Digital Security Systems Protecting Society from Potential Threats

● Yuji Yamaoka

As the value brought about by data utilization is attracting attention, the global big data market has been making growth at an annual rate of more than 10% and is estimated to reach a scale of about 20 trillion yen by 2020. In line with this trend, development of laws for promoting proper data distribution and utilization has progressed on a global basis, and technologies to meet the requirements of those laws are beginning to come into practical use. However, there are also anxieties expressed arising from the inability to make decisions regarding risks in personal data distribution; individuals may agree to data distribution without realizing how high the risk is, or business owners may cause privacy issues by distributing personal data with low anonymity, possibly resulting in major losses such as compensation for damages. To deal with these issues, Fujitsu Laboratories has developed a technology to quantify privacy risks from personal data disclosures in terms of monetary value. We have also developed a model for calculating the identifiability (how low anonymity is) of data after anonymization, which was insufficient in the past, and confirmed that these are applicable to real data. Furthermore, we have developed a high-speed identifiability calculation technology that allows for the calculation of data sets on a scale of 1 million people in about an hour with a general performance PC, confirming adequate practicability. This paper describes the technology that allows for risk evaluations regarding personal data and the concept of realizing a society that can better extract the value of data by utilizing this technology.

1. Introduction

Recently, as advances are seen in technologies such as cloud computing, AI, and IoT, the value brought about by data utilization is attracting attention.

The global big data market continues to grow at an annual rate of 11.9% and is expected to reach about 20 trillion yen by 2020.[1)] In particular, personal data was compared to the new oil and currency by the World Economic Forum in 2011[2)] and this value is drawing attention. In view of this trend, laws are being developed around the world to ensure proper data distribution and utilization.

In Japan, data utilization is said to be effective for solving various issues in this super-aging society, and laws intended to expand data distribution are under development. In 2015, the Personal Information Protection Act was amended (fully enforced in 2017) and a system allowing for the free utilization of anonymously processed information (personal data processed so that a specific individual cannot be identified by following the rules) was established. In 2016, the Basic Act on the Advancement of Public and Private Sector Data Utilization was enforced, and the distribution and utilization of data are promoted by both the public and private sectors while the people's rights and interests are protected.

In Europe, rights to the protection of personal data were positioned as fundamental human rights, and a law for their protection called the General Data Protection Regulation (GDPR) was enforced on May 25, 2018. Under this law, in principle all organizations must explicitly obtain the consent of the individual regarding the purpose of collection and use of the personal data.

Against this backdrop, Fujitsu is promoting Connected Services[3)] that maximize the value from the data owned by customers. To maximize value, often data must be analyzed in combination with data

from other organizations. For example, in response to a request for predicting the effect of setting up a shop in a certain location, the prediction accuracy can be increased if customer data can be analyzed using information such as consumption trends in the area around the location. Preferably, information such as the age bracket, gender, time slot, consumption expenditure, and number of consumers should be obtained from other shops in the area around the location. The smooth transfer of such information requires vehicles for the distribution of data between organizations through the buying and selling of data and so on. Fujitsu is also working on the realization of such vehicles.

In data distribution between organizations, compliance with laws and regulations is required when handling personal data. Personal data distribution methods that are legal in Japan and many other countries include the following two methods: a method for distribution only within the scope to which the person's consent has been obtained (informed consent method) and a method for distribution of data subjected to anonymization into data that do not allow for the identification of individuals (anonymization method). These methods are chosen and used according to the purpose of data distribution.

So far, Fujitsu has been working on developing technologies for both methods and is leading the expansion of data distribution. We provide the FUJITSU Cloud Service for OSS Personium Service[4] for the informed consent method and the FUJITSU Business Application NESTGate Anonymization[5] and the FUJITSU Software Symfoware Analytics Server[6] for the anonymization method.

However, some individuals and business owners have expressed their concern over privacy. For example, individuals may agree to data distribution without realizing how high the risk is, and business owners may cause privacy issues by distributing personal data with low anonymity, possibly resulting in major losses such as compensation for damages. In this way, the inability to make decisions regarding the risk levels causes free-floating anxiety.

Fujitsu Laboratories has developed a technology to quantify privacy risks from personal data leakage in terms of monetary value in order to eliminate this anxiety and further expand data distribution. This technology allows for the evaluation of risks regardless of whether anonymization is applied, and the above-mentioned anxiety can be visualized as a specific value with either the informed consent method or the anonymization method. This enables business owners to distribute data with a sense of security. Other business owners can combine the data with data they own to extract further value from it.

This paper details the technology that allows for risk evaluations regarding personal data.

2. Conventional evaluation technologies: anonymization not supported

Conventional technologies for quantifying risks related to personal data include the JNSA Damage Operation Model for Individual Information Leak (JO Model)[7] and k-anonymity.[8] The former does not support data subjected to anonymization, and the latter has an issue in the inability to evaluate the method of deciding on the target of quantification itself.

The JO Model is a formula for calculating the projected compensation for damages related to the leakage of personal information, and was created by experts on insurance and information security based on multiple precedents. This model features the capability to represent risk per person in monetary value, and a major part of it can be shown by equation below.

Value of personal information leaked =
Degree of information sensitivity
× Degree of ease in identifying the individual
× 500 yen (1)

Degree of information sensitivity is the result of the quantification of the impact on a person caused by a leakage of personal information leading to the identification of that person. Degree of ease in identifying the individual is the result of quantification of how easily a person can be identified when the relevant personal information is leaked.

However, this model is not designed assuming anonymization. Accordingly, degree of ease in identifying the individual often remains unchanged even if anonymization is applied, and reductions in the risk by anonymization cannot be evaluated. For example, anonymization includes the deletion of identifiers such as the item "E-mail address," classification of the item "Age," and generalization of the item "Occupation" as shown in **Figure 1**. This makes identification of the

individual more difficult, but the value calculated by the JO Model does not change between before and after anonymization.

Degree of information sensitivity is specified for each type of representative item. Furthermore, there is a calculation formula to deal with combinations of items, and it is generally approximate to the maximum value of each item. For example, the value is 2 for the item "Occupation" and 101 for the item "Religion" and 101 for the combination of these.

k-anonymity is an indicator for anonymity of personal data used around the world. It is based on a concept that, for a certain person X, the larger the number of people with the same data as X, the more difficult it is to identify X. For a group of multiple items (called quasi-identifiers) specified by the user of this indicator as the target of evaluation, the number of people with the same data is used to quantify the anonymity.

The number of people with the same data is often counted only within the target data set. This is because the number of people can be said to always exist. The idea is compatible with the consideration of differences within a data set (Article 19, Paragraph 5 of the Enforcement Rules for the Act on the Protection of Personal Information), which is emphasized in the requirements for anonymously processed information.

While k-anonymity is not an indicator for evaluating risk, this can be used for degree of ease in identifying the individual in the JO Model to produce a risk calculation model that supports data after anonymization. However, simply combining does not provide an appropriate model. For example, the decision of which types of information to use as quasi-identifiers is generally difficult.

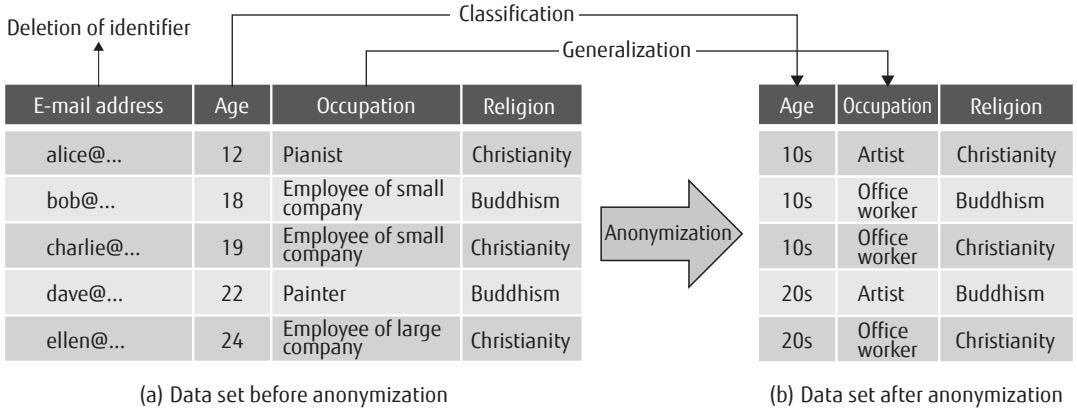
This model also has a problem in the inability to identify the risk that specification of quasi-identifiers is inappropriate. This problem can be avoided by using all items as quasi-identifiers, but anonymously processed information does not usually require anonymity of that level. Accordingly, the effect of anonymization for individual items such as age classification must also be evaluated, and practical evaluations cannot be conducted if all items are used as quasi-identifiers.

For example, consider the case where all items are used as quasi-identifiers to evaluate k-anonymity within the data in Figure 1. In this case, no one has the same data as anyone else before or after anonymization, and the calculated monetary value does not change between before and after anonymization. However, while data before anonymization clearly violate the rules of anonymously processed information, they do not violate the rules after anonymization. Therefore, it is inappropriate to decide that they have the same risk before and after anonymization.

To solve these problems, a model that does not require specification of quasi-identifiers by the user is necessary.

3. Risk evaluation technology

Fujitsu Laboratories have developed a technology capable of evaluating risk from the leakage of personal data before and after anonymization by applying a developed high-speed anonymization technology to exhaustively quantify anonymity at high speed. This



E-mail address	Age	Occupation	Religion
alice@...	12	Pianist	Christianity
bob@...	18	Employee of small company	Buddhism
charlie@...	19	Employee of small company	Christianity
dave@...	22	Painter	Buddhism
ellen@...	24	Employee of large company	Christianity

(a) Data set before anonymization

Age	Occupation	Religion
10s	Artist	Christianity
10s	Office worker	Buddhism
10s	Office worker	Christianity
20s	Artist	Buddhism
20s	Office worker	Christianity

(b) Data set after anonymization

Figure 1
Example of Anonymization.

technology calculates identifiability (how low anonymity is) by extending the concept of k-anonymity to make it applicable to more general anonymization.

The following subsections describe two features of this technology.

3.1 Identifiability calculation model

The first feature is the use of a model capable of more appropriate quantification of risk from the leakage of personal data before and after anonymization. This model quantifies identifiability of data for each person with the combination of items allowing the easiest identification within the data set. This is used to replace degree of ease in identifying the individual in the JO Model for representation in terms of monetary value.

Ease of identification models the following two properties.

- Property 1: Ease of acquisition of information

The easier it is to acquire information required for identification, the easier it is to identify. For example, the first person in Figure 1 (b) can be identified by age and occupation, and the second person by age and religion. In many cases, occupation information is easier to acquire than religion information, and the former is easier to identify if other conditions are the same.

- Property 2: Fewness of items

The fewer the items required for identification, the easier it is to identify. For example, the third person in Figure 1 (b) cannot be identified unless the three items—age, occupation, and religion—are combined. In contrast, the other four people can be identified by combinations of two out of the three items, such as age and occupation or age and religion, and are easier to identify.

Generally, more sensitive information is more difficult to acquire, and we have decided that information with higher degree of information sensitivity in the JO Model is more difficult to acquire. For example, religion information has higher degree of information sensitivity and is more difficult to acquire.

Figure 2 shows how this model is used to calculate identifiability (range of value: 0 to 1). The artist in their teens, artist in their 20s, and office worker in their 20s all have an identifiability value of 0.9, but assuming that there are more office workers in their 20s among the general public, this evaluation may seem odd. However, the possibility must be considered that individuals whose data may be registered within this data set do not include many office workers in their 20s, and the result is assumed to be reasonable.

3.2 High-speed search

The second feature is that identifiability in this model can be calculated at high speed.

In the calculation of identifiability, the combination of items of data for each person allowing the easiest identification within the data set is efficiently searched for. Based on the two properties described above, combinations providing higher identifiability are examined to judge whether each data can be specified, which allows for the elimination of unnecessary judgments as well as high-speed calculation. For example, for data that can be identified by occupation, judgment of whether it can be identified by religion can be eliminated based on Property 1, and identification by occupation is searched for before the religion. For data that can be identified only by age and occupation,

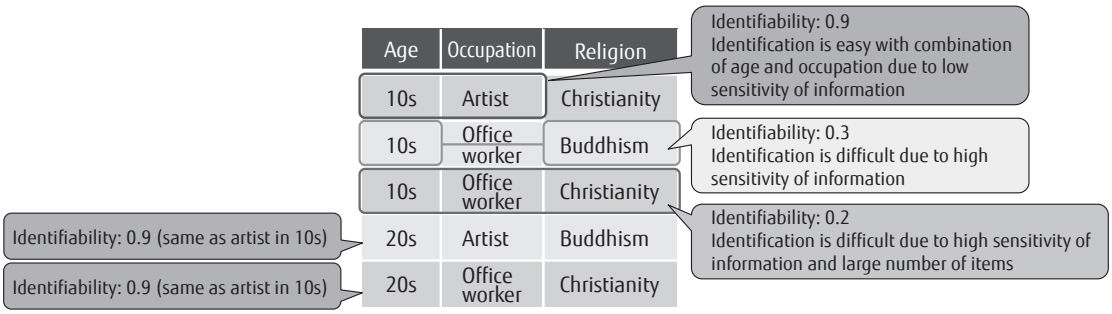


Figure 2
Example of identifiability calculation.

searching by age, occupation, and religion can be eliminated based on Property 2.

4. Performance evaluation

To verify that, unlike the conventional technologies, this technology allows for the quantification of risks before and after anonymization and that the processing time is practical, we conducted an experiment by using actual data.

4.1 Verification of calculation model

To verify risk quantification, we applied the technology to the Adult Data Set,[9] which is used as the benchmark for anonymization. For a total of 48,842 persons, we used nine items (age, workclass, education, marital status, occupation, race, sex, native country, and income) of data often used in prior research. For the data set, we applied three types of anonymization as shown below to create data sets and, together with the one before anonymization, we quantified risks for the four types of data sets.

(1) Classification of the item "age" at intervals of 10 years
(2) k-anonymization (processing for achieving k-anonymity) with the items other than "occupation" and "income" as quasi-identifiers (k=3)
(3) k-anonymization with the items other than "income" as quasi-identifiers (k=3)

Of these, anonymization (2) and (3) are types of processing often seen in prior research and, theoretically, (2), which has less quasi-identifiers, should have higher risk than (3).

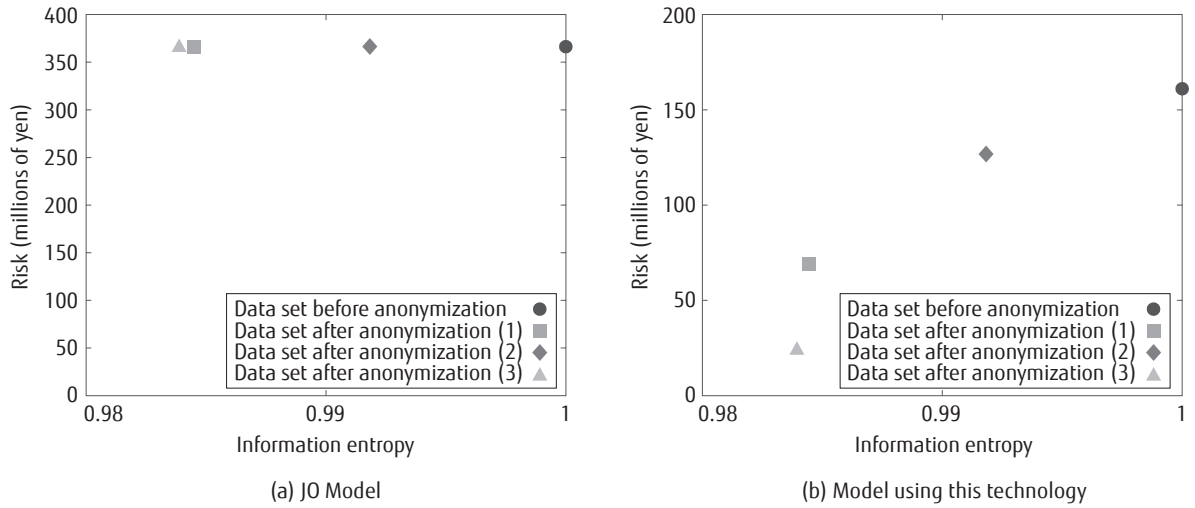
We also calculated the information entropy of the individual data sets. Generally, a smaller information entropy means lower risk, and validity of the model can be determined to some extent by observing the distribution. We calculated the information entropy based on the information gain (Kullback-Leibler divergence) often used in the statistics community.

The results of application of (1) to (3) are shown in **Figure 3**. The risk is a total of the values of personal information leaked for all persons in the JO Model shown in Equation (1). A larger information entropy and lower risk are better, and a graph plotted closer to the bottom right is closer to ideal anonymization.

As shown in Figure 3 (a), with the JO Model, the risk does not change between before and after anonymization. Meanwhile, Figure 3 (b) shows a trade-off between lowness of the risk and largeness of the information entropy, and anonymization (2) has higher risk than that of anonymization (3), in line with the theory. In this way, it has been verified that, unlike the conventional technologies, this technology is capable of quantifying risks before and after anonymization.

4.2 Verification of processing time

The processing time was verified by applying the technology to multiple data sets actually handled by



(a) JO Model

(b) Model using this technology

Figure 3
Relationship between information entropy and risks in the two calculation models.

Table 1
Processing time with actual data sets.

Data set	No. of records	No. of items	Size	Processing time
A	6,000	49	1 MB	2 sec
B	500,000	27	107 MB	30 min
C	940,000	29	241 MB	40 min
D	2.02 million	22	376 MB	1 hr
E	2.9 million	49	865 MB	1 hr

business owners.

The results of processing by using a general performance PC are shown in **Table 1**. Data sets on the scale of 1 million people are processed in about 1 hour, which means that data sets for the entire Japanese population can be processed in a few days. Generally, big data analysis requires trial and error and is said to take several weeks or longer. On the other hand, risk quantification using this technology is completed in a few hours to a few days, which accounts for only a small proportion of the entire work time from data distribution to utilization, and is assumed to offer high practicability.

5. Conclusion

This paper described the risk evaluation technology developed by Fujitsu Laboratories that allows for the distribution of personal data with a sense of security.

This technology enables business owners who had concerns over privacy up until now to start new businesses via the distribution of personal data with more peace of mind. For example, some business owners may present estimated risks to consumers when obtaining informed consent to data distribution from them, which offers a sense of security and makes it easier to obtain consent. Other business owners may develop their businesses by taking risks to the allowable upper limit to create anonymously processed data, whose demand from other business owners is strong.

In the future, we intend to expand secure distribution of personal data by commercializing this technology and aim to realize a society in which required data are distributed for the purpose of extracting more value from data owned by business owners. In this way, we wish to further improve the value offered by Connected Services.

References

1) IDC: Big Data and Business Analytics Revenues Forecast to Reach $150.8 Billion This Year, Led by Banking and Manufacturing Investments, According to IDC. *https://www.idc.com/getdoc.jsp?containerId=prUS42371417*
2) World Economic Forum: Personal Data: The Emergence of a New Asset Class. *http://www3.weforum.org/docs/WEF_ITTC_PersonalDataNewAsset_Report_2011.pdf*
3) Fujitsu: Fujitsu Technology and Service Vision. *http://www.fujitsu.com/global/vision/*
4) Fujitsu: Personium Service (in Japanese). *http://jp.fujitsu.com/solutions/cloud/k5/function/paas/personium/*
5) Fujitsu: FUJITSU Business Application NESTGate Anonymization (in Japanese). *http://www.fujitsu.com/jp/solutions/business-technology/intelligent-data-services/bigdata/ba-solutions/nestgate/anony/*
6) Fujitsu: Symfoware Analytics Server. *http://www.fujitsu.com/global/products/software/middleware/database/symfoware/products/analytics-server/*
7) NPO Japan Network Security Association: 2016 Survey Report of Information Security Incident–Personal Information Leakage– (in Japanese). *http://www.jnsa.org/result/incident/2016.html*
8) L. Sweeney: k-anonymity: a model for protecting privacy. Int. J. Uncertainty Fuzziness Knowl.-Based Syst., Vol. 10, pp. 557–570, October 2002.
9) C. Blake et al.: UCI Machine Learning Repository. *http://archive.ics.uci.edu/ml/*

Yuji Yamaoka
Fujitsu Laboratories Ltd.
Mr. Yamaoka is currently engaged in research and development for data privacy protection technology.