymization |
| Generalization | N/A | N/A | Anonymization | Anonymization |
| Differential privacy | N/A | N/A | Anonymization | N/A |

### Table 3  Comparison of key terms between ISO/IEC 19944 and ISO/IEC 20889 [6].

| ISO/IEC 20889[6] | ISO 19944[11] |
|-----------------|----------------|
| original, unprocessed data containing identifiers; in other words, no de-identification techniques are applied yet; for the other qualifiers, the identifiers are removed (masked) | Identified data |
| data processed using pseudonymization techniques with controlled re-identification possible/implemented | Pseudonymized data |
| data processed using pseudonymization techniques with no controlled re-identification allowed | Unlinked pseudonymized data |
| data proceed using generalization and/or randomization techniques | Anonymized data |
| Data processed using aggregation techniques | Aggregated data |
Table 4  Comparison of data release models [14].

|                              | Public release model                                                                 | Semi-public release model                                                                 | Non-public release model                                                                 |
|------------------------------|--------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------|
| **Access rights**            | • Any one can access to released dataset                                               | • Access to data (or subset) from restrictive individual or organization                  | • Access to data is to a subset of individuals or organizations                          |
| **Use cases**                | • Unrestricted data access through Web portal, i.e., freely available to anyone       | • On-site safe setting                                                                    | • Sharing within and among organizations                                                 |
| **Privacy risk level**       | • Maximum risk                                                                        | • Maximum risk                                                                            | • Strictly average risk                                                                    |
| **Rights**                   | • Unlimited rights to reuse and redistribute data                                      | • Available to authorized individual or organization                                       | • Re-use, republishing, or distribution of data is forbidden                             |
| **Re-identification attack**| • A demonstration attack for publicity                                               | • Deliberate insider attack                                                               |                                                                                          |
|                              |                                                                                      | • Inadvertent recognition of an individual in the dataset by an acquaintance               |                                                                                          |
|                              |                                                                                      | • Data leakage                                                                           |                                                                                          |

2.5 De-Identification Process

There are typical two types of de-identification process models: a 4 stage process model [14], [72] and a 9 stage process model [9]. The main difference between two de-identification processes is that the former process is based on adequacy assessment of re-identification risk in step 3, a simplified model, and the latter is based on re-identification risk assessment with the assurance levels.

This first model of the de-identification process in [14], [72] can be used for providing de-identified PII as shown in Fig. 2. Four steps are included in the de-identification operating procedures.

(Step 1) Preliminary review: This step involves verifying whether the target data is PII or not. If the target data does not include PII, the data can be used in a limitless manner. If the data does include PII, proceed to Step 2. De-identification is needed.

(Step 2) De-identification: This step involves de-identifying data to prevent inferences of specific individual information from the target dataset. This step invokes methods for removing or transforming PII elements, either in total or partially. PII elements include identifiers, quasi-identifiers and sensitive attributes.

(Step 3) Adequacy assessment: This step involves assessing the adequacy of the de-identified dataset including PII elements. Relevant considerations include whether the target dataset still contains PII, re-identification risks, or probabilities of re-identification. The K-anonymity model is used among other privacy protection models when assessing adequacy. The K-anonymity model is a basic means of assessment. Additional assessment models (L-diversity, T-closeness, differential privacy (DP), etc.) may be applied if necessary.
(Step 4) Follow-up management: Step 4 involves measuring managerial and technical safety to prevent re-identification.

The second model of the de-identification process in [9] can be used for providing de-identified PII as shown in Fig. 3. Nine steps are included in the de-identification process as follows:

(Step 1) Determine the release models: This step is to determine the data release model among released publicly, semi-publicly or non-publicly. Each release model allows for different levels of availability and protection of information. Depending on the purposes and/or legislative requirements of the data release, the suitability of each model may vary.

(Step 2) Classify variables: This step is to classify variables among two kinds of such variables: direct identifiers and indirect or quasi-identifiers. Depending on the type of information, some variables may be used to identify individuals, either directly or indirectly, while others may not. De-identification is only concerned with variables that may be used to identity individuals. As noted above, there are two kinds of such variables: direct identifiers and indirect or quasi-identifiers.

(Step 3) Determine an acceptable re-identification risk threshold: This step is to determine an acceptable level of re-identification risk (or threshold) for a dataset. To protect personal privacy, the amount of de-identification that is required to be applied is proportional to the level of re-identification risk. The higher the re-identification risk of a data release, the greater the amount of de-identification required.

(Step 4) Measure the data risk: This step is to measure the amount of re-identification risk in the dataset itself. The data risk is used to determine the level of re-identification risk involved in the release. Measuring the amount of re-identification risk in a dataset is a two-step process. The probability of re-identification of each row should be (1) calculated, and (2) the appropriate risk measurement method should be applied based on the release model used.

(Step 5) Measure the context risk: This step is to measure the context risk, while the risk from the dataset plays an important role in determining the level of re-identification risk involved in the release of a dataset. The re-identification risk is also a function of the kinds of re-identification attacks that are possible on the dataset given the release model used.

(Step 6) Calculate the overall risk: This step is to calculate the overall risk of re-identification once the data risk and the context risk have been measured.

(Step 7) De-identify the data: This step is to remove any identifiable information, you should 1) mask direct identifiers, 2) modify the size of equivalence classes, and 3) ensure that the overall risk is less than or equal to the re-identification risk threshold.

(Step 8) Assess data utility: This step is to assess data utilization. There may be a trade-off between the amount of de-identification applied to a dataset and the utility of the resulting information. The more the variables that qualify as quasi-identifiers are de-identified using techniques such as generalization and suppression, the higher the potential for a corresponding loss in the utility of the dataset.

(Step 9) Document the process: This step is to document the process. That is, a report documenting the process and its results should be produced. There are a number of benefits to this best practice, including:

- the ability to demonstrate due diligence and evidence of compliance, which may be important in the event of a privacy breach or complaint to the Information and privacy commissioner.
- confidence (of individuals, other institutions, partners and your own management) that best practices are being followed.
- increased transparency, awareness, understanding and trust in your organization’s information management practices.

3. De-Identification Techniques and Privacy Measurement Models

This chapter describes various de-identification techniques [6]. A de-identification technique is a method for transforming a dataset with the objective of reducing the extent to which information is able to be associated with individual data principals. A re-identification is a process of associating data in a de-identified dataset with the original data principal which is a reverse process of a de-identification process using de-identification techniques. Privacy measurement models provide a means of assessing the effectiveness of de-identification.
3.1 De-Identification Techniques

3.1.1 Statistical Tools

There are two types of statistical tools: sampling and aggregation. Sampling is a process in which a sample of an entire dataset is released, instead of releasing an entire dataset. If a subsample is released, the probability of re-identification can decrease. Aggregation is a set of statistical functions that produce the represented value of an entire dataset [25].

3.1.2 Suppression Techniques

Suppression techniques involve removing selected attributes across all records (e.g., masking), selected attribute values (e.g., local suppression), or selected records from a dataset (e.g., record suppression). The example of suppression technique includes replacing a cellular phone number with asterisks or a randomly generated pseudonym. The quasi-identifier can be suppressed or removed [6], [25]. Removing the data maximizes privacy protection, but may decrease the utility of the dataset [25]. Local suppression is a process that suppresses or removes specific values of attributes from selected records. Record suppression is a process that involves removing an entire record or records from a dataset.

3.1.3 Pseudonymization Techniques

Pseudonymization is a particular type of de-identification that both removes the association with a data subject and adds an association between a particular set of characteristics relating to the data subject and one or more pseudonyms [10]. Typically, pseudonymization is implemented by replacing direct identifiers with a pseudonym, such as a randomly generated value using cryptographic techniques such as a hash function [25]. Examples of direct identifiers include names, email addresses, and government issued numbers. All direct identifiers and potentially additional or all remaining identifying attributes are replaced with a pseudonym.

3.1.4 Generalization Techniques

Generalization can be applied to the entire dataset or to specific records [6], [25]. There are two types of generalization techniques: rounding and top and bottom coding. Rounding is a process of putting numerical value by another value that is approximately equal but has a shorter, simpler, or more explicit representation. Top and bottom coding is a process for which attributes whose values are above an upper bound (or lower bound) are set as a threshold on the largest (or smallest) value possible. Specific quasi-identifier values can be reported as being within a given range or as a member of a set [25]. For example, the ZIP code 12345 could be generalized to a ZIP code between 12000 and 12999.

3.1.5 Randomization Techniques

Noise addition is a process in which a random value that cannot be predicted is added to a selected attribute of a dataset [6]. Permutation is a process for exchanging the values of a selected attribute across the records in a dataset without modification. Micro-aggregation is a process in which all values of continuous attributes are replaced with their averages computed in a certain algorithmic way.

3.1.6 Cryptographic Tools for De-Identification Techniques

There are five kinds of cryptographic tools that are used for de-identification: determined encryption, order-preserving encryption, homomorphic encryption, format-preserving encryption, and homomorphic secret sharing. Deterministic encryption [18]: an encryption scheme that always produces the same ciphertext for a given plaintext and key over separate executions of the encryption algorithm. Order-preserving encryption is an encryption scheme in which numerical ordering of the plaintexts is preserved. Homomorphic encryption [77] is an encryption scheme that allows computations to be carried out on ciphertext, thus generating an encrypted result which matches the result of operations to be performed on the plaintext, when decrypted. Format-preserving encryption [26] is an encrypting scheme in which the ciphertext is in the same format as the plaintext. Homomorphic secret sharing [77] is a type of secret sharing algorithm in which the secret is encrypted using homomorphic encryption.

3.1.7 Synthetic Data

Synthetic data is an approach that artificially generates micro data to represent a predefined statistical data model [6]. By definition, synthetic datasets do not contain data collected from existing data subjects, but they look real for their intended purpose.

3.2 Formal Privacy Measurement Models

A formal privacy measurement model is an approach to the application of data de-identification techniques that enables the calculation of re-identification risk and, in some cases, provides mathematical guarantees against re-identification risk [6]. A formal privacy measurement model reflects the context of the use case.

3.2.1 K-Anonymity Model

K-anonymity [41], [51], [52], [54], [63] is a technique to release data such that the ability to link to other information using the quasi-identifier is limited. K-anonymity achieves this through suppression of identifiers and output perturbation. K-anonymity is a formal privacy measurement model
that ensures that for each identifier there is a corresponding equivalence class containing at least K records. While the resulting dataset has limited (i.e., 1/K) linkability, it does not contain measures designed to prevent potential inference attempts.

L-diversity [37], [42], [50]–[52] is a refinement to the K-anonymity approach which assures that groups of records specified by the same identifiers have sufficient diversity to prevent inferential disclosure [51]. It is designed to protect against deterministic inference attempts by ensuring that each equivalence class has at least L well-represented values for each sensitive attribute. L-diversity is not a single model but a group of models. Each model has diversity defined slightly differently, e.g., by counting distinct values or by entropy.

T-closeness [7], [39] is an enhancement to L-diversity for datasets with attributes that are unevenly distributed, belong to a small range of values, or are categorical. It is designed to protect against statistical inference attempts, as it ensures that the distance between the distribution of a sensitive attribute in any equivalence class and the distribution of the attribute in the overall dataset is less than a threshold T. This technique is useful when it is important for the resulting dataset to remain as close as possible to the original one.

3.2.2 Differential Privacy Model

Differential Privacy is a set of techniques based on a mathematical definition of identity disclosure and information leakage from operations on a dataset [25]. Differential privacy prevents disclosure by adding non-deterministic noise (usually small random values) to the results of mathematical operations before the results are reported. Differential privacy [7], [9], [15], [35], [44], [45], [50] is a formal privacy measurement model that, if incorporated in the design of a particular statistical analysis, provides mathematical guarantees that the probability distribution of the output of this analysis differs by a factor no greater than a specified parameter regardless of whether any particular data principal is included in the input dataset. The specified parameter can be used to measure the “privacy loss” that the analysis incurs every time it provides an output. The term “privacy loss” is conventionally used in the differential privacy discipline. It does not refer to actual loss of privacy but instead to a reduction in the probability that privacy is maintained.

4. Current Standardization Status and Future Standardization Directions

4.1 Current Typical Standardization Activities

Several Standardization groups also known as standard development organizations (SDOs) such as ITU-T, ISO/IEC JTC 1/SCs, and ISO/TC have been working on standardization of de-identification as described in Fig. 4.

De-identification techniques can be applied to specific sector (for example, health). Each standardization group takes different roles of application for de-identification standardization work. ITU-T SG17 [73] is focusing on development of de-identification which is applicable to all sectors including the telecommunication sector. ISO/IEC JTC 1/SC 27/WG 5 is working on de-identification techniques and framework which are applicable to general sectors. ISO/IEC JTC 1/SC 38 has developed data categories and use related to de-identification in the context of cloud computing.

ISO TC/215 developed de-identification techniques in the health informatics. The regulations in the health sector require controlled re-identification which leads to reversible pseudonym. Due to some similar area, ITU-T SG17 has a collaboration mechanism known as a common text or twin text with ISO/IEC JTC 1/SC 27, which are defined in [79], [81]. A common text is a standard which was developed jointly by ITU-T and ISO/IEC and has identical text,
while a twin text is a standard which was developed in close collaboration between ITU-T and ISO/IEC, and whose texts are technically aligned but not identical. Some results have already been achieved [6]–[25]. For example, the Question 3 in ITU-T SG17 [73] developed ITU-T X.1058|ISO/IEC 29151, “code of practice for Personally Identifiable Information protection” in 2017 [12] together with ISO/IEC JTC 1/SC 27/WG 5. The Question 7 in ITU-T SG17 [73] is working on development for two draft Recommendations: X.fdip, “Framework of de-identification processing service for telecommunication service providers” under development [14] and X.rdda, “Requirements for data de-identification assurance” [15] under development. ISO/TC 215 [75] developed ISO 25237, “Health informatics — Pseudonymization” in 2017. ISO/IEC JTC 1/SC 27 [74] has developed two standards, ISO/IEC 29100, “Privacy frame-

| International standard Title | Scope | Applications /sectors | Status |
|------------------------------|-------|-----------------------|--------|
| ISO/IEC 20889 [6]            | Privacy enhancing data de-identification terminology and classification of techniques | - provide a description of privacy-enhancing data de-identification techniques, to be used to describe and design de-identification measures in accordance with the privacy principles in ISO/IEC 29100 [7].<br> - specify terminology, a classification of de-identification techniques according to their characteristics, and their applicability for reducing the risk of re-identification. | Generic area | Published November 2018 |
| ISO 25237 [10]              | Health informatics - Pseudonymization | - define one basic concept for pseudonymization.<br> - defines one basic methodology for pseudonymization services including organizational, as well as technical aspects.<br> - specify a policy framework and minimal requirements for controlled re-identification.<br> - give an overview of different use cases for pseudonymization that can be both reversible and irreversible.<br> - give a guide to risk assessment for re-identification.<br> - provide an example of a system that uses de-identification.<br> - provide informative requirements to an interoperability to pseudonymization services.<br> - specify a policy framework and minimal requirements for trustworthy practices for the operations of a pseudonymization service. | Health sector | Published January 2017 |
| ISO/IEC 19944 [11]          | Information technology — Cloud computing — Cloud services and devices: Data flow, data categories and data use | - extend the existing cloud computing vocabulary and reference architecture in ISO/IEC 17788 and ISO/IEC 17789 to describe an ecosystem involving devices using cloud services.<br> - describe the various types of data flowing within the devices and cloud computing ecosystem.<br> - describe the impact of connected devices on the data that flow within the cloud computing ecosystem.<br> - describe flows of data between cloud services, cloud service customers and cloud service users.<br> - provide foundational concepts, including a data taxonomy.<br> - identify the categories of data that flow across the cloud service customer devices and cloud services. | Cloud computing application | Published July 2017 |
| NISTIR 8053 [25]           | De Identification of Personal Information | - provides an overview of de-identification issues and terminology.<br> - summarize significant publications to date involving de-identification and re-identification.<br> - does not make recommendations regarding the appropriateness of de-identification or specific de-identification algorithms. | All sectors | Published October 2015 |
| ITU-T X.fdip [14]          | Framework of de-identification process for telecommunication service providers | - provides an overview of de-identification process, including the data lifecycle model, roles of stakeholders in the de-identification process, data release models and data storage in the de-identification process, and de-identification approaches.<br> - describes the de-identification process service model for telecom operators, including the architecture and de-identification process operations and procedures. | Generic area focusing on telecommunication sector | To be published in 2020. |
work” [7] in 2011 and ISO/IEC 20889, “Privacy enhancing data de-identification terminology and classification of techniques” [6] in 2018.

ISO/IEC JTC 1/SC 38 [76] has developed ISO/IEC 19944, “Cloud services and devices: Data flow, data categories and data use” [11] in 2017. ISO/IEC 27701 [8] specifies requirements and provides guidance for establishing, implementing, maintaining and continually improving a Privacy Information Management System (PIMS) for privacy management within the context of the organization. However, there are many remaining items for further standardizations, which are identified in the following subsection 4.3 (as Table 7). Each document has been developed to meet the sector specific requirement. Table 5 provides a summary of current standardization activities focusing international standards already developed in the area of de-identification techniques including the scope, application area, and current document status, which have been carried out or underway by many SDOs, for example, ITU-T. It is provided as a basis to give an overview of standardization activities and to identify what are the remaining work items.

4.2 National Policy and Its Documents for De-Identification

De-identification heavily relies on national policy and regulation since it related to the privacy of individuals. Some jurisdictions specify the obligations of organizations in the national regulations while other countries implement these de-identification techniques as guidelines. Each jurisdiction has a different de-identification policy and its documents. Table 6 presents some countries’ national policy and its regulation as well as their documents regarding de-identification.

4.3 Future Work Items and Standardization Directions

The basic standardization directions are as follows:

- Relevant standardization groups such as ITU-T SG17 and ISO/IEC SC27 should continue to standard this subject matter. ITU-T refers to its approved documents as a Recommendation while ISO and IEC refer to their approved texts as international standards.
- The collaboration to develop a common text or twin text according to ITU-T A.23 [78] should be encouraged to extend their application to 193 countries of ITU-T member states.
- Two groups need to continue to develop new work items to meet the market need.
- Two groups need to develop the standardization roadmap in this area.
- Two groups should use increase collaboration and cooperation for further work.

The ten work items are identified as well as their scope described in Table 7. The appropriate standardization groups also are selected taking into account their nature of work item and application area. It is noted that some of the work items have been addressed in the existing work. However, most of work items should be carried out in the future. For example, an item on use cases for specific sector needs new work item proposal (NWIP) that needs further development. Table 7 could be used as guidance for identifying further work items in standardization groups. Specifically,

| Country | document | Published timing | Description |
|---------|----------|------------------|-------------|
| EU      | GDPR article 4, clause 5 [27] | May 25, 2018 | ‘Pseudonymisation’ means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information. |
| Japan   | Act on Protection of Personal Information [80] | September 2015 | Concept of anonymized data introduced, “Anonymized information”, referring to any information about individuals from which all personal information, and all personal identifier codes, have been removed, making it impossible to re-identify the data subject. |
| Korea   | Guidelines for De-identification of personal Data [28, 72] | June 30, 2016 | Provides detailed standards, procedures, and methods of de-identification for big data utilization. |
| UK      | UK ICO, Anonymisation: managing data protection risk code of practice [29] | July 2014 | Provides an overview of the privacy issues which arise from using big data and suggests how to comply with the Data Protection Act 1998. Anonymize data in a big data project. |
| USA     | NISTIR 8053, De-Identification of Personal Information [25] | October 2015 | Provides an overview of de-identification issues and terminology. It summarizes significant publications to date involving de-identification and re-identification. |
Table 7  Potential new work items and general standardization directions.

| Work items                              | Scope of work items                                                                 | Appropriate standard group | Type of documents | Potential new work items |
|----------------------------------------|--------------------------------------------------------------------------------------|---------------------------|-------------------|--------------------------|
| Use cases (for various sectors)        | To identify various use cases, services model based on de-identification in the various sectors, such as health, telecommunication, and banks. | ITU-T SG17                | Recommendation    | Potential NWIP (new work item proposal) |
| Security and privacy threats           | To identify security threats and risks based on risk assessment (identification, analysis and evaluation) related to de-identification based services | ITU-T SG17                | Recommendation    | Potential NWIP           |
| Terminology and techniques             | To identify common terminology and de-identification techniques that can be applied to various sector | ISO/IEC SC27              | International Standard | ISO/IEC 20889 + potential NWIP |
| Security and privacy requirements and architecture | To identify security and privacy requirements to mitigate the risks and to develop de-identification framework and architecture | ISO/IEC JTC 1/SC 27       | International standard | Potential NWIP           |
| Security and privacy controls         | To develop security and privacy controls to meet the security and privacy requirements | ISO/IEC JTC 1/SC 27       | International standard | Potential NWIP           |
| Security and privacy mechanisms       | To develop the detailed security and privacy mechanisms to implement the security controls | ITU-T SG17                | Recommendation    | Potential NWIP           |
| Assurance framework                   | To provide de-identification assurance level and criteria for each assurance level | ITU-T SG17, ISO/IEC JTC 1/SC 27 | Common text [81] | ITU-T X.rda [15] + potential NWIP |
| De-identification process and framework | To develop the de-identification process and framework/architecture | ITU-T SG17, ISO/IEC JTC 1/SC 27 | Common text [81] | ITU-T X.fdp [14] + Potential NWIP [82] |
| Organizational security and privacy management | To develop the organizational security management aspect’s measures | ITU-T SG17                | Recommendation    | Potential NWIP           |
| Operational measures                  | To develop the operational measures for de-identification | ITU-T SG17                | Recommendation    | Potential NWIP           |

the item 4 should focus on high level requirements for de-identification and architecture consisting of major components for implementing the requirements, the item 5 should focus on all kind of controls such as physical and administrative controls, and the item 6 should focus on detailed technical mechanisms to implement each function or component.

5. Conclusion

De-identification is very important for government agencies, companies, and other organizations that seek to make data to be released to outsiders [6]. For example, significant medical research resulting in societal benefit is made possible by the sharing of de-identified patient information under the framework established by the regulation providing for the privacy of medical records. Data subjects need evidence from organizations that when processing their personal data, organizations are compliant with privacy regulations such as requirements for de-identification techniques. Certifications for the privacy management system are important. The international standard ISO/IEC 27701 [8] can be used by any organization that acts as a PII controller or processor to develop a Privacy Information Management System (PIMS). Certification to ISO/IEC 27701 proves that organizations have appropriate controls that address the requirements of privacy regulations. The role of standards is changing from a basis of certification to a way to ensure the right evidence of organizations is available. Personal data in government or organization dataset should be removed prior to publication, using de-identification techniques. Data published by organizations, which could be used to identify individuals
or small groups in the population, needs to be anonymized using de-identification techniques. If anonymization cannot safeguard PII or data can be used to identify individuals or small population groups, data cannot be published without additional technical measures and legal advice.

In the paper, we present the overview of de-identification techniques and future standardization directions for this important area. This paper could be used for standard community to identify the future directions and to identify new work items for this de-identification techniques and their applications.

Acknowledgments

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-00660). It was also supported by Soonchunhyang University, Republic of Korea.

References

[1] L. Sweeney, “Simple demographics often identify people uniquely,” Data Privacy Working, p.3, 2000.
[2] Y.-A. de Montjoye, C.A. Hidalgo, M. Verleysen, and V.D. Blondel, “Unique in the crowd: The privacy bounds of human mobility,” Scientific Reports, vol.3, Article number 1376, 2013.
[3] Y.-A. de Montjoye, L. Radaelli, V.K. Singh, and A.S. Pentland, “Unique in the shopping mall: On the reidentifiability of credit card metadata,” Science, vol.347, no.6221, pp.536–539, Jan. 2015.
[4] A. Narayanan, “How to break anonymity of the Netflix prize dataset,” Cornell University, 2006, https://arxiv.org/abs/cs/0610105, accessed 14 Nov. 2019.
[5] K. El Emam, E. Jonker, L. Arbuckle, B. Malin, and R.W. Scherer, “A systematic review of re-identification attacks on health data,” PLOS ONE, vol.10, no.4, e28071, 2011.
[6] ISO/IEC 20889:2018, Privacy enhancing data de-identification terminology and de-identification techniques.
[7] ISO/IEC 29100:2011, Information technology — Security techniques — Privacy framework.
[8] ISO/IEC 27701:2019, Security techniques — Extension to ISO/IEC 27001 and ISO/IEC 27002 for privacy information management — Requirements and guidelines.
[9] Information and Privacy Commissioner of Ontario, De-identification Guidelines for Structured Data, June 2016.
[10] ISO 25237:2017, Health informatics — Pseudonymization.
[11] ISO/IEC 19944, Information technology — Cloud computing — Cloud services and devices: Data flow, data categories and data use.
[12] ITU-T X.1058 | ISO/IEC 29151:2017, Information technology — Security techniques — Code of practice for personally identifiable information protection, 2017.
[13] ISO/IEC 27000, Information technology — Security techniques — Information security management systems — Overview and vocabulary.
[14] ITU-T X.6dip, Framework of de-identification process for telecommunication service providers (under development).
[15] ITU-T X.rdda, Requirements for data de-identification assurance (under development).
[16] ISO/IEC 9797-2, Information technology — Security techniques — Message Authentication Codes (MACs) — Part 2: Mechanisms using a dedicated hash-function.
[17] ISO/IEC 10118 (all parts), Information technology — Security techniques — Hash-functions.
[18] ISO/IEC 11770 (all parts), Information technology — Security techniques — Key management.
[19] ISO/IEC 18033 (all parts), Information technology — Security techniques — Encryption algorithms.
[20] ISO/IEC 19592 (all parts), Information technology — Security techniques — Secret sharing.
[21] ISO/IEC 27038, Information technology — Security techniques — Specification for digital redaction.
[22] ISO/IEC 29101, Information technology — Security techniques — Privacy architecture framework.
[23] ISO/IEC 29134, Information technology — Security techniques — Guidelines for privacy impact assessment.
[24] ISO 31000, Risk management — Principles and guidelines.
[25] NIST, NISTIR 8053: De-identification of personal information, Oct. 15, 2015.
[26] NIST, NIST Special Publication 800-38G, Recommendation for block cipher modes of operation: Methods for format-preserving encryption, March 2016.
[27] GDPR, General Data Protection Regulation, EU, 2016.
[28] Article 29 Data Protection Working Party, Opinion 05/2014 on Anonymisation Techniques, April 2014.
[29] Information Commissioner Office (ICO), Anonymisation: Managing data protection risk code of practice, July 2014.
[30] B.C.M. Fung, K. Wang, R. Chen, and P.S. Yu, “Privacy-preserving data publishing: A survey on recent developments,” ACM Computing Surveys, vol.42, no.4, Article No.14, June 2010.
[31] A. Blum, K. Ligett, and A. Roth, “A learning theory approach to non-interactive database privacy,” Proc. 40th ACM Symposium on Theory of Computing (STOC 08), pp.609–618, ACM, 2008.
[32] M. Bun and T. Steinke, “Concentrated differential privacy: Simplifications, extensions and lower bounds,” 14th International Conference, Theory of Cryptography Conference, TCC 2016-B, Beijing, China, Oct. 31-Nov. 3, 2016, Lecture Notes in Computer Science, vol.9985, pp.635–658, Springer-Verlag, 2016.
[33] S. Chawla, C. Dwork, F. McSherry, A. Smith, and H. Wee, “Towards privacy in public,” Proc. 2nd International Conference on Theory of Cryptography (TCC 05), Lecture Notes in Computer Science, vol.3378, pp.363–385, Springer, Berlin, Heidelberg, 2005.
[34] T. Dalenius, “Towards a methodology for statistical disclosure control,” Statistik Tidskrift, vol.15, pp.429–222, 1977.
[35] J. Domingo-Ferrer and J. Soria-Comas, “From t-closeness to differential privacy,” Proc. 2009 European Conference on Data Privacy, (Eurocrypt 2006, Lecture Notes in Computer Science, vol.4004, pp.486–503, Springer-Verlag, 2006.
[36] C. Dwork and A. Roth, “The algorithmic foundations of differential privacy,” Foundations and Trends in Theoretical Computer Science, vol.9, no.3–4, pp.211–407, 2014.
[37] C. Dwork and G.N. Rothblum, “Concentrated differential privacy,” arXiv: 1603.01887, 2016.
[38] K. El Emam and F.K. Dankar, “Protecting privacy using k-anonymity,” J. Amer. Med. Info. Assoc., vol.15, no.5, pp.627–637, 2008.
[39] K. El Emam and B. Malin, “Appendix B: Concepts and methods
Heung Youl Youm received his PhD degree in Electronics Engineering from Hanyang University, Seoul, Korea in 1990. He received his Master and Bachelor degree in Electronics Engineering from Hanyang University, Seoul, Korea, in 1981 and 1983, respectively. Currently, he is a Professor in the department of information security engineering at the Soonchunhyang University, Korea. He is a director of SCH cybersecurity research center since December 2014. He is the emeritus president of KIISC (Korea Institute of information security and cryptology) and was the president of KIISC in 2011. He is a chairman of ITU-T SG17 during 2009 to 2016. He is a member of self-evaluation committee for ministry of Science and ICT since 2013. He was a chairman of ITU-T WP 3/SG 17 during 2013 to 2016 and a chairman of ITU-T WP 2/SG 17 during 2009 to 2012. He had been working for the former MIC (Ministry of Information and Communication), Korea as a Project Manager for information security, from November 2006 to February 2008. His current interest includes theoretical and practical study on various security technologies/protocols such as IPTV/USN/NGN/IoT security. He had been an editor-in-chief for the KIISC Journal for KIISC (Korea Institute of Information Security and Cryptology) from January 2008 to December 2009, respectively. Since 2005, he has contributed to ITU-T by serving as an editor of more than 20 approved ITU-T Recommendations or Supplements such as Recommendation X.1034 (Guideline on extensible authentication protocol based authentication and key management in a data communication network), X.1111, X.1311 (Information technology – Security framework for ubiquitous sensor networks), X.1151 (Guideline on secure password-based authentication protocol with key exchange), X.1158 (Multi-factor authentication mechanisms using a mobile device), X.1191 (Functional requirements and architecture for IPTV security aspects), X.1193 (Key management for IPTV services), X.1196 (Framework for the downloadable service and content protection system in the mobile Internet Protocol television environment), X.1197 (Guidelines on criteria for selecting cryptographic algorithms for IPTV service and content protection), X.1208 (A cybersecurity indicator of risk to enhance confidence and security in the use of telecommunication/information and communication technologies), ITU-T X.1210 (Overview of source-based security troubleshooting mechanisms for Internet protocol-based networks), ITU-T X.1361 (Security framework for the Internet of things based on the gateway model), X.1127 (Functional security requirements and architecture for mobile phone anti-theft measures) and X Suppl. 10 (ITU-T X.1205 – Supplement on usability of network traceback).