A Proposal of a Simple and Secure Statistical Processing System using Secret Sharing

Kiyomi Shirakawa[*], Ken Nakamatsu[**], Koji Chida[***], Satoshi Takahashi[***] and Yutaka Abe[****]

[*] Hitotsubashi University / National Statistics Center, kshirakawa@ier.hit-u.ac.jp
[**] Takumi Information Technology Corporation, ken.nakamatsu@takumi-it.co.jp
[***] NTT Secure Platform Laboratories, {chida.koji, takahashi.s}@lab.ntt.co.jp
[****] National Statistics Center, yabe3@nstac.go.jp

Abstract: Utilization of individual data in official statistics requires strict and secure data management. It also requires inspection of data access methods and confidentiality of aggregated results.
However, there is no system that enables utilization of individual data satisfying these requirements in Japan. We propose a new system that satisfies these requirements in this paper.
In our system, individual data stored in databases is impossible to illegally restore due to a secret sharing scheme. When providing statistical data to data analysts, the system reconstructs only the necessary minimum attribute values and performs statistical processing in memory. Furthermore, it adopts statistical disclosure control functions used in such as τ-ARGUS, and outputs only safe results into a file.
As a result, our system expands a way of providing individual data and improves convenience of users.

1 Introduction

Utilization of individual data in official statistics requires strict and secure data management. It is necessary to consider the method of providing and using data securely. Until now, data has been provided and used by magneto-optical media. However, with these methods, it is impossible for the data provider to manage the data securely. Remote access is a novel way to use and manage individual data securely. In the field of official statistics in Japan, remote access is used not via the Internet but only at the on-site facility. To improve the convenience to users, it is desired that we can use remote access via the Internet. In addition, the provider must check confidentiality of results produced by the user in the remote access environment. However, it is not a trivial task. We propose a new system that satisfies these requirements in this paper.

In our system, individual data stored in databases is impossible to illegally restore due to a secret sharing scheme (Shamir, 1979). When providing statistical data to data analysts, the system reconstructs only the necessary minimum attribute values and performs statistical processing in memory. Furthermore, it adopts statistical disclosure control functions used in such as τ-ARGUS, and outputs only safe results into a file.

As a result, our system expands a way of providing individual data securely and improves convenience of users. The data confidentiality is also improved by the statistical disclosure control functions.

Section 2 describes an outline of a secure computation system and τ-ARGUS. Section 3 presents the approach of this research, and Section 4 discusses the result of empirical analysis. Section 5 describes our conclusion and Section 6 presents future tasks.

2 Outline of Secure Computation System and τ-ARGUS

2.1 Outline of Secure Computation System

2.1.1 Secure computation

Secure computation (Ben-Or, Goldwasser and Wigderson, 1988; Chaum, Crépeau and Damgard, 1988; Yao, 1986) is capable of processing data in an encrypted state, as the technology for making secondary use of official statistics. Because the secure computation protects the data during use, it can provide high-level data privacy. However, this feature gives rise to the issue of tremendous processing times required compared with original computation. Thus, we focus on a secret-sharing-based secure computation to implement a versatile and fast system capable of computing various statistics.

2.1.2 Secret sharing

Secret sharing is a technique of encrypting the original data into multiple pieces of data so that the original data cannot be reconstructed unless at least a certain number of pieces of data are obtained. In a well-known secret sharing scheme known as the (k, n)-threshold scheme, where k and n are integers with value of 2 or greater satisfying $n \geq k$, n pieces (fragments) of encrypted data are created from the original data. If arbitrary k of n fragments are obtained, the original data can be reconstructed. However, the original data cannot be reconstructed even partially with fewer than k fragments.

2.1.3 Related works

Secure computation is an approach being researched in cryptography. The concept was introduced by Yao (1986). Assuming that two parties hold secret values x and y, respectively, Yao's protocol makes it possible to compute an arbitrary function $f(x, y)$ without each party revealing its secret value to the other party. Ben-Or et al. and Chaum et al. then independently proposed secret-sharing-based secure computation schemes that allow function values to be computed when secret values are held by three or more parties without revealing the secret held by any party. To compute a function, both schemes require the parties' computers to communicate with one another. Thus, this type of secure computation is also called multi-party computation.

An issue of multi-party computation is the tremendous processing time required compared with original processing. Research on improved processing time is still being vigorously pursued today. For example, Bogdanov, Laur and Willemson (2008) proposed a fast multi-party computation system using a simple secret sharing scheme that generates three fragments called Sharemind. Kamm, Bogdanov, Laur and Vilo (2013) applied this system and demonstrated the genome-wide analysis. Many development projects for implementing practical multi-party computation systems such as SEPIA (Burkhart, Strasser, Many and Dimitropoulos, 2010), TASTY (Henecka, Sadeghi, Schneider and Wehrenberg, 2010), and VIFF (Geisler, 2010) have been ongoing.

Meanwhile, secure computation research based on approaches different from multi-party computation is also being advanced. In particular, schemes based on fully homomorphic encryption, proposed by Gentry (2009), are receiving attention. The distinction of Gentry's scheme is its ability to allow an arbitrary third party to compute by itself a function with the encrypted input data remaining in the encrypted state and return the true function value only to the owner of the decryption key. A software implementation of the scheme has been released (Halevi, 2017); however, currently the processing time is a major issue. To realize performance at a practical level for large-scale statistical analysis and complex processing, we would need further research breakthroughs. Also, MONOMI (Tu, Kaashoek, Madden and Zeldovich, 2013) and similar systems, which can process various requests (SQL commands) over an encrypted database, are at a near-practical stage.

2.2 Outline of τ-ARGUS

τ-ARGUS is a general-purpose confidential processing software that can protect statistical tables by GUI operation. The name comes from the giant of ancient Greek mythology, and the function is to protect the data. Sometimes, it is described as an initial letter of the Anti-Reidentification General Utility System as a joke. It was developed by Statistics Netherlands as closed source software for a long time, thereafter it was open-sourced with the support of Eurostat. There are two European projects in the fields of SDC (Statistical Disclosure Control) and CASC (Computational Aspects of Statistical Confidentiality). Details of the projects can be found at http://neon.vb.cbs.nl/casc/.

With the contribution of the CENEX project on SDC in 2006, further expansion of the ARGUS software has been realized. Furthermore, the ESSnet project, which is the successor to the CENEX project, is progressing the development of τ-ARGUS.

The latest τ-ARGUS is version 4.1.5. Features added in the latest major update τ-ARGUS version 4.1 are as follows.

- New GUI with statistical table in the center window
- Controlled aggregate adjustment

- Open source code was rewritten in Java.
- The C++ DLL for data manipulation and modular methods was modified to correspond to an open source compiler.
- It is usable with open solvers as with commercial solvers.

The primary and secondary suppression method of this software, and the information loss of cell suppression are as follows.

- The primary confidential method
  1. (n, k) rule
  2. $p\%$ rule
  3. Minimum frequency rule
- the secondary suppression method
  1. Hypercube/GHMITER method
  2. Optimal method
  3. Modular approach
  4. Network
- The information loss of cell concealment
  1. The number of concealed cells
  2. The frequency of concealed cells
  3. The value of concealed cells
  4. Optional value set by the user

After the secondary suppression, all the cells are verified by the inspector based on SDC rules.

3 Approach

We review the method to make a table without looking at individual data and verify the confidentiality of the table, which is by connecting a secure computation system and τ-ARGUS.

3.1 Utilization of Secure Computation System

The data is registered to the secure computation system in the following flow. The Registered data's safety is ensured by the secret sharing scheme and does not cause encryption compromise.

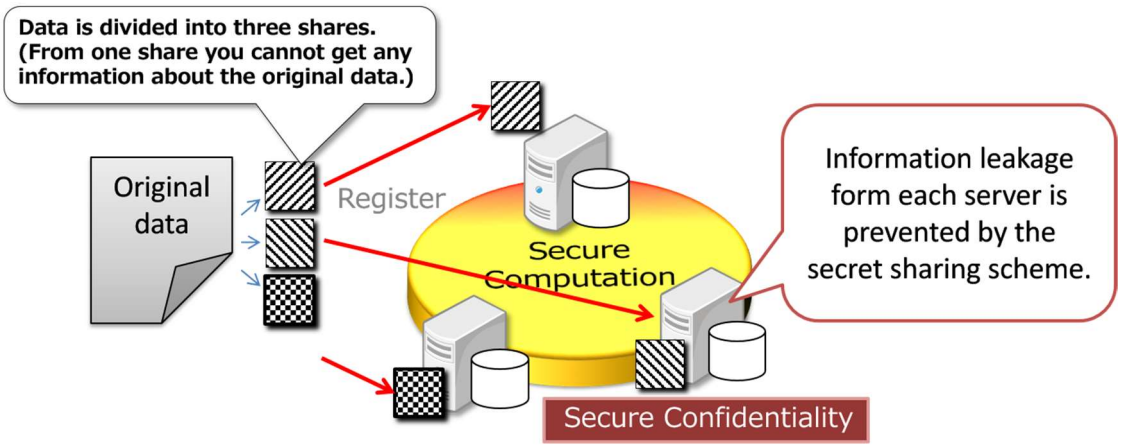

Fig 1 Cooperation between τ-ARGUS and secure computation, "Data registration".

The flow in the following is for outputting the summary table with SDC from the secure computation system in which the data is registered safely. Addition of SDC using τ-ARGUS is carried out in the "semi-trust area".

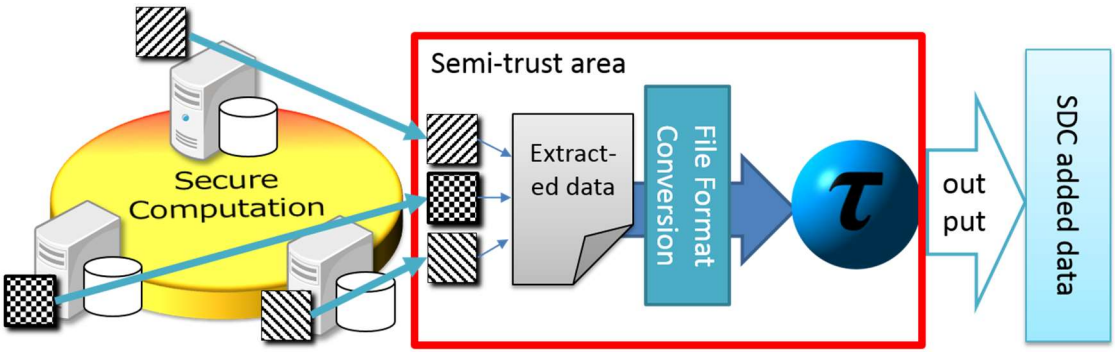

Fig 2 Cooperation between τ-ARGUS and secure computation, "Data output".

For cooperation between secure computation and the τ-ARGUS system, it is required that only barebones data are decrypted and stored in a semi-trust area, because there is a possibility that the file in the semi-trust area may be stolen due to external attack.

3.2 Utilization of τ-ARGUS

3.2.1 Data form

τ-ARGUS has two kinds of readable data formats. One is a micro data format and the other is a tabulation form. When using the micro data format, it is possible to tabulate by the aggregation function of τ-ARGUS. If there are multiple response variables, it is necessary to group them into one variable. If using the tabulation form, it is necessary to total counts such as frequency, total value of each cell, and so on in addition to the

aggregated numerical value. The primary suppression processing corresponds to two data formats. Furthermore, it is possible to use original confidentiality rules.

In this research, since there are restrictions on the use of individual data, we use the tabulation form data. The table used for the verification is as shown in Table 1.

Item(row)	Item(col)	N	SUM	TOP 1	TOP 2	Risk
IN	20	10	1000	150	130	S
IN	21	10	1000	200	150	U
IN	22	20	2400	200	170	S
IN	23	25	3250	250	200	S
IN	24	20	2800	300	250	S
IN	25	30	4500	400	320	S
…	…	…	…	…	…	…

Table 1 Data for verification.

We assume that items in Table 1 have a hierarchical structure including sums and subtotals. The item name shall be a name indicating a clear hierarchical structure. Specific item name structure is as follows:

"IN": income, "IN01": real income, "IN0101": ordinary income

The first and second largest values are mentioned in the "TOP 1" and "TOP 2" columns. Those values are mandatory when using the $p\%$ rule. The result of the primary suppression is mentioned in the "Risk" column. Beforehand, classification into "S" (Safe) and "U" (Unsafe) using an original evaluation method makes primary suppression processing unnecessary.

In addition, τ-ARGUS needs a metadata file for reading into. The metadata is a configuration such as "delimiter character or length", "variable name and role", "hierarchical structure", and so on.

3.2.2 Procedure for using τ-ARGUS

In pre-processing, it is necessary to set the variable names to make the hierarchical information easy to understand. Also, it is necessary to add "rounding processing", "converting to list format", "calculating frequency N and total value SUM" and "TOP 1 and TOP 2".

For τ-ARGUS operation, both primary and secondary suppression are performed on the GUI screen. The default method of primary suppression is the $p\%$ rule. After the verification of primary and secondary suppression, the operator performs "conversion to tabular format", "calculation of average" and "hide cell from suppression processing result" as post-processing.

4 Empirical analysis

Using the secure computation system, create encrypted data of the magnitude table. In the empirical analysis of this research, primary suppression of the table was done using the minimum frequency rule. The verification subject is Table #2 (Two-or-more-person Households and Workers' Households from the 2009 National Survey of Family Income and Expenditure) in the special tabulation by the Institute of Economic Research, Hitotsubashi University (Kinoshita and Sakashita, 2014). This table summarizes 41 variables on income and expenditure at each age from 16 to 89 years old. Since it is a sampling survey, some ages are not included in it. This table has 2,788 cells.

The verification subjects by minimum frequency rule are the cells of 16, 83, 85 and 89 years old where frequency N is less than 3. This processing suppressed 164 cells. Furthermore, when using the p% rule ($p = 10$) and minimum frequency rule, suppressed cells increased to 318 cells (Fig 3). Note that suppressed cells by p% rule ($p = 10$) is 154 cells.

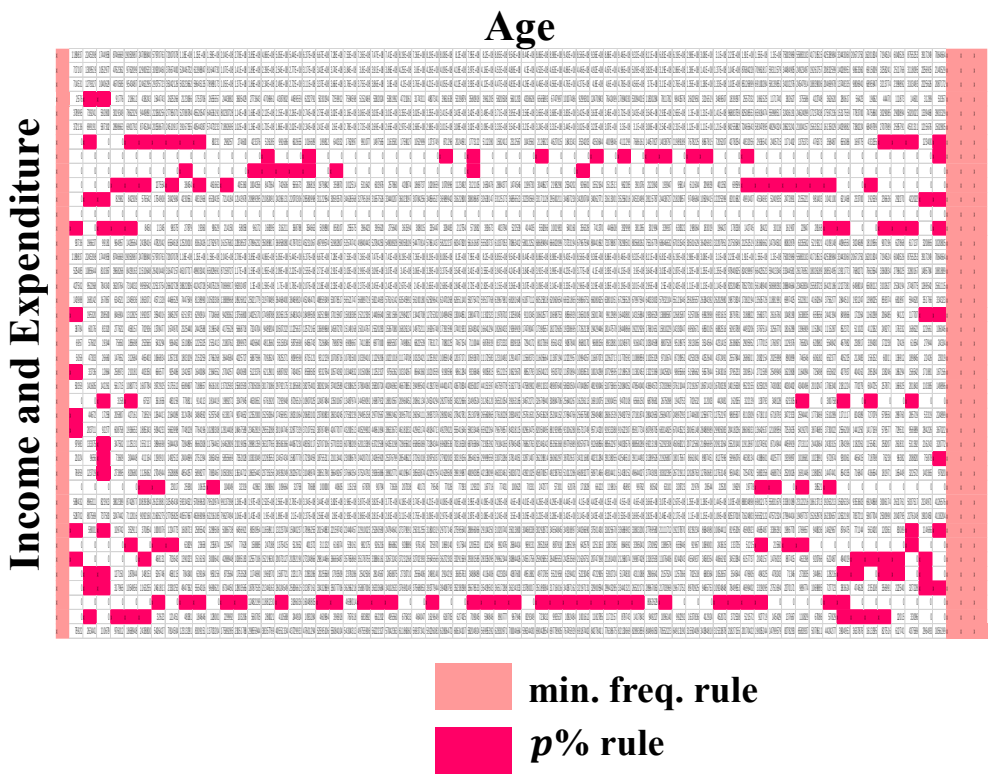

Fig 3 Results of primary suppression.

An example of suppression by p% rule is a cell which is the crossover point of the 24-year-old column and the property purchases row. The column has frequency 102 and

the cell has weighted mean 188, nevertheless, TOP 1 is 34,214 and TOP 2 is 0. In this case, the reason for suppression is that the cell with the value is only Top 1.

Next, we used Hyper Cube method in the secondary suppression processing. As a result, 106 cells were suppressed by secondary suppression. An example of secondary suppression is a cell of 20-year-old ordinary income, because 20-year-old special income was suppressed by primary suppression. It was suppressed because it can be easily calculated from real income and special income.

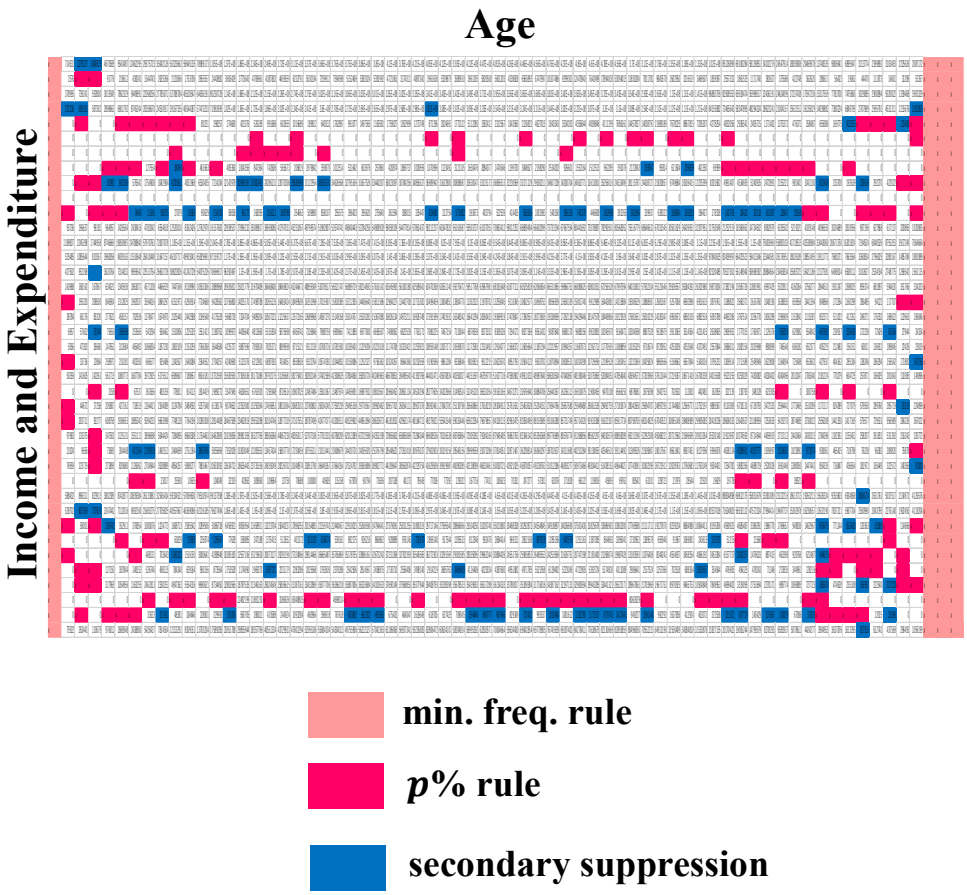

Fig 4 Result of secondary suppression.

5 Conclusion

In this research, we performed the making of the table using encrypted files and the suppression verification by the function of τ-ARGUS. As a result, the usefulness of τ-ARGUS by applying the suppression rules suitable for the statistical tables was confirmed.

Specifically, we used major methods such as the frequency rule and $p\%$ rule for primary suppression. As another method, it is possible to use the "interval rule", for example. For secondary suppression, it is possible to verify confidentiality by diverse options which can be operated by GUI. Thus, it is indispensable to concatenate with the secure computation system to verify by GUI and to move to CUI based on that result. Processing in GUI corresponds to a statistical table of complicated layout. Therefore, aggregating total counts and/or grouping response variables in advance will contribute to improving accuracy and efficiency of output checking.

6 Future tasks

In this paper, we conducted empirical analysis based on the suppression rule of the statistical table by connecting a secure computation system and τ-ARGUS, using micro data of official statistics. This analysis was carried out while confirming the result by GUI as to whether individual processing is completed.

One of future task is to create a Japanese manual showing these operations and procedures by referring to the published manual Version 4.1 (November 2014) and checking the update history of later versions.

Furthermore, the feature of this research is to make automatic processing by a concatenate secure computation system with τ-ARGUS which has a tabulation function. For this reason, it is required to execute in CUI, not in GUI, to enable the execution of suppression processing only in the semi-trust area.

To automate batch processing, we will check the operation of τ-ARGUS by GUI to make batch processing without execution errors. However, there are also many disadvantages in the English version for Japanese users. Hereafter, we will consider changing the language of τ-ARGUS into Japanese.

References

De Wolf, P. P. (2017). *τ-ARGUS home page*, Retrieved June 28, 2017 from http://neon.vb.cbs.nl/casc/tau.htm.

Yao, A. C. C. (1986, October). How to generate and exchange secrets. In *Foundations of Computer Science, 1986., 27th Annual Symposium on* (pp. 162-167). IEEE.

Ben-Or, M., Goldwasser, S., & Wigderson, A. (1988). Completeness theorems for non-cryptographic fault-tolerant distributed computation. In *Proceedings of the twentieth annual ACM symposium on Theory of computing* (pp. 1-10). ACM.

Chaum, D., Crépeau, C., & Damgard, I. (1988, January). Multiparty unconditionally secure protocols. In *Proceedings of the twentieth annual ACM symposium on Theory of computing* (pp. 11-19). ACM.

Shamir, A. (1979). How to share a secret. *Communications of the ACM*, 22(11), 612-613.

Bogdanov, D., Laur, S., & Willemson, J. (2008). Sharemind: A framework for fast privacy-preserving computations. *Computer security-ESORICS 2008*, 192-206.

Kamm, L., Bogdanov, D., Laur, S., & Vilo, J. (2013). A new way to protect privacy in large-scale genome-wide association studies. *Bioinformatics*, 29(7), 886-893.

Burkhart, M., Strasser, M., Many, D., & Dimitropoulos, X. (2010). SEPIA: Privacy-preserving aggregation of multi-domain network events and statistics. *USENIX Security Symposium 2010*, 223-240

Henecka, W., Sadeghi, A. R., Schneider, T., & Wehrenberg, I. (2010, October). TASTY: tool for automating secure two-party computations. In *Proceedings of the 17th ACM conference on Computer and communications security* (pp. 451-462). ACM.

Geisler, M. J. B. (2010). *Cryptographic Protocols: Theory and Implementation*. Doctoral dissertation, Aarhus University.

Gentry, C. (2009, May). Fully homomorphic encryption using ideal lattices. In *STOC* (Vol. 9, No. 2009, pp. 169-178).

Halevi, S. (2017). *HElib*. Retrieved June 28, 2017 from https://github.com/shaih/HElib.

Tu, S., Kaashoek, M. F., Madden, S., & Zeldovich, N. (2013, March). Processing analytical queries over encrypted data. In *Proceedings of the VLDB Endowment* (Vol. 6, No. 5, pp. 289-300). VLDB Endowment.

Kinoshita, K. & Sakashita, K. (2014). Family Income and Expenditure by Age Group of Household Head –Result of the Special Tabulation from 2009 National Survey of Family Income and Expenditure in Japan–. *Statistical Data and Documentation Series*, 72. Hitotsubashi University (in Japanese).