

Reviewing Privacy-Enhanced Social Survey System that Employs Combinatorial Anonymity Measure

A. IWAI

Abstract— Although conventional electronic systems for a social survey offer various levels of privacy protection, patterns in the input data itself can accidentally lead to the leakage of personal information. Recently, a prototype of a survey system that can automatically prevent such unintended information leakage has been proposed. The basic design consists of a framework that analyzes the input data to find elements that can lead to information leakage and a mechanism to correct such flaws by modifying the questionnaire design in the database. In this paper, we present a review of the abovementioned prototype system, particularly from the viewpoint of its usage of a combinatorial anonymity measure. The dichotomy aspect that divides alternatives into two parts seems to be reasonable, although it leads to limitations on the usage of the anonymity measure.

Index Terms— social survey, personal information, combinatorial anonymity measure, privacy preservation

I. INTRODUCTION

ONE of the most common requirements of social surveys is to ensure the privacy of respondents; this is particularly true with evaluation surveys undertaken by any official authority.

Even with a survey where no identification of respondents is required, there are many cases where data patterns give rise to concerns regarding privacy. For example, if an evaluation survey is conducted in a small class of 3 male and 15 female students, a question about the gender of the respondent would be harmful to the privacy of male students and may result in the deterioration of the quality of the obtained data.

Notice that this type of risk of accidental leakage of personal information also exists in relatively large-scale surveys, as cross tabulation of personal attribution data such as gender, age, or major may yield particular cells where only a small number of respondents are classified. Although in this study, we use course evaluation as a simple example, the aim of this study includes support for relatively large-scale surveys.

A prototype of the survey system that can automatically prevent such unintended information leakage has been

proposed in earlier studies (Iwai (2012a, 2013)). The basic design consists of a framework that analyzes the input data to find elements that can lead to information leakage and a mechanism to correct such flaws by modifying the questionnaire design in the database. It employs a mechanism that modifies the questionnaire design in the database after the respondents answer the questions and before any assessors see this information.

As technical tools for enhancing privacy, k-anonymity by Sweeney (2002) and l-diversity by Machanavajjhala et al. (2006) have been widely known, but these researchers aimed at concealing information from the public. In contrast, the aim of the present research is to conceal information from the survey assessors.

However, prototype development still needs a technical review. One critical point to review is its usage of an anonymity measure. The combinatorial anonymity measure used in the prototype system has been studied for more than a decade (Iwai (2003, 2012b, 2013), Edman et al. (2008), Gierlichs et al. (2008), Bagai et al. (2011)). However, most previous studies have focused on its usage in the context of voting privacy or network security. Social surveys where ordinal scales are often used are a new field of application.

In this paper, we present a review of the prototype system, particularly from the viewpoint of its usage of the combinatorial anonymity measure. We focus on the dichotomy aspect that divides alternatives into two parts and its implication on the usage of the anonymity measure.

The rest of this paper is structured as follows: Section II discusses the design and implementation of the prototype system. Next, Section III describes the review of the usage of the anonymity measure. Finally, Section IV presents the concluding remarks.

II. PROTOTYPE SYSTEM

This section discusses the prototype system rather precisely. Although the main topics of this section have been discussed in earlier study (Iwai (2012a, 2013)), this is the first English publication that addresses them.

A. Basic Approach

The prototype system consists of a framework that analyzes the input data to find elements that can lead to information leakage and a mechanism to correct such flaws by modifying the questionnaire design in the database.

Let us take a look at the first example of a course

Parts of this study were supported by grants from the Japan Society for the Promotion of Science, KAKENHI25282088.

A. Iwai is with the Faculty of Social and Information Studies, Gunma University, Maebashi city, 371-8510 Japan. (corresponding author to provide phone: +81-27-220-7440; fax: +81-27-220-7440; e-mail: iwai@si.gunma-u.ac.jp).

evaluation in a small class of 3 male and 15 female students. In this class, if a single question sheet contains a question concerning gender and the other questions about course evaluation, this would be harmful to male student privacy and can result in a deterioration of the quality of the obtained data. However, if the question sheet is divided into two parts, with one part including only the gender question and the other part only the course evaluation questions, then no privacy problem will arise to compromise the quality of the students' answers (Figure 1).

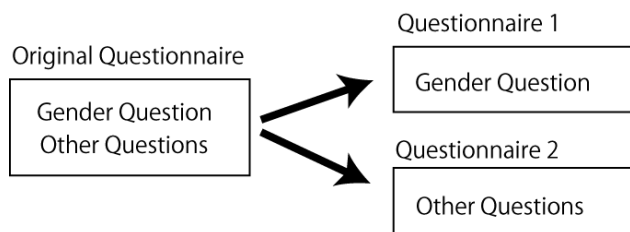


Figure 1 Modification of Questionnaire Design

The target system processes this division operation after all students have finished responding to the questionnaire, and when it finds problematic questions that can lead to information leakage. The division process is realized as a database operation of modifying the table structure related to the questionnaire design. As the computational process is triggered automatically and perfectly terminates before a lecturer sees the output of the system, no information leakage is possible.

The main topics of the system design are i) classification of questions and ii) threshold setting to determine elements that can lead to information leakage. The basic approach of each topic is discussed next.

B. Classification of Questions

This system design is based on the hypothesis that all questions on a question sheet can be divided into two categories, X and Y. X is defined as a category of individual attributes, such as gender or age. Y is defined as a category of individual attitudes such as course evaluation.

There is usually no reason to conceal an answer to a category X question; for example, the answers given to a gender question in a classroom can be confirmed at any time, as everybody in the classroom knows who is male or female. A category X question, however, can cause a privacy problem when it is asked along with a category Y question, as in the above example of a small class.

For each Y category question, a cross tabulation of many X category questions is likely to yield special cells where only a small number of respondents exist, and these cells are likely to cause some unintended information leakage.

When the system finds a certain level of risk of unintended information leakage, the system automatically corrects this situation.

C. Threshold Setting

Simply having a large number of respondents in each cell is also not sufficient to prevent unintended information leakage. For example, in the small class example, if all 15

female students and one of the 3 male students evaluate a lecture as poor (which implies that the other two male students have evaluated it as not poor), the privacy of the female students will be compromised and any estimated danger to them is more than that to the male students.

To establish a technical method of finding problematic cells of information leakage, Iwai (2012a) employs an anonymity metric as follows:

$$\log \frac{N!}{M!(N-M)!}$$

In the context of course evaluation, N represents the number of all respondents and M denotes the number of respondents who answered positively (or not positively). When this metric is applied to the above example, the anonymity level for female students is calculated as $\log(1) = 0$ and that for male students is calculated as $\log(1/6)$. This reflects the fact that all female students replied negatively, and the anonymity level for female students is evaluated as zero. Meanwhile, the anonymity level for male students is higher than that for the female students.

Although setting the threshold values can be arbitrary, $\log({}_5C_2) = 1$ will be one of the reasonable assignments. The base of the logarithm is set to 10 in this study.

D. Implementation

This subsection illustrates a prototype system implemented on the basis of the design described above.

The system was implemented for course evaluation in one department of a national university in Japan. The question sheet in the department consists of 11 Japanese questions. Three of them are X category questions (gender, grade, and number of absent days), and the remainder are Y category questions. The content of the question sheet is precisely implemented in the system interface. As the threshold, the value of $\log({}_5C_2)$ was assigned.

Figure 2 shows the webpage of the questionnaire.



Figure 2: Interface of Course Evaluation System

For each Y category question, a cross tabulation of many X category questions is likely to yield special cells where only a small number of respondents exist, and these cells are likely to cause some unintended information leakage.

A formal description of the system implementation is as follows:

[Sets of Respondents and Questions]

P denotes the set of all respondents. Using the respondent number $i(i \leq N)$, we can define it as follows:

$$P = \{1, 2, \dots, N\}$$

Q represents the set of all questions in a question form. Each element of Q is classified into X items and Y items.

X items (x_1, x_2, \dots, x_n) represent individual attributes that are observable by outsiders.

Y items (y_1, y_2, \dots, y_m) represent individual attitudes that are not observable by outsiders.

The index numbers of X items x_1, x_2, \dots, x_n denote the priority (an item with a relatively large index number is eliminated faster in the database).

$$X = \{x_1, x_2, \dots, x_n\}$$

$$Y = \{y_1, y_2, \dots, y_m\}$$

$$Q = X \cup Y$$

[Respondents' Answers]

For $\forall q \in Q$, $D(q)$ represents the domain of the answer to the question q . The 3-tuple (i, q, a) indicates that a respondent $i \in P$ selected the answer $a \in D(q)$ for the question $q \in Q$. T_0 denotes the set of all such 3-tuples and contains the information of all the answers provided by all the respondents. For $\forall q \in Y$, $D_c(q)$ is defined as the set of all $D(q)$ elements that are sensitive alternatives requiring their selectors to be concealed. That is, $D(q)$ represents the set of the alternatives of a negative evaluation.

[Question Block]

A question block is defined as a non-empty subset of Y . Different questions that belong to a question block are expected to be analyzed together by using multivariate analysis methods. Each question belongs to a different question block, and all questions belong to some question blocks. When the number of question blocks is M in total ($B_j(1 \leq j \leq M)$), for all $j(1 \leq j \leq M)$, the following holds.

$$B_j \neq \phi$$

$$a \neq b \rightarrow B_a \cap B_b = \phi$$

$$Y = \bigcup_{1 \leq j \leq M} B_j$$

For each $B_j(1 \leq j \leq M)$, (the initial value of) the set of answer $AB_j(1 \leq j \leq M)$ can be defined as follows:

$$AB_j = Delete(T_0, 2, Y - B_j)$$

Here, $Delete(S_1, j, S_2)$ returns a subset of the 3-tuple set S_1 . This calculation eliminates the 3-tuple of S_1 , when the j -th element of the 3-tuple belongs to set S_2 . Similarly, $Select(S_1, j, S_2)$ is defined as a function that returns a subset

of the 3-tuple set S_1 , and this calculation selects the 3-tuple of S_1 when the j -th element of the 3-tuple belongs to set S_2 . $Project(S, j)$ denotes a function that returns the set of all j -th elements of the 3-tuples that belong to S . Similarly, $Random(S, j)$ represents a function that swaps the j -th elements of all 3-tuples of S and returns the set that can be obtained as a result of the calculation.

[Grouping of Respondents by Individual Attributes]

$$\text{For } Project(AB_j, 2) \cap X = \{x_1, x_2, \dots, x_k\},$$

$$Dx(AB_j) = D(x_1) \times D(x_2) \times \dots \times D(x_k)$$

is defined. As each element of $\{x_1, x_2, \dots, x_k\}$ represents a question about individual attributes, the answer of $i \in Project(AB_j, 1)$ is related to one point of $Dx(AB_j)$. (In the case of $Dx(AB_j) = \phi$, we consider this point to be ϕ .) The respondent group that relates to point x of $Dx(AB_j)$ is denoted as $G(AB_j, x) (\subset P)$. (In the case of $Dx(AB_j) = \phi$, $G(AB_j, x) = P$.)

[Finding Risky Elements]

$L(AB_j, x, q)$, which represents the anonymity level observed at $x \in Dx(AB_j)$ and $q \in B_j$ with AB_j , is defined as follows:

$$L(AB_j, x, q) = \log\left(\frac{|DS|!}{|DS - DSc|! \times |DSc|!}\right)$$

Here, DS and DSc are the abbreviations of $DS(AB_j, x, q)$ and $DSc(AB_j, x, q)$ that are defined as follows:

$$DS(AB_j, x, q) = Select(AB_j, 1, G(AB_j, x)) \cap Select(AB_j, 2, \{q\})$$

$$DSc(AB_j, x, q) = DS(AB_j, x, q) \cap Select(AB_j, 3, D_c(q))$$

$DS(AB_j, x, q)$ represents the set of all data at $x \in Dx(AB_j)$ and $q \in B_j$ with AB_j . $DSc(AB_j, x, q)$ represents the set of $DS(AB_j, x, q)$ elements whose answer part belongs to $D_c(q)$.

$Flag(AB_j)$ denotes the function to determine whether the operation of modifying the database is needed for protecting privacy with AB_j . That is, it returns a value of 1 if the following condition holds, and 0 if the condition does not hold.

$$\min_{x \in Dx(AB_j), q \in B_j, |DSc| > 0} (L(AB_j, x, q)) < Ta$$

Here, Ta represents the threshold value to find a risk.

[Main Routine to Enhance Privacy]

For each $AB_j(1 \leq j \leq M)$, perform the following procedure:
STEP 1)

If $Project(AB_j, 2) \cap X = \phi$ or $Flag(AB_j) = 0$, then go to

STEP 2.

If $Project(AB_j, 2) \cap X \neq \phi$ and $Flag(AB_j) = 1$,
 $AB_j = Delete(AB_j, 2, \{x_{Project(AB_j, 2) \cap X}\})$ and do STEP 1
again.

STEP 2)

Perform the following procedure:

i) $A_{Attributes} = Random(Delete(T_0, 2, Y), 1)$

ii) For $k(0 \leq k \leq n = |X|)$,

$$A_k = Random\left(\bigcup_{|Project(AB_j, 2) \cap X|=k} (AB_j), 1\right)$$

iii) Delete data except of $A_{Attributes}, A_0, A_1, A_2, \dots, A_n$

iv) Output $A_{Attributes}, A_0, A_1, A_2, \dots, A_n$

As the questionnaire of the focused department has three items of attribute questions, the number of final output tables is expected to be five.

E. Simple Evaluation

A simple evaluation experiment was conducted in a class titled "Programming I" in the department described in the previous subsection. After explaining the course evaluation system and the actual evaluation practice using the system, a survey to evaluate the system was conducted.

The following are the major questions for evaluating the course evaluation system and the answers given by students. (The question numbers are relabeled for this paper.) The number of respondents is 34, and the date of the survey is December 14, 2012. The alternatives a, b, c, and d represent "Strongly agree," "Agree," "Disagree," and "Strongly disagree" respectively, except for Q6 where a, b, c and d represent "It is very promising," "It is promising," "It is not promising," and "It is not promising at all" respectively.

Q1) You think that a procedure to enhance privacy, which is the aim of this system, would be valuable in course evaluation.

- a) 21 [61.76%] b) 13 [38.24%]
c) 0 [0.00%] d) 0 [0.00%]

Q2) You could understand how the system works to make input data more anonymous.

- a) 21 [61.76%] b) 13 [38.24%]
c) 0 [0.00%] d) 0 [0.00%]

Q3) You think that this system can contribute to the collection of more accurate course evaluation data.

- a) 16 [47.06%] b) 16 [47.06%]
c) 2 [5.88%] d) 0 [0.00%]

Q4) You think that the lecturer gave you an accurate explanation of this system and did not deceive you.

- a) 15 [44.12%] b) 18 [52.94%]
c) 0 [0.00%] d) 1 [2.94%]

Q5) You think that the processing of this system is correct and will not cause information leakage.

- a) 9 [26.47%] b) 14 [41.18%]

- c) 11 [32.35%] d) 0 [0.00%]

Q6) Please evaluate and classify this system into one of these four levels.

- a) 10 [29.41%] b) 22 [64.71%]
c) 1 [2.94%] d) 1 [2.94%]

Q7) You think that a course evaluation contributes to an improvement of your lectures in general.

- a) 9 [26.47%] b) 11 [32.35%]
c) 11 [32.35%] d) 3 [8.82%]

Q8) You think that using this system, instead of the conventional paper-based course evaluation system, is a good idea.

- a) 11 [32.35%] b) 14 [41.18%]
c) 8 [23.53%] d) 1 [2.94%]

According to the above student answers, the prototype system is evaluated to have offered improved privacy. Although more than 10 students showed a skeptical attitude in Q5 and Q7, Q5 actually reflects doubts about completeness (the system was introduced as a "prototype") and Q7 reflects doubts about the effectiveness of the course evaluation in general. That is, the answers are not an evaluation of the system itself. As the answers to Q1, Q2, Q3, Q4, and Q6 are very positive, students seem to have evaluated the system as basically valuable.

Q8, which compares the implemented electronic system and the conventional paper-based system, shows that the majority supported the electronic system. However, 23.53% of the respondents disagreed with the use of the electronic system. Although, according to the students' answers, the prototype system did offer improved privacy, there may be other advantages of using a paper-based system.

III. REVIEW OF USAGE OF ANONYMITY MEASURE

This section reviews the prototype system described above section, particularly from the viewpoint of its usage of the combinatorial anonymity measure.

The formula of the anonymity level used in the prototype system can be re-written as follows:

$$\log(N C_M)$$

In the original simple voting context, N represents the number of all respondents and M denotes the number of respondents who answered positively (or not positively).

The measurement technique was proposed to evaluate the level of voting anonymity by a summation of the informational values of the votes of all the voters. The informational value of each vote is evaluated as $-\log(\text{generation probability})$, following the concept of self-information based on the information theory proposed by Shannon (1948).

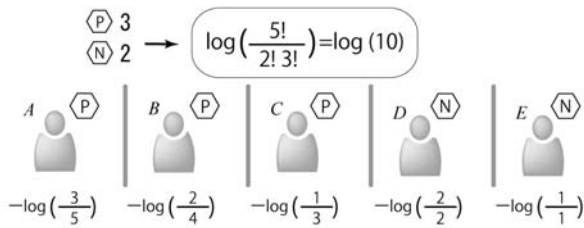


Figure 3: Example of Anonymity Level Calculation

Figure 3 demonstrates the calculation of the sum of the self-information for a vote. In this example, each piece of self-information is added in the order from A to E. When A, B, and C are known to be positive in this order, the proportion of positive voters is $3/5$, $2/4$, and $1/3$, respectively, and the calculus described in the figure is based on these numbers. After the votes by A, B, and C are known, it is obvious that the remaining members are all negative. Reflecting this fact, each of the two following terms equals zero:

The sum obtained in this procedure is independent of the order of calculation. In fact, the calculation only needs the number of total voters and supporters (or opponents). In general, if the number of total voters and supporters is N and M , respectively, the anonymity level of voting for an outside observer is defined as per the above formula.

In the prototype system, N represents the number of all respondents and M denotes the number of *all* respondents who answered positively (or not positively). Suppose five students are asked about the quality of the course with the alternatives of {a: unacceptable, b: needs improvement, c: satisfactory, d: excellent} and M represents the number of respondents who answered not positively. If one chose a, two chose b, and the other two chose c, then M equals 3 and $N - M$ equals 2.

The summation technique used in the prototype seems to be unnatural as the formula can be arranged for n -multiple alternatives as follows:

$$\log({}_N C_{r_1} \times {}_{N-r_1} C_{r_2} \times \dots \times {}_{N-r_1-r_2-\dots-r_{(n-2)}} C_{r_{(n-1)}})$$

where N denotes the number of total members and each of $r_1, r_2, \dots, r_{(n-1)}$ represents the number of members who chose the alternative 1, 2, ..., $n-1$, respectively.

However, the dichotomy approach that divides the alternatives into two parts and calculates the sums of positive and negative members can be regarded to be reasonable when we focus on the difference among the alternatives. Suppose that five students are asked about the quality of the course with the alternatives of {a: unacceptable, b: needs improvement, c: satisfactory, d: excellent}, and one chose c and the other four chose d (Case 1). The result of the anonymity level calculation would be the same as the result for another case in which one respondent chose b and the other four chose c (Case 2). However, Case 2 would be a situation that should be cared, as the alternatives a and b are the critical choices in this example.

The dichotomy approach brings about some limitations in an analysis using the combinatorial anonymity measure. One of the most significant difficulties is the inconsistency in the difference calculation of the anonymity levels between two time points.

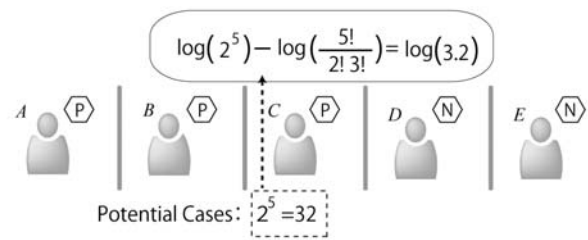


Figure 4: Example of Difference Calculation

Figure 4 illustrates this difference calculation, which was introduced in a voting context (Iwai (2014)). The situation in this example is the same as that shown in Figure 3; that is, three among the five members are positive, and the other two are negative. After the voting, the anonymity level is found to be $\log(10) = 1$. Now, as the antilogarithm of the anonymity level formula is the combination number that reflects the number of possible voting patterns, it can also be applied to a pre-voting scenario. As the number of possible voting patterns is 32 before the voting, the anonymity level is equal to $\log(32)$, and the difference between $\log(32)$ and $\log(10)$ can be regarded as the loss of the anonymity level for the voting members. The difference, however, can also be regarded as the amount of information that the voting process extracted from the members. Thus, we can determine the amount of information that a social process extracts from people by using the combinatorial anonymity measure in general.

For example, it will be helpful to use the difference calculation in the context of social choice theory, where a qualitative analysis such as possibility and impossibility is mainly discussed (See Arrow (1951) and Sen (1970)). The difference calculation may make some qualitative analyses of these social choice procedures possible.

The usage of the dichotomy approach, however, does not seem to be consistent with the difference calculation. If we calculate the difference values for Case 1 and Case 2 discussed above, the result values for both cases will be $\log(4^5) - \log(5)$. That is, the same amount of individual information remains in both cases. However, with the dichotomy calculation, the anonymity level of Case 2 ($\log(5)$) is larger than that of Case 1 ($\log(1) = 0$). (Notice that the case in which all respondents reply positively is not treated as a problem case in the implementation of the prototype system discussed in the previous section. It is designed as an exception in the system design. However, there can be a case in which all respondents reply negatively, and the inconsistency holds here.)

From this point of view, although the dichotomy approach is reasonable, it leads to certain limitations to the usage of the combinatorial anonymity measure.

IV. CONCLUDING REMARKS

Although conventional electronic systems for a course evaluation offer various levels of privacy protection, patterns in the input data itself can accidentally lead to the leakage of personal information. In this paper, we presented a review of a prototype survey system that can automatically prevent such unintended information leakage, particularly from the

viewpoint of its usage of the combinatorial anonymity measure.

This paper concludes that the dichotomy approach that divides alternatives into two parts seems to be reasonable, although it leads to certain limitations to the usage of the anonymity measure.

As mentioned earlier, the risk of the accidental leakage of personal information does exist outside the classroom as well. In any relatively large-scale survey, similar problems can arise, as the cross tabulation of the personal attribution data may yield special cells where only a small number of respondents are classified. In this sense, the core approach of this research may contribute to the improvement of the privacy protection levels of general surveys.

However, more precise design and experiments are needed for developing a practical survey system. As seen in the simple evaluation discussed in Section II, some users still find the paper-based survey method more adequate. For example, as evident by the responses to Q8, there may be other advantages of using a paper-based system. This is an area for examination and a task for the next stage of this study.

REFERENCES

- [1] Arrow, K. J. (1951): *Social Choice and Individual Values*, Wiley, New York.
- [2] Bagai, R., Lu, H., Li R., and Tang, B. (2011): An Accurate System-Wide Anonymity Metric for Probabilistic Attacks, in *Proceedings of the 11th Privacy Enhancing Technologies Symposium (PETS-2011)*, pp. 117-133.
- [3] Edman, M., Sivrikaya, F., and Yener, B. (2007): A combinatorial approach to measuring anonymity. In *Intelligence and Security Informatics*, 2007 IEEE, pp. 356–363.
- [4] Gierlichs, B., Troncoso, C., Diaz C., and Preneel, B. (2008): Revisiting a Combinatorial Approach Toward Measuring Anonymity, in *Proceedings of WPES*, 2008, pp. 111-116.
- [5] Iwai, A. (2003), Tohyô Kôî ni okeru Tokumeisei Gainen no Keishikika, *Preprints of the 35th Conference of the Japanese Association for Mathematical Sociology*, pp. 92-93.
- [6] Iwai, A. (2012a), A Framework of Social Survey System that Prevents Personal Information Leakage by Automatic Modification of Questionnaire Design, in *Proceedings of 18th symposium on socio-information systems*, pp. 127-132.
- [7] Iwai, A. (2012b), Evaluation of an Anonymity Measure as an Index of Voting Privacy, *Journal of Socio-Informatics*, Vol.5, No.1, pp11-25.
- [8] Iwai, A. (2013), Development of a Robust Course Evaluation System That Prevents Individual Information Leakage By Employing Input Data Analysis, Kagaku Kenkyuhi Josei Jigyô (Gakujutsu Kenkyu Josei Kikin Joseikin) Kenkyu Seika Houkokusho (KAKENHI 23650526). <https://kaken.nii.ac.jp/pdf/2012/seika/F-19/12301/23650526seika.pdf>
- [9] Iwai, A. (2014), Seisaku Kettei no tamenô Kofuku Shihyô ha Jitsugen Suruka, *Synergy Shakai Ron* (Imada, T. and Tateoka, Y. ed.), Tokyo University Press, pp.73-85..
- [10] Machanavajjhala, A., Gehrke, J., Kifer, D., and Venkatasubramanian, M., ℓ -diversity: Privacy beyond k-anonymity (2006), in *Proceedings of IEEE International Conference on Data Engineering (ICDE)*, pp. 24–35.
- [11] Sen, A. K. (1970): *Collective Choice and Social Welfare*, Holden-Day.
- [12] Serjantov, A. and Danezis, G. (2002): Towards an information theoretic metric for anonymity, in *Proceedings of Privacy Enhancing Technologies Workshop (PET2002)*, Dingledine, R. and Syverson, P. Eds. Springer-Verlag, LNCS 2482, pp. 41-53.
- [13] Shannon, C. E. (1948): A Mathematical Theory of Communication, *The Bell System Technical Journal*, Vol.27, pp379-423, pp. 623-656.
- [14] Sweeney, L., k-Anonymity (2002): A Model for Protecting Privacy, *International Journal of Uncertainty Fuzziness and Knowledge Based Systems*, Vol. 10, No. 5, pp. 557-570.