NESTGate—Realizing Personal Data Protection with *k*-Anonymization Technology

● Yoshihiro Morisawa ● Shinji Matsune

Anonymization is attracting attention as a technology that can prevent the identification of a specific individual from data representing personal information (name, address, age, etc.) and private information (location information, route information, purchase history, etc.). NESTGate is a personal data protection tool implementing Fujitsu Laboratories' version of *k*-anonymization technology. This technology, which is being heavily researched around the world, can process data representing personal information so that at least *k* individual records have the same attributes. NESTGate provides a user interface that can easily handle *k*-anonymization in business operations, and it processes data to make it difficult to identify individuals from a large volume of information. NESTGate also features an authentication function to control access to personal information that includes sensitive information and a job management function to monitor the state of job execution and prevent the simultaneous execution of multiple anonymization processes. NESTGate is a product that can be rapidly incorporated and used in cloud computing and business systems handling personal data. This paper discusses points of concern in handling personal data and describes NESTGate functions.

1. Introduction

Driven by the spread of cloud computing, the expansion of inter-industry business, and the diversification of telecommunication devices such as smartphones and tablets, a new era has arrived in which information that has traditionally been stored in a company's in-house system is also coming to be stored, accumulated, and used outside the company. In this era, the formation of new markets is highly anticipated thanks to advances in big data technology for collecting and analyzing highly diverse and massive amounts of information. Concurrently, 2015 marks the tenth year since the Act on the Protection of Personal Information (commonly known as the Personal Information Protection Act) came into effect in Japan. A bill to revise this act was approved by the Diet in 2015 in response to increasing public concern about the handling of personal information due to many news reports of companies leaking personal information. As a result, we have entered an era in which there has never been a greater need for technology for appropriately protecting and using personal data, which represents personal information (name, address, age, etc.) and private information (location information, route information, purchase history, etc.).

Anonymization technology is attracting attention as a means of reducing the risk of a specific individual being identified from large amounts of diverse information that includes personal data. Anonymization is the processing of information so that a specific individual cannot be identified from the processed information. A number of anonymization algorithms are available. Generally, the party that owns the data chooses which algorithm to use in accordance with the reason for using the information as well as the range of its release.

It is not easy for a company in need of anonymization in daily operations to understand such algorithms and to systematize operations for anonymizing information. Consequently, it has become routine for some companies to make the information itself partially meaningless by removing certain attributes or redacting certain data. However, from the viewpoint of putting information to good use, such processing

methods limit the meaningful use of the information. Thus, there is growing interest in *k*-anonymization as it does not impose restrictions on the use of the processed information. This technology processes information so as to reduce the risk of identifying an individual while preserving as much value of the information as possible.

In this paper, we describe *k*-anonymization technology and the functions of NESTGate, a tool based on this technology. We also discuss some points of concern regarding the use of anonymization technology.

2. Personal data and risk of individual identification

The Personal Information Protection Act, which went into effect in April 2005, designates a business operator who possesses the personal information of 5,000 persons or more in a database or similar data structure and who uses that information for business purposes as an entity handling personal data. All business operators with this designation are obliged to protect personal rights and interests while considering how to make use of such personal information. In the ten-year period since this act went into effect, advances in information and communications technology (ICT) have highlighted problems that could not have been imagined in past data usage scenarios. Consequently, an amendment to this act was submitted to the Diet in 2015 and approved in September of that year. The major outcomes of this amendment were as follows (**Figure 1**).[1)]

1) Clarification of the definition of personal information
2) Ensurance of the usefulness of personal information in accordance with appropriate regulations
3) Enhanced protection of personal information
4) Establishment of the Personal Information Protection Commission and a definition of its authority
5) Development of procedures for handling personal information worldwide
6) Other revisions

This amendment recognizes the corporate right to freely use information that has been subjected to a process that anonymizes the information and thereby prevents specific individuals from being identified. Taking into account the economic revitalization from the use of personal information in this way reflects a new approach that embraces the use of big data technology for collecting and analyzing massive amounts of personal data.

The Personal Information Protection Act defines

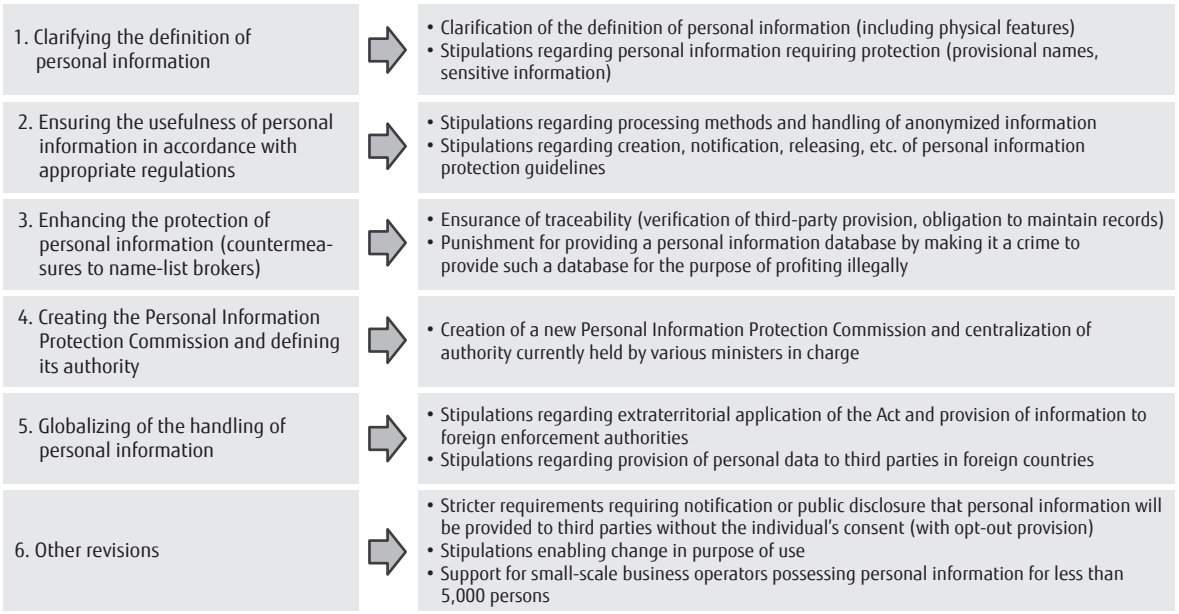

Figure 1
Revisions to Personal Information Protection Act.

personal information as information regarding a living individual—including name, date of birth, and other descriptors—that can be used to identify that individual and information regarding a living individual that can be easily compared with other information to identify that individual. Personal data, on the other hand, includes data such as location information and purchase history that, while insufficient for identifying that individual, could be used with other information to do so. We define personal data as shown in **Figure 2**.

There is thus a risk that an individual may be identified from only their personal data, such as their medical information. **Figure 3 a)** shows medical history information with the personal information (names) removed. This type of processing to remove the names of individuals is commonly performed and is considered to produce data that does not include personal information. However, if a person who knows that Mr. Nakamura, a 65-year-old male, is on this list, they will likely be able to identify his record. In this example, although the names have been removed as prescribed by the Personal Information Protection Act, there is still the risk that an individual can be identified by comparing known information with other attributes on the list.

This illustrates the need for a technology that prevents the identification of an individual by combining attributes that are not directly related to personal information. To realize this, privacy preserving data mining (PPDM) is being researched with the aim of applying it to the processing of big data. In PPDM, the focus is on anonymization technology as a means of reducing the risk of an individual being identified. *k*-Anonymization is one type of technology that accomplishes this.

Figure 3 b) shows the results of applying *k*-anonymization processing to the data shown in Figure 3 a). The age field value has been converted from units of one year to units of ten years. As a result, Mr. Nakamura's record is no longer unique, and therefore it cannot be identified as in the previous manner.

k-Anonymization is a technology for processing information so that a record cannot be distinguished from that of $k-1$ individuals even by combining attributes. This technology is currently being researched by a variety of institutions.

3. NESTGate anonymization processing

NESTGate is a software package that implements *k*-anonymization, a promising technology for protecting personal data. NESTGate also incorporates anonymization processing based on statistical methods presented in the Guidelines for the Preparation and Provision of Anonymous Data[2)] (established in February 2010 by the Ministry of Internal Affairs and Communications) and which can be used in a variety scenarios.

In this section, we outline the various types of anonymization processing implemented by NESTGate. Each type of processing can be used alone or in combination with other types.

1) *k*-Anonymization

NESTGate includes three *k*-anonymization algorithms.

- Anonymization by generalization

This algorithm achieves *k*-anonymization by generalizing quasi-identifiers (QIs). In contrast to an identifier (information such as a name that can directly identify an individual), a combination of QIs constitute

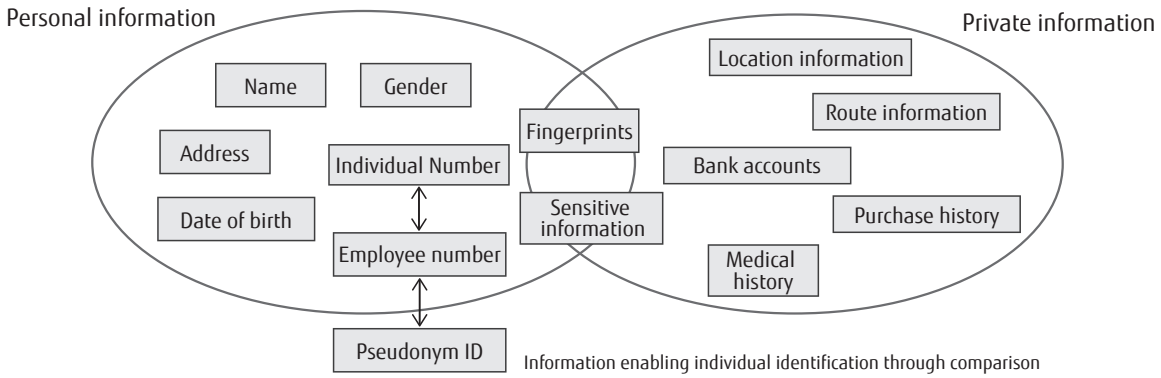

Figure 2
Personal data.

information that can potentially be used to identify an individual. In the example in **Figure 3**, age and gender are QIs. Anonymization by generalization protects information that can directly identify an individual by adding ambiguity to QIs.

- Anonymization by prioritizing existence of information

This algorithm adds noise to QIs but places priority on the existence of information. No information is lost, but the values of the QIs different from those of the original information.

- Anonymization by item deletion

This algorithm deletes QIs to eliminate identifiability. Although those QIs can no longer be used, the risk of an individual being identified is minimized.

For all of these algorithms, specifying a value for *k* results in the processing of information such that the number of rows with the same attributes is equal to at least *k* records. The user decides whether to prioritize information existence or to protect information through deletion. The algorithm used depends on the reason for using the information, the scope of its release, and/or the nature of the information itself (sensitive information, etc.).

2) Sorting

If the order of rows after anonymization processing happens to be the same as that of the original information, it may be simple for a person who knows the order of the original information to infer which record belongs to which individual even after anonymization. To mitigate this risk, NESTGate recommends sort processing at the end of anonymization processing. This sorting is generally done by rearranging the data in dictionary order, but sort processing in NESTGate rearranges the rows so that rows with the same attributes are output together. This produces a shuffling effect that makes the order of rows different from that of the original information.

3) Item conversion

This process replaces the value of an attribute

a) Name removal

Name	Telephone	Gender	Age	Conditions	Income (million yen)
Yamamoto	03-...	Male	40	Diabetes	6.0
Nakamura	03-...	Male	65	Diabetes	4.0
Koyama	03-...	Male	60	Diabetes	5.0
Yamamoto	03-...	Male	43	Diabetes	7.0
Sato	03-...	Male	48	None	8.0
Tanaka	03-...	Female	26	Diabetes	4.0
Nakamoto	03-...	Female	22	None	3.0

b) *k*-Anonymization

	Telephone	Gender	Age	Conditions	Income (million yen)
Yamamoto	03-...	Male	40s	Diabetes	6.0
Nakamura	03-...	Male	60s	Diabetes	4.0
Koyama	03-...	Male	60s	Diabetes	5.0
Yamamoto	03-...	Male	40s	Diabetes	7.0
Sato	03-...	Male	40s	None	8.0
Tanaka	03-...	Female	20s	Diabetes	4.0
Nakamoto	03-...	Female	20s	None	3.0

Figure 3
Example of *k*-anonymization.

with another value. A regular expression may be specified as a substitute character string. This enables the usage of a wide variety of replacement character strings. For example, the following types of conversion are possible.

- Attribute top coding

The value of the attribute is replaced by another value if it exceeds a certain threshold. For example, if the age of the individual is 50 or more, the expression "50 or older" can replace the original value of this attribute.

- Attribute bottom coding

The value of the attribute is replaced by another value if it falls below a certain threshold. For example, if the age of the individual is less than 50, the expression "under 50" can replace the original value of this attribute.

- Attribute grouping

Attribute values are placed into a certain group, and the value of any one attribute is replaced by another value accordingly. For example, the value of the age attribute can be replaced by an expression such as "10s" or "20s".

4) Swapping

The value of a certain attribute is replaced with a value from a different row. The replacement value is selected so as to be close to the original value.

5) Resampling

The original information is thinned. The ratio of remaining rows and the records to be deleted can be specified randomly.

6) Error introduction

Noise is randomly added to specified attributes.

7) Attribute deletion

Rows or columns are deleted from the original information. It is assumed that this process will be used to delete identifiers directly related to the individuals, such as the name attribute.

4. Function configuration of NESTGate

The functions in NESTGate are divided into the functional blocks shown in **Figure 4**. Here, we describe the user interface, job management function, and authentication function.

1) User interface

This functional block provides an interface for executing the NESTGate anonymization functions. NESTGate includes a server for performing anonymization and client functions executed by the user. The Representational State Transfer (REST) message protocol is used between the server and client as a mechanism for performing HTTP message-based communication.

Web-based systems have recently come into use for a variety of tasks and operations. These systems can be linked to and integrated with existing systems. It is easy to customize this user interface as the need arises. Microsoft's smart client technology is used to implement the standard client functions.

2) Job management function

Anonymization in NESTGate is accomplished by batch processing using comma separated value

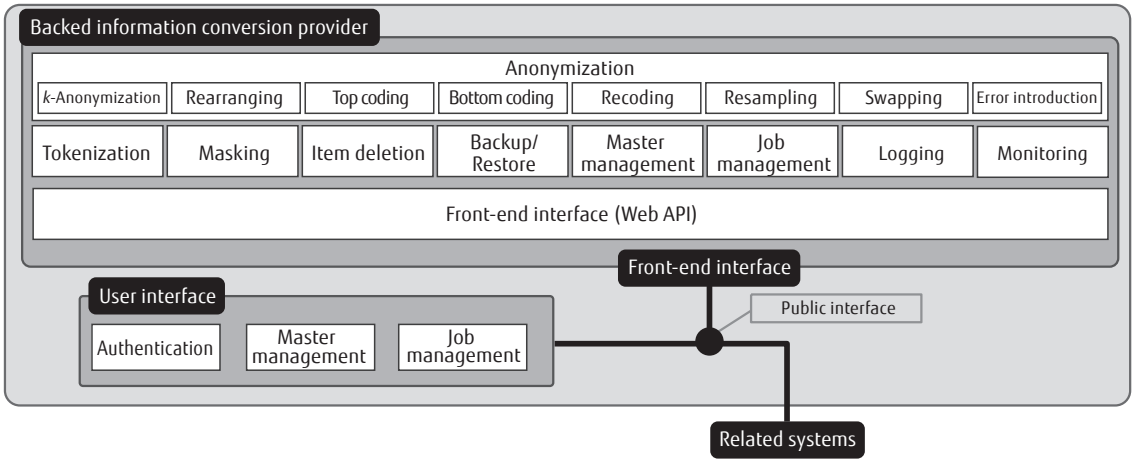

Figure 4
NESTGate *k*-anonymization function configuration.

(CSV) files for both input and output. Anonymization processing is performed one job at a time—no parallel processing is performed on the same server. NESTGate provides a job management function to place a scheduled anonymization process into standby if the processing of another job is still in progress.

The job management function executes a series of jobs; it receives requests for anonymization, performs anonymization processing, and returns the results to the client. It also includes other functions, such as a job cancellation function, all of which are performed through a Web service API (application programming interface) provided by the server.

3) Authentication function

The authentication function is extremely important when the client is using NESTGate anonymization functions from a web system or similar resource. It can be used to control access so that only authorized users are allowed to use NESTGate's functions. If the information targeted for anonymization happens to include sensitive information, access to such information may need to be restricted. NESTGate's authentication function enables this.

5. Points of concern regarding use of anonymization

Further evolution of ICT is expected to support and enrich our lives in the years to come. At the same time, the owners of personal data will be required to take great care in the way they handle that data. Revisions to the Personal Information Protection Act have helped clarify stipulations for handling anonymized information. The anonymization technology described in this paper is useful for both protecting and using personal data.

Nevertheless, there are cases in which the premise that anonymized information cannot be used to identify individuals is false. Let's look at such a case, which provides an example demonstrating the difficulty of anonymization.

In 1997, information on the governor of Massachusetts in the United States was identified from anonymized medical information This information was released after names identifying individuals had been deleted; however, along with his name, comparison of the medical information with a commercially available voter registration list made it possible to identify the governor's information. This is a famous example of how an individual can be identified by comparing anonymized information with publically released information.

NESTGate is a personal data protection tool equipped with a number of anonymization algorithms. Nevertheless, users should first give careful consideration to the purpose of use, the scope of release, and the risk of leakage with respect to identifiers and quasi-identifiers prior to anonymizing personal data.

6. Conclusion

In this paper, we described the use of anonymization to protect personal data and the functions of the NESTGate tool provided by Fujitsu. *k*-Anonymization implemented in NESTGate is a technology attracting growing interest for facilitating the handling of personal data. We are currently developing practical applications for *k*-anonymization and examining the use of *l*-diversity and homomorphic encryption technologies, which are currently being developed.

References

1) Act on the Protection of Personal Information (Act No. 57 of 2003).
http://www.cas.go.jp/jp/seisaku/hourei/data/APPI.pdf
2) Ministry of Internal Affairs and Communications (MIC): Guidelines for the Preparation and Provision of Anonymous Data (in Japanese).
http://www.soumu.go.jp/main_content/000398971.pdf

Yoshihiro Morisawa
Fujitsu Systems West Ltd.
Mr. Morisawa is currently engaged in the development of business solutions focused on security.

Shinji Matsune
Fujitsu Systems West Ltd.
Mr. Matsune is currently engaged in the development of NESTGate.