Atsuko Miyaji
Tomoaki Mimoto *Editors*

Security Infrastructure Technology for Integrated Utilization of Big Data

Applied to the Living Safety and Medical Fields

Springer Open

Security Infrastructure Technology for Integrated Utilization of Big Data

Atsuko Miyaji · Tomoaki Mimoto
Editors

Security Infrastructure Technology for Integrated Utilization of Big Data

Applied to the Living Safety and Medical Fields

Springer Open

Editors

Atsuko Miyaji
Osaka University
Suita, Osaka, Japan

Tomoaki Mimoto
KDDI Research, Inc.
Fujimino, Japan

ISBN 978-981-15-3653-3 ISBN 978-981-15-3654-0 (eBook)
https://doi.org/10.1007/978-981-15-3654-0

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Foreword

The Japan Science and Technology Agency (JST) is an independent public body of the Ministry of Education, Culture, Sports, Science and Technology (MEXT). JST plays a key role in implementing science and technology policies formulated in line with the nation's Science and Technology Basic Plan. The Basic Research Programs at JST focus on fundamental research areas that help developing technological breakthroughs, which in turn lead to the advance of S&T and creation of new industries. The programs also encourage researches that trigger, through innovations, reformation of social and economic structures. Core Research for Evolutionary Science and Technology (CREST) program is one of the Basic Research Programs at JST. With an aim to promote and encourage the development of breakthrough technologies that contribute to the attainment of the country's strategic objectives, JST provides a variety of research funding programs for promising research projects. CREST is one of JST's major undertakings for stimulating achievement in fundamental science fields. In addition, returning the fruits of such research to society through innovations is another important responsibility of JST.

"Advanced Core Technologies for Big Data Integration" study area will aim for the creation, advancement, and systematization of next-generation core technology solving of essential issues common among a number of data domains, and integrated analysis of big data in a variety of fields. Specific development targets include technology for stable operation of large-scale data management systems that compress, transfer, and store big data, technology for efficiently retrieving truly necessary knowledge by means of search, comparison, and visualization across diverse information, and the mathematical methods and algorithms enabling such services. In pursuing these studies, with a view to overall system design up to the creation of value for society from big data, the creation, advancement, and systematization of next-generation common core technology highly acceptable to the public will be undertaken, through active efforts at fusion with fields outside of information and communication technology. There are total 11 projects. Especially, "The Security Infrastructure Technology for Integrated Utilization of Big Data," by Atsuko Miyaji (Research director), focuses on secure well-balanced utilization of big data. Many existing security researches focus on technologies of "fast encrypted calculation" since they focus on statistical computation such as sum and average. However, the big data are varied, and thus, there are many usages. It cannot be said that use only for statistical data such as sum and average is enough. It would not be limited to statistical data in the case of medical image data, picture data, etc. What should be the security infrastructure for the utilization of such a wide variety of big data? In addition, extremely secure technologies often may give any benefit to neither the data owner nor the data user. Her project builds a technology to realize balanced security and utilization of big data from the viewpoint of three organizations of the data owner, analyst, and user. Their technology can be combined with fast encrypted calculation, which is a typical target of existing cryptographic researches. We really hope that their concept of security infrastructure technology for the utilization of big data would open up the world of big data utilization in various fields such as the medical and living safety field.

January 2020

Prof. Masaru Kitsuregawa
The University of Tokyo
Tokyo, Japan

Preface

A project of "The Security Infrastructure Technology for Integrated Utilization of Big Data" started in October 2014. Our team consists of four groups: security primitive group under the guidance of Atsuko Miyaji at Osaka university, security management primitive group under Kiyomoto at KDDI Laboratory, the living safety field under Kitamura at AIST and Nishida at Tokyo Institute of Technology, and the medical field under Tanaka at the National Cancer Center and Yamamoto at MEDIS. Concretely, both Kiyomoto and Miyaji have investigated the security infrastructure necessary for the utilization of big data. Based on this security infrastructure, Kitamura and Nishida made testbed systems in the living safety field; Tanaka and Yamamoto made testbed systems in the medical field. All studies combined aim to ensure the good working of the security infrastructure in the real world. Furthermore, after both Kitamura and Nishida will integrate the necessary big data excluding privacy information using our security infrastructure, they will analyze why serious injuries occur at elementary schools. In contrast, both Tanaka and Yamamoto have made an open medical network using our security infrastructure, which enables patients to check the usage of their medical records distributed in different hospitals.

One of the features of our project is that it builds security infrastructure for big data utilization based not on security researchers but on issues from the living safety and medical fields that actually use big data. In other words, it is an important feature that the required specifications do not deviate from actual problems. In addition, we report the results of actual research in both fields using the security infrastructure constructed according to their requirements. Thus, the analysis has been performed on only the available and acceptable data from the point of view of privacy policy until our security infrastructure was realized. Furthermore, the evaluation or analysis of security primitives is often based on dummy data. However, our security primitives have been evaluated by researchers who actually use big data. Furthermore, we clarify how to introduce such security solutions into living safety and medical fields. We also provide guidance on how to use the security infrastructure. We hope that this book will be used by companies, schools, and public organizations that are considering using big data.

Acknowledgements Finally, we would like to thank Prof. Masaru Kitsuregawa at the University of Tokyo who is a research supervisor of “Advanced Core Technologies for Big Data Integration” at JST. We would like to appreciate valuable comments given by Prof. Etsuya Shibayama at the University of Tokyo who is a deputy research supervisor. We also would like to extend our gratitude to useful advice by advisors in the research area: Prof. Kaoru Arakawa (School of Interdisciplinary Mathematical Sciences, Meiji University), Prof. Mitsuru Ishizuka (The University of Tokyo), Naonori Ueda Fellow, NTT Communication Science Laboratories, Hidehiko Tanaka, Director, IWASAKI GAKUEN, Jun’ichi Tsujii, Fellow of Advanced Industrial Science and Technology (AIST), Hideyuki Tokuda, President, National Institute of Information and Communication Technology, Prof. Takeshi Tokuyama, Kwansei Gakuin University, Prof. Teruo Higashino, Osaka University, Prof. Koichi Hori, the University of Tokyo, Prof. Hiroyuki Kitagawa, University of Tsukuba, Prof. Kenji Yamanishi, the University of Tokyo, Prof. Calton Pu, Georgia Institute of Technology, and Nozha Boujemaa, Median Technologies. Finally, we express thanks to our project members: Yuuki Takano, Shinya Okumura, Chen-Mou Cheng, Akinori Kawachi, Sinsaku Kiyomoto, Tomoaki Mimoto, Touru Nakamura, Yoshifumi Nishida, Koji Kitamura, Mikiko Oono, Katsuya Tanaka, and Ryuichi Yamamoto.

Osaka, Japan Prof. Atsuko Miyaji
January 2020

Contents

Chapter 1
Introduction

Atsuko Miyaji, Shinsaku Kiyomoto, Katsuya Tanaka, Yoshifumi Nishida, and Koji Kitamura

1.1 Purpose of Miyaji-CREST

Recently, big data analysis results are expected to be used in various situations such as medical or industrial fields for new medicine or product development. For this reason, it is important to establish a secure infrastructure of the collection, analysis, and use of big data. We need to consider mainly three entities for the infrastructure: data owner, analysis institutions, and users. This research pays attention to a balance between privacy and utilization and also realizes appropriate reduction and feedback of the data analysis results to the data owners.

To build a secure big data infrastructure that connects data owners, analysis institutions, and user institutions in a circle of trust, we construct security technologies necessary for big data utilization. Our main security technologies are oblivious RAM (ORAM), private set intersection (PSI), privacy-preserving classification,

A. Miyaji (✉)
Osaka University, 1-1 Yamadaoka, Suita, Osaka 565-0871, Japan
e-mail: miyaji@comm.eng.osaka-u.ac.jp

S. Kiyomoto
KDDI Research, Inc., 2-1-15 Ohara, Fujimino-shi, Saitama 356-8502, Japan
e-mail: kiyomoto@kddi-research.jp

K. Tanaka
National Cancer Center Japan, 5-1-1 Tsukiji, Chuo-ku, Tokyo 104-0045, Japan
e-mail: katstana@ncc.go.jp

Y. Nishida
Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro-ku, Tokyo 152-8550, Japan
e-mail: nishida.y.af@m.titech.ac.jp

K. Kitamura
National Institute of Advanced Industrial Science and Technology,
2-4-7, Aomi, Koto, Tokyo 135-0064, Japan
e-mail: k.kitamura@aist.go.jp

A. Miyaji and T. Mimoto (eds.), *Security Infrastructure Technology for Integrated Utilization of Big Data*,
https://doi.org/10.1007/978-981-15-3654-0_1

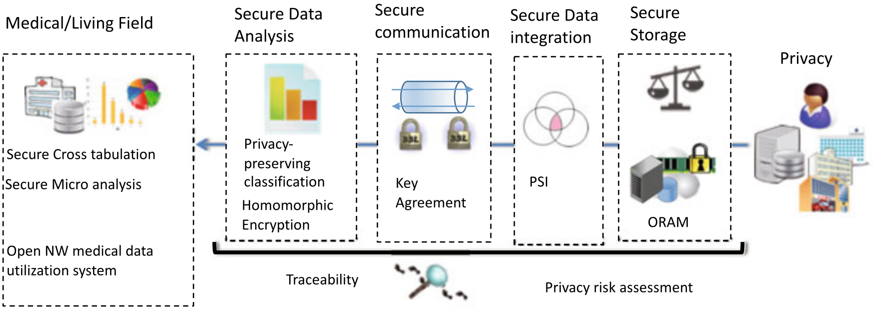


Fig. 1.1 Overview of security infrastructure from data collections to utilization

privacy configuration support, privacy risk assessment, and traceability. Furthermore, we consider the robustness against various attacks such as cyber attacks and post-quantum security.

We construct a safe and privacy-preserving big data distribution platform that realizes the collection, analysis, utilization, and return of owners of big data in a secure and fair manner.

In addition, we demonstrate our secure big data infrastructure in a medical and living safety field. Figure 1.1 shows an overview of our research.

1.2 Roles of Each Group

1.2.1 Security Core Group

We constructed security primitives in the following fields with the aim of realizing an infrastructure for big data utilization that conducts collection, analysis, and utilization of big data securely: 1. Analysis of security basis: Any security primitive which is used for an infrastructure of big data utilization is based on cryptology algorithms. That is, a security primitive becomes compromised if the underground cryptology algorithm is attacked. Therefore, security analysis on cryptographic primitives is important. In this research, we focus on elliptic curve cryptosystems, which achieve a compact public key cryptosystem, and learning with error (LWE)-based cryptosystems, which are types of post-quantum cryptosystems. 2. Privacy-preserving data integration among databases distributed in different organizations: This primitive integrates the same data among databases kept in different organizations while keeping any different data in an organization secret to other organizations. 3. A privacy-preserving classification: This primitive executes a procedure for the server's classification rule to the client's input database and outputs only a result to the client while keeping client's input database secret to the server and server's classification rule to the client.

1.2.2 Security Management Group

Our group focuses on research on data anonymization techniques. First, we analyze the existing anonymization techniques and adversary models for the techniques and clarify our research motivation. Then, we propose our adversary model applicable to several anonymization methods and propose a novel privacy risk analysis method. An implementation of our data anonymization tool based on the risk analysis method is introduced in the chapter.

1.2.3 Living Safety Testbed Group

The Living Safety Group deals with developing new technologies for injury prevention in daily environments such as school safety and home safety based on the security platform developed by the Security Core Group and the Security Management Group. This group has devoted itself to not only developing technology for handing the big data related to injury but also empowering practitioners through social implementation utilizing the developed technologies in cooperation with multiple stakeholders.

1.2.4 Health Testbed Group

Health Testbed Group is focused on implementing a secure clinical data collection and analysis infrastructure for clinical research using the cloud by applying the security primitives developed by the Security Core Group and Security Management Group. This group is working on standardization of data storage, cross-institutional collection, and analysis for electronic medical record data, management mechanism of patient consent information, and traceability for secondary use of medical data, for the development of our health testbed.

Chapter 2
Cryptography Core Technology

Chen-Mou Cheng, Kenta Kodera, Atsuko Miyaji, and Shinya Okumura

Abstract In this chapter, we describe the analysis of security basis. One is the analysis of elliptic curve discrete logarithm problem (ECDLP). ECDLP is one of the public-key cryptosystems that can achieve a short key size but it is not a post-quantum cryptosystem. Another is analysis to learning with error (LWE), which is a post-quantum cryptosystem and has the functionality of *homomorphic encryption*. These two security bases have important roles in each protocol described in Sect. 2.2.4.2

2.1 Analysis on ECDLP

2.1.1 Introduction

In recent years, elliptic curve cryptography is gaining momentum in deployment because it can achieve the same level of security as RSA using much shorter keys and ciphertexts. The security of elliptic curve cryptography is closely related to the computational complexity of the elliptic curve discrete logarithm problem (ECDLP). Let p be a prime number and E, a nonsingular elliptic curve over $\mathbb{F}_{p^n}$, which is a finite field of p^n elements. That is, E is a plane algebraic curve defined by the equation $y^2 = x^3 + ax + b$ for $a, b \in \mathbb{F}_{p^n}$ such that $\Delta = -16(4a^3 + 27b^2) \neq 0$. Along with a point O at infinity, the set of rational points $E(\mathbb{F}_{p^n})$ forms an abelian group with O as the identity. Given $P \in E(\mathbb{F}_{p^n})$ and Q in the subgroup generated by P, ECDLP is the problem of finding an integer α such that $Q = \alpha P$.

C.-M. Cheng · K. Kodera · A. Miyaji (✉) · S. Okumura
Osaka University, Suita, Japan
e-mail: miyaji@comm.eng.osaka-u.ac.jp

C.-M. Cheng
e-mail: ccheng@cy2sec.comm.eng.osaka-u.ac.jp

K. Kodera
e-mail: kodera@cy2sec.comm.eng.osaka-u.ac.jp

S. Okumura
e-mail: okumura@comm.eng.osaka-u.ac.jp

A. Miyaji and T. Mimoto (eds.), *Security Infrastructure Technology for Integrated Utilization of Big Data*,
https://doi.org/10.1007/978-981-15-3654-0_2

Today, the best practical attacks against ECDLP are exponential-time, generic discrete logarithm algorithms such as Pollard's rho method [34]. However, recently, a line of research has been dedicated to the index calculus for ECDLP which was started by Semaev, Gaudry, and Diem [25, 30, 35]. Under certain heuristic assumptions, such algorithms could lead to subexponential attacks to ECDLP in some cases [27, 31, 33]. The interested reader is referred to a survey paper by Galbraith and Gaudry for a more comprehensive and in-depth account of the recent development of ECDLP algorithms along various directions [28].

In this section, we investigate the computational complexity of ECDLP for elliptic curves in various forms—including Hessian [36], Montgomery [32], (twisted) Edwards [23, 24], and Weierstrass, using index calculus. Recently, elliptic curves of various forms such as Curve25519 [22] have been drawing considerable attention in deployment partly because some of them allow fast implementation and security against timing-based side-channel attacks. Furthermore, we can construct these curves not only over prime fields (such as the field of $2^{255} - 19$ elements as used in Curve25519) but also over extension fields. In this section, we will focus on curves over optimal extension fields (OEFs) [21]. An OEF is an extension field from a prime field $\mathbb{F}_p$ with p close to 2^8, 2^{16}, 2^{32}, 2^{64}, etc. Such primes fit nicely into the processor words of 8-, 16-, 32-, or 64-bit microprocessors and hence are particularly suitable for software implementation, allowing efficient utilization of fast integer arithmetic on modern microprocessors [21]. As we will see, our experimental results show considerably significant differences in the computational complexity of ECDLP for elliptic curves in various forms over OEFs.

2.1.2 *Previous Works*

2.1.2.1 Index Calculus for ECDLP

Let E be an elliptic curve defined over a finite field $\mathbb{F}_{p^n}$. For cryptographic applications, we are mostly interested in a prime-order subgroup generated by a rational point $P \in E(\mathbb{F}_{p^n})$. Here, we first give a high-level overview of a typical index-calculus algorithm for finding an integer α such that $Q = \alpha P$ for $Q \in \langle P \rangle$.

1. Determine a *factor base* $\mathcal{F} \subset E(\mathbb{F}_{p^n})$.
2. Collect a set $\mathcal{R}$ of *relations* by decomposing random points $a_i P + b_i Q$ into a sum of points from $\mathcal{F}$, i.e.,

$$\mathcal{R} = \left\{ a_i P + b_i Q = \sum_j P_{i,j} : P_{i,j} \in \mathcal{F} \right\}.$$

3. When $|\mathcal{R}| \approx |\mathcal{F}|$, eliminate the right-hand side using linear algebra to obtain an equation of the form $aP + bQ = O$ and $\alpha = -a/b \bmod \operatorname{ord} P$.

The last step of linear algebra is relatively well studied in the literature, so we will focus on the subproblem in the second step, namely, the point decomposition problem (PDP) on an elliptic curve in the rest of this section.

Definition 2.1 (*Point Decomposition Problem of mth Order*) Given a rational point $R \in E(\mathbb{F}_{p^n})$ on an elliptic curve E and a factor base $\mathcal{F} \subset E(\mathbb{F}_{p^n})$, find, if they exist, $P_1, \dots, P_m \in \mathcal{F}$ such that

$$R = P_1 + \cdots + P_m.$$

2.1.2.2 Semaev's Summation Polynomials

We can solve PDP by considering when the sum of a set of points becomes zero on an elliptic curve. It is straightforward that if two points sum to zero on an elliptic curve $E : y^2 = x^3 + ax + b$ in Weierstrass form, then their x-coordinates must be equal. Let us now consider the simplest yet nontrivial case where three points on E sum to zero. Let

$$Z = \left\{ \begin{array}{r} (x_1, y_1, x_2, y_2, x_3, y_3) \in \mathbb{F}_{p^n}^6 : (x_i, y_i) \in E(\mathbb{F}_{p^n}), i = 1, 2, 3; \\ (x_1, y_1) + (x_2, y_2) + (x_3, y_3) = O \end{array} \right\}.$$

Clearly, Z is in the variety of the ideal $I \subset \mathbb{F}_{p^n}[X_1, Y_1, X_2, Y_2, X_3, Y_3]$ generated by

$$\left\{ \begin{array}{l} Y_i^2 - (X_i^3 + aX_i + b), i = 1, 2, 3; \\ (X_3 - X_1)(Y_2 - Y_1) - (X_2 - X_1)(Y_3 - Y_1) \end{array} \right\}.$$

Now let $J = I \cap \mathbb{F}_{p^n}[X_1, X_2, X_3]$. Using MAGMA's `EliminationIdeal` function, we find that J is actually a principal ideal generated by the polynomial $(X_2 - X_3)(X_1 - X_3)(X_1 - X_2) f_3$, where

$$\begin{aligned} f_3 =& X_1^2X_2^2 - 2X_1^2X_2X_3 + X_1^2X_3^2 - 2X_1X_2^2X_3 - 2X_1X_2X_3^2 - 2aX_1X_2 - 2aX_1X_3 \\ & - 4bX_1 + X_2^2X_3^2 - 2aX_2X_3 - 4bX_2 - 4bX_3 + a^2. \end{aligned}$$

Clearly, the linear factors of this generator correspond to the degenerated case where two or more points are the same or of opposite signs, and f_3 is the 3rd *summation polynomial*, that is, the summation polynomial for three distinct points summing to zero.

Starting from the 3rd summation polynomial, we can recursively construct the subsequent summation polynomials f_m for $m > 3$ by taking resultants. As a result, the degree of each variable in f_m is 2^{m-2}, which grows exponentially as m. This is the observation Semaev made in his seminal work [35]. In short, his proposal is to consider factor bases of the following form:

$$\mathcal{F} = \left\{ (x, y) \in E(\mathbb{F}_{p^n}) : x \in V \subset \mathbb{F}_{p^n} \right\},$$

where V is a subset of $\mathbb{F}_{p^n}$. Then, we solve PDP of mth order by solving the corresponding $(m+1)$th summation polynomial $f_{m+1}(X_1, \ldots, X_m, \tilde{x}) = 0$, where $\tilde{x}$ is the x-coordinate of the point to be decomposed.

Note that this factor base is naturally invariant under point negation. That is, $P_i \in \mathcal{F}$ implies $-P_i \in \mathcal{F}$. In this case, we have about $|\mathcal{F}|/2$ (trivial) relations $P_i + (-P_i) = O$ for free, so we only need to find the other $|\mathcal{F}|/2$ nontrivial relations. In general, we will only discuss factor bases that are invariant under point negation, so by abuse of language, both $\mathcal{F}$ and $\mathcal{F}$ modulo point negation may be referred to as a factor base in the rest of this section.

2.1.2.3 Weil Restriction

Restricting the x-coordinates of the points in a factor base to a subset of $\mathbb{F}_{p^n}$ is important from the viewpoint of polynomial system solving. Take f_3 as an example. When decomposing a random point $aP + bQ$, we first substitute its x-coordinate into say X_3, projecting the ideal onto $\mathbb{F}_{p^n}[X_1, X_2]$. The dimension of the variety of this ideal is nonzero. Therefore, we would like to pose some restrictions on X_1 and X_2 to reduce the dimensions to zero so that the solving time can be more manageable.

When looking for solutions to a polynomial $f = \sum a_i X^i \in \mathbb{F}_{p^n}[X]$ in $\mathbb{F}_{p^n}$, we can view $\mathbb{F}_{p^n}[X]$ as a commutative affine algebra $\mathcal{A} = \mathbb{F}_{p^n}[X]/(X^{p^n} - X) \cong \mathbb{F}_{p^n}[X_1, \ldots, X_n]/(X_1^p - X_1, \ldots, X_n^p - X_n)$. This can be done by identifying the indeterminate X as $X_1\theta_1 + \cdots + X_n\theta_n$, where $(\theta_1, \ldots, \theta_n)$ is a basis for $\mathbb{F}_{p^n}$ over $\mathbb{F}_p$. Hence, f can be identified as a polynomial $f_1\theta_1 + \cdots + f_n\theta_n$, where $f_1, \ldots, f_n \in \mathcal{A}' = \mathbb{F}_p[X_1, \ldots, X_n]/(X_1^p - X_1, \ldots, X_n^p - X_n)$, by appropriately sending each coefficient $a_i \in \mathbb{F}_{p^n}$ to $a_i^{(1)}\theta_1 + \cdots + a_i^{(n)}\theta_n$ for $a_i^{(1)}, \ldots, a_i^{(n)} \in \mathbb{F}_p$. Therefore, an equation $f = 0$ over $\mathbb{F}_{p^n}$ will give rise to a system of equations $f_1 = \cdots = f_n = 0$ over $\mathbb{F}_p$. This technique is known as the *Weil restriction* and is used in the Gaudry–Diem attack, where the factor base is chosen to consist of points whose x-coordinates lie in a subspace V of $\mathbb{F}_{p^n}$ over $\mathbb{F}_p$ [25, 30].

2.1.2.4 Exploiting Symmetry

Naturally, the symmetric group S_m acts on a point decomposition $P_1 + \cdots + P_m$ because elliptic curve groups are abelian. As noted by Gaudry in his seminal work [30], we can therefore rewrite the variables $x_1, \ldots, x_m \in \mathbb{F}_{p^n}$ by elementary symmetric polynomials $e_1, \ldots, e_m$, where $e_1 = \sum x_i$, $e_2 = \sum_{i \neq j} x_i x_j$, $e_3 = \sum_{i \neq j, i \neq k, j \neq k} x_i x_j x_k$, etc. Such rewriting can reduce the degree of summation polynomials and significantly speed up point decomposition [27, 31].

We might be able to exploit additional symmetry brought by actions of other groups, e.g., when the factor base is invariant under addition of small torsion points. For example, consider a decomposition of a point R under the action of addition of a 2-torsion point T_2:

$$R = P_1 + \cdots + P_n = (P_1 + u_1 T_2) + \cdots + (P_{n-1} + u_{n-1} T_2) + \left(P_n + \left(\sum_{i=1}^{n-1} u_i \right) T_2 \right).$$

Clearly, this holds for any $u_1, \ldots, u_{n-1} \in \{0, 1\}$, so a decomposition can give rise to $2^{n-1} - 1$ other decompositions. Similar to rewriting using the elementary symmetric polynomials for the action of S_m, we can also take advantage of this additional symmetry by appropriately rewriting [26].

Naturally, such speedup is curve-specific. Furthermore, even if the factor base is invariant under additional group actions, we may or may not be able to exploit such symmetry to speed up the point decomposition depending on whether the action is "easy to handle in the polynomial system solving process" [26].

2.1.2.5 PDP on (Twisted) Edwards Curves

Faugère, Gaudry, Hout, and Renault studied PDP on twisted Edwards, twisted Jacobi intersections, and Weierstrass curves [26]. For the sake of completeness, we include some of their results here. An Edwards curve over $\mathbb{F}_{p^n}$ for $p \neq 2$ is defined by the equation $x^2 + y^2 = 1 + dx^2y^2$ for certain $d \in \mathbb{F}_{p^n}$ [24]. A twisted Edwards curve $tE_{a,d}$ over $\mathbb{F}_{p^n}$ for $p \neq 2$ is defined by the equation $ax^2 + y^2 = 1 + dx^2y^2$ for certain $a, d \in \mathbb{F}_{p^n}$ [23]. A twisted Edwards curve is a quadratic twist of an Edwards curve by $a_0 = 1/(a - d)$. For $P = (x, y) \in tE_{a,d}$, $-P = (-x, y)$. Furthermore, the addition and doubling formulae for $(x_3, y_3) = (x_1, y_1) + (x_2, y_2)$ are given as follows:

$$\text{When } (x_1, y_1) \neq (x_2, y_2) : \begin{cases} x_3 = \dfrac{x_1 y_2 + y_1 x_2}{1 + d x_1 x_2 y_1 y_2}, \\ y_3 = \dfrac{y_1 y_2 - a x_1 x_2}{1 - d x_1 x_2 y_1 y_2}. \end{cases}$$

$$\text{When } (x_1, y_1) = (x_2, y_2) : \begin{cases} x_3 = \dfrac{2 x_1 y_1}{1 + d x_1^2 y_1^2}, \\ y_3 = \dfrac{y_1^2 - a x_1^2}{1 - d x_1^2 y_1^2}. \end{cases}$$

The 3rd summation polynomial for twisted Edwards curves is [26]:

$$\begin{aligned} f_{tE,3}(Y_1, Y_2, Y_3) = &\left(Y_1^2 Y_2^2 - Y_1^2 - Y_2^2 + \frac{a}{d} \right) Y_3^2 \\ &+ 2\frac{d - a}{d} Y_1 Y_2 Y_3 + \frac{a}{d} \left(Y_1^2 + Y_2^2 - 1 \right) - Y_1^2 Y_2^2. \end{aligned}$$

Again, the subsequent summation polynomials are obtained by taking resultants.

2.1.2.6 Symmetry and Decomposition Probability

Symmetry brought by group action on point decomposition will inevitably be accompanied by a *decrease in decomposition probability*. For example, if a factor base $\mathcal{F}$ is invariant under addition of a 2-torsion point, then the decomposition probability for PDP of the mth order should decrease by a factor of 2^{m-1}. This is due to the same reason that the decomposition probability decreases by a factor of $m!$ because the symmetric group S_m acts on $\mathcal{F}$.

However, this simple fact seems to have been largely ignored in the literature. For example, Faugère, Gaudry, Hout, and Renault explicitly stated in Sect. 5.3 of their study that "[the] probability to decompose a point [into a sum of n points from the factor base] is $\frac{1}{n!}$" for twisted Edwards or twisted Jacobi intersections curves, despite the fact that the factor base is invariant under the addition of 2-torsion points [26]. At first glance, this may not seem a problem, as we would expect to obtain 2^{n-1} solutions if we can successfully solve a PDP instance. (Unfortunately, this is also *not true* in general. We will return to it in more detail in Sect. 2.1.5.3.) However, when estimating the cost of a complete ECDLP attack, they proposed to *collapse* these 2^{n-1} relations into one to reduce the size of the factor base and thus the cost of the linear algebra, cf. Remark 5 of the paper. In this case, the decrease in decomposition probability *does* have an adverse effect, and their estimation for the overall ECDLP cost ended up being overoptimistic by a factor of at least 2^{n-1}.

2.1.3 Montgomery and Hessian Curves

2.1.3.1 Montgomery Curves

A Montgomery curve $M_{A,B}$ over $\mathbb{F}_{p^n}$ for $p \neq 2$ is defined by the equation

$$By^2 = x^3 + Ax^2 + x \tag{2.1}$$

for $A, B \in \mathbb{F}_{p^n}$ such that $A \neq \pm 2$, $B \neq 0$, and $B(A^2 - 4) \neq 0$ [32]. For $P = (x, y) \in M_{A,B}$, $-P = (x, -y)$. Furthermore, the addition and doubling formulae for $(x_3, y_3) = (x_1, y_1) + (x_2, y_2)$ are given as follows. When $(x_1, y_1) \neq (x_2, y_2)$:

$$\begin{cases} x_3 = B\left(\dfrac{y_2 - y_1}{x_2 - x_1}\right)^2 - A - x_1 - x_2 = \dfrac{B(x_2 y_1 - x_1 y_2)^2}{x_1 x_2 (x_2 - x_1)^2}, \\ y_3 = \dfrac{(2x_1 + x_2 + A)(y_2 - y_1)}{x_2 - x_1} - \dfrac{B(y_2 - y_1)^3}{(x_2 - x_1)^3} - y_1. \end{cases}$$

When $(x_1, y_1) = (x_2, y_2)$:

$$\begin{cases} x_3 = \dfrac{(x_1^2 - 1)^2}{4x_1(x_1^2 + Ax_1 + 1)}, \\ y_3 = \dfrac{(2x_1 + x_1 + A)(3x_1^2 + 2Ax_1 + 1)}{2By_1} - \dfrac{B(3x_1^2 + 2Ax_1 + 1)^3}{(2By_1)^3} - y_1. \end{cases}$$

It was noted by Montgomery himself in his original paper that such curves can give rise to efficient scalar multiplication algorithms [32]. That is, consider a random point $P \in M_{A,B}(\mathbb{F}_{p^n})$ and $nP = (X_n : Y_n : Z_n)$ in projective coordinates for some integer n. Then

$$\begin{cases} X_{m+n} = Z_{m-n}[(X_m - Z_m)(X_n + Z_n) + (X_m + Z_m)(X_n - Z_n)]^2, \\ Z_{m+n} = X_{m-n}[(X_m - Z_m)(X_n + Z_n) - (X_m + Z_m)(X_n - Z_n)]^2. \end{cases}$$

In particular, when $m = n$

$$\begin{cases} X_{2n} = (X_n + Z_n)^2(X_n - Z_n)^2, \\ Z_{2n} = (4X_nZ_n)\left((X_n - Z_n)^2 + ((A+2)/4)(4X_nZ_n)\right), \\ 4X_nZ_n = (X_n + Z_n)^2 - (X_n - Z_n)^2. \end{cases}$$

In this way, scalar multiplication on the Montgomery curve can be performed without using y-coordinates, leading to fast implementation.

2.1.3.2 Summation Polynomials for Montgomery Curves

Following Semaev's approach [35], we can construct summation polynomials for Montgomery curves. Like Weierstrass curves, the 2nd summation polynomial for Montgomery curves is simply $f_{M,2} = X_1 - X_2$. Now, we consider $P, Q \in M_{A,B}$ for $P = (x_1, y_1)$ and $Q = (x_2, y_2)$. Let $P + Q = (x_3, y_3)$ and $P - Q = (x_4, y_4)$. By the addition formula, we have

$$x_3 = \frac{B(x_2y_1 - x_1y_2)^2}{x_1x_2(x_2 - x_1)^2}, x_4 = \frac{B(x_2y_1 - x_1y_2)^2}{x_1x_2(x_2 + x_1)^2}.$$

It follows that

$$\begin{cases} x_3 + x_4 = \dfrac{2\left((x_1 + x_2)(x_1x_2 + 1) + 2Ax_1x_2\right)}{(x_1 - x_2)^2}, \\ x_3x_4 = \dfrac{(1 - x_1x_2)^2}{(x_1 - x_2)^2}. \end{cases}$$

Using the relationship between the roots of a quadratic polynomial and its coefficients, we obtain

$$(x_1 - x_2)^2 x^2 - 2\left((x_1 + x_2)(x_1x_2 + 1) + 2Ax_1x_2\right)x + (1 - x_1x_2)^2.$$

From here, we can obtain for Montgomery curve which is the 3rd summation polynomial:

$$\begin{aligned} f_{M,3}(X_1, X_2, X_3) = (X_1 - X_2)^2 X_3^2 - 2((X_1 + x_2)(X_1X_2 + 1) \\ + 2AX_1X_2)X_3 + (1 - X_1X_2)^2, \end{aligned}$$

as well as the subsequent summation polynomials by taking resultants:

$$\begin{aligned} f_{M,m}(X_1, \ldots, X_m) = \mathrm{Res}_X \big(f_{M,m-k}(X_1, \ldots, X_{m-k-1}, X), \\ \times\ f_{M,k+2}(X_{m-k}, \ldots, X_m, X)\big)\,. \end{aligned}$$

2.1.3.3 Small Torsion Points on Montgomery Curves

A Montgomery curve always contains an affine 2-torsion point T_2. Because $T_2 + T_2 = 2T_2 = \mathcal{O}$, $-T_2 = T_2$. If we write $T_2 = (x, y)$, then we can see that $y = 0$ in order for $-T_2 = T_2$ as $p \neq 2$. Substituting $y = 0$ into Eq. (2.1), we get an equation $x^3 + Ax^2 + x = 0$. The left-hand side factors into $x(x^2 + Ax + 1) = 0$, so we get

$$x = 0, \frac{-A \pm \sqrt{A^2 - 4}}{2}.$$

Therefore, the set of rational points over the definition field F_{p^n} of a Montgomery curve includes at least two 2-torsion points, namely $\mathcal{O}$ and $(0, 0)$. The other 2-torsion points may or may not be rational, so we will focus on $(0, 0)$ in this section. Substituting $(x_2, y_2) = (0, 0)$ into the addition formula for Montgomery curves, we get that for any point $P = (x, y) \in M_{A,B}$, $P + (0, 0) = (1/x, -y/x^2)$.

To be able to exploit the symmetry of addition of $T_2 = (0, 0)$, we need to choose the factor base $\mathcal{F} = \{(x, y) \in E(\mathbb{F}_{p^n}) : x \in V \subset \mathbb{F}_{p^n}\}$ invariant under addition of T_2. This means that V needs to be closed by undertaking multiplicative inverses. In other words, V needs to be a *subfield* of $\mathbb{F}_{p^n}$, i.e., $V = \mathbb{F}_{p^\ell}$ for some integer ℓ that divides n. In this case, f_m is invariant under the action of $x_i \mapsto 1/x_i$. Unfortunately, such an action is not linear and hence not easy to handle in polynomial system solving. How to take advantage of such kind of symmetry in PDP is still an open research problem.

2.1.3.4 Hessian Curves

A Hessian curve H_d over $\mathbb{F}_{p^n}$ for $p^n = 2 \bmod 3$ is defined by the equation

$$x^3 + y^3 + 1 = 3dxy \tag{2.2}$$

for $d \in \mathbb{F}_{p^n}$ such that $27d^3 \neq 1$ [36]. For $P = (x, y) \in H_d$, $-P = (y, x)$. Furthermore, the addition and doubling formulae for $(x_3, y_3) = (x_1, y_1) + (x_2, y_2)$ are given as follows.

$$\text{When } (x_1, y_1) \neq (x_2, y_2) : \begin{cases} x_3 = \dfrac{y_1^2 x_2 - y_2^2 x_1}{x_2 y_2 - x_1 y_1}, \\ y_3 = \dfrac{x_1^2 y_2 - x_2^2 y_1}{x_2 y_2 - x_1 y_1}. \end{cases}$$

$$\text{When } (x_1, y_1) = (x_2, y_2) : \begin{cases} x_3 = \dfrac{y_1(1 - x_1^3)}{x_1^3 - y_1^3}, \\ y_3 = \dfrac{x_1(y_1^3 - 1)}{x_1^3 - y_1^3}. \end{cases}$$

2.1.3.5 Summation Polynomials for Hessian Curves

Following a similar approach outlined by Galbraith and Gebregiyorgis [29], we can construct summation polynomials for Hessian curves. First, we introduce a new variable $T = X + Y$, which is invariant under point negation. The 2nd summation polynomial for Hessian curves is simply $f_{H,2} = T_1 - T_2$. Now let

$$Z = \left\{ \begin{matrix} (x_1, y_1, t_1, x_2, y_2, t_2, x_3, y_3, t_3) \in \mathbb{F}_{p^n}^9 : (x_i, y_i) \in H_d(\mathbb{F}_{p^n}), i = 1, 2, 3; \\ (x_1, y_1) + (x_2, y_2) + (x_3, y_3) = O; x_i + y_i = t_i, i = 1, 2, 3 \end{matrix} \right\}.$$

Clearly, Z is in the variety of the ideal $I \subset \mathbb{F}_{p^n}[X_1, Y_1, T_1, X_2, Y_2, T_2, X_3, Y_3, T_3]$ generated by

$$\left\{ \begin{array}{l} X_i^3 + Y_i^3 + 1 - 3dX_iY_i, i = 1, 2, 3; \\ (X_3 - X_1)(Y_2 - Y_1) - (X_2 - X_1)(Y_3 - Y_1); \\ X_i + Y_i - T_i, i = 1, 2, 3 \end{array} \right\}.$$

Again, we compute the elimination ideal $I \cap \mathbb{F}_{p^n}[T_1, T_2, T_3]$ and obtain a principal ideal generated by some polynomial. After removing the degenerate factors, we can obtain for Hessian curve the 3rd summation polynomial:

$$\begin{aligned} f_{H,3}(T_1, T_2, T_3) =& T_1^2T_2^2T_3 + dT_1^2T_2^2 + T_1^2T_2T_3^2 + dT_1^2T_2T_3 + dT_1^2T_3^2 - T_1^2 + \\ & T_1T_2^2T_3^2 + dT_1T_2^2T_3 + dT_1T_2T_3^2 + 3d^2T_1T_2T_3 + 2T_1T_2 + 2T_1T_3 + \\ & 2dT_1 + dT_2^2T_3^2 - T_2^2 + 2T_2T_3 + 2dT_2 - T_3^2 + 2dT_3 + 3d^2, \end{aligned}$$

as well as the subsequent summation polynomials by taking resultants:

$$f_{H,m}(T_1,\ldots,T_m) = \mathrm{Res}_T\left(f_{H,m-k}(T_1,\ldots,T_{m-k-1},T), f_{H,k+2}(T_{m-k},\ldots,T_m,T)\right).$$

2.1.3.6 Small Torsion Points on Hessian Curves

As we shall see in Sect. 2.1.4.1, we will compare elliptic curves in various forms that are isomorphism to one another over the same definition field. As a result, we will only experiment with those Hessian curves that include 2-torsion points like Montgomery or (twisted) Edwards curves. Because $T_2 + T_2 = 2T_2 = \mathcal{O}$, it follows that $-T_2 = T_2$. If we write $T_2 = (x, y)$, then we can see that $x = y$ in order for $-T_2 = T_2$ as $-T_2 = (y, x)$. Substituting $x = y$ into Eq. (2.2), we get an equation $2x^3 - 3dx^2 + 1 = 0$. Therefore, a Hessian curve $H_d(\mathbb{F}_{p^n})$ has a 2-torsion point (ζ, ζ) if the polynomial $2X^3 - 3dX^2 + 1$ has a root ζ in $\mathbb{F}_{p^n}$. In this case, the addition of this 2-torsion point to a point (x, y) would give a point (x', y'), where

$$\begin{cases} x' = \dfrac{\zeta y^2 - \zeta^2 x}{\zeta^2 - xy}, \\ y' = \dfrac{\zeta x^2 - \zeta^2 y}{\zeta^2 - xy}. \end{cases}$$

Obviously, the typical factor bases are not invariant under addition of this 2-torsion point in general.

A Hessian curve always contains a 3-torsion point T_3 such that $3T_3 = \mathcal{O}$ [36]. If we let $T_3 = (x, y)$, then we see that $2(x, y) = -(x, y) = (y, x)$, substituting which into the doubling formula, we get

$$\begin{cases} \dfrac{y(1 - x^3)}{x^3 - y^3} = y, \\ \dfrac{x(y^3 - 1)}{x^3 - y^3} = x. \end{cases}$$

Because x and y cannot be zero at the same time, we have $x^3 - y^3 = 1 - x^3 = y^3 - 1$, or $x^3 = y^3 = 1$. Now because $p^n = 2 \bmod 3$, $\mathbb{F}_{p^n}$ does not have any primitive cubic roots of unity, $x = y = 1$ and $T_3 = (1, 1)$. By the addition formula, if $P = (x, y)$, then

$$P + T_3 = (x, y) + (1, 1) = \left(\frac{y^2 - x}{1 - xy}, \frac{x^2 - y}{1 - xy}\right).$$

However, for $P \in \mathcal{F}$, we only know that $t = x + y \in V \subset \mathbb{F}_{p^n}$, but we know nothing about $1 - xy$, which can lie outside of V. Therefore, again, typical factor bases are not invariant under addition of this 3-torsion point in general. Therefore, it is not

Fig. 2.1 Experimental results on PDP solving for the case of $n = 5$

m	p	Curve	Time	Dreg	Matcost	Rank
3	239	Hessian	0	6	42336.8	1
		Weierstrass	0	6	41259.0	1
		Montgomery	0	6	61239.0	4
		tEdwards	0	6	6308.4	4
	251	Hessian	0	6	41420.4	1
		Weierstrass	0	6	42132.0	1
		Montgomery	0	6	61127.9	4
		tEdwards	0	6	6308.4	4
4	239	Hessian	3.990	19	12066100000	1
		Weierstrass	3.680	19	12064700000	1
		Montgomery	3.489	18	11399100000	5
		tEdwards	0.150	18	54093000	5
	251	Hessian	3.459	19	12069800000	1
		Weierstrass	3.659	19	12066400000	1
		Montgomery	3.280	18	11401700000	5
		tEdwards	0.119	18	54102900	5

clear how to exploit such symmetry brought by addition of small torsion points for Hessian curves.

2.1.4 *Experiments on PDP Solving*

This section shows the results of our experiments conducted to compare the computational complexity of PDP on four different curves: Hessian(H), Weierstrass(W), Montgomery(M), and twisted Edwards(tE).

2.1.4.1 Experimental Setup

As explained in Sect. 2.1.2.1, we focus on PDP in these experiments as the linear algebra step is already well understood. Furthermore, we focus on the bottleneck computation in PDP, namely, the cost of the F4 algorithm for computing Gröbner bases of the polynomial systems obtained after rewriting using the elementary symmetric polynomials and applying the Weil restriction technique to summation polynomials. This way we will be taking advantage of the symmetry of S_m acting on point decompositions. However, we *did not* exploit symmetry of any other group actions. This is because we want to compare the *intrinsic* computational complexity of PDP and hence only consider the symmetry that is present in *all* curves. Exploiting further curve-specific symmetry whenever possible will result in a further speedup, but it would be independent of our findings here.

2.1.4.2 Experimental Results

Figure 2.1 presents our experimental results for the case of $n = 5$. Here, we choose our factor base by taking V as the base field $\mathbb{F}_p$ of $\mathbb{F}_{p^n}$. All our experiments were performed using the MAGMA computation algebra system (version 2.23-1) on a single core of an Intel Xeon CPU E7-4830 v4 running at 2 GHz. Comparisons to solve each PDP were performed by running time (in second), Dreg, Matcost, and Rank. The "Dreg" is the maximum step degree reached during the execution of the F4 algorithm, which is referred to as the "degree of regularity" in the literature [29] and provides an upper bound for the sizes of the Macaulay submatrices involved in the computation, the "Matcost" is a number output by the MAGMA implementation of the F4 algorithm and provides an estimate of the linear algebra cost during the execution of the F4 algorithm, and finally, the "Rank" is the number of linearly independent relations we obtain once successfully solving a PDP instance. It is an important factor to consider, as it determines how many PDP instances we need to successfully solve to have enough relations for a complete ECDLP attack using index calculus. We can clearly see that the PDP solving time and Matcost for twisted Edwards curves are much smaller than those for the other curves. In contrast, the degrees of regularity for Montgomery and twisted Edwards curves are smaller than those of the other curves in the case of $m = 4$. In addition, we can see that the rank for Hessian and Weierstrass curves is 1 in all cases, whereas for Montgomery and twisted Edwards curves, it is 4 and 5 in the case of $m = 3$ and $m = 4$, respectively. Last but not least, although we only present the results for small p (around 8-bit long), here, we have some preliminary results for larger p (around 16-bit and 32-bit long). Apart from the slight difference in the absolute running time, all other results such as Dreg, Matcost, and Rank are similar, so we do not repeat them here.

2.1.5 *Analysis*

2.1.5.1 Revisit Summation Polynomial in Each Form

As we have seen in Sect. 2.1.4.2, PDP on (twisted) Edwards curves seems easier to solve than on other curves. The explanation offered by Faugère, Gaudry, Hout, and Renault is "due to the smaller degree appearing in the computation of Gröbner basis of $\mathscr{S}_{D_n}$ in comparison with the Weierstrass case," cf. Sect. 4.1.1 of their paper [26]. Unfortunately, this *cannot* explain the difference between (twisted) Edwards and Montgomery curves as the highest degrees appearing in the computation of Gröbner bases are *the same* for these two curves. Therefore, there must be other reasons. We have found that the total number of terms for twisted Edwards curves is significantly lower than that for the other curves in all cases. Naturally, this could lead to faster solving time with the F4 algorithm. We also note that, except for the twisted Edwards curves, the summation polynomials before Weil restriction for the other curves are all 100% dense without any missing terms.

2.1.5.2 Missing Terms of Summation Polynomials in (Twisted) Edwards Curves

In this section, we will show that the summation polynomials for (twisted) Edwards curves *mainly* have terms of *even* degrees. The set of terms of even degrees is closed under multiplication, so intuitively, such polynomials are easier to solve, which can be the main reason for the efficiency gain observed in the case of (twisted) Edwards curves.

We shall make this intuition precise in Theorem 2.1, but before we state the main result, we need to clarify our terminology for ease of exposition. When a multivariate polynomial is regarded as a univariate polynomial in one of its variables T, we say that the coefficient a_i of a term $a_i T^i$ is an *even or odd-degree coefficient* depending on whether i is even or odd, respectively. Note that these coefficients are themselves multivariate polynomials in one fewer variable.

We say that a monomial $m = \prod_{i=1}^{n} x_i^{e_i}$, $e_i \geq 0$ in a multivariate polynomial in n variables is *of even degree* or simply an *even-degree monomial* if $\sum_i e_i$ is even; that it is *of odd degree* or simply an *odd-degree monomial* otherwise. In contrast, a monomial is *of (homogeneous) even parity* if all e_i are even; it is *of (homogeneous) odd parity* if all e_i are odd. A monomial is *of homogeneous parity* if it is either of homogeneous even or odd parity. Note that the definition of monomials of odd parity depends on the total number of variables in the polynomial, which is not the case for monomials of even parity because we regard 0 as even. For example, the monomial $x_1 x_2$ is a monomial of odd parity in a polynomial in x_1 and x_2 but not so in another polynomial in $x_1, \ldots, x_n$ for $n > 2$.

By abuse of language, we say that a polynomial is *of even or odd parity* if it is a linear combination of monomials of even or odd parity, respectively; that a polynomial is *of homogeneous parity* if it is a linear combination of monomials of homogeneous parity. The set of polynomials of even parity is closed under polynomial addition and multiplication and hence forms a subring. In contrast, a polynomial f in $x_1, \ldots, x_n$ of odd parity must have the form $\sum_i c_i \left(\prod_{j=1} x_j^{e_{ij}} \right)$, for e_{ij} odd. Therefore, if f is a polynomial of odd parity and g, a polynomial of even parity, then fg must be of odd parity.

Theorem 2.1 *Let $\mathcal{E}$ be a family of elliptic curves such that its 3rd summation polynomial $f_{\mathcal{E},3}(X_1, X_2, X_3)$ is of degree 2 in each variable X_i and of homogeneous parity. Let $g_{\mathcal{E},m}$ be the polynomial corresponding to the PDP of mth order for $\mathcal{E}$ as described in Sect. 2.1.2.2. That is, $g_{\mathcal{E},m}(X_1, \ldots, X_m) = f_{\mathcal{E},m+1}(X_1, \ldots, X_m, x)$, where x is a constant depending on the point to be decomposed.*

1. *If m is even, then $g_{\mathcal{E},m}$ has no monomials of odd degrees.*
2. *If m is odd, then $g_{\mathcal{E},m}$ has some but not all monomials of odd degrees.*

Among the four forms of elliptic curves that we investigated in this section, only the (twisted) Edwards form satisfies the premises of Theorem 2.1. As we have seen in Sect. 2.1.4, the PDP solving time for the (twisted) Edwards form is thus significantly faster than that for the other forms.

We will prove Theorem 2.1 in the rest of this section, for which we will need the following lemmas.

Lemma 2.1 *Let $f_1(T_1, \ldots, T_r, T) = a_0 + a_1 T + \cdots + a_m T^m$ and $f_2(T_1, \ldots, T_r, T) = b_0 + b_1 T + \cdots + b_n T^n$ be two polynomials in $r + 1$ variables, where a_i and b_i are polynomials in $T_1, \ldots, T_r$. Let $f(T_1, \ldots, T_r) = \mathrm{Res}_T(f_1, f_2)$ be the resultant of f_1 and f_2 regarded as two univariate polynomials in T. If both m and n are even, then every monomial of f is a product of an even number or none of the odd-degree coefficients of f_1 and f_2 and some or none of the even-degree coefficients of f_1 and f_2. Specifically, the odd-degree coefficients a_{2k+1} and b_{2k+1} of f_1 and f_2, respectively, appear in total an even number of times in each monomial of f.*

Proof The resultant $\mathrm{Res}_T(f_1, f_2)$ of f_1 and f_2 is the determinant of the following $(m + n) \times (m + n)$ matrix S:

$$S = \begin{bmatrix} a_m & a_{m-1} & \cdots & a_0 & & & \\ & a_m & a_{m-1} & \cdots & a_0 & & \\ & & \ddots & & & \ddots & \\ & & & a_m & a_{m-1} & \cdots & a_0 \\ b_n & b_{n-1} & \cdots & b_0 & & & \\ & b_n & b_{n-1} & \cdots & b_0 & & \\ & & \ddots & & & \ddots & \\ & & & b_n & b_{n-1} & \cdots & b_0 \end{bmatrix} \begin{array}{l} \left.\vphantom{\begin{matrix} a \\ a \\ a \\ a \end{matrix}}\right\} n \\ \left.\vphantom{\begin{matrix} b \\ b \\ b \\ b \end{matrix}}\right\} m \end{array}. \tag{2.3}$$

We denote s_{ij} as the entry at the ith row and jth column of S for $1 \leq i, j \leq m + n$. Because both m and n are even, an even-degree coefficient a_{2k} or b_{2k} will appear in s_{ij} for which the sum of indices $i + j$ is even. Similarly, an odd-degree coefficient a_{2k+1} or b_{2k+1} will appear in s_{ij} for which the sum of indices $i + j$ is odd. Now recall that the determinant of S is defined as

$$\sum_{\sigma \in S_{n+m}} \mathrm{sgn}(\sigma) s_{1,\sigma(1)} \cdot s_{2,\sigma(2)} \cdots s_{m+n,\sigma(m+n)}.$$

We note that the sum of the indices of any summand is

$$\sum_{i}^{m+n} i + \sigma(i) = (m + n)(m + n + 1),$$

which is always even. Therefore, the odd-degree coefficients must appear an even number of times, thus completing the proof.

Lemma 2.2 *Let $\mathcal{E}$ be a family of elliptic curves such that its 3rd summation polynomial $f_{\mathcal{E},3}(X_1, X_2, X_3)$ is of degree 2 in each variable X_i and of homogeneous parity. Then, any subsequent summation polynomial $f_{\mathcal{E},m}(X_1, \ldots, X_m)$ for $m > 3$ is of homogeneous parity.*

Proof As the summation polynomial $f_{\mathcal{E},m+1}$ for $m \geq 3$ is defined recursively from $f_{\mathcal{E},m}$ and $f_{\mathcal{E},3}$ by taking resultants

$$f_{\mathcal{E},m+1}(X_1, \ldots, X_{m+1}) = \mathrm{Res}_X \left(f_{\mathcal{E},m}(X_1, \ldots, X_{m-1}, X), f_{\mathcal{E},3}(X_m, X_{m+1}, X) \right),$$

we shall prove this lemma by induction on m. Let $f_{\mathcal{E},m}(X_1, \ldots, X_{m-1}, X) = a_{2^{m-2}} X^{2^{m-2}} + \cdots + a_1 X + a_0$ and $f_{\mathcal{E},3}(X_m, X_{m+1}, X) = b_2 X^2 + b_1 X + b_0$. By the premise that $f_{\mathcal{E},3}$ is of homogeneous parity, b_0 and b_2 must consist only of monomials (in X_m and X_{m+1}) of even parity. Furthermore, $b_1 = c X_m X_{m+1}$ for some constant c. This is because $f_{\mathcal{E},3}$ is of degree 2 in each variable, for which the only monomial of odd parity is $X_m X_{m+1} X$.

Now consider a term $c_k X_{m+1}^k$ of

$$f_{\mathcal{E},m+1}(X_1, \ldots, X_m, X_{m+1}) = c_{2^{m-1}} X_{m+1}^{2^{m-1}} + \cdots + c_1 X_{m+1} + c_0$$

as a univariate polynomial in X_{m+1}. Again as $f_{\mathcal{E},3}$ is of degree 2 in X, we have the case of $n = 2$ in Eq. 2.3. Now X_{m+1} must come from b_1, so we can conclude that

$$c_k X_{m+1}^k = \sum_i \alpha_i a_{\beta_i} a_{\gamma_i} b_0^{\delta_i} b_2^{\epsilon_i} X_m^k X_{m+1}^k,$$

where α_i a constant, $\beta_i, \gamma_i \in \{0, \ldots, 2^{m-2}\}$, and δ_i, ϵ_i nonnegative integers such that $\delta_i + \epsilon_i + k = 2^{m-2}$. We will complete the proof by showing that $c_k X_{m+1}^k$ is a polynomial in $X_1, \ldots, X_{m+1}$ of homogeneous parity for all k as follows.

1. If k is even, then by Lemma 2.1, β_i and γ_i are both even or both odd in each summand. In either case, the product $a_{\beta_i} a_{\gamma_i}$ is a polynomial in $X_1, \ldots, X_{m-1}$ of even parity. It follows that each summand is a polynomial of even parity because it is a product of polynomials of even parity. Hence, $c_k X_{m+1}^k$ is a polynomial of even parity.
2. If k is odd, the situation is similar but slightly more complicated. By Lemma 2.1, exactly one of β_i and γ_i is odd in each summand, say β_i. By induction hypothesis, a_{β_i} is a polynomial in $X_1, \ldots, X_{m-1}$ of odd parity because it comes from $a_{\beta_i} X^{\beta_i}$ in $f_{\mathcal{E},m}$. It follows that each summand is a polynomial of odd parity because it is a product of a polynomial of even parity $a_{\gamma_i} b_0^{\delta_i} b_2^{\epsilon_i}$ and a polynomial of odd parity $a_{\beta_i} X_m^k X_{m+1}^k$. Hence, $c_k X_{m+1}^k$ is a polynomial of odd parity.

By Lemma 2.2, $g_{\mathcal{E},m}(X_1, \ldots, X_m) = f_{\mathcal{E},m+1}(X_1, \ldots, X_m, x)$ is of homogeneous parity. Obviously, the monomials of even parity will remain of even degree after x is substituted. If m is even, then the monomials of odd parity in $f_{\mathcal{E},m+1}$ will become of even degree after x is substituted because an even number of odd numbers sum to an even number. Similarly, if m is odd, then the monomials of odd parity in $f_{\mathcal{E},m+1}$ will become of odd degree after x is substituted. However, those odd-degree monomials that are *not* of homogeneous parity, e.g., $X_1^2 X_2$, cannot appear in $g_{\mathcal{E},m}$ by Lemma 2.2. This completes the proof of Theorem 2.1.

2.1.5.3 What Price for a Highly Symmetric Factor Base?

Last but not least, we discuss the price needed to pay to have a highly symmetric factor base $\mathcal{F}$ that is invariant under more group actions in addition to that of the symmetric group S_m. As previewed in Sect. 2.1.2.6, we would expect that the effect of the decrease in decomposition probability due to additional symmetry in $\mathcal{F}$ could be offset by that of the increase in number of solutions. For example, let us reconsider the group action of addition of T_2 in Sect. 2.1.2.4. If we could get 2^{m-1} solutions, then the loss of the factor of 2^{m-1} in decomposition probability would be compensated. This way everything would be the same as if there were no such symmetry, and we could exploit the additional symmetry at no cost.

Unfortunately, this proposition is *false* in general. Consider an example of $m = 4$. Let $Q_i = P_i + T_2$ for $i = 1, 2, 3, 4$. We can write down all $2^{m-1} = 8$ possible ways of a point decomposition under this group action:

$$\begin{aligned} P_1 + P_2 + P_3 + P_4 =& Q_1 + Q_2 + P_3 + P_4 \\ = Q_1 + P_2 + Q_3 + P_4 =& Q_1 + P_2 + P_3 + Q_4 \\ = P_1 + Q_2 + Q_3 + P_4 =& P_1 + Q_2 + P_3 + Q_4 \\ = P_1 + P_2 + Q_3 + Q_4 =& Q_1 + Q_2 + Q_3 + Q_4. \end{aligned}$$

It is easy to find that we have only five linearly independent relations from these eight relations, as there are nontrivial linear combinations summing to zero, e.g.:

$$\begin{aligned} &(P_1 + P_2 + P_3 + P_4) - (Q_1 + Q_2 + P_3 + P_4) - (P_1 + P_2 + Q_3 + Q_4) \\ &\quad + (Q_1 + Q_2 + Q_3 + Q_4) = O. \end{aligned}$$

As explained in Sect. 2.1.4.1, the factor bases for Montgomery and twisted Edwards curves are invariant under addition of 2-torsion points. For $m = 3$, we achieve maximum rank of $2^{m-1} = 4$. For $m = 4$, as we have explained above, we can only have rank 5, which is strictly less than the maximum possible rank $2^{m-1} = 8$.

Finally, we note that we have not exploited any symmetry for Hessian curves in our experiments. However, the rank for Hessian curves is always 1 in all our experiments. This shows that the factor base we have chosen for Hessian curves is *not* invariant under addition of small torsion points, as the rank would be > 1 otherwise.

2.1.6 Concluding Remarks

In this section, we experimentally explored index-calculus attack on ECDLP over different forms such as twisted Edwards, Montgomery, Hessian, and Weierstrass curves under the totally fair conditions as they are isomorphic to each other over the same definition field $\mathbb{F}_{p^n}$ and showed that twisted Edwards curves are clearly faster than others. We investigated the summation polynomials of all forms in detail, found that big differences exist in the number of terms, and proved that monomials of odd degrees in summation polynomials on twisted Edwards curves do not exist. We showed that this difference causes less solving time of index-calculus attack on ECDLP over twisted Edwards than others.

2.2 Analysis on Ring-LWE over Decomposition Fields

2.2.1 Introduction

The ring variant of learning with errors (Ring-LWE) based cryptography [15, 16] is one of the most attractive research areas in cryptography. Ring-LWE has provided efficient and provably secure post-quantum cryptographic protocols, which include homomorphic encryption (HE) schemes [4, 5, 9]. The development of the efficiency and security of both post-quantum cryptography and HE is strongly desirable. In fact, the standardization of post-quantum cryptography is under development by the National Institute of Standards and Technology. Moreover, HE schemes that enable us to execute the computation on encrypted data without decryption have many applications in cloud computing.

Ring-LWE is characterized by two probabilistic distributions, modulus parameters (integers) and number fields, as detailed in Sect. 2.2.2.4. Usually, cyclotomic fields are used as the underlying number fields to increase efficiency and security [17]. However, especially in the case of HE schemes, improving the efficiency of the encryption/decryption procedures and homomorphic arithmetic operations on encrypted data while ensuring security remain important tasks.

To construct an HE scheme that can simultaneously encrypt many plaintexts efficiently, Arita and Handa proposed the use of a decomposition field, which is contained in a cyclotomic field with prime conductors, as an underlying number field for Ring-LWE [1]. (Sect. 2.2.3 presents the details of decomposition fields and of Arita and Handa's idea.) Arita and Handa's HE scheme, which is called the subring HE scheme, is indistinguishably secure under a chosen-plaintext attack if the decision variant of Ring-LWE over the decomposition fields is computationally infeasible. Arita and Handa's experiments [1, Sect. 5] showed that the performance of the subring HE scheme is much better than that of the FV scheme based on Ring-LWE over ℓth cyclotomic fields with prime numbers ℓ, as implemented in HElib [11].

As for the security of the subring HE scheme, Arita and Handa remarked that in the case of decomposition fields, some of the security properties of Ring-LWE in the case of cyclotomic fields are also satisfied. More concretely, there exists a quantum polynomial-time reduction from the approximate shortest vector problem on certain ideal lattices to Ring-LWE over decomposition fields, and the equivalence between the decision and search variants of Ring-LWE over decomposition fields is satisfied.

However, solving Ring-LWE is reduced to solving certain problems on lattices, such as the closest vector problem (CVP) and the shortest vector problem, and the difficulty of problems on lattices depends heavily on the structure and given bases of the underlying lattices. For example, if the shortest vector is much shorter than the second shortest vector in a certain lattice $\mathcal{L}$, then the shortest vector problem for lattice $\mathcal{L}$ would be easy. This means that the underlying number fields affect the difficulty of lattice problems arising in Ring-LWE. Hence, to ensure the security of the subring HE scheme, experimental or theoretical analyses of (lattice) attacks should be performed. However, [1] does not provide any such analysis.

In this study, we provide an experimental analysis of the security of Ring-LWE over decomposition fields. More precisely, we compare the security of Ring-LWE over decomposition fields and of Ring-LWE over the ℓth cyclotomic fields with some prime numbers ℓ. In our experiments, we reduce the search Ring-LWE to the (approximate) CVP on certain lattices in the same way as Bonnoron et al.'s analysis [3] because the target of Bonnoron et al.'s analysis is Ring-LWE optimized for HE. We use Babai's nearest plane algorithm [2] and Kannan's embedding technique [12] to solve the CVP. We then compare the running times, success rates, and Hermite root factors. (The root Hermite factor [10] is usually used to evaluate the quality of lattice attacks.) We also compare the experimental results of lattice attacks against Ring-LWE over various decomposition fields to find those fields that provide weak Ring-LWE.

Our experimental results indicate that the success rates and Hermite root factors for the decomposition fields are almost the same as those for the cyclotomic fields. However, the running time for decomposition fields is longer than that for cyclotomic fields. Moreover, the difference in running time increases as the rank of the lattices increases.

Therefore, we believe that Ring-LWE over decomposition fields is more secure against the above lattice attacks than that over cyclotomic fields because the ranks of the lattices occurring in our experiments are much lower than the ranks of the lattices used in practice. This means that to construct HE schemes (or schemes of other types), fewer parameters are needed for Ring-LWE over decomposition fields than for Ring-LWE over cyclotomic fields. Therefore, as a result of our analysis, we believe that Ring-LWE over decomposition fields can be used to construct more efficient HE schemes.

2.2.2 Preliminaries

In this section, we briefly review the notation of lattices, Galois theory, number fields, and Ring-LWE. Throughout this study, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ denote the ring of (rational) integers, field of rational numbers, field of real numbers, and field of complex numbers, respectively. For a positive integer $m \in \mathbb{Z}$, we suppose that any element of $\mathbb{Z}/m\mathbb{Z}$ is represented by an integer contained in the interval $(-m/2, m/2] \cap \mathbb{Z}$.

2.2.2.1 Lattices

An m-dimensional lattice is defined as a discrete additive subgroup of $\mathbb{R}^m$. It is well known that for any lattice $\mathcal{L} \subset \mathbb{R}^m$, there exist $\mathbb{R}$-linearly independent vectors $\mathbf{b}_1, \ldots, and\, \mathbf{b}_n \in \mathbb{R}^m$ such that $\mathcal{L} = \sum_{1 \le i \le n} \mathbb{Z}\mathbf{b}_i := \{\sum_{1 \le i \le n} a_i \mathbf{b}_i \mid a_i \in \mathbb{Z}\ \}$. In other words, for a matrix $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_n)$ whose ith column vector is $\mathbf{b}_j$, we have $\mathcal{L} = \{\mathbf{Bx} \mid \mathbf{x} \in \mathbb{Z}^n\}$. Then, we say that $\{\mathbf{b}_1, \ldots, \mathbf{b}_n\}$ is a lattice basis of $\mathcal{L}$, and $\mathbf{B}$ is the basis matrix of $\mathcal{L}$ with respect to $\{\mathbf{b}_1, \ldots, \mathbf{b}_n\}$. The value n is called the rank of $\mathcal{L}$, and it is denoted by $\mathrm{rank}(\mathcal{L})$. There are infinite bases for a lattice. In fact, for any unimodular matrix $\mathbf{U}$, all column vectors of $\mathbf{UB}$ also form a basis of $\mathcal{L}$. An important invariant of $\mathcal{L}$ is the determinant defined as $\det(\mathcal{L}) := \sqrt{\det(\mathbf{BB}^t)}$. This determinant is independent of basis.

There are various computationally hard problems on lattices. Here, we explain the CVP, which is a well-known problem on lattices. Given a lattice $\mathcal{L}$ and target vector $\mathbf{t} \in \mathbb{R}^m \setminus \mathcal{L}$, the CVP on $(\mathcal{L}, \mathbf{t})$ is the problem of finding a vector $\mathbf{x} \in \mathcal{L}$ such that for all vectors $\mathbf{y} \in \mathcal{L}$, we have $\|\mathbf{t} - \mathbf{x}\| \le \|\mathbf{t} - \mathbf{y}\|$. For a real number $\gamma > 1$, the approximate CVP on $(\mathcal{L}, \mathbf{t}, \gamma)$ is the problem of finding a vector $\mathbf{x} \in \mathcal{L}$ such that for all vectors $\mathbf{y} \in \mathcal{L}$, we have $\|\mathbf{t} - \mathbf{x}\| \le \gamma \|\mathbf{t} - \mathbf{y}\|$. Babai's nearest plane algorithm and Kannan's embedding technique are basic algorithms for solving the approximate CVP. Almost all known problems on lattices that are useful for constructing cryptographic protocols become more difficult as the ranks of the underlying lattices increase, and the quality of the two algorithms mentioned earlier depends on ranks of input lattices.

Breaking some cryptographic protocols can be reduced to solving certain computational problems on lattices, including the (approximate) CVP [3, 8]. To solve such problems on lattices, we usually use lattice basis reduction algorithms, which transform a given basis of a lattice into a basis of the same lattice that consists of nearly orthogonal and relatively short vectors. In fact, an input of Babai's nearest plane algorithm is an (LLL) reduced basis, and Kannan's embedding technique outputs an appropriate vector from the reduced basis. In our experiments, to solve CVP using Babai's nearest plane algorithm and Kannan's embedding technique, we use the LLL algorithm [13] and BKZ algorithm [7, 19], which are well-known algorithms for computing such bases.

The quality of basis reduction algorithms is usually estimated by the root Hermite factor, which is defined as follows: Let $\mathbf{b}$ be the shortest vector of a basis of a lattice $\mathcal{L}$ with rank n, which has been reduced by a basis reduction algorithm $\mathcal{A}$. Then, the root Hermite factor $\delta_{\mathcal{A},\mathcal{L}}$ is defined as a constant satisfying $\delta_{\mathcal{A},\mathcal{L}}^n := \|\mathbf{b}\| / \det(\mathcal{L})^{1/n}$. Better basis reduction algorithms provide smaller Hermite root factors.

2.2.2.2 Galois Theory

To describe decomposition fields, we need to describe Galois theory.

Let K be a field and L an extension field of K; we denote this situation by L/K. The field L is a K-vector space, and the degree of extension of L/K, denoted by

$[L : K]$, is defined as the dimension of L as K-vector space. If M is a subfield of L containing K as a subfield, i.e., $K \subset M \subset L$, then we call M an intermediate field of L/K. If L/K satisfies $[L : K] < \infty$, then L/K is called a finite extension of K. If M is an intermediate field of L/K with $[L : K] < \infty$, then we have $[L : K] = [L : M][M : K]$. If for any $\alpha \in L$, there exists a nonzero polynomial $f(x) \in K[x]$ such that $f(\alpha) = 0$, then L/K is called an algebraic extension of K. It is known that all finite extensions are algebraic extensions.

From now on, we suppose that L/K is a finite algebraic extension. For any $\alpha \in L$, the minimal polynomial over K of α is defined as the monic polynomial $f(x) \in K[x]$ with the lowest degree of all polynomials in $K[x]$ that vanish at α. We denote $\mathrm{Irr}(\alpha, K)(x)$ as the minimal polynomial over K of α. Note that the minimal polynomial over K of α coincides with the monic irreducible polynomial over K that vanishes at α. For a subset $S \subset L$, we denote $K(S)$ as the smallest subfield of L among subfields containing K and S. We call $K(S)$ the field generated by S over K. If L is generated by one element $\theta \in L$ over K, i.e., $L = K(\theta)$, then we have an isomorphism $L \cong K[x]/(\mathrm{Irr}(\theta, K)(x))$ by $\theta \mapsto x$ (mod. $(\mathrm{Irr}(\theta, K)(x))$. This implies that $[K(\theta) : K] = \deg \mathrm{Irr}(\alpha, K)$.

Next, we describe separable, normal, and Galois extensions of fields. If $\mathrm{Irr}(\alpha, K)(x)$ for any α that has no multiple roots, then L/K is called a separable extension of K. If L contains all roots of $\mathrm{Irr}(\alpha, K)(x)$ for any $\alpha \in L$, then L/K is called a normal extension of K. If all algebraic extensions of K, including infinite algebraic extensions, are separable, then K is called a perfect (field). It is known that fields with characteristic zero and any finite field are perfect, and that any finite separable extension field can be generated by one element. If L/K is a separable and normal extension of K, then L/K is called a Galois extension of K. Let Ω be a sufficiently large field containing K such that any ring-homomorphism ϕ fixing K, i.e., $\phi(a) = a$ for any $a \in K$, to L satisfies $\phi(L) \subset \Omega$. We define the set of all ring-homomorphisms by fixing K to the range L to Ω as follows:

$$\mathrm{Hom}_K(L, \Omega) := \{\sigma : L \hookrightarrow \Omega \mid \sigma(a) = a, \forall a \in K\}.$$

(Note that any nonzero ring-homomorphism between fields is injective.) Let L/K be separable with $[L : K] = n$ and $L = K(\theta)$. Let $\theta = \theta_1, \ldots, \theta_n$ be all roots of $\mathrm{Irr}(\theta, K)(x)$. For any $\sigma \in \mathrm{Hom}_K(L, \Omega)$, we have $\sigma\,(\mathrm{Irr}(\theta, K)(\theta)) = \mathrm{Irr}(\theta, K)\,(\sigma(\theta)) = 0$. This means that $\sigma(\theta) = \theta_i$ for some $i = 1, \ldots, n$. This then implies $\#\mathrm{Hom}_K(L) = n$. (Any $\tau \in \mathrm{Hom}_K(L, \Omega)$ is completely determined by the image of θ under τ because τ fixes K.)

Moreover, if L/K is normal, then σ induces an isomorphism $L \cong L$. Note that $L = K(\theta) \cong K(\theta_i)$ for any $i = 1, \ldots, n$ because these fields are isomorphic to $K[X]/(\mathrm{Irr}(\theta, K))$. Therefore, we may take L as Ω and can write $\mathrm{Aut}_K(L) = \mathrm{Hom}_K(L, \Omega)$.

Now, we can describe the fundamental theorem of Galois theory (for finite field extensions). Let L/K be a finite Galois extension of K. Then, we can write $\mathrm{Gal}(L/K) = \mathrm{Aut}_K(L)$. For any subgroup $H \subset \mathrm{Gal}(L/K)$ and an intermediate field M of L/K, we define

$$L^H := \{a \in L \mid \sigma(a) = a, \forall \sigma \in H\},$$
$$G_M := \{\sigma \in \mathrm{Gal}(L/K) \mid \sigma(a) = a, \forall a \in M\}.$$

We note that L/M is a Galois extension with $\mathrm{Gal}(L/M) = G_M$. It is not difficult to see that L^H is an intermediate field of L/K and that G_M is a subgroup of $\mathrm{Gal}(L/K)$. We can define two maps with respect to L/K. One is a map Φ from $A := \{M \subset L \mid M$ is an intermediate field of $L/K\}$ to $B := \{H \subset \mathrm{Gal}(L/K) \mid H$ is a subgroup of $\mathrm{Gal}(L/K)\}$ by $M \mapsto G_M$. The other is a map Ψ from B to A by $H \mapsto L^H$. The fundamental theorem of Galois theory is as follows:

Theorem 2.2 *Let L/K, A, B, Φ, and Ψ be as above. Then, the following statements are true:*

(1) There is a one-to-one correspondence between A and B. More precisely, Φ and Ψ are inverse maps of each other.
(2) If M_1 and M_2 are intermediate fields of L/K with $M_1 \subset M_2$, then we have $\Phi(M_2) \subset \Phi(M_1)$. Similarly, if H_1 and H_2 are subgroups of $\mathrm{Gal}(L/K)$ *with $H_1 \subset H_2$, then we have $\Psi(H_2) \subset \Psi(H_1)$.*
(3) Let M_1, M_2, H_1 and H_2 be as in (2). Then, we have $(H_2 : H_1) = \#H_2/H_1 = [\Psi(H_1) : \Psi(H_2)]$ and $[M_2 : M_1] = (\Phi(M_1) : \Phi(M_2))$.
(4) A subfield M of L/K is a Galois extension of K if and only if $G_M = \Phi(M)$ is a normal subgroup of $\mathrm{Gal}(L/K)$. *Moreover, if $G_M =$* $\mathrm{Gal}(L/M)$ *is a normal subgroup of* $\mathrm{Gal}(L/K)$, *then we have*

$$\mathrm{Gal}(L/K)/\mathrm{Gal}(L/M) \cong Gal(M/K).$$

In particular, if $\mathrm{Gal}(L/K)$ *is an abelian group, then all subfields of L/K are Galois extensions of K.*

For a proof of Theorem 2.2, see [18] for example. (It is easy to prove (2) of Theorem 2.2 from the definitions of Φ and Ψ.)

2.2.2.3 Number Fields

To describe Ring-LWE and decomposition fields, which play central roles in this paper, we need some notations from algebraic number theory.

An (algebraic) number field is a finite extension field of $\mathbb{Q}$. Let K be a number field with extension degree $[K : \mathbb{Q}] = n$. An element $a \in K$ is called an algebraic integer if there exists a monic polynomial $f \in \mathbb{Z}[x]$ such that $f(a) = 0$. The ring of integers O_K of K is defined as a subring of K consisting of all algebraic integers of K. The ring O_K has an integral basis ($\mathbb{Z}$-basis) $\{u_1, \ldots, u_n\}$, i.e., for any element $u \in O_K$, there exist integers $a_1, \ldots, a_n$ such that u is uniquely written as $u = \sum_{1 \le i \le n} a_i u_i$. It is well known that any (integral) ideal I of O_K is uniquely factored into products of some prime ideals, i.e., there exist prime ideals $\mathcal{P}_1, \ldots, \mathcal{P}_m$ satisfying $I = \mathcal{P}_1^{e_1} \cdots \mathcal{P}_m^{e_m}$ for $e_i \geq 1$. If $I = pO_K$ for a prime number p and K is a Galois extension of $\mathbb{Q}$, then we have $O_K/\mathcal{P}_i = \mathbb{F}_{p^d}$ for some $d \in \mathbb{N}$ and all e_i's are mutually equal. Moreover, we have $med = n$, where $e := e_i$, and if all e_i's are equal to 1 (resp. all e_i's and d are equal to 1), then we say that p is unramified (resp. splits completely) in K. Any prime ideal of O_K is a maximal ideal in O_K, and thus we have $P_i + P_j = O_K$ for any $i \neq j$. This induces an isomorphism of rings $O_K/\mathcal{P}_1 \cdots \mathcal{P}_m \cong O_K/\mathcal{P}_1 \times \cdots \times O_K/\mathcal{P}_m$.

2.2.2.4 Ring-LWE Problem

Let K and O_K be as above. Let χ_{secret} and χ_{error} be probabilistic distributions on O_K and let p be an integer. We denote by $O_{K,p}$ the residue ring O_K/pO_K. For a probabilistic distribution χ on a set X, we write $a \leftarrow \chi$ when $a \in X$ is chosen according to χ. We denote $U(X)$ as the uniform distribution on X. The Ring-LWE distribution on $O_{K,p}$, denoted by $\text{RLWE}_{K,p,\chi_{\text{error}},\chi_{\text{sec}}}$, is defined as a probabilistic distribution that takes elements of the form $(a, as + e)$ with $a \leftarrow U(O_{K,p})$, $s \leftarrow \chi_{\text{secret}}$, and with $e \leftarrow \chi_{\text{error}}$. The Ring-LWE problem has two variants. One is the problem of distinguishing $\text{RLWE}_{K,p,\chi_{\text{error}},\chi_{\text{sec}}}$ from $U(O_{K,p} \times O_{K,p})$, which is called the decision Ring-LWE problem. The other is a problem of finding $s \in O_{K,p}$, given arbitrarily many samples $(a_i, a_i s + e_i) \in O_{K,p} \times O_{K,p}$ chosen according to $\text{RLWE}_{K,p,\chi_{\text{error}},\chi_{\text{sec}}}$, which is called the search Ring-LWE problem.

The Ring-LWE problem is expected to be computationally difficult even with quantum computers. It is proved that the decision Ring-LWE problem is equivalent to the search problem if K is a cyclotomic field and if p is a prime number and (almost) splits completely in K [16]. In addition, this equivalence is generalized to the cases where $K/\mathbb{Q}$ is a Galois extension and where p is unramified in K [6]. Moreover, there is a quantum polynomial-time reduction from the search Ring-LWE to the shortest vector problem on certain ideal lattices.

2.2.3 Ring-LWE over Cyclotomic and Decomposition Fields

In this section, we describe why Arita and Handa proposed the use of decomposition fields as the underlying number fields of Ring-LWE to construct efficient HE schemes.

2.2.3.1 Cyclotomic Fields and Decomposition Fields

First, we briefly review cyclotomic fields. For a positive integer m, let $\zeta_m \in \mathbb{C}$ be a primitive mth root of unity and $n = \varphi(m)$, where $\varphi(\cdot)$ denotes Euler's totient function. Then, $K := \mathbb{Q}(\zeta_m)$ is called the mth cyclotomic field. The ring of integers of K coincides with $R := \mathbb{Z}[\zeta_m]$. Any prime number p that does not divide m is unramified in K, and if $p \equiv 1 \pmod{m}$, then p splits completely in K. Here, $K/\mathbb{Q}$

is a Galois extension of degree $[K : \mathbb{Q}] = n$, and its Galois group $\mathrm{Gal}(K/\mathbb{Q})$ is isomorphic to $(\mathbb{Z}/m\mathbb{Z})^*$.

Next, we describe the decomposition fields of number fields. Let L be a number field, and suppose that $L/\mathbb{Q}$ is a Galois extension and that its Galois group $G := \mathrm{Gal}(L/\mathbb{Q})$ is a cyclic group. Let p be a prime number that is unramified in L and satisfies $pO_L = \mathcal{P}_1 \cdots \mathcal{P}_g$, where the $\mathcal{P}_i$'s are the prime ideals of O_L. Let G_Z be a subgroup of G that consists of all elements ρ fixing all $\mathcal{P}_i$, i.e., $\rho(\mathcal{P}_i) = \mathcal{P}_i$ for $1 \le i \le g$, and Z is the fixed field of G_Z. Then, we call Z the decomposition field with respect to p. The field Z is a number field and the ring of integers of Z is $O_Z = O_L \cap Z$. Suppose $\mathfrak{p}_i := O_Z \cap \mathcal{P}_i$. Then, we have $pO_Z = \mathfrak{p}_1 \cdots \mathfrak{p}_g$. A generator σ of G_Z acts on $O_L/\mathcal{P}_i \cong \mathbb{F}_{p^d}$ as the pth Frobenius map, i.e., $\sigma(x) \equiv x^p \pmod{\mathcal{P}_i}$ for all $x \in O_L$ and for $1 \le i \le g$. Therefore, we have $O_Z/\mathfrak{p}_i \cong \mathbb{F}_p$ and $[Z : \mathbb{Q}] = g$, i.e., p splits completely in Z.

2.2.3.2 Cyclotomic Fields Versus Decomposition Fields

Let K, L, and Z be as above and p be a prime number that is unramified in K and splits completely in Z. Assume that L is the ℓth cyclotomic field with a prime number ℓ. As we mentioned in Sect. 2.2.1, cyclotomic fields are usually used as the underlying number fields of Ring-LWE. From the viewpoint of the efficiency of Ring-LWE based schemes, there are good $\mathbb{Z}$-bases of the rings of integers of K and Z [1, 17]. As for the security of the Ring-LWE, in the cases of K and Z, both the equivalence and the reduction mentioned in Sect. 2.2.2.4 are satisfied because both $K/\mathbb{Q}$ and $Z/\mathbb{Q}$ are Galois extensions.

The main difference between K and Z is the algebraic structures of their rings of integers modulo p. Because p is unramified in K, we have $O_{K,p} \cong O_K/\mathcal{P}_1 \times \cdots \times O_K/\mathcal{P}_k$ and $O_K/\mathcal{P}_i \cong \mathbb{F}_{p^d}$ for $1 \le i \le k$ and for $d > 1$, where the $\mathcal{P}_i$'s are prime ideals in O_K lying over p, i.e., $pO_K = \mathcal{P}_1 \cdots \mathcal{P}_k$. The FV scheme [9], which is an HE scheme based on Ring-LWE, uses $O_{K,p}$ as its plaintext space, and thus, the FV scheme (or any HE scheme with the same plaintext space) can encrypt and execute several additions of $dk = n = [K : \mathbb{Q}]$ plaintexts in $\mathbb{F}_p$ simultaneously. However, the FV scheme cannot execute the multiplication of the same number of plaintexts in $\mathbb{F}_p$ simultaneously. To execute the multiplication of plaintexts in $\mathbb{F}_p$, we can only use $\mathbb{F}_p \times \cdots \times \mathbb{F}_p$ (the direct product of k finite fields) as the plaintext space.

In contrast, because p splits completely in Z, we have $O_{Z,p} \cong O_Z/\mathfrak{p}_1 \times \cdots \times O_Z/\mathfrak{p}_g$ and $O_Z/\mathfrak{p}_i \cong \mathbb{F}_p$ for any $1 \le i \le g$, where the $\mathfrak{p}_i$'s are prime ideals in O_Z lying over p. This means that one can encrypt $g = [Z : \mathbb{Q}]$ plaintexts simultaneously. Moreover, one can execute additions and multiplications of the same number of plaintexts in $\mathbb{F}_p$ simultaneously. Because the extension degrees g and n are directly related to the ranks of the lattices occurring in known lattice attacks, we should set $g \approx n$ to compare the security of Ring-LWE over these fields. Therefore, the HE scheme over Z can encrypt and operate d times as many plaintexts as the FV scheme over K simultaneously.

Remark 2.1 1. If $p \equiv 1 \pmod{m}$, then p splits completely in K (recall that K is the mth cyclotomic field), and then there is no advantage to using decomposition fields. However, for some cryptographic applications, we want to use a small p, e.g., $p = 2$ [1]. Moreover, to avoid lattice attacks, the extension degree $[K : \mathbb{Q}]$ must be large, as we discussed above. Thus, we cannot expect $p \equiv 1 \pmod{m}$ for practical parameters in some applications.

2. By the Hensel lifting technique, for $r > 1$ and $q := p^r$, we have $O_{Z,q} \cong \mathbb{Z}/q\mathbb{Z} \times \cdots \times \mathbb{Z}/q\mathbb{Z}$.

2.2.4 Our Experimental Analysis

In this section, we present our experimental results on lattice attacks against Ring-LWE over decomposition fields and cyclotomic fields. First, we explain lattice attacks in our experiments.

2.2.4.1 Lattice Attack in Our Experiments

In our experiments, we reduce the search Ring-LWE to a CVP (or approximate CVP) in the same way as Bonnoron et al.'s analysis [3] because the target of Bonnoron et al.'s analysis is Ring-LWE optimized for HE. We describe this approach briefly in the case of decomposition fields. Let O_Z and p be as in Sect. 2.2.3.1. Set $q := p^r$ for $r > 1$. Let $\{\mu_1, \dots, \mu_g\}$ be a $\mathbb{Z}$-basis of O_Z, which is a good basis, as shown in [1, Lemma 3]. We sample vectors $\mathbf{a} = (a_1, \dots, a_g)$, $\mathbf{s} = (s_1, \dots, s_g)$ and $\mathbf{e} = (e_1, \dots, e_g)$ from $U(\mathbb{Z}^g)$, $D_{\mathbb{Z}^g,\sigma_{\mathrm{s}}}$, and $D_{\mathbb{Z}^g,\sigma_{\mathrm{e}}}$, respectively, where $D_{\mathbb{Z}^g,\sigma}$ denotes the discrete Gaussian distribution with mean 0 and variance σ^2.

We put $a := \sum_{1 \le i \le g} a_i \mu_i$, $s := \sum_{1 \le i \le g} s_i \mu_i$, $e := \sum_{1 \le i \le g} e_i \mu_i$, and $b := as + e = \sum_{1 \le i \le g} b_i \mu_i \pmod{q}$. Then, (a, b) is a Ring-LWE instance over Z. Note that to use Ring-LWE to construct HE schemes, the value σ_{s} and σ_{e} should be sufficiently small because the ℓ_∞-norm $\|\mathbf{s}\|_\infty$ directly affects the growth of noise after multiplication. In our experiments, we set $\sigma_{\mathrm{s}} = 1$ and $\sigma_{\mathrm{e}}^2 = 8$ according to [14]. By comparing all coefficients of both sides, we get $\mathbf{As} + \mathbf{e} = (b_1, \dots, b_g)^t = \mathbf{b}$, where $\mathbf{A}$ is a matrix. (For any vector $\mathbf{v}$, $\mathbf{v}^t$ means its transpose.) If we set $\mathbf{A}'$ as $(\mathbf{A} \;\; \mathbf{I})$, then we have $\mathbf{A}'(\mathbf{s} \;\; \mathbf{e})^t = \mathbf{b} \pmod{q}$, where $\mathbf{I}$ denotes the $g \times g$ identity matrix. From the choice of s_i's and e_i's, our target vector $(\mathbf{s} \;\; \mathbf{e})^t$ is a very short vector from among all solutions to $A'\mathbf{y} = \mathbf{b}$, and thus, we can expect that our target vector can be found by solving the (approximate) CVP on the lattice $\mathcal{L} = \{\mathbf{x} \in \mathbb{Z}^{2g} \mid \mathbf{A}'\mathbf{x} = \mathbf{0} \pmod{q}\}$ and on $\mathbf{w} := (\mathbf{0} \;\; \mathbf{b})^t$, which is a solution to $\mathbf{A}'\mathbf{y} = \mathbf{b}$.

We take

$$\mathbf{B} = \begin{pmatrix} \mathbf{I} & \mathbf{0}_{g,g} \\ -\mathbf{A} & q\mathbf{I} \end{pmatrix}$$

as a basis matrix of $\mathcal{L}$, where $\mathbf{0}_{g,g}$ denotes the $g \times g$ zero matrix. We reduce the basis matrix $\mathbf{B}$ using the LLL and BKZ algorithms with block size $\beta = 10$. (In practice, β should be 10 or 20.) Let $\mathbf{B}_{\text{red}}$ be a reduced basis of $\mathbf{B}$. We input $\mathbf{B}_{\text{red}}$ and $\mathbf{w}$ to Babai's nearest plane algorithm. The quality of the results of Babai's nearest plane algorithm depends on the quality of the basis reduction algorithms used to compute the reduced input bases, and thus, we compute the root Hermite factor for $\mathbf{B}_{\text{red}}$.

In contrast, Kannan's embedding technique takes a basis matrix

$$\mathbf{C} = \begin{pmatrix} \mathbf{B} & -\mathbf{w} \\ \mathbf{0}_{1\times 2g} & M \end{pmatrix}$$

as input, and we set $M = 1$ according to the result of an experimental study on Kannan's embedding technique for LWE [20]. We also use the LLL and BKZ algorithms with $\beta = 10$ to reduce the above basis matrix.

Remark 2.2 In the case of ℓ-cyclotomic fields with prime numbers ℓ, we use $\{1, \zeta_\ell, \dots, \zeta_\ell^{\ell-2}\}$ as a $\mathbb{Z}$-basis, which is also a good basis [17].

Remark 2.3 For $1 \le r' < r$ and $q' := p^{r'}$, we can obtain samples of $\text{RLWE}_{K,q',\chi_{\text{error}},\chi_{\text{sec}}}$ from samples of $\text{RLWE}_{K,q,\chi_{\text{error}},\chi_{\text{sec}}}$ by a natural projection $O_{Z,q} \to O_{Z,q'}$ by $a \mapsto a \pmod{q'}$. In our experiments, we use a small r' to reduce running times. In our experimental results, we only show r'.

2.2.4.2 Experimental Results

We used a computer with 2.00 GHz CPUs (Intel(R) Xeon(R) CPU E7-4830 v4 (2.00GHz)x111) and 3 TB memory to conduct the experiments. The OS was Ubuntu 16.04.4. We implemented the code for sampling Ring-LWE instances in SageMath version 7.5.1. We also used Magma V2.23-1 to execute lattice attacks. We took 100 samples and performed lattice attacks on them.

We show our experimental results in Tables 2.1 and 2.2 for $p = 2$. Table 2.1 shows that there is not a considerable difference between the experimental results of cyclotomic fields and those for decomposition fields. In contrast, Table 2.2 shows that Kannan's embedding technique is much faster than Babai's nearest plane algorithm.

This implies that the behaviors of the basis reduction algorithms heavily depend on the structure of the input lattices. This is a reason why experimental analyses are necessary for ensuring the security of lattice-based schemes (or other problems). Table 2.2 also shows that the running times for the decomposition fields become longer than those for cyclotomic fields as g (or $\ell - 1$) increases. Therefore, we can expect that decomposition fields provide Ring-LWE that is more secure against the lattice attacks described in Sect. 2.2.4.1 than ℓth cyclotomic fields because the ranks of the lattices occurring in our experiments are very low compared to the ranks of lattices used in practice. This means that we can use decomposition fields with lower extension degrees than would be needed for ℓth cyclotomic fields, and the use of such number fields makes Ring-LWE-based schemes more efficient. Therefore, as a

Table 2.1 Experimental results on Babai's nearest plane algorithm for $p = 2$

ℓ	59	16183	73	2089	83	4051	131	5419	173	14449	227	9719
g	–	58	–	72	–	81	–	129	–	172	–	226
Lattice rank	118	116	146	144	166	162	262	258	346	344	454	452
r'	20	20	20	20	20	20	30	30	30	30	30	30
Number of samples	93	100	100	100	100	100	100	100	40	37	15	14
Success rate (%)	100	100	100	100	100	100	100	100	100	89	0	0
Average root Hermite factor	1.014	1.014	1.014	1.014	1.014	1.014	1.020	1.020	1.020	1.020	1.089	1.021
Average running time (s)	72.22	88.97	218.4	238.2	443.3	456.1	12790.5	11744.6	54763.0	57862.3	231816.1	237846.9
Ratio of running times (%)	–	123.2	–	109.0	–	102.9	–	91.8	–	105.7	–	102.6

The columns for which the values g are indicated show the results for decomposition fields; the other columns show the results for cyclotomic fields

The "ratio of running times" is the ratio of the average of running time for a decomposition field to that of a cyclotomic field for each g

Table 2.2 Experimental results on Kannan's embedding technique for $p = 2$

ℓ	59	16183	73	2089	83	4051	131	5419	173	14449	227	9719
g	–	58	–	72	–	81	–	129	–	172	–	226
Lattice rank	119	117	147	145	167	163	263	259	347	345	455	453
r'	20	20	20	20	20	20	30	30	30	30	40	40
Number of samples	100	100	100	100	100	100	100	100	100	100	23	21
Success rate (%)	100	100	100	100	100	100	100	100	100	100	100	100
Average running time (s)	10.4	10.7	36.7	41.4	92.3	97.6	4714.6	5556.7	19387.5	25138.7	136978.2	159772.6
Ratio of running times (%)	–	103.5	–	112.7	–	105.7	–	117.9	–	129.7	–	116.6

We computed the root Hermite factor for the reduced bases, but we do not show them because the success rates in these results are 100%

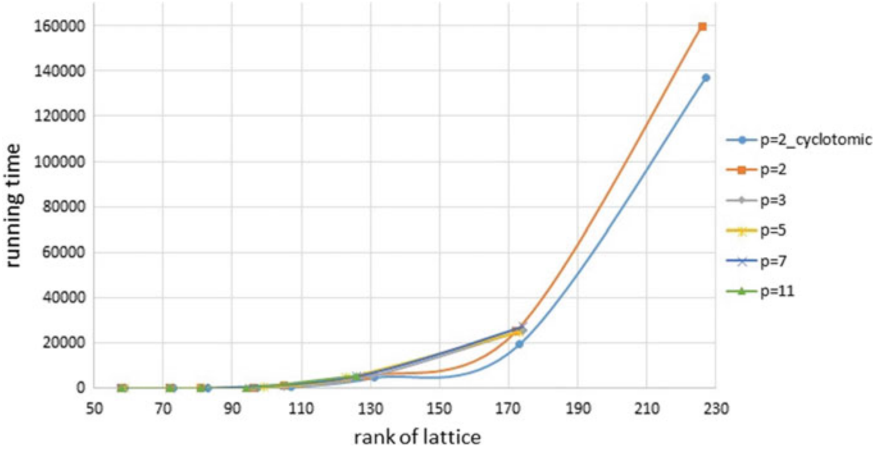


Fig. 2.2 Average running times of Kannan's embedding technique for cyclotomic and decomposition fields with respect to $p = 2, 3, 5, 7, 11$. The label "$p = 2$_cyclotomic" indicates the results of the cyclotomic fields shown in Table 2.2, and the other labels indicate the results for decomposition fields with respect to the corresponding prime numbers p. We set modulus parameter $q = p^{r'}$ so that these moduli have the almost same bit sizes. We only show the average results on at least 10 samples

result of our analysis, we believe that Ring-LWE over decomposition fields can be used to construct more efficient HE schemes.

We also conducted experiments for decomposition fields with respect to $p = 3, 5, 7, 11$ to find decomposition fields that provide weak Ring-LWE instances (Fig. 2.2). In these experiments, we could not find decomposition fields that provide weak Ring-LWE.

References

1. S. Arita, S. Handa, Subring homomorphic encryption, in *Proceedings of ICISC 2017*. LNCS, vol. 10779 (Springer, Cham, 2018), pp. 112–136
2. L. Babai, On Lovász' Lattice reduction and the nearest lattice point problem. Combinatorica **6**(1), 1–13 (1986). Springer (Preliminary version in STACS 1985)
3. G. Bonnoron, C. Fontaine, A note on ring-LWE security in the case of fully homomorphic encryption, in *Proceedings of INDOCRYPT 2017*. LNCS, vol. 10698 (Springer, Cham, 2017), pp. 27–43
4. Z. Brakerski, C. Gentry, V. Vaikuntanathan, (Leveled) fully homomorphic encryption without bootstrapping, in *Proceedings of ITCS 2012* (ACM New York, NY, USA, 2012), pp. 309–325
5. Z. Brakerski, V. Vaikuntanathan, Fully homomorphic encryption from ring-LWE and security for key dependent messages, in *Proceedings of CRYPTO 2011*. LNCS, vol. 6841 (Springer, Berlin, Heidelberg, 2011), pp. 505–524
6. H. Chen, K. Lauter, K.E. Stange, Security considerations for Galois non-dual RLWE families, in *Proceedings of SAC 2016*. LNCS, vol. 10532 (Springer, Cham, 2016), pp. 443–462

7. Y. Chen, P.Q. Nguyen, BKZ 2.0: better lattice security estimates, in *Proceedings of ASIACRYPT 2011*. LNCS, vol. 7073 (Springer, Berlin, Heidelberg, 2011), pp. 1–20
8. D. Coppersmith, Small solutions to polynomial equations, and low exponent RSA vulnerabilities. J. Cryptol. **10**(4), 233–260 (1997). Springer
9. J. Fan, F. Vercauteren, Somewhat practical fully homomorphic encryption. Cryptology ePrint Archive, Report 2012/144 (2012)
10. N. Gama, P.Q. Nguyen, Predicting lattice reduction, in *Proceedings of EUROCRYPT 2008*. LNCS, vol. 4965. Springer, Berlin, Heidelberg, 2008), pp. 31–51
11. S. Halevi, V. Shoup, Algorithms in HElib, in *Proceedings of CRYPTO 2014*. LNCS, vol. 8616. (Springer, Berlin, Heidelberg, 2014), pp. 554–571
12. R. Kannan, Minkowski's Convex body theorem and integer programming, *Mathematics of Operations Research*, vol. 12 (3), pp. 415–440, INFORMS, Linthicum, Maryland, USA, (1987)
13. A.K. Lenstra, H.W. Lenstra Jr., L. Lovász, Factoring polynomials with rational coefficients, Math. Ann. **261**(4), 515–534 (1982). Springer
14. T. Lepoint, M. Naehrig, A comparison of the homomorphic encryption schemes FV and YASHE, in *Proceedings of AFRICACRYPT 2014*. LNCS, vol 8469. (Springer, Cham, 2014), pp. 318–335
15. V. Lyubashevsky, C. Peikert, O. Regev, On ideal lattices and learning with errors over rings, in *Proceedings of EUROCRYPT 2010*. LNCS, vol. 6110 (Springer, Berlin, Heidelberg, 2010), pp. 1–23
16. V. Lyubashevsky, C. Peikert, O. Regev, On ideal lattices and learning with errors over rings. J. ACM (JACM) **60**(6), 43:1–43:35 (2013), ACM New York, NY, USA
17. V. Lyubashevsky, C. Peikert, O. Regev, A toolkit for ring-LWE cryptography, in *Proceedings of EUROCRYPT 2013*. LNCS, vol. 7881 (Springer, Berlin, Heidelberg, 2013), pp. 35–54
18. P. Morandi, Field and galois theory, *Graduate Texts in Mathematics*, vol. 167 (Springer-Verlag, New York, 1996)
19. C.P. Schnorr, M. Euchner, Lattice basis reduction: improved practical algorithms and solving subset sum problems. Math. Progr. **66**(1-3), 181–199 (1994). Springer
20. Y. Wang, Y. Aono, T. Takagi, An experimental study of Kannan's embedding technique for the search LWE problem, in: *Proceedings of ICICS 2017*. LNCS, vol. 10631 (Springer, Cham, 2018), pp. 541–553
21. D.V. Bailey, C. Paar, Optimal extension fields for fast arithmetic in public-key algorithms, in *Advances in Cryptology - CRYPTO '98, 18th Annual International Cryptology Conference, Santa Barbara, California, USA, August 23-27, 1998, Proceedings* (Springer, 1998), pp. 472–485
22. D.J. Bernstein, Curve25519: new diffie-hellman speed records. in *Public Key Cryptography - PKC 2006, 9th International Conference on Theory and Practice of Public-Key Cryptography, New York, NY, USA, April 24-26, 2006, Proceedings* (Springer, 2006) pp. 207–228
23. D.J. Bernstein, P. Birkner, M. Joye, T. Lange, C. Peters, Twisted Edwards curves. IACR Cryptology ePrint Archive **2008**, 13 (2008)
24. D.J. Bernstein, T. Lange, Faster addition and doubling on elliptic curves. IACR Cryptology ePrint Archive **2007**, 286 (2007)
25. C. Diem, On the discrete logarithm problem in class groups of curves. Math. Comput. **80**(273), 443–475 (2011)
26. J. Faugère, P. Gaudry, L. Huot, G. Renault, Using symmetries in the index calculus for elliptic curves discrete logarithm. J. Cryptol. **27**(4), 595–635 (2014)
27. J. Faugère, L. Perret, C. Petit, G. Renault, Improving the complexity of index calculus algorithms in elliptic curves over binary fields. in *Advances in Cryptology - EUROCRYPT 2012 - 31st Annual International Conference on the Theory and Applications of Cryptographic Techniques, Cambridge, UK, April 15-19, 2012. Proceedings* (Springer, 2012) pp. 27–44
28. S.D. Galbraith, P. Gaudry, Recent progress on the elliptic curve discrete logarithm problem. Des. Codes Cryptogr. **78**(1), 51–72 (2016)
29. S.D. Galbraith, S.W. Gebregiyorgis, Summation polynomial algorithms for elliptic curves in characteristic two. in *Progress in Cryptology - INDOCRYPT 2014 - 15th International*

Conference on Cryptology in India, New Delhi, India, December 14-17, 2014, Proceedings (Springer, 2014), pp. 409–427

30. P. Gaudry, Index calculus for abelian varieties of small dimension and the elliptic curve discrete logarithm problem. J. Symb. Comput. **44**(12), 1690–1702 (2009)
31. Y. Huang, C. Petit, N. Shinohara, T. Takagi, Improvement of Faugère et al.'s Method to Solve ECDLP, in *Advances in Information and Computer Security - 8th International Workshop on Security, IWSEC 2013, Okinawa, Japan, November 18-20, 2013, Proceedings* (Springer, 2013), pp. 115–132
32. P.L. Montgomery, Speeding the Pollard and elliptic curve methods of factorization. Math. Comput. **48**, 243–264 (1987). URLhttp://links.jstor.org/sici?sici=0025-5718(198701)48:177<243:STPAEC>2.0.CO;2-3
33. C. Petit, J. Quisquater, On polynomial systems arising from a weil descent. IACR Cryptology ePrint Archive **2012**, 146 (2012)
34. J.M. Pollard, Monte Carlo methods for index computation mod p. Math. Comput. **32**, 918–924 (1978)
35. I.A. Semaev, Summation polynomials and the discrete logarithm problem on elliptic curves. IACR Cryptology ePrint Archive **2004**, 31 (2004)
36. N.P. Smart, The hessian form of an elliptic curve, in *Cryptographic Hardware and Embedded Systems - CHES 2001, Third International Workshop, Paris, France, May 14-16, 2001, Proceedings*, number Generators. (Springer, 2001), pp. 118–125

Chapter 3
Secure Primitive for Big Data Utilization

Akinori Kawachi, Atsuko Miyaji, Kazuhisa Nakasho, Yiying Qi, and Yuuki Takano

Abstract In this chapter, we describe two security primitives for big data utilization. One is a privacy-preserving data integration among databases distributed in different organizations. This primitive integrates the same data among databases kept in different organizations while keeping any different data in an organization secret to other organizations. Another is a privacy-preserving classification. This primitive executes a procedure for server's classification rule to client's input database and outputs only the result to the client while keeping the client's input database secret to the server and server's classification rule to the client. These primitives can be executed not only independently but also jointly. That is, after we integrate databases from distributed organization by executing the privacy-preserving data integration, we can execute a privacy-preserving classification.

3.1 Privacy-Preserving Data Integration

3.1.1 Introduction

Medical organizations often store the data accumulated through medical analyses. However, detailed data analysis sometimes requires separate datasets to be integrated without violating patient or commercial privacy. Consider the scenario in which the

A. Kawachi
Mie University, 1577 Kurimamachiya-cho, Tsu City, Mie 514-8507, Japan
e-mail: kawachi@cs.info.mie-u.ac.jp

A. Miyaji (✉) · Y. Qi · Y. Takano
Osaka University, 1-1 Yamadaoka, Suita, Osaka 565-0871, Japan
e-mail: miyaji@comm.eng.osaka-u.ac.jp

Y. Takano
e-mail: ytakano@cy2sec.comm.eng.osaka-u.ac.jp

K. Nakasho
Yamaguchi University, 1677-1 Yoshida, Yamaguchi City, Yamaguchi 753-8511, Japan
e-mail: nakasho@yamaguchi-u.ac.jp

A. Miyaji and T. Mimoto (eds.), *Security Infrastructure Technology for Integrated Utilization of Big Data*,
https://doi.org/10.1007/978-981-15-3654-0_3

occurrence of similar accidents can be attributed to a particular defective product. Such defective products should be identified as quickly as possible. However, the databases related to accidents are maintained separately by different organizations. Thus, investigating the causes of accidents is often time-consuming. For example, assume child A has broken her/his leg at school, but it is not clear whether the accident was caused by defective equipment. In this case, information relating to A's injury, such as the patient's name and type of injury, is stored in hospital database S_1. Information pertaining to A's accident, such as their name and the location of the swing at the school, is stored in database S_2, which is held by the fire department. Finally, information relating to the insurance claim following A's accident, such as the name and medical costs, is maintained in the insurance company's database, S_3. Computing the intersection of these databases, $S_1 \cap S_2 \cap S_3$, without compromising privacy would enable us to combine the separate sets of information, which may allow the cause of the accident to be identified. Let us consider another situation. Several clinics, denoted as P_i, maintain separate databases, represented as S_i. The clinics wish to know the patients they have in common to enable them to share treatment details; however, P_i should not be able to access any information about patients not stored in their own dataset. In this case, the intersection of the set must not reveal private information.

These examples illustrate the need for the Multiparty Private Set Intersection (MPSI) protocol [1–4]. MPSI is executed by multiple parties who jointly compute the intersection of their private datasets. Ultimately, only designated parties can access this intersection. Previous protocols are impractical because the bulk of the computation depends on the number of players. One previous study required the size of the datasets maintained by the different players to be equal [1, 2]. Another study [3] computed only the approximate number of intersections, whereas other researchers [4] required more than two trusted third-parties.

In this section, we propose a practical MPSI with the following features:

1. The size of the datasets maintained by each party is independent of those maintained by the other parties.
2. The computational complexity for each party is independent of the number of parties. This is accomplished by introducing an outsourcing provider, O. In fact, all computations related to the number of parties are carried out by O. Thus, the number of parties is irrelevant.

3.1.2 Preliminaries

In this section, we summarize the DDH assumption, Bloom filter, and ElGamal encryption. We consider security according to the honest-but-curious model [5]: all players act according to their prescribed actions in the protocol. A protocol that is secure in an honest-but-curious model does not allow any player to gain information about other players' private input sets, besides that that can be deduced from the result of the protocol. Note that the term *adversary* here refers to insiders, i.e., protocol participants. Outsider adversaries are not considered. In fact, behavior by outsider adversaries can be mitigated via standard network security techniques.

Our protocol is based on the following security assumption.

Definition 3.1 (*DDH Assumption*) Let t be a security parameter. A decisional Diffie–Hellman (DDH) parameter generator $\mathcal{IG}$ is a probabilistic polynomial time (PPT) algorithm, a finite field $\mathbb{F}_p$, and a basepoint $g \in \mathbb{F}_p$ with prime order q. We say that $\mathcal{IG}$ satisfies the *DDH assumption* if $|p_1 - p_2|$ is negligible (in κ) for all PPT algorithms A, where $p_1 = \Pr[(\mathbb{F}_p, g) \leftarrow \mathcal{IG}(1^\kappa); y_1 = g^{x_1}, y_2 = g^{x_2} \leftarrow \mathbb{F}_p : A(\mathbb{F}_p, g, y_1, y_2, g^{x_1 x_2}) = 0]$ and $p_2 = \Pr[(\mathbb{F}_p, g) \leftarrow \mathcal{IG}(1^\kappa); y_1 = g^{x_1}, y_2 = g^{x_2}, z \leftarrow \mathbb{F}_p : A(\mathbb{F}_p, g, y_1, y_2, z) = 0]$.

A Bloom filter [6], denoted by BF, consists of m arrays and has a space-efficient probabilistic data structure. The BF can check whether an element x is included in a set S by encoding S with at most w elements. The encoded Bloom filter of S is denoted by $\mathsf{BF}(S)$.

The BF uses a set of k independent uniform hash functions $\mathcal{H} = \{H_0, \ldots, H_{k-1}\}$, where $H_i : \{0, 1\}^* \longrightarrow \{0, 1, \ldots, m-1\}$ for $0 \leq \forall i \leq k-1$. The BF consists of two functions: Const embeds a given set S into $\mathsf{BF}(S)$ and $\mathsf{ElementCheck}$ checks whether an element x is included in S. $\mathsf{SetCheck}$, an extension of $\mathsf{ElementCheck}$, checks whether an element x in S' is in $S' \cap S$ (see Algorithm 3.3). In Const (see Algorithm 3.1), $\mathsf{BF}(S)$ is constructed for a given set S by first setting all bits in the array to 0. To embed an element $x \in S$ into the filter, the element is hashed using k hash functions to obtain k index numbers, and the bits at these indexes are set to 1, i.e., set $\mathsf{BF}[H_i(x)] = 1$ for $0 \leq i \leq k-1$. In $\mathsf{ElementCheck}$ (see Algorithm 3.2), we check all locations where x is hashed; x is considered to be not in S if any bit at these locations is 0; otherwise, x is probably in S.

Some false positive matches may occur, i.e., it is possible that all $\mathsf{BF}[H_i(y)]$ are set to 1, but y is not in S. The false positive rate `FPR` is given by $\mathtt{FPR} = \left\{1 - \left(1 - \frac{1}{m}\right)^{kw}\right\}^k \approx \left\{1 - e^{-kw/m}\right\}^k$ [7]. However, false negatives are not possible, and so Bloom filters have a 100% recall rate.

Algorithm 3.1 $\mathsf{Const}(S)$

```
Input: A set $S$
Output: A Bloom filter $\mathsf{BF}(S)$
for $i = 0$ to $m - 1$ do
    $\mathsf{BF}(S)[i] \leftarrow 0$
end for
for all $x \in S$ do
    for $i = 0$ to $k - 1$ do
        $j = H_i(x)$
        if $\mathsf{BF}(S)[j] = 0$ then
            $\mathsf{BF}(S)[j] \leftarrow 1$
        end if
    end for
end for
output $\mathsf{BF}(S)$. stop.
```

Algorithm 3.2 $\mathsf{ElementCheck}(\mathsf{BF}, x)$

```
Input: A Bloom filter $\mathsf{BF}(S)$, an element $x$
Output: 1 if $x \in S$ and 0 if $x \notin S$
for $i = 0$ to $k - 1$ do
    $j = H_i(x)$
    if $\mathsf{BF}(S)[j] = 0$ then
        output 0. stop.
    end if
end for
output 1. stop.
```

Algorithm 3.3 $\mathsf{SetCheck}(\mathsf{BF}, S')$

```
Input: A Bloom filter $\mathsf{BF}(S)$, a set $S'$
Output: A set $S_\cap (= S \cap S')$
$S_\cap \leftarrow \{\}$
for all $x \in S'$ do
    for $i = 0$ to $k - 1$ do
        $j = H_i(x)$
        if $\mathsf{BF}[j] = 0$ then
            go to next $x$.
        end if
    end for
    add $x$ to the set $S_\cap$
end for
output $S_\cap$. stop.
```

Homomorphic encryption under addition is useful for processing encrypted data. A typical homomorphic encryption under addition was proposed by Paillier [8]. However, because Paillier encryption cannot reduce the order of a composite group, it is computationally expensive compared with the following ElGamal encryption. Our protocol requires matching without revealing the original messages, for which exponential ElGamal encryption (exElGamal) is sufficient [9]. In fact, the decrypted results of exElGamal encryption can distinguish whether two messages m_1 and m_2 are equal, although the exElGamal scheme cannot decrypt messages itself. Furthermore, exElGamal can be used in (n, n)-threshold distributed decryption [10], where the decryption must be performed by *all players acting together*. An exElGamal encryption with (n, n)-threshold distributed decryption consists of three functions:

Key generation:
Let $\mathbb{F}_p$ be a finite field, $g \in \mathbb{F}_p$, with prime order q. Each player P_i chooses $x_i \in \mathbb{Z}_q$ at random and computes $y_i = g^{x_i} \pmod p$. Then, $y = \prod_{i=1}^{n} y_i \pmod p$ is a public key and each x_i is a share for each player to decrypt a ciphertext.

Encryption: $\mathsf{thrEnc}[m] \to (u, v)$
Let $m \in \mathbb{Z}_q^*$ be a message. Choose $r \in \mathbb{Z}_q$ at random, and compute both $u = g^r \pmod p$ and $v = g^m y^r \pmod p$ for the input message $m \in \mathbb{Z}_q$ and a public key y. Output (u, v) as a ciphertext of m.

Decryption: $\mathsf{thrDec}[(u, v)] \to g^m$
Each player P_i computes $z_i = u^{x_i} \pmod p$. All players then compute $z = \prod_{i=1}^{n} z_i \pmod p$ jointly.[1] Finally, each player can decrypt the ciphertext as $g^m = v/z \pmod p$.

ExElGamal encryption with (n, n)-threshold decryption has the following features:
(1) homomorphic under addition: $\mathsf{Enc}(m_1)\mathsf{Enc}(m_2) = \mathsf{Enc}(m_1 + m_2)$ for messages $m_1, m_2 \in \mathbb{Z}_p$.
(2) homomorphic under scalar operations: $\mathsf{Enc}(m)^k = \mathsf{Enc}(km)$ for a message m and $k \in \mathbb{Z}_q$.

3.1.3 Previous Work

This section summarizes prior works on PSI between a server and a client and MPSI among n players. In PSI, let $S = \{s_1, \ldots, s_v\}$ and $C = \{c_1, \ldots, c_w\}$ be server and client datasets, respectively, where $|S| = v$ and $|C| = w$. In MPSI [1], we assume that each player holds the same number of datasets.

PSI protocol based on polynomial representation: The main idea is to represent the elements in C as the roots of a polynomial. The encrypted polynomial is sent to the server, where it is evaluated on the elements in S, as originally proposed by

[1] The computational complexity of z for each player can be made independent of the number of players in various ways. For example, set $z = 1$. P_1 computes $z = z \cdot z_1$ and sends z to P_2, P_2 computes $z = z \cdot z_2$ and sends z to P_3, and, finally, P_n computes $z = z \cdot z_n$ and shares z among all players. If we place all players in a binary tree, the communication complexity can be reduced, but each player's computational complexity is still independent of the number of players.

Freedman [11]. This is secure against honest-but-curious adversaries under secure public key encryption. The computational complexity is $O(vw)$ exponentiations, and the communication overhead is $O(v + w)$. The computational complexity can be reduced to $O(v \log \log w)$ exponentiations using the balanced allocation technique [12]. Kissner and Song extended this protocol to MPSI [1], which requires $O(nw^2)$ exponentiations and $O(nw)$ communication overhead. The MPSI version is secure against honest-but-curious and malicious adversaries (in the random oracle model) using generic zero-knowledge proofs.

PSI protocol based on DH-key agreement: The main objective here is to apply the DH-key agreement protocol [13]: after representing the server and client datasets as hash values $\{h(s_i)\}$ and $\{h(c_i)\}$, respectively, the client encrypts the dataset as $\{h(c_i)^{r_i}\}$ using a random number r_i and sends the encrypted set to the server. The server encrypts the client set $\{h(c_i)^{r_i}\}$ and the server set $\{h(s_i)\}$ using a random number r, which gives $\{h(c_i)^{rr_i}\}$ and $\{h(s_i)^r\}$, respectively, and returns these sets to the client. Finally, the client evaluates $S \cap C$ by decrypting to $\{h(c_i)^r\}$. This is secure against honest-but-curious adversaries under the DDH assumption. The total computational complexity is $O(v + w)$ exponentiations, and the total communication overhead is $O(v + w)$. The security of this approach can be enhanced against malicious adversaries in the random oracle model [14] by using a blind signature. However, no extensions to MPSI based on the DH-key agreement protocol have been proposed.

PSI protocol based on BF: This protocol was originally proposed in [4]. As the Bloom filter itself reveals information about the other player's dataset, the set of players is separated into two groups: input players who have datasets and privacy players who perform private computations under shared secret information. In [15], the privacy of each player's dataset is protected by encrypting each array of the Bloom filter using Goldwasser–Micali encryption [16]. In an honest-but-curious version, the computational complexity is $O(kw)$ hash operations and $O(m)$ public key operations, and the communication overhead is $O(m)$, where m and k are the number of arrays and hash functions, respectively, used in the Bloom filter. The Bloom filter is used in the Oblivious transfer extension [17, 18] and the newly constructed garbled Bloom filter [19]. The main novelty in the garbled Bloom filter is that each array requires λ bits rather than the single bit needed for the conventional Bloom filter. To embed an element $x \in S$ to a garbled Bloom filter, x is split into k shares with λ bits using XOR-based secret sharing ($x = x_1 \bigoplus \cdots \bigoplus x_k$). The x_i are then mapped to an index of $H_i(x)$. An element y is queried by subjecting all bit strings at $H_i(y)$ to an XOR operation. If the result is y, then y is in S; otherwise, y is not in S. The client uses a Bloom filter $\mathsf{BF}(C)$, and the server uses a garbled Bloom filter $\mathsf{GBF}(S)$. If x is in $C \cap S$, then for every position i it hashes to, $\mathsf{BF}(C)[i]$ must be 1 and $\mathsf{GBF}(S)[i]$ must be x_i. Thus, the client can compute $C \cap S$. The computational complexity of this method is $O(kw)$ hash operations and $O(m)$ public key operations, and the communication overhead is $O(m)$. The number of public key operations can be changed to $O(\lambda)$ using the Oblivious transfer extension. This is secure against honest-but-curious adversaries if the Oblivious transfer protocol is secure. Finally, some researchers have computed the approximate number of multiparty set unions [3].

3.1.4 Practical MPSI

This section presents a practical MPSI that is secure under the honest-but-curious model.

3.1.4.1 Notation and Privacy Definition

In the remainder of this paper, the following notations are used.

- P_i: ith player, $i = 1, \ldots, n$
- O: outsourcing provider with no knowledge of the inputs or outputs
- $S_i = \{s_{i,1}, s_{i,2}, \ldots, s_{i,w_i}\}$: dataset held by P_i, where $|S_i| = \omega_i$
- $\cap S_j$: intersection of all n players
- thrEnc and thrDec: (n, n)-threshold exElGamal encryption and decryption, respectively
- m and k: number of arrays and hashes used in BF
- $\boldsymbol{\ell} = [\ell, \ldots, \ell]$ $(1 \leq \ell \leq n)$: an n-dimensional array, where all strings in the array are set to ℓ
- $\mathsf{BF}(S_i) = [\mathsf{BF}_i[0], \ldots, \mathsf{BF}_i[m-1]]$: Bloom filter applied to a set S_i
- $\mathsf{IBF}(\cap S_i) = [\sum_{i=1}^{n} \mathsf{BF}_i[0], \ldots, \sum_{i=1}^{n} \mathsf{BF}_i[m-1]]$: integrated Bloom filter of n sets $\{S_i\}$, where $\sum_{i=1}^{n} \mathsf{BF}_i[j]$ is the sum of all players' arrays

We introduce an outsourcing provider O to reduce the computational burden on all players. The dealer has no information regarding the elements of any player's set. The privacy issues faced by MPSI with an outsourcing provider can be informally written as follows.

Definition 3.2 (*MPSI privacy*) An MPSI scheme with an outsourcing provider O is player-private if the following two conditions hold:

- P_i does not learn anything about the elements of other players' datasets except for the elements that P_i originally possesses.
- the outsourcing provider O does not learn anything about the elements of any player's set.

3.1.4.2 Proposed MPSI

Our MPSI comprises four phases: (i) initialization, (ii) Bloom filter construction and the encryption of P_i data, (iii) the O's randomization of $\mathsf{thrEnc}(\mathsf{IBF}(\cup S_i) - \mathbf{n})$, and (iv) the computation of $\cap \mathsf{P}_i$. The computation of $\cap \mathsf{P}_i$ consists of three steps: (a) joint decryption of an (n, n)-threshold exElGamal among n players, (b) Bloom filter check, and (c) output intersection.

Figure 3.1 shows an overview of our protocol after the initialization phase. The system parameters of a finite field $\mathbb{F}_p$ and a basepoint $g \in \mathbb{F}_p$ with order q for an

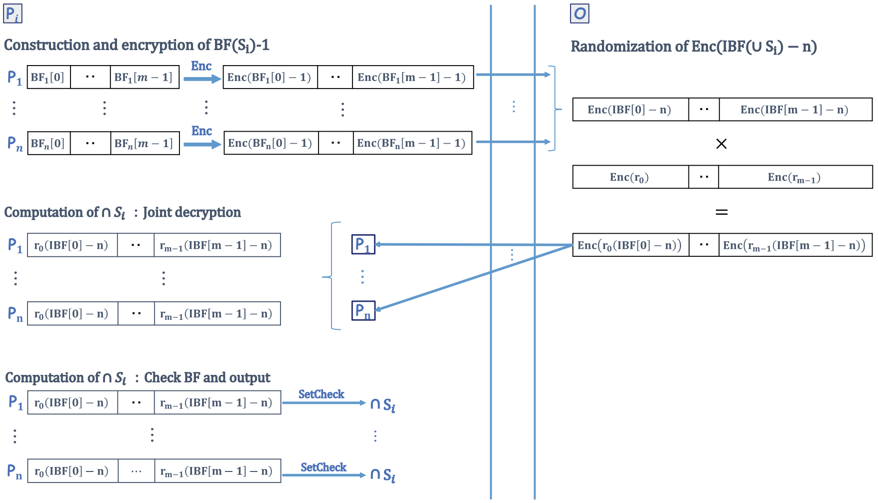


Fig. 3.1 Overview of our MPSI

(n, n)-threshold exElGamal encryption (thrEnc, thrDec) are provided to both P_i and $\mathcal{O}$. For the Bloom filter, $\mathsf{Const}(S)$ and $\mathsf{SetCheck}(\mathsf{BF}, S')$ are only provided to P_i, where the array size is m and k independent hash functions are used.

To encrypt, randomize, or subtract a vector such as a Bloom filter $\mathsf{BF} = [a_0, \ldots, a_{m-1}]$, each location is encrypted, randomized, or subtracted independently:

$$\mathsf{thrEnc}(\mathsf{BF}) = [\mathsf{thrEnc}(a_0), \ldots, \mathsf{thrEnc}(a_{m-1})],$$
$$\mathbf{r}\mathsf{BF} = [r_0 a_0, \ldots, r_{m-1} a_{m-1}], \text{ or}$$
$$\mathsf{BF} - \mathbf{r} = [a_0 - r_0, \ldots, a_{m-1} - r_{m-1}]$$

for $\mathbf{r} = [r_0, \ldots, r_{m-1}] \in \mathbb{Z}_q^m$.

Our protocol proceeds as follows.

Initialization:

1. P_i generates $x_i \in \mathbb{Z}_q$, computes $y_i = g^{x_i} \in \mathbb{Z}_q$, and publishes y_i to the other players as a public key, where the corresponding secret key is x_i.
2. P_i computes $y = \prod_i y_i$, where y is the n-player public key. Note that no player knows the corresponding secret key $x = \sum x_i$ before executing the joint decryption.

Construction and encryption of $\mathsf{BF}(S_i) - \mathbf{1}$:

1. P_i executes $\mathsf{Const}(S_i) \longrightarrow \mathsf{BF}(S_i) = [\mathsf{BF}_i[0], \ldots, \mathsf{BF}_i[m-1]]$ (Algorithm 3.1).
2. P_i encrypts $\mathsf{BF}(S_i) - \mathbf{1}$ using thrEnc_y:

$$\mathsf{thrEnc}_y(\mathsf{BF}(S_i) - \mathbf{1}) = [\mathsf{thrEnc}_y(\mathsf{BF}_i[0] - 1), \ldots, \mathsf{thrEnc}_y(\mathsf{BF}_i[m-1] - 1)],$$

where y is an n-player public key.

3. P_i sends $\mathsf{thrEnc}_y(\mathsf{BF}(S_i) - \mathbf{1})$ to O.

Randomization of $\mathsf{thrEnc}(\mathsf{IBF}(\cap S_i) - \mathbf{n})$:

1. O encrypts $\mathsf{IBF}(\cap S_i) - \mathbf{n}$ without knowing $\mathsf{IBF}(\cap S_i)$ using an additive homomorphic feature and multiplying by $\mathsf{thrEnc}_y(\mathsf{BF}(S_i) - \mathbf{1})$ as follows:

$$\mathsf{thrEnc}_y(\mathsf{IBF}(\cap S_i) - \mathbf{n}) = \prod_{i=1}^{n} \mathsf{thrEnc}_y(\mathsf{BF}(S_i) - \mathbf{1}).$$

2. O randomizes $\mathsf{thrEnc}_y(\mathsf{IBF}(\cap S_i) - \mathbf{n})$ by $\mathbf{r} = [r_0, \ldots, r_{m-1}] \in \mathbb{Z}_q^m$:

$$\mathsf{thrEnc}_y(\mathbf{r}(\mathsf{IBF}(\cap S_i) - \mathbf{n})) = (\mathsf{thrEnc}_y(\mathsf{IBF}(\cup S_i) - \mathbf{n}))^{\mathbf{r}}.$$

3. O broadcasts $\mathsf{thrEnc}_y(\mathbf{r}(\mathsf{IBF}(\cap S_i) - \mathbf{n}))$ to P_i.

Computation of $\cap S_i$:

1. All players decrypt $\mathsf{thrEnc}_y(\mathbf{r}(\mathsf{IBF}(\cap S_i) - \mathbf{n}))$ jointly.
2. P_i computes $\mathsf{SetCheck}(\mathbf{r}(\mathsf{IBF}(\cap S_i) - \mathbf{n}), S_i)$ and obtains $\cap S_i$.

The above protocol satisfies the correctness requirement. This is because each array position of $\mathsf{thrEnc}_y(\mathbf{r}(\mathsf{IBF}(\cap S_i) - \mathbf{n}))$ is decrypted to 1, where $x \in \cap S_i$ is embedded by each hash function; however, each array position for which $x \notin \cap S_i$ is embedded by each hash function is decrypted to a random value.

3.1.4.3 Security Proof

The security of our MPSI protocol is as follows.

Theorem 3.1 *For any coalition of fewer than n players, the MPSI is player-private against an honest-but-curious adversary under the DDH assumption.*

Proof The views of P_i and O, that is,

$$\mathsf{thrEnc}_y(\mathsf{BF}_{m,k}(S_i)) = [\mathsf{thrEnc}_y(\mathsf{BF}_i[0]), \ldots, \mathsf{thrEnc}_y(\mathsf{BF}_i[m-1])],$$

are shown to be indistinguishable from a random vector $\mathbf{r} = [r_0, \ldots, r_{m-1}] \in \mathbb{Z}_q^m$. Assume that a polynomial-time distinguisher $\mathcal{D}$ outputs 0 when the views are presented as a random vector and outputs 1 when they are constructed in MPSI, $\mathsf{thrEnc}(\mathsf{BF}_i[0]), \ldots, \mathsf{thrEnc}(\mathsf{BF}_i[m-1])$. We show that a simulator $\overline{\mathsf{SIM}}$ that solves the DDH assumption can be constructed as follows.

Upon receiving a DDH challenge $(\overline{g}, \overline{g}^{\alpha}, \overline{g}^{\beta}, \overline{g}^{\gamma})$, $\overline{\mathsf{SIM}}$ executes the following:

1. Set n-player public key $y = \overline{g}^{\beta}$ and choose random numbers $d_0, \ldots, d_{m-1}$ and $r_1, \ldots, r_{m-1}$ from $\mathbb{Z}_q$.

2. Send $[(\overline{g}^{\alpha}, \overline{g}^{d_0} \cdot \overline{g}^{\gamma}), ((\overline{g}^{\alpha})^{r_1}, \overline{g}^{d_1} \cdot (\overline{g}^{\gamma})^{r_1}), \ldots, ((\overline{g}^{\alpha})^{r_{m-1}}, \overline{g}^{d_{m-1}} \cdot (\overline{g}^{\gamma})^{r_{m-1}})]$ as $\mathsf{thrEnc}_y(\mathsf{BF}_{m,k}(S_i))$ to $\mathcal{D}$.

If $(\overline{g}, \overline{g}^{\alpha}, \overline{g}^{\beta}, \overline{g}^{\gamma})$ is a DH-key-agreement-protocol element, i.e., $\gamma = \alpha\beta$, then $\mathsf{thrEnc}_y(\mathsf{BF}_{m,k}(S_i))$ is distributed in the same way as when constructed by the MPSI scheme. Thus, $\mathcal{D}$ must output 1. If $(\overline{g}, \overline{g}^{\alpha}, \overline{g}^{\beta}, \overline{g}^{\gamma})$ is not a DH tuple, then $\mathsf{thrEnc}_y(\mathsf{BF}_{m,k}(S_i))$ is randomly distributed, and $\mathcal{D}$ has to output 0. Therefore, SIM can use the output of $\mathcal{D}$ to respond to the DDH challenge correctly. Therefore, $\mathcal{D}$ can answer correctly with negligible advantage over random guessing. Furthermore, as all inputs of each player are encrypted until the decryption is performed, and decryption cannot be performed by fewer than n players, nothing can be learned by any player prior to decryption.

As for the views of $\mathsf{thrEnc}_y(\mathbf{r}(\mathsf{IBF}_{m,k}(\cap S_i) - \mathbf{n}))$, the same argument holds. Therefore, for any coalition of fewer than n players, MPSI is player-private under the honest-but-curious model.

Next, we present d-and-over MPSI. The procedures of d-and-over MPSI are the same as those of MPSI until $\mathcal{O}$ computes $\mathsf{thrEnc}_y(\mathsf{IBF}(\cap S_i))$. Thus, we describe the procedure after $\mathcal{O}$ computes $\mathsf{thrEnc}_y(\mathsf{IBF}(\cap S_i))$.

Encryption of ℓ-subtraction of $\mathsf{IBF}(\cap S_i)$: $\mathcal{O}$ executes the following:

1. Encrypt $\mathsf{IBF}(\cap S_i) - \boldsymbol{\ell}$ randomized by $\mathbf{r} = [r_0, \ldots, r_{m-1}] \in \mathbb{Z}_q^m (d \leq \ell \leq n)$: $\mathsf{thrEnc}_y(\mathbf{r}(\mathsf{IBF}(\cap S_i) - \boldsymbol{\ell})) = (\mathsf{thrEnc}_y(\mathsf{IBF}(\cap S_i)) \cdot \mathsf{thrEnc}_y(-\boldsymbol{\ell}))^{\mathbf{r}}$.
2. Broadcast $\{\mathsf{thrEnc}_y(\mathbf{r}(\mathsf{IBF}(\cap S_i) - \boldsymbol{\ell}))\}_{\ell}$ $(d \leq \ell \leq n)$ to P_i.

d-and-over MPSI computation: P_i executes the following:

1. All P_i jointly decrypt $\{\mathsf{thrEnc}_y(\mathbf{r}(\mathsf{IBF}(\cap S_i) - \boldsymbol{\ell}))\}_{\ell}$.
2. Let CBF_{ℓ} be an m-array for $d \leq \ell \leq n$, where an array is set to 1 if and only if the corresponding array of $\mathbf{r}\mathsf{IBF}(\cap S_i) - \boldsymbol{\ell}$ is 1, and others are set to 0.
3. Set $\mathsf{CBF} = \mathsf{CBF}_{\ell} \vee \cdots \vee \mathsf{CBF}_n$.
4. Execute $\mathsf{SetCheck}_{m,k}(\mathsf{CBF}, S_i) \longrightarrow \cap^{\geq d} S[i]$ and output $\cap^{\geq d} S[i]$.

The correctness of d-and-over MPSI follows from the fact that if an element $x \in \cap^{\ell} S$ for $d \leq \exists \ell \leq n$, the corresponding array locations in $\mathsf{IBF}(\cap S_i) - \mathbf{j}$ for $\ell \leq \exists j \leq n$, where x is mapped by k hashes, are an encryption of 0, which are decrypted to 1; otherwise, it is an encryption of randomized value.

3.1.5 *Efficiency*

Although many PSI protocols have been proposed, to the best of our knowledge, relatively few consider the multiparty scenario [1–4]. Our target is multiparty private set intersection, and the final result must be obtained by *all* players acting together, without a trusted third-party (TTP). Among previous MPSI protocols, the approach in [3] computes only the approximate number of intersections, and that in [4] requires

Table 3.1 Efficiency of [1] and the proposed protocol

	[1]	Ours
Computational complexity	$O(n\omega^2)$	$\mathsf{P}_i : O(\omega_i), O : O(n\omega)$
Communication overhead	$O(n\omega)$	$\mathsf{P}_i : O(\omega + n), O : O(n\omega)$
Restriction on set size	$\|S_1\| = \cdots = \|S_n\|$	None
Protected values	$S_i (\forall i \in [1, n])$	$S_i, \|S_i\| (\forall i \in [1, n])$

more than two TTPs. In contrast, [2] follows almost the same method as [1] and thus has a similar complexity. The only difference exists in the security model. Hence, we only compare our scheme with that of [1].

The computational and communication efficiency of the proposed protocol and [1] are compared in Table 3.1. These approaches are secure against honest-but-curious adversaries without a TTP under exElGamal encryption (DDH security) and Paillier encryption (Decisional Composite Residue (DCR) security), respectively. The Bloom filter parameters (m, k) used in our protocol are set as follows: $k = 80$ and $m = 80\omega/\ln 2$, where ω is the maximum $|S_i| = \omega_i$. Then, the probability of false positives is given by $p = 2^{-80}$.

Our MPSI uses the Bloom filter for the computations performed by P_i and the integrations performed by the O. The use of a Bloom filter eliminates the restriction on set size. Thus, in our MPSI, the set size of each player is flexible. However, P_i's computations consist of Bloom filter construction, joint decryption, and Bloom filter check. Neither the computations related to the Bloom filter nor the joint decryption depends on the number of players, as shown in Sect. 3.1.2. In summary, the computational complexity of operations performed by P_i is $O(\omega_i)$. All player-dependent data are sent to O, who integrates $\prod_{i=1}^{n} \mathsf{thrEnc}_y(\mathsf{IBF}(\cap S_i))$ without decryption. Therefore, the computational complexity of operations performed by O is $O(n\omega)$.

3.1.6 System and Performance

PSI or MPSI implicitly assumes that every attendee can provide data, any attendee can retrieve data from the shared data, and all attendees can communicate with each other. If PSI or MPSI is implemented straightforwardly, such implementation should become a system like a peer-to-peer (P2P) network system. Although a fully distributed system like P2P network has attractive features, such as high availability and scalability, it incurs some unfavorable features.

The network address and port translation (NAPT) is a major obstacle for P2P network systems. Modern P2P network systems take advantage of NAPT traversal technologies to overcome NAPT, but it should be costly to make the architecture complex. The absence of trusted node is also an obstacle for attendee or group management. Making consensus on a P2P network system is difficult or highly

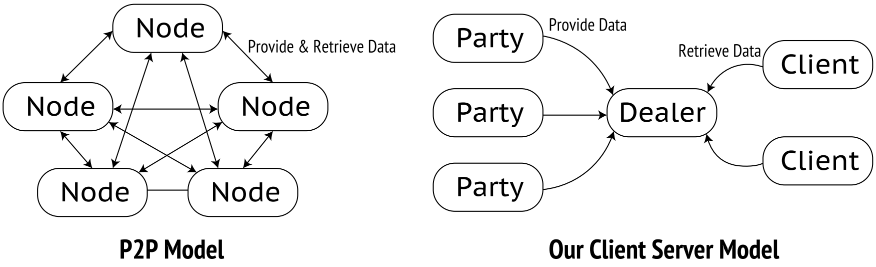


Fig. 3.2 P2P and client server model

costly. Additionally, unpredictable node joining and leaving are reasons that make the P2P network systems complex. To avoid the complexities of P2P networks, we designed a system based on the client server model.

Then, we discuss the design of PSI or MPSI's client server model. There are 2 main functionalities of PSI or MPSI: (1) First, the data sharing is a functionality for sharing data among attendees. (2) Next, the data retrieving from the shared data is a functionality. Any attendee can retrieve data from the shared data, but the retrieving avoids correcting privacy sensitive data by using privacy preserving techniques described above.

However, we do not assume that every attendee provides and retrieves data. Imagine that an incident analysis situation in which data are provided by several organizations which employ labor and operate some machines, and a research institute collects data from the organizations and analyzes it. In such a situation, data providers do not need the data retrieving functionality, and data analysts do not need the data sharing functionality.

Therefore, we define 3 roles for our MPSI application design as follows.

- Parties: entities for data providing
- Clients: entities for data retrieving
- Dealer: an entity for forwarding requests between parties and clients

From the perspective of privilege separation, defining and separating roles are significant. Figure 3.2 shows a P2P network model and our client server model. As show in this figure, every P2P network node is connected to each other and can provide and retrieve data, but parties only provide data and clients only retrieve data in the client server model. The dealer forwards requests from parties and clients and provides other functionalities that are not specified by PSI or MPSI. For example, attendee or group management, user authentication, and data logging should be performed by the dealer.

Figure 3.3 shows an example sequence diagram of our MPSI application. In this figure, there are 2 parties, 1 client, and 1 dealer. First of all, parties 1 and 2 join the dealer (join p1 and p2). A party must join before providing data, and it must be performed only once at initialization. After that, the client sends a request of data retrieval to the dealer (cl req), and parties send a request to confirm whether the dealer

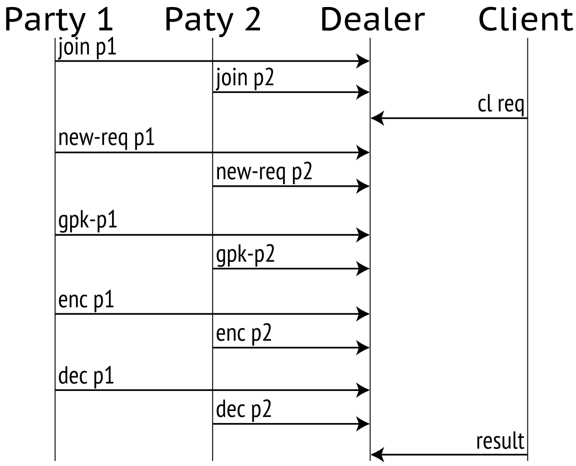


Fig. 3.3 Sequence diagram of MSPI application

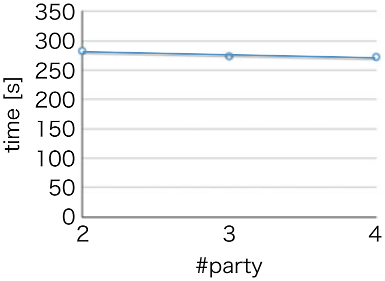


Fig. 3.4 Performance

received data retrieval requests by clients (new-req p1 and p2). Then, the parties and the dealer generate keys, share the keys, encrypt data, and decrypt data (gpk p1 and p2, enc p1 and p2, and dec p1 and p2). Finally, the client gets the result from the dealer.

We measured performance of our MPSI application written in Python language on an Amazon's EC2 server (2.4 GHz CPU, 1 GB Memory). Figure 3.4 shows the results when there are from 2 to 4 parties which provide data including 10,000 entries. The results show that it takes approximately 280 s to accomplish data retrieval and that the computational amount does not depend on the number of parties.

3.2 Classification

In this section, we present a secure classification protocol, a type of secure computation protocols. We assume two participants Alice and Bob of the protocol. Alice has private data x, and Bob has a classification model C. The task is that Alice learns $C(x)$ at the end of the protocol while preserving the privacy of x and C. That is, Alice can learn only $C(x)$ and Bob can learn nothing. Our construction is based on a code-based public-key encryption scheme called HQC [20], which is a candidate of NIST's Post-Quantum Cryptography standardization [21].

3.2.1 Error-Correcting Code

We start with several fundamental notions for error-correcting codes.

Definition 3.3 (*Linear code*) A code $\mathbb{C}$ such that $c_1 + c_2 \in \mathbb{C}$ always holds for any codeword $c_1, c_2 \in \mathbb{C}$ is called a linear code. The code $\mathbb{C}$ of code length n and information bit number k is described as "a" code.

Definition 3.4 (*Generation matrix*) For matrices $\mathbb{G} \in \mathbb{F}^{k \times n}$,$\mathbb{G}$ that satisfy

$$\mathbb{C} = \{\boldsymbol{m} \cdot \mathbb{G} | \boldsymbol{m} \in \mathbb{F}^k\} \tag{3.1}$$

is called a generator matrix. The generator matrix is the basis of linear codes and generates all codewords.

Definition 3.5 (*Parity check matrix*) For a matrix $\mathbf{H} \in \mathbb{F}^{(n-k) \times n}$, $\mathbf{H}$ that satisfies

$$\mathbb{C} = \{\boldsymbol{x} \in \mathbb{F}^n | \mathbf{H} \cdot \boldsymbol{x}^\top = \mathbf{0}\} \tag{3.2}$$

is called a parity check matrix.

Definition 3.6 (*Cyclic matrix*) When $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{F}^n$, the circulant matrix for $\boldsymbol{x}$ is defined as

$$\mathbf{rot}(\boldsymbol{x}) = \begin{pmatrix} x_1 & x_n & \cdots & x_2 \\ x_2 & x_1 & \cdots & x_3 \\ \vdots & \vdots & \ddots & \vdots \\ x_n & x_{n-1} & \cdots & x_1 \end{pmatrix} \in \mathbb{F}^{n \times n} \tag{3.3}$$

In addition, the multiplication of two polynomials x, y has the following properties:

$$\begin{aligned} \boldsymbol{x} \cdot \boldsymbol{y} &= \boldsymbol{x} \times \mathbf{rot}(\boldsymbol{y})^\top \\ &= (\mathbf{rot}(\boldsymbol{x}) \times \boldsymbol{y}^\top)^\top \\ &= \boldsymbol{y} \times \mathbf{rot}(\boldsymbol{x})^\top \\ &= \boldsymbol{y} \cdot \boldsymbol{x}. \end{aligned} \tag{3.4}$$

Definition 3.7 (*Cyclic shift*) The operation of shifting $(c_0, \ldots, c_{n-1})$ to the right by one position with respect to n-dimensional vector c_i $(i = 0, \ldots, n-2)$ and moving c_{n-1} to the beginning of the vector is called cyclic shift. That is, for any n dimensional vector $(c_0, \ldots, c_{n-1})$, it is a mapping $\sigma : (c_0, c_1, \ldots, c_{n-1}) \mapsto (c_{n-1}, c_0, \ldots, c_{n-2})$.

Definition 3.8 (*Quasi-cyclic code*) Let $\boldsymbol{c} = (\boldsymbol{c}_0, \ldots, \boldsymbol{c}_{s-1}) \in (\mathbb{F}_2^n)^s$ be an arbitrary codeword of code $\mathbb{C}$ and let σ be a cyclic shift operation. If $(\sigma(\boldsymbol{c}_0), \ldots, \sigma(\boldsymbol{c}_{s-1}) \in \mathbb{C}$, $\mathbb{C}$ is called the s-quasi-cyclic code. In particular, when $s = 1$, $\mathbb{C}$ is called a cyclic code.

Definition 3.9 (*Systematic quasi-cyclic code*) An s-quasi-cyclic $[sn, n]$ code is called a systematic quasi-cyclic code if it has a parity check matrix of the form.

$$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{I}_n & 0 & \cdots & 0 & \boldsymbol{A}_1 \\ 0 & \boldsymbol{I}_n & & & \boldsymbol{A}_2 \\ & & \ddots & & \vdots \\ 0 & & \cdots & \boldsymbol{I}_n & \boldsymbol{A}_{s-1} \end{bmatrix} \tag{3.5}$$

Here, $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_{s-1}$ is an $n \times n$ circulant matrix.

3.2.2 Security Assumptions

As mentioned above, the security of the public-key cryptosystem HQC is based on the computational difficulty of the quasi cyclic syndrome decoding problem. More specifically, its security is proved under the following quasi cyclic syndrome decoding decision assumptions.

Definition 3.10 (*quasi-cyclic syndrome decoding assumption*) The quasi-cyclic syndrome decoding decision problem of a s-quasi-cyclic code in which n and w are integers and the number of blocks is $s \geq 2$ is $(\mathbf{H}, \boldsymbol{y}^\top)$ when the parity check matrix $\mathbf{H} \xleftarrow{\$} \mathbb{F}^{(sn-n)\times sn}$ and the matrix $\boldsymbol{y} \xleftarrow{\$} \mathbb{F}^{sn-n}$ of random systematic quasi-cyclic code are given, every efficient algorithm distinguish only with negligible probability whether it is quasi-cyclic syndrome decoding distribution or the uniform distribution over $\mathbb{F}^{(sn-n)\times sn} \times \mathbb{F}^{(sn-n)}$.

As will be described later, since the security of the secure computation protocol proposed in this section is reduced to the security of HQC, the secure computation protocol of this section is proved to be secure under this assumption as well as under HQC.

3.2.3 Security Requirements for 2PC

Secure two-party computation is a subproblem of multi-party secure computation. The studies have been conducted by many researchers since it is closely related to many cryptographic protocols. The purpose of 2PC is to construct a general-purpose protocol so that arbitrary functions can be jointly computed without sharing the input values of the two parties with the other. One of the best-known examples of 2PCs is the millionaire problem [22] in Yao, where Alice and Bob do not reveal their money and decide who is richer. Specifically, suppose that Alice has a yen, and Bob has b yen. The problem is to decide whether $a \geq b$ or not while keeping each other secret. Generally speaking, the security requirement of 2PC is that the computation of any function is performed using a protocol without leaking the two inputs to the other, and only the computation result is known.

A two-party linear function evaluation is a kind of 2PC that satisfies the 2PC security requirements. In other words, the participants perform the evaluation without notifying the other party of their input. In addition, the function of the protocol is the evaluation of linear functions. Specifically, linear function secure computation protocol computes $f(m) = a \cdot m + b$. The participants in the protocol are called Alice and Bob. Alice's input is m, and Bob's input is linear function parameters a, b. Alice gets only the result of $f(m) = a \cdot m + b$ through the protocol, and Bob gets nothing.

Below we define the security requirements for two-party linear function secure computation.

Definition 3.11 (*Security against semi-honest adversaries*) Let $f = (f_A, f_B)$ be the function that maps the input x of Alice(A) and the input y of Bob(B) to $f_A(x, y)$,$f_B(x, y)$. A aims to obtain $f_A(x, y)$ and B aims to obtain $f_B(x, y)$.

Let $f = (f_A, f_B)$ be a function of probabilistic polynomial time, and π be a two-way protocol for computing function f. Let the view of A with (x, y) execution $\pi(x, y)$ and the security parameter n be $\text{view}_A^\pi(x, y, n)$ and the view of B be $\text{view}_B^\pi(x, y, n)$. The output of A is $\text{output}_A^\pi(x, y, n)$ and the output of B is $\text{output}_B^\pi(x, y, n)$. In addition, the joint output of the two is denoted as output^π $(x, y, n) = (\text{output}_A^\pi(x, y, n), \text{output}_B^\pi(x, y, n))$.

For semi-honest adversaries, we say that the protocol $\pi(x, y)$ can securely compute the function f if there are probabilistic polynomial-time algorithms S_A and S_B that satisfy the following equations. For any x, y that satisfy $|x| = |y| = n, n \in \mathbb{N}$, the following holds:

$$\begin{aligned}
&\{(S_A(1^n, x, f_A(x, y)), f(x, y))\}_{x,y,n} \\
\overset{c}{\equiv}&\{(\text{view}_A^\pi(x, y, n), \text{output}^\pi(x, y, n))\}_{x,y,n}, \\
&\{(S_B(1^n, x, f_B(x, y)), f(x, y))\}_{x,y,n} \\
\overset{c}{\equiv}&\{(\text{view}_B^\pi(x, y, n), \text{output}^\pi(x, y, n))\}_{x,y,n}.
\end{aligned}$$

3.2.4 HQC Encryption Scheme

The protocols proposed in this section are based on the Hamming Quasi-Cyclic cryptosystem of Gaborit et al. First, we introduce the cryptosystem proposed by Gaborit et al. [20], which is a public key cryptosystem based on the quasi-cyclic syndrome decoding problem. In this cryptosystem, two kinds of codes quasi-cyclic code and error-correcting code $\mathbb{C}$ are used. The error-correcting code $\mathbb{C}$ is an arbitrary linear code (such as a BCH code) used for message encoding and decoding and with sufficient error correction capability. A quasi-cyclic code is used for a security requirement of this public key cryptosystem to generate noise that an adversary cannot decrypt.

The participants of the HQC cryptosystem are Alice (A) and Bob (B), and B aims to send the input message $\boldsymbol{m}$ securely to A. The cryptosystem is performed as follows:

1. Global parameter settings:
   Parameters param = $(n, k, \delta, w_x, w_r, w_e)$ and the sign $\mathbb{C}$ generation matrix $\mathbb{G} \in \mathbb{F}^{k \times n}$.
2. Key generation:
   A generates random $\boldsymbol{h} \xleftarrow{\$} \mathbb{R}$.
   Furthermore, $(\boldsymbol{x}, \boldsymbol{y}) \xleftarrow{\$} \mathbb{R}^2$ is generated, and the Hamming weight of $\boldsymbol{x}, \boldsymbol{y}$ is w_x.
   Secret information sk = $(\boldsymbol{x}, \boldsymbol{y})$ Public information pk = $(\boldsymbol{h}, \boldsymbol{s} = \boldsymbol{x} + \boldsymbol{h} \cdot \boldsymbol{y})$. A sends public information pk to B.
3. Encryption:
   B generates a random $\boldsymbol{e} \xleftarrow{\$} \mathbb{R}$, $(\boldsymbol{r_1}, \boldsymbol{r_2}) \xleftarrow{\$} \mathbb{R}^2$.
   The Hamming weight of $\boldsymbol{e}$ is w_e, and the Hamming weight of $\boldsymbol{r_1}$ and $\boldsymbol{r_2}$ is w_r.
   Then, we compute $\boldsymbol{u} = \boldsymbol{r_1} + \boldsymbol{h} \cdot \boldsymbol{r_2}$ and $\boldsymbol{v} = \boldsymbol{m} \cdot \mathbb{G} + \boldsymbol{s} \cdot \boldsymbol{r_2} + \boldsymbol{e}$ on input $\boldsymbol{m}$. B sends the ciphertext $\boldsymbol{u}, \boldsymbol{v}$ back to A.
4. Decryption:
   A uses the decoding function $\mathbb{C}$.Decode$(\boldsymbol{v} - \boldsymbol{u} \cdot \boldsymbol{y})$ of the error-correcting code $\mathbb{C}$ to recover the message $\boldsymbol{m}$ of B.

In the HQC cryptosystem, public information $\boldsymbol{s}$ is added to the message $\boldsymbol{m}$ encoded by the error-correcting code when it is encrypted. Since $\boldsymbol{s}$ is noise with a large Hamming weight generated by the quasi-cyclic code, security is guaranteed by the quasi-cyclic syndrome decoding decision assumption introduced above. In addition, A can use the secret key for the encrypted error-protected ciphertext in the decryption stage, and can remove a large amount of noise from $\boldsymbol{s}$. However, some noise of $\boldsymbol{x} \cdot \boldsymbol{r_2} - \boldsymbol{r_1} \cdot \boldsymbol{y} + \boldsymbol{e}$ remains. If the weight of this noise is smaller than the maximum number of correctable errors δ of the error-correcting code, correct decoding is possible. Hamming weights $w, w_r, w_e = O(\sqrt{n})$ are assumed and analyzed. Moreover, the conclusion that the probability of becoming $\omega(\boldsymbol{x} \cdot \boldsymbol{r_2} + \boldsymbol{e} - \boldsymbol{y} \cdot \boldsymbol{r_1}) \leq \delta$ increases as the code space n becomes larger is shown in the paper of Gaborit et al. In addition, the HQC cryptosystem is IND-CPA secure under the quasi-cyclic syndrome decoding decision assumption.

3.2.5 Proposed Protocol

3.2.5.1 Linear Function Evaluation

We introduce the secure evaluation protocol of the linear functions between two parties.

We use two codes, quasi-cyclic code and arbitrary error-correcting code $\mathbb{C}$, based on Gaborit's HQC cryptosystem. The participants in the protocol are Alice (A) and Bob (B). A's input is $m \in \mathbb{F}_2$, B's input is $a, b \in \mathbb{F}_2$, B's output is nothing, and A's output is $a \cdot m + b$. The protocol is given in Protocol 3.2.5.1.

Protocol Linear function evaluation protocol

input	A: $m \in \mathbb{F}_2$
	B: $a, b \in \mathbb{F}_2$
output	A: $a \cdot m + b$
	B: $\perp$

1. Global parameter param = $(n, k, \delta, w_x, w_r, w_e)$ and the sign $\mathbb{C}$ generation matrix $\mathbb{G} \in \mathbb{F}^{k \times n}$ are chosen.
2. A generates the random $\boldsymbol{h} \xleftarrow{\$} \mathbb{R}$. Furthermore, $(\boldsymbol{x}, \boldsymbol{y} \xleftarrow{\$} \mathbb{R}^2)$ is generated, and the Hamming weight of $\boldsymbol{x}$ and $\boldsymbol{y}$ is w. Secret information sk = $(\boldsymbol{x}, \boldsymbol{y})$, Public information pk = $(\boldsymbol{h}, \boldsymbol{s} = \boldsymbol{x} + \boldsymbol{h} \cdot \boldsymbol{y})$.
3. By padding the input m with 0, A makes $\boldsymbol{m} = (m, 0, \ldots, 0)$ of dimension k. A generates a random $\boldsymbol{r_A}, \boldsymbol{r_u}, \boldsymbol{r_v} \xleftarrow{\$} \mathbb{R}$. Here, the Hamming weight of $\boldsymbol{r_A}, \boldsymbol{r_u}, \boldsymbol{r_v}$ is w_r. Then, we compute $(\boldsymbol{u} = \boldsymbol{h} \cdot \boldsymbol{r_A} + \boldsymbol{r_u}, \boldsymbol{v} = \boldsymbol{m} \cdot \mathbb{G} + \boldsymbol{s} \cdot \boldsymbol{r_A} + \boldsymbol{r_v})$. A sends public information $\boldsymbol{h}, \boldsymbol{s}$ and ciphertext pair $\boldsymbol{u}, \boldsymbol{v}$ to B.
4. Let B be $\boldsymbol{b} = (b, 0, \ldots, 0)$. Generate $\boldsymbol{r_B} \xleftarrow{\$} \mathbb{R}$ and $(\boldsymbol{e_u}, \boldsymbol{e_v}) \xleftarrow{\$} \mathbb{R}^2$. Here, the Hamming weight of $\boldsymbol{r_B}$ is w_r, and the Hamming weight of $\boldsymbol{e_u}$ and $\boldsymbol{e_v}$ is w_e. B computes $\boldsymbol{u}' = a \cdot \boldsymbol{u} + \boldsymbol{h} \cdot \boldsymbol{r_B} + \boldsymbol{e_u}$ and $\boldsymbol{v}' = a \cdot \boldsymbol{v} + \boldsymbol{b} \cdot \mathbb{G} + \boldsymbol{s} \cdot \boldsymbol{r_B} + \boldsymbol{e_v}$. B sends $\boldsymbol{u}', \boldsymbol{v}'$ back to A.
5. A uses $\mathbb{C}$. Decode$(\boldsymbol{v'} - \boldsymbol{u'} \cdot \boldsymbol{y})$ to decode the error-correcting code $\mathbb{C}$, and recovers $a \cdot m + b$ by taking the first bit of the result.

First, we set global parameters. n is the code length of the code, k is the number of information bits, δ is the maximum number of correctable errors in the error-correcting code, and w_x, w_r, w_e are Hamming weights set in advance. For example, it is half the weight of $O(\sqrt{n})$ assumed by Gaborit et al. The public parameter $\mathbb{G}$ is a generator matrix of error-correcting code $\mathbb{C}$, which maps messages and codewords as $\mathbb{F}_2^k \to \mathbb{F}_2^n$.

A generates random $\boldsymbol{h} \xleftarrow{\$} \mathbb{R}$ and $(\boldsymbol{x}, \boldsymbol{y}) \xleftarrow{\$} \mathbb{R}^2$ and computes $\mathbf{s} = \mathbf{x} + \mathbf{h} \cdot \mathbf{y}$. Here,

$$\begin{aligned} \boldsymbol{s} &= \boldsymbol{x} + \boldsymbol{h}\cdot\boldsymbol{y} \\ &= \boldsymbol{x} + \boldsymbol{y}\cdot\mathbf{rot}(\boldsymbol{h})^\top \\ &= (\boldsymbol{x}\ \ \boldsymbol{y})(\boldsymbol{I}_n\ \mathbf{rot}(\boldsymbol{h}))^\top. \end{aligned} \tag{3.6}$$

It can be converted to and can be reduced to the quasi cyclic syndrome decoding problem. Then, A sets secret information sk as $(\boldsymbol{x}, \boldsymbol{y})$ and public information pk as $(\boldsymbol{h}, \boldsymbol{s})$.

A pads the input m with 0, making $\boldsymbol{m} = (m, 0, \ldots, 0)$ with dimension k. A generates $\boldsymbol{r_A}, \boldsymbol{r_u}, \boldsymbol{r_v} \xleftarrow{\$} \mathbb{R}$, encodes the value of $\boldsymbol{m}$ with an error-correcting code, and re-randomizes it. A generates a ciphertext pair of $(\boldsymbol{u} = \boldsymbol{h} \cdot \boldsymbol{r_A} + \boldsymbol{r_u}, \boldsymbol{v} = \boldsymbol{m} \cdot \mathbb{G} + \boldsymbol{s} \cdot \boldsymbol{r_A} + \boldsymbol{r_v})$ and send it to B. As for B, $\boldsymbol{v}$ has a noise $\boldsymbol{s}$ that cannot be decoded, and has no secret information that can be removed, so B cannot learn $\boldsymbol{m}$.

B sets $\boldsymbol{b} = (b, 0, \ldots, 0)$ and generates $\boldsymbol{r_B} \xleftarrow{\$} \mathbb{R}$ and $(\boldsymbol{e_u}, \boldsymbol{e_v}) \xleftarrow{\$} \mathbb{R}^2$. B produces $\boldsymbol{u}' = a \cdot \boldsymbol{u} + \boldsymbol{h} \cdot \boldsymbol{r_B} + \boldsymbol{e_u}, \boldsymbol{v}' = a \cdot \boldsymbol{v} + \boldsymbol{b} \cdot \mathbb{G} + \boldsymbol{s} \cdot \boldsymbol{r_B} + \boldsymbol{e_v}$ and re-randomize $\boldsymbol{u}$ and $\boldsymbol{v}$ after updating. Since the error-correcting code is a linear code, $\boldsymbol{u'}$ and $\boldsymbol{v'}$ after update are

$$\boldsymbol{u}' = \begin{cases} \boldsymbol{h} \cdot \boldsymbol{r_B} + \boldsymbol{e_u} & \text{(In the case of a = 0)} \\ \boldsymbol{u} + \boldsymbol{h} \cdot \boldsymbol{r_B} + \boldsymbol{e_u} & \text{(In the case of a = 1).} \end{cases} \tag{3.7}$$

$$\boldsymbol{v}' = \begin{cases} \boldsymbol{b} \cdot \mathbb{G} + \boldsymbol{s} \cdot \boldsymbol{r_B} + \boldsymbol{e_v} & \text{(In the case of a = 0)} \\ \boldsymbol{v} + \boldsymbol{b} \cdot \mathbb{G} + \boldsymbol{s} \cdot \boldsymbol{r_B} + \boldsymbol{e_v} & \text{(In the case of a = 1).} \end{cases} \tag{3.8}$$

Finally, A uses his secret information to decrypt $\boldsymbol{v}' - \boldsymbol{u}' \cdot \boldsymbol{y}$. The result is

$$\begin{aligned} &\boldsymbol{v}' - \boldsymbol{u}' \cdot \boldsymbol{y} \\ =&(a\boldsymbol{m} + \boldsymbol{b})\mathbb{G} + \boldsymbol{x}(a\boldsymbol{r_A} + \boldsymbol{r_B}) - \boldsymbol{y}(a\boldsymbol{r_u} + \boldsymbol{e_u}) + (a\boldsymbol{r_v} + \boldsymbol{e_v}) \\ =&\begin{cases} \boldsymbol{b}\mathbb{G} + \boldsymbol{x}\boldsymbol{r_B} - \boldsymbol{y}\boldsymbol{e_u} + \boldsymbol{e_v} & \text{(in the case of a = 0)} \\ (\boldsymbol{m} + \boldsymbol{b})\mathbb{G} + \boldsymbol{x}(\boldsymbol{r_A} + \boldsymbol{r_B}) - \boldsymbol{y}(\boldsymbol{r_u} + \boldsymbol{e_u}) + (\boldsymbol{r_v} + \boldsymbol{e_v}) & \\ & \text{(in the case of a = 1).} \end{cases} \end{aligned} \tag{3.9}$$

As shown by the Eq. (3.9), the result of $\boldsymbol{v}' - \boldsymbol{u}' \cdot \boldsymbol{y}$ is the result of removing $\boldsymbol{h}$ and $\boldsymbol{s}$. Taking the first bit makes $a \cdot m + b$ available to A.

3.2.5.2 Correctness and Security of the Proposed Protocol

The correctness of the two-way linear function evaluation protocol proposed in this study obviously depends on the decoding ability of the code $\mathbb{C}$. Specifically, assuming that $\mathbb{C}$. Decode decodes $\boldsymbol{v} - \boldsymbol{u} \cdot \boldsymbol{y}$ correctly, the following equation is satisfied:

$$\text{Decrypt}(sk, \text{Encrypt}(pk, a \cdot \boldsymbol{m} + \boldsymbol{b})) = a \cdot \boldsymbol{m} + \boldsymbol{b}. \tag{3.10}$$

Also, let ϵ be the error of $\boldsymbol{v} - \boldsymbol{u} \cdot \boldsymbol{y}$. The error is

$$\epsilon = \begin{cases} \boldsymbol{x}\boldsymbol{r_B} - \boldsymbol{y}\boldsymbol{e_u} + \boldsymbol{e_v} & \text{(In the case of a = 0)} \\ \boldsymbol{x}(\boldsymbol{r_A} + \boldsymbol{r_B}) - \boldsymbol{y}(\boldsymbol{r_u} + \boldsymbol{e_u}) + (\boldsymbol{r_v} + \boldsymbol{e_v}) & \\ & \text{(In the case of a = 1)} \end{cases} \tag{3.11}$$

for the error correction capability of the code $\mathbb{C}$. In the paper of Gaborit et al., $\mathbb{C}$.Decode can work correctly when $\omega(\boldsymbol{x} \cdot \boldsymbol{r_2} + \boldsymbol{e} - \boldsymbol{y} \cdot \boldsymbol{r_1}) \leq \delta$ is satisfied, and w_r and w_e have the same value when actually evaluated. If the Hamming weight of $\boldsymbol{r_0}, \boldsymbol{r_1}, \boldsymbol{r_u}, \boldsymbol{r_v}, \boldsymbol{r_B}$ of the protocol proposed in this section is set to $1/2$ of w_r of Gaborit et al., then, the Hamming weight of $\boldsymbol{e_u}, \boldsymbol{e_v}$ is set to $1/2$ of w_e of Gaborit et al. The Hamming weight of the error Eq. (3.11) is less than or equal to the Hamming weight of errors in Gaborit et al.'s setting. Therefore, the conclusion of the paper of Gaborit et al. also holds for the proposed protocol. As the code length n increases, the decoding failure rate of the error-correcting code decreases. If the appropriate code space size n and noise Hamming weights w_r and w_e are set, the decoding failure rate approaches 0.

The security requirements of the proposed protocol are described above. In this section, we prove the security against semi-honest adversaries.

Theorem 3.2 *Under the quasi-cyclic syndrome decoding assumption, the 2PC protocol securely computes linear functions for semi-honest adversaries.*

Proof First, consider the semi-honest adversary A. With the global parameter omitted, the view of A is $\text{view}_A = (\boldsymbol{m}; \boldsymbol{h}, \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{r_0}, \boldsymbol{r_1}, \boldsymbol{r_u}, \boldsymbol{r_v}; \boldsymbol{u'}, \boldsymbol{v'})$. We construct a simulator $S_A(\boldsymbol{m}, \boldsymbol{x}, \boldsymbol{y})$ as follows:

1. Generate $\widetilde{\boldsymbol{h}}, \widetilde{\boldsymbol{r_0}}, \widetilde{\boldsymbol{r_A}}, \widetilde{\boldsymbol{r_u}}, \widetilde{\boldsymbol{r_v}}, \widetilde{\boldsymbol{u'}}, \widetilde{\boldsymbol{v'}} \xleftarrow{\$} \mathbb{R}$ randomly.
   Here, the Hamming weight of $\widetilde{\boldsymbol{r_A}}, \widetilde{\boldsymbol{r_u}}, \widetilde{\boldsymbol{r_v}}$ is w_r.
2. Output $(\boldsymbol{m}, \boldsymbol{x}, \boldsymbol{y}; \widetilde{\boldsymbol{h}}, \widetilde{\boldsymbol{r_A}}, \widetilde{\boldsymbol{r_u}}, \widetilde{\boldsymbol{r_v}}; \widetilde{\boldsymbol{u'}}, \widetilde{\boldsymbol{v'}})$.

Since, $\boldsymbol{h}, \boldsymbol{r_A}, \boldsymbol{r_u}, \boldsymbol{r_v}$ and $\widetilde{\boldsymbol{h}}, \widetilde{\boldsymbol{r_A}}, \widetilde{\boldsymbol{r_u}}, \widetilde{\boldsymbol{r_v}}$ follow the same distribution, the following equation holds:

$$\begin{aligned} &(\boldsymbol{m}, \boldsymbol{x}, \boldsymbol{y}; \widetilde{\boldsymbol{h}}, \widetilde{\boldsymbol{r_A}}, \widetilde{\boldsymbol{r_u}}, \widetilde{\boldsymbol{r_v}}; \widetilde{\boldsymbol{u'}}, \widetilde{\boldsymbol{v'}}) \\ \equiv_s\ &(\boldsymbol{m}, \boldsymbol{x}, \boldsymbol{y}; \boldsymbol{h}, \boldsymbol{r_A}, \boldsymbol{r_u}, \boldsymbol{r_v}; \widetilde{\boldsymbol{u'}}, \widetilde{\boldsymbol{v'}}). \end{aligned} \tag{3.12}$$

At view_A, $\boldsymbol{u'} = a \cdot \boldsymbol{u} + \boldsymbol{h} \cdot \boldsymbol{r_B} + \boldsymbol{e_u}$, $\boldsymbol{v'} = a \cdot \boldsymbol{v} + \boldsymbol{b} \cdot \mathbb{G} + \boldsymbol{s} \cdot \boldsymbol{r_B} + \boldsymbol{e_v}$, and it holds

$$\begin{bmatrix} \boldsymbol{h} \cdot \boldsymbol{r_B} + \boldsymbol{e_u} \\ \boldsymbol{s} \cdot \boldsymbol{r_B} + \boldsymbol{e_v} \end{bmatrix} = \begin{bmatrix} \boldsymbol{I}_n & 0 & \text{rot}(\boldsymbol{h}) \\ 0 & \boldsymbol{I}_n & \text{rot}(\boldsymbol{s}) \end{bmatrix} \begin{bmatrix} \boldsymbol{e_u} \\ \boldsymbol{e_v} \\ \boldsymbol{r_B} \end{bmatrix}. \tag{3.13}$$

Therefore, the adversary of probabilistic polynomial time cannot distinguish between $(\boldsymbol{h} \cdot \boldsymbol{r_B} + \boldsymbol{e_u}, \boldsymbol{s} \cdot \boldsymbol{r_B} + \boldsymbol{e_v})$ and uniform random numbers under the assumption of 3-quasi-cyclic syndrome decoding of quasi-cyclic code. Since $\boldsymbol{u}$ and $\boldsymbol{v}$ are also under the 3-quasicyclic syndrome decoding decision assumption, they cannot distinguish between $\boldsymbol{u}$ and $\boldsymbol{v}$ and uniform random numbers. Thus, the distribution of $\boldsymbol{u}'$ and $\boldsymbol{v}'$ also approaches uniform random numbers and satisfies the following equation:

$$\begin{aligned} &(\boldsymbol{m}, \boldsymbol{x}, \boldsymbol{y}; \boldsymbol{h}, \boldsymbol{r_A}, \boldsymbol{r_u}, \boldsymbol{r_v}, \widetilde{\boldsymbol{u}'}, \widetilde{\boldsymbol{v}'}) \\ \equiv_c\ &(\boldsymbol{m}, \boldsymbol{x}, \boldsymbol{y}; \boldsymbol{h}, \boldsymbol{r_A}, \boldsymbol{r_u}, \boldsymbol{r_v}, \boldsymbol{u}', \boldsymbol{v}'). \end{aligned} \tag{3.14}$$

Thus, the distributions of the view view$_A$ of A and the simulator S_A are indistinguishable against polynomial-time adversaries:

$$\begin{aligned} &S_A(\boldsymbol{m}, \boldsymbol{x}, \boldsymbol{y}) \\ \equiv_c\ &\mathrm{view}_A(\boldsymbol{m}, \boldsymbol{x}, \boldsymbol{y}; \boldsymbol{h}, \boldsymbol{r_A}, \boldsymbol{r_u}, \boldsymbol{r_v}; \boldsymbol{u}', \boldsymbol{v}'). \end{aligned} \tag{3.15}$$

Next, consider the semi-honest adversary B. With the global parameter omitted, the view of B is view$_B = (a, b; \boldsymbol{h}, \boldsymbol{s}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{r_B}, \boldsymbol{e_u}, \boldsymbol{e_v})$. Configure the simulator $S_B(a, b)$ as follows:

1. Randomly generate $\widetilde{\boldsymbol{h}}, \widetilde{\boldsymbol{s}}, \widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{v}}, \widetilde{\boldsymbol{r_B}}, \widetilde{\boldsymbol{e_u}}, \widetilde{\boldsymbol{e_v}} \xleftarrow{\$} \mathbb{R}$. Here, the Hamming weight of $\widetilde{\boldsymbol{r_B}}$ is w_r, and the Hamming weight of $\widetilde{\boldsymbol{e_u}}$ and $\widetilde{\boldsymbol{e_v}}$ is w_e
2. Output $(a, b; \widetilde{\boldsymbol{h}}, \widetilde{\boldsymbol{s}}, \widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{v}}, \widetilde{\boldsymbol{r_B}}, \widetilde{\boldsymbol{e_u}}, \widetilde{\boldsymbol{e_v}})$.

Since, $\boldsymbol{h}, \boldsymbol{r_B}, \boldsymbol{r_u}, \boldsymbol{r_v}$ and $\widetilde{\boldsymbol{h}}, \widetilde{\boldsymbol{r_B}}, \widetilde{\boldsymbol{r_u}}, \widetilde{\boldsymbol{r_v}}$ follow the same distribution, the following equation holds:

$$\begin{aligned} &(a, b; \widetilde{\boldsymbol{h}}, \widetilde{\boldsymbol{s}}, \widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{v}}, \widetilde{\boldsymbol{r_B}}, \widetilde{\boldsymbol{e_u}}, \widetilde{\boldsymbol{e_v}}) \\ \equiv_s\ &(a, b; \boldsymbol{h}, \widetilde{\boldsymbol{s}}, \widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{v}}, \boldsymbol{r_B}, \boldsymbol{e_u}, \boldsymbol{e_v}). \end{aligned} \tag{3.16}$$

Note that $\boldsymbol{s}$ can be reduced to 2-cyclic syndrome decoding decision, and the distribution cannot be distinguished from uniform random numbers for the adversary in polynomial time. Therefore, the following equation is satisfied.

$$\begin{aligned} &(a, b; \boldsymbol{h}, \widetilde{\boldsymbol{s}}, \widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{v}}, \boldsymbol{r_B}, \boldsymbol{e_u}, \boldsymbol{e_v}) \\ \equiv_c\ &(a, b; \boldsymbol{h}, \boldsymbol{s}, \widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{v}}, \boldsymbol{r_B}, \boldsymbol{e_u}, \boldsymbol{e_v}). \end{aligned} \tag{3.17}$$

Moreover, since $\boldsymbol{u}$ and $\boldsymbol{v}$ are indistinguishable between $(\boldsymbol{h} \cdot \boldsymbol{r_B} + \boldsymbol{e_u}, \boldsymbol{s} \cdot \boldsymbol{r_B} + \boldsymbol{e_v})$ and uniform random numbers based on the assumption of quasi-cyclic syndrome decoding and the adversary of probabilistic polynomial time cannot be distinguished, the following holds:

$$\begin{aligned} &(a, b; \boldsymbol{h}, \boldsymbol{s}, \widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{v}}, \boldsymbol{r_B}, \boldsymbol{e_u}, \boldsymbol{e_v}) \\ \equiv_c\ &(a, b; \boldsymbol{h}, \boldsymbol{s}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{r_B}, \boldsymbol{e_u}, \boldsymbol{e_v}). \end{aligned} \tag{3.18}$$

Therefore, the distributions of the view view$_B$ of B and the simulator S_B cannot be distinguished against the adversary of polynomial time:

$$\begin{aligned} &S_B(a, b) \\ \equiv_c\ &\mathrm{view}_B(a, b; \boldsymbol{h}, \boldsymbol{s}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{r_B}, \boldsymbol{e_u}, \boldsymbol{e_v}). \end{aligned} \tag{3.19}$$

□

The above protocol works over $\mathbb{F}_2$, but one can see that this can be easily extended to a larger field $\mathbb{F}_q$ by using appropriate error-correcting linear codes over $\mathbb{F}_q$.

3.2.5.3 Secure Comparison

Two-party secure comparison protocol proposed in this section is based on the size comparison method used in the secure decision tree classification protocol of Wu et al. [23]. In this section, we used the following criteria given in Proposition 3.1 for comparison.

Proposition 3.1 *For a t-bit x, y, if there is an $i \in [t]$ such that the following expression holds, then $x < y$.*

$$x_i - y_i + 1 + 3\sum_{j<i}(x_j \oplus y_j) = 0.$$

In this section, we introduce the proposed protocol for two-party secret comparison protocol. The proposed protocol for two-party secret comparison protocol uses a quasi-cyclic code and an arbitrary error-correcting code (For example, Reed-Solomon code) on $\mathbb{F}_q$. The participants in the protocol are Alice (A) and Bob (B). The input of A is $c \in \mathbb{N}$, and the input of B is $d \in \mathbb{N}$. The output of A is the result of the comparison between c and d, and the output of B is none.

The flow of two-party secret comparison is shown as follows:

Protocol Two-party secret comparison protocol

Input	A : $c \in \mathbb{N}$
	B : $d \in \mathbb{N}$
Output	A : Comparison result of c and d
	B : $\perp$

1. A and B perform binary expansion of c and d for each input so that $\boldsymbol{c} = c_1c_2\dots c_l$, $\boldsymbol{d} = d_1d_2\dots d_l$. Then, each bit c_i, d_i is padded to make $\boldsymbol{c_i}, \boldsymbol{d_i}, i \in [l]$ of k bits. In addition, they set the global parameter param = (n, k, δ, w_x, w_r) and the generator matrix $\mathbb{G} \in \mathbb{F}_q^{k\times n}$ of code $\mathbb{C}$.
2. A generates random $\boldsymbol{h} \xleftarrow{\$} \mathbb{R}$. Furthermore, $(\boldsymbol{x}, \boldsymbol{y} \xleftarrow{\$} \mathbb{R}^2)$ with Hamming weight w_x is generated. Private key $sk = (\boldsymbol{x}, \boldsymbol{y})$, and public key $pk = (\boldsymbol{h, s = x + h \cdot y})$.
3. A generates a random $\boldsymbol{r_{Ai}}, \boldsymbol{r_{ui}}, \boldsymbol{r_{vi}} \xleftarrow{\$} \mathbb{R}, i \in [l]$ with Hamming weight w_r. Then, A computes $\boldsymbol{u_i = h \cdot r_{Ai} + r_{ui}}$ and $\boldsymbol{v_i = c_i \cdot G + s \cdot r_{Ai} + r_{vi}}$ for l pairs and sends l pairs of ciphertext $\boldsymbol{u_i, v_i}$ to B.
4. B generates $(\boldsymbol{r_{Bi}}, \boldsymbol{e_{ui}}, \boldsymbol{e_{vi}}) \xleftarrow{\$} \mathbb{R}^3$ with Hamming weight w_r^* and computes the expression $c_i - d_i + 1 + 3\sum_{w<i}(c_w \oplus d_w)$ for c_i. Specifically, B substitutes plaintext d_i for $i \in [l]$ in the above formula and sets appropriate $a_{1i}, a_{2i}, \dots, a_{li}$, $\boldsymbol{b_i}$. B computes $\boldsymbol{u_i}' = a_{1i} \cdot \boldsymbol{u_1} + \cdots + \boldsymbol{h \cdot r_{Bi} + e_{ui}}$ and $\boldsymbol{v_i}' = a_{1i} \cdot \boldsymbol{v_1} + \cdots + \boldsymbol{b_i} \cdot \mathbb{G} + \boldsymbol{s \cdot r_{Bi} + e_{vi}}$ for l pairs. Then, the order of $(\boldsymbol{u_i}', \boldsymbol{v_i}')$ of l pairs is randomly replaced and sent to A in a random order.

5. A computes $\boldsymbol{v_i}' - \boldsymbol{u_i}' \cdot \boldsymbol{y}$ for each $i \in [l]$ and decrypts the result. If there is 0 in the first bit of the decoded results, $c < d$ is output. Conversely, if there is no 0, $c \geq d$ is output.

Protocol Description

1. In step 1, A and B expand c and d of each input to l-bit binary input, so that $\boldsymbol{c} = c_1 c_2 \dots c_l$ and $\boldsymbol{d} = d_1 d_2 \dots d_l$. Where $c_i, d_i, i \in [l]$ is the ith digit of $\boldsymbol{c}, \boldsymbol{d}$, and l is the bit length. To encode, pad each input to $\boldsymbol{c_i}, \boldsymbol{d_i}, i \in [l]$ with bit length k.
   In addition, set global parameters. n is the code length, k is the number of information bits, δ is the maximum number of errors that can be corrected by the error-correcting code, and w_x and w_r are the Hamming weights set in advance. The public parameter $\mathbb{G}$ is the generator matrix(For example, the Reed-Solomon code generator matrix) of the error-correcting code $\mathbb{C}$, which maps the message and code length as $\mathbb{F}_q^k \to \mathbb{F}_q^n$.
2. In step 2, A generates a private key and public key for HQC encryption scheme.
3. In step 3, A uses the public key and encrypts each of the $\boldsymbol{c_i}$ pieces. Send $(\boldsymbol{u_i}, \boldsymbol{v_i}), i \in [l]$ of the encrypted result to B.
4. Step 4 uses Proposition 3.1 for the evaluation of $c_i - d_i + 1 + 3\sum_{w<i}(c_w \oplus d_w)$. In other words, $c < d$ if $i \in [l]$ exists such that

$$c_i - d_i + 1 + 3\sum_{w<i}(c_w \oplus d_w) = 0. \tag{3.20}$$

   In particular, since B has plaintext d_i and encrypted c_i, Eq. (3.20) can be regarded as an equation with c_i as an unknown and can be computed. In addition, for XOR operations, B can transform $x_i \oplus y_i$ into

$$x_i \oplus y_i = \begin{cases} x_i & (y_i = 0) \\ 1 - x_i & (y_i = 1). \end{cases} \tag{3.21}$$

   Therefore, the XOR operation requires only the additive homomorphism of HQC encryption scheme.
   That is, B substitutes plaintext $d_i, i \in [l]$ into the above equation, sets the appropriate $a_{1i}, a_{2i}, \dots, a_{li}, \boldsymbol{b_i}$, and computes as follows:

$$\boldsymbol{u_i}' = a_{1i} \cdot \boldsymbol{u_1} + \dots + a_{li} \cdot \boldsymbol{u_l} + \boldsymbol{h} \cdot \boldsymbol{r_{Bi}} + \boldsymbol{e_{ui}}. \tag{3.22}$$

$$\boldsymbol{v_i}' = a_{1i} \cdot \boldsymbol{v_1} + \dots + a_{li} \cdot \boldsymbol{v_l} + \boldsymbol{b_i} \cdot \boldsymbol{G} + \boldsymbol{s} \cdot \boldsymbol{r_{Bi}} + \boldsymbol{e_{vi}}. \tag{3.23}$$

   Here, the Hamming weight of $\boldsymbol{r_{Bi}}, \boldsymbol{e_{ui}}, \boldsymbol{e_{vi}}, i \in [l]$ is w_r^*.
   Furthermore, to not leak the information about which bits are different to A, B needs to replace the order of each $(\boldsymbol{u_i}', \boldsymbol{v_i}')$ computed at random.

5. In step 5, A computes $\boldsymbol{v_i}' - \boldsymbol{u_i}' \cdot \boldsymbol{y}, i \in [l]$. The result is

$$\begin{aligned}
&\boldsymbol{v_i}' - \boldsymbol{u_i}' \cdot \boldsymbol{y} \\
&= (a_{1i} \cdot \boldsymbol{m_1} + \cdots + a_{li} \cdot \boldsymbol{m_l}) \cdot \mathbb{G} \\
&\quad + \boldsymbol{x} \cdot (a_{1i} \cdot \boldsymbol{r_{A1}} + \cdots + a_{li} \cdot \boldsymbol{r_{Al}} + \boldsymbol{r_{Bi}}) \\
&\quad - \boldsymbol{y} \cdot (a_{1i} \cdot \boldsymbol{r_{u1}} + \cdots + a_{li} \cdot \boldsymbol{r_{ul}} + \boldsymbol{e_{ui}}) \\
&\quad + (a_{1i} \cdot \boldsymbol{r_{v1}} + \cdots + a_{li} \cdot \boldsymbol{r_{vl}} + \boldsymbol{e_{vi}}).
\end{aligned} \tag{3.24}$$

Then, the evaluation result is decoded by the error-correcting code. A takes out the first 1 bit of each of l decoding results, and outputs $c < d$ if there is 0 in it. If there is no 0, $c \geq d$ is output.

3.2.5.4 Correctness and Security of the Proposed Protocol

Correctness

First, we explain step 4 w_r^*. The Hamming weight of the polynomial coefficient vector $\boldsymbol{x}, \boldsymbol{y}$ is w_x, and the Hamming weight of $\boldsymbol{r_{Ai}}, \boldsymbol{r_{ui}}, \boldsymbol{r_{vi}}, i \in [l]$ is w_r. Since each is selected uniformly and independently, the probability of each bit value of the vector is expressed as follows:

$$x_i = y_i = \begin{cases} 0 & \text{w.p. } 1 - p \\ 1 & \text{w.p. } p = \frac{w_x}{n}. \end{cases} \tag{3.25}$$

Similarly,

$$r_{Ai,j} = r_{ui,j} = r_{vi,j} = \begin{cases} 0 & \text{w.p. } 1 - p_r \\ 1 & \text{w.p. } p_r = \frac{w_r}{n}. \end{cases} \tag{3.26}$$

Let L be the set of $a_{1i}, a_{2i}, \ldots, a_{li} \neq 0$ in each $a_{1i} \cdot \boldsymbol{r_{A1}} + a_{2i} \cdot \boldsymbol{r_{A2}} + \cdots + a_{li} \cdot \boldsymbol{r_{Al}}$ for the expression $i \in [l]$.

$$L = \{a_{ki} | a_{ki} \neq 0\}$$

Let $|L|$ be the number of elements in set L. Set the Hamming weights w_r^* for $\boldsymbol{r_{Bi}}, \boldsymbol{e_{ui}}, \boldsymbol{e_{vi}}$ be as follows:

$$w_r^* = (n - |L| + 1) w_r.$$

Thus, the value of each w_r^* can be determined based on the nonzero numbers in a_i and $i \in [l]$.

Next, we analyze the validity of the proposed protocol.

The legitimacy of the proposed bilateral linear function secure computation protocol clearly depends on the decoding ability of $\mathbb{C}$. Set the $\boldsymbol{v'} - \boldsymbol{u'} \cdot \boldsymbol{y}$ error to ϵ. For the error correction capability of code $\mathbb{C}$, the error is

$$\begin{aligned} \epsilon = \; & \boldsymbol{x} \cdot (a_{1i} \cdot \boldsymbol{r_{A1}} + \cdots + a_{li} \cdot \boldsymbol{r_{Al}} + \boldsymbol{r_{Bi}}) \\ & - \boldsymbol{y} \cdot (a_{1i} \cdot \boldsymbol{r_{u1}} + \cdots + a_{li} \cdot \boldsymbol{r_{ul}} + \boldsymbol{e_{ui}}) \\ & + (a_{1i} \cdot \boldsymbol{r_{v1}} + \cdots + a_{li} \cdot \boldsymbol{r_{vl}} + \boldsymbol{e_{vi}}). \end{aligned} \tag{3.27}$$

In other words, if $\epsilon < \delta$, decoding is successful. Here, δ is the maximum number of errors that can be corrected by error-correcting code $\mathbb{C}$. In addition, in order to analyze the validity of the proposed protocol, we generalize the validity of the HQC encryption scheme proved by Gaborit et al. [20].

The following proposition holds for the Hamming weight of the error.

Proposition 3.2 *There are polynomial coefficient vectors* $\boldsymbol{x} = (X_1, \ldots, X_n)$ *and* $\boldsymbol{r} = (R_1, \ldots, R_n)$, *and* $\boldsymbol{y} = \boldsymbol{x} \cdot \boldsymbol{r} = (Y_1, \ldots, Y_n)$. *The probability that the sum of the random variables* $Y_i, i \in [n]$ *on* $\mathbb{F}_q$ *is 0 is*

$$\Pr[Y_1 + \cdots + Y_n = 0] = \frac{1}{q}\{1 + (1 - \frac{q}{q-1}p)^n \cdot (q-1)\}. \tag{3.28}$$

Where the probability distribution of the random variable Y_i *is*

$$Y_i = \begin{cases} 0 & \text{w.p. } p_0 = 1 - p \\ 1 & \text{w.p. } p_1 = \frac{p}{q-1} \\ 2 & \text{w.p. } p_1 = \frac{p}{q-1} \\ \vdots & \\ q-1 & \text{w.p. } p_1 = \frac{p}{q-1}. \end{cases} \tag{3.29}$$

Proof For Y_i, the following equation holds:

$$\begin{aligned} & \Pr[Y_1 + \cdots + Y_n = 0] \\ = & \sum_{\substack{i_0 + i_1 + \cdots + i_{q-1} = n \\ i_0 \cdot 0 + i_1 \cdot 1 + \cdots + i_{q-1} \cdot (q-1) = 0}} \left(\frac{n!}{i_0! \cdots i_{q-1}!} \right) p_0^{i_0} \cdots p_{q-1}^{i_{q-1}}, \end{aligned} \tag{3.30}$$

where $i_0, \ldots, i_{q-1}$ is the number of times the corresponding $0, \ldots, q-1$ appears. From the polynomial theorem, the following equation holds:

$$\begin{aligned}
&\{p_0+p_1+\ldots+p_{q-1}\}^n+\{p_0+(\omega_q)p_1+\cdots+(\omega_q^{q-1})p_{q-1}\}^n\\
&+\cdots+\{p_0+(\omega_q)^{q-1}p_1+\cdots+(\omega_q^{q-1})^{q-1}p_{q-1}\}^n\\
=&\sum_{i_0+\cdots+i_{q-1}=n}\left(\frac{n!}{i_0!\cdots i_{q-1}!}\right)p_0^{i_0}\cdots p_{q-1}^{i_{q-1}}\\
&\quad\{1+(\omega_q)^{i_1}(\omega_q^2)^{i_2}\cdots(\omega_q^{q-1})^{i_{q-1}}+\cdots\\
&\quad+(\omega_q)^{(q-1)i_1}(\omega_q^2)^{(q-1)i_2}\cdots(\omega_q^{q-1})^{(q-1)i_{q-1}}\}\\
=&\sum_{i_0+\cdots+i_{q-1}=n}\left(\frac{n!}{i_0!\cdots i_{q-1}!}\right)p_0^{i_0}\cdots p_{q-1}^{i_{q-1}}\\
&\quad\{1+\omega_q^{i_1+2i_2+\cdots+(q-1)i_{q-1}}+\cdots\\
&\quad+\omega_q^{(q-1)\{i_1+2i_2+\cdots+(q-1)i_{q-1}\}}\}.
\end{aligned}\tag{3.31}$$

Where ω_q is the q root of 1 and has the following properties:

$$1+\omega_q+\omega_q^2+\cdots+\omega_q^{q-1}=0\tag{3.32}$$

Substituting $i_0\cdot 0+i_1\cdot 1+\cdots+i_{q-1}\cdot(q-1)=0$ into Eq. 3.31 can be transformed as follows:

$$\begin{aligned}
&\{p_0+p_1+\cdots+p_{q-1}\}^n\\
&+\{p_0+(\omega_q)p_1+\cdots+(\omega_q^{q-1})p_{q-1}\}^n+\cdots\\
&+\{p_0+(\omega_q)^{q-1}p_1+\cdots+(\omega_q^{q-1})^{q-1}p_{q-1}\}^n\\
=&\sum_{\substack{i_0+\cdots+i_{q-1}=n\\ i_0\cdot 0+\cdots+i_{q-1}\cdot(q-1)=0}}\left(\frac{n!}{i_0!\cdots i_{q-1}!}\right)p_0^{i_0}\cdots p_{q-1}^{i_{q-1}}\cdot q.
\end{aligned}\tag{3.33}$$

Substituting Eq. (3.33) into Eq. (3.30), the proposition holds:

$$\begin{aligned}
&\Pr[Y_1+\cdots+Y_n=0]\\
=&\frac{1}{q}\{(p_0+p_1+\cdots+p_{q-1})^n+\cdots\\
&+(p_0+(\omega_q)^{q-1}p_1+\cdots+(\omega_q^{q-1})^{q-1}p_{q-1})^n\}\\
=&\frac{1}{q}\{1^n+(1-p+\frac{p}{q-1}(\omega_q+\omega_q^2+\cdots+\omega_q^{q-1}))^n\cdot(q-1)\}\\
=&\frac{1}{q}\left\{1+\left(1-\frac{q}{q-1}p\right)^n\cdot(q-1)\right\}.
\end{aligned}\tag{3.34}$$

□

In addition, the following analysis is the same as the validity analysis in Gaborit et al. [20]. According to the analysis result of [20], in the case of $\mathbb{F}_2$, the decoding failure rate can be controlled by setting an appropriate code space size n and noise Hamming weights w_x and w_r. Therefore, in the case of $\mathbb{F}_q$, it can be expected that the decoding failure rate can be controlled by setting the appropriate parameters.

Security

This section describes the security of the proposed secret comparison protocol.

First, consider semi-honest adversaries A and output$_\mathrm{A}$ = (c < d). Omitting global parameters, A's view is $\mathrm{view}_A = (c, \boldsymbol{x}, \boldsymbol{y}; \boldsymbol{h}, \{\boldsymbol{r_{Ai}}\}_{i=1}^{l}, \{\boldsymbol{r_{ui}}\}_{i=1}^{l}, \{\boldsymbol{r_{vi}}\}_{i=1}^{l}, \{\boldsymbol{u_i}'\}_{i=1}^{l}, \{\boldsymbol{v_i}'\}_{i=1}^{l})$. However, the first bit is 0 only for $\boldsymbol{u_{i*}}' - \boldsymbol{v_{i*}}' \cdot \boldsymbol{y}$ with index $i*$. The simulator $S_A(c, \boldsymbol{x}, \boldsymbol{y})$ is configured as follows:

1. Generates $\widetilde{\boldsymbol{h}}, \{\widetilde{\boldsymbol{r_{Ai}}}\}_{i=1}^{l}, \{\widetilde{\boldsymbol{r_{ui}}}\}_{i=1}^{l}, \{\widetilde{\boldsymbol{r_{vi}}}\}_{i=1}^{l}, \{\widetilde{\boldsymbol{u_i}'}\}_{i=1}^{l}, \{\widetilde{\boldsymbol{v_i}'}\}_{i=1}^{l} \xleftarrow{\$} \mathbb{R}$ at random. Here, the Hamming weight of $\{\widetilde{\boldsymbol{r_{Ai}}}\}_{i=1}^{l}, \{\widetilde{\boldsymbol{r_{ui}}}\}_{i=1}^{l}, \{\widetilde{\boldsymbol{r_{vi}}}\}_{i=1}^{l}$ is w_r. It also selects random $i* \in [l]$, the first bit of $\boldsymbol{u_{i*}}' - \boldsymbol{v_{i*}}' \cdot \boldsymbol{y}$ is 0, and the first bit of other $\{\widetilde{\boldsymbol{u_i}}' - \widetilde{\boldsymbol{v_i}}' \cdot \boldsymbol{y}\}_{i=1, i \neq i*}^{l}$ is non-zero.
2. This replaces $\{\widetilde{\boldsymbol{u_i}'}\}_{i=1}^{l}, \{\widetilde{\boldsymbol{v_i}'}\}_{i=1}^{l}$ at random to make $\{\widetilde{\boldsymbol{u_j}'}\}_{j=1}^{l}, \{\widetilde{\boldsymbol{v_j}'}\}_{j=1}^{l}$ in random order.
3. This outputs $(c, \boldsymbol{x}, \boldsymbol{y}; \widetilde{\boldsymbol{h}}, \{\widetilde{\boldsymbol{r_{Ai}}}\}_{i=1}^{l}, \{\widetilde{\boldsymbol{r_{ui}}}\}_{i=1}^{l}, \{\widetilde{\boldsymbol{r_{vi}}}\}_{i=1}^{l}, \{\widetilde{\boldsymbol{u_j}'}\}_{j=1}^{l}, \{\widetilde{\boldsymbol{v_j}'}\}_{j=1}^{l})$.

Since $\boldsymbol{h}, \{\boldsymbol{r_{Ai}}\}_{i=1}^{l}, \{\boldsymbol{r_{ui}}\}_{i=1}^{l}, \{\boldsymbol{r_{vi}}\}_{i=1}^{l}$ and $\widetilde{\boldsymbol{h}}, \{\widetilde{\boldsymbol{r_{Ai}}}\}_{i=1}^{l}, \{\widetilde{\boldsymbol{r_{ui}}}\}_{i=1}^{l}, \{\widetilde{\boldsymbol{r_{vi}}}\}_{i=1}^{l}$ follow the same distribution, the following equation holds:

$$\begin{aligned} &(\widetilde{\boldsymbol{h}}, \{\widetilde{\boldsymbol{r_{Ai}}}\}_{i=1}^{l}, \{\widetilde{\boldsymbol{r_{ui}}}\}_{i=1}^{l}, \{\widetilde{\boldsymbol{r_{vi}}}\}_{i=1}^{l}) \\ \equiv_s &(\boldsymbol{h}, \{\boldsymbol{r_{Ai}}\}_{i=1}^{l}, \{\boldsymbol{r_{ui}}\}_{i=1}^{l}, \{\boldsymbol{r_{vi}}\}_{i=1}^{l}). \end{aligned} \tag{3.35}$$

From the assumption of quasi-cyclic syndrome decoding of quasi-cyclic codes, the probabilistic polynomial time adversary cannot distinguish between $\boldsymbol{u_j}', \boldsymbol{v_j}', j \in [l]$ and uniformly random ones. Furthermore, since $\{\widetilde{\boldsymbol{u_i}'}\}_{i=1}^{l}$ and $\{\widetilde{\boldsymbol{v_i}'}\}_{i=1}^{l}$ are replaced randomly, the first bit is 0, and the index of $\widetilde{\boldsymbol{u_{i*}}} - \widetilde{\boldsymbol{v_{i*}}} \cdot \boldsymbol{y}$ where the index $i*$ is a uniformly random one satisfying the following expression:

$$\left(\{\widetilde{\boldsymbol{u_j}'}\}_{j=1}^{l}, \{\widetilde{\boldsymbol{v_j}'}\}_{j=1}^{l}\right) \equiv_c \left(\{\boldsymbol{u_i}'\}_{i=1}^{l} \{\boldsymbol{v_i}'\}_{i=1}^{l}\right). \tag{3.36}$$

Therefore, the distribution of the view view_A and simulator S_A when A is $\mathrm{output}_A = (c < d)$ is indistinguishable against polynomial time opponents.

Semi-honest adversary A and $\mathrm{output}_A = (c \geq d)$ are the same as the security proof in the case of $\mathrm{output}_A = (c < d)$, so details are omitted.

Next, we consider semi-honest adversary B. Omitting the global parameters, B's view is $\mathrm{view}_B = (d; \boldsymbol{h}, \boldsymbol{s}, \{\boldsymbol{u_i}\}_{i=1}^{l}, \{\boldsymbol{v_i}\}_{i=1}^{l}, \{\boldsymbol{r_{Bi}}\}_{i=1}^{l}, \{\boldsymbol{e_{ui}}\}_{i=1}^{l}, \{\boldsymbol{e_{vi}}\}_{i=1}^{l})$. Configure simulator $S_B(d)$ as follows:

1. Generates $\widetilde{\boldsymbol{h}}, \widetilde{\boldsymbol{s}}, \{\widetilde{\boldsymbol{u}}_i\}_{i=1}^{l}, \{\widetilde{\boldsymbol{v}}_i\}_{i=1}^{l}, \{\widetilde{\boldsymbol{r_{Bi}}}\}_{i=1}^{l}, \{\widetilde{\boldsymbol{e_{ui}}}\}_{i=1}^{l}, \{\widetilde{\boldsymbol{e_{vi}}}\}_{i=1}^{l} \xleftarrow{\$} \mathbb{R}$ at random. Here, the Hamming weight of $\{\widetilde{\boldsymbol{r_{Bi}}}\}_{i=1}^{l}, \{\widetilde{\boldsymbol{e_{ui}}}\}_{i=1}^{l}, \{\widetilde{\boldsymbol{e_{vi}}}\}_{i=1}^{l}$ is w_r^*.

2. This outputs $(\boldsymbol{d}; \tilde{\boldsymbol{h}}, \tilde{\boldsymbol{s}}, \{\widetilde{\boldsymbol{u}_i}\}_{i=1}^{l}, \{\widetilde{\boldsymbol{v}_i}\}_{i=1}^{l}, \{\widetilde{\boldsymbol{r}_{\boldsymbol{B}i}}\}_{i=1}^{l}, \{\widetilde{\boldsymbol{e}_{\boldsymbol{u}i}}\}_{i=1}^{l}, \{\widetilde{\boldsymbol{e}_{\boldsymbol{v}i}}\}_{i=1}^{l})$.

Since $\boldsymbol{h}$, $\{\boldsymbol{r}_{\boldsymbol{B}i}\}_{i=1}^{l}$, $\{\boldsymbol{e}_{\boldsymbol{u}i}\}_{i=1}^{l}$, $\{\boldsymbol{e}_{\boldsymbol{v}i}\}_{i=1}^{l}$ and $\tilde{\boldsymbol{h}}$, $\{\widetilde{\boldsymbol{r}_{\boldsymbol{B}i}}\}_{i=1}^{l}$, $\{\widetilde{\boldsymbol{e}_{\boldsymbol{u}i}}\}_{i=1}^{l}$, $\{\widetilde{\boldsymbol{e}_{\boldsymbol{v}i}}\}_{i=1}^{l}$ follow the same distribution, the following equation holds:

$$\begin{aligned}&(\boldsymbol{h}, \{\boldsymbol{r}_{\boldsymbol{B}i}\}_{i=1}^{l}, \{\boldsymbol{e}_{\boldsymbol{u}i}\}_{i=1}^{l}, \{\boldsymbol{e}_{\boldsymbol{v}i}\}_{i=1}^{l})\\ \equiv_s &(\tilde{\boldsymbol{h}}, \{\widetilde{\boldsymbol{r}_{\boldsymbol{B}i}}\}_{i=1}^{l}, \{\widetilde{\boldsymbol{e}_{\boldsymbol{u}i}}\}_{i=1}^{l}, \{\widetilde{\boldsymbol{e}_{\boldsymbol{v}i}}\}_{i=1}^{l}).\end{aligned} \tag{3.37}$$

$\boldsymbol{s}$ can be reduced to a 2-quasi-cyclic syndrome decoding decision assumption, and the distribution is indistinguishable from uniform random numbers for probabilistic polynomial-time adversaries. Thus, $\tilde{\boldsymbol{s}} \equiv_c \boldsymbol{s}$ holds.

In addition, since $\boldsymbol{u}_i, \boldsymbol{v}_i, i \in [l]$ are based on the assumption of quasi-cyclic syndrome decoding, an adversary in probabilistic polynomial time cannot distinguish between $\boldsymbol{u}_i, \boldsymbol{v}_i, i \in [l]$ and uniform random numbers.

$$(\{\widetilde{\boldsymbol{u}_i}\}_{i=1}^{l}, \{\widetilde{\boldsymbol{v}_i}\}_{i=1}^{l}) \equiv_c (\{\boldsymbol{u}_i\}_{i=1}^{l}, \{\boldsymbol{v}_i\}_{i=1}^{l}). \tag{3.38}$$

Therefore, the distribution of B's view view_B and simulator S_B is indistinguishable against polynomial time adversaries.

3.2.6 Support Vector Machine from Secure Linear Function Evaluation and Secure Comparison

We can construct a code-based protocol for a support vector machine from the protocols for evaluation of linear functions and comparison described above. Note that the result of secure evaluation of linear function is in $\mathbb{F}_q$ while that of secure composition is a bit string. Therefore, we need to provide secure bit-decomposition protocol. The bit-decomposition protocols have been already studied well in the research area of secure computation, and indeed, we can use the bit-decomposition protocol given in [24] with secure computation protocol from a threshold homomorphic encryption [25]. (It is straightforward to construct a threshold version of HQC scheme by setting $sk_A = (\boldsymbol{x}_1, \boldsymbol{y}_1)$ and $sk_B = (\boldsymbol{x}_2, \boldsymbol{y}_2)$ as distributed decryption keys for A and B. Then, the encryption key is $(\boldsymbol{h}, (\boldsymbol{x}_1 + \boldsymbol{x}_2) + \boldsymbol{h} \cdot (\boldsymbol{y}_1 + \boldsymbol{y}_2))$.

We describe the overview of the protocol below. For simplification, we denote $[m]$ as the ciphertext for m under HQC encryption scheme over $\mathbb{F}_q$.

Protocol

Input	A : $m \in \mathbb{F}_q$
	B : $a, b, t \in \mathbb{F}_q$
Output	A : $a \cdot m + b > t$ or not
	B : $\perp$

1. A and B perform the secure linear evaluation protocol over $\mathbb{F}_q$. Then, B sends A $[a \cdot m + b]$ at step 4 in the original protocol.
2. A and B start the secure bit-decomposition protocol on $[a \cdot m + b]$.
3. From the result of the bit-decomposition protocol, B obtains the binary representation $[(a \cdot m + b)_1], \ldots, [(a \cdot m + b)_\ell]$.
4. A and B perform the secure comparison protocol from step 4.

References

1. L. Kissner, D. Song, Privacy-preserving set operations, in *CRYPTO 2005*. LNCS, vol. 3621 (Springer, Berlin, 2005), pp. 241–257
2. Y. Sang, H. Shen, Efficient and secure protocols for privacy-preserving set operations. ACM Trans. Inf. Syst. Secur. **13**(1), 9:1–9:35 (2009)
3. R. Egert, M. Fischlin, D. Gens, S. Jacob, M. Senker, J. Tillmanns, Privately computing set-union and set-intersection cardinality via bloom filters, in *ACISP 2015*. LNCS, vol. 9144 (Springer, Berlin, 2015), pp. 413–430
4. D. Many, M. Burkhart, X. Dimitropoulos, Fast private set operations with sepia. Technical Report, 345 (2012)
5. O. Goldreich, Secure multi-party computation. Manuscript, Preliminary version (1998)
6. B.H. Bloom, Space/time trade-offs in hash coding with allowable errors. Commun. ACM **13**(7), 422–426 (1970)
7. A. Broder, M. Mitzenmacher, Network applications of bloom filters: a survey. Internet Math. **1**(4), 485–509 (2004)
8. P. Paillier, Public-key cryptosystems based on composite degree residuosity classes, in *EUROCRYPT 1999*. LNCS, vol. 1592 (Springer, Berlin, 1999), pp. 223–238
9. R. Cramer, R. Gennaro, B. Schoenmakers, A secure and optimally efficient multi-authority election scheme. Eur. Trans. Telecommun. **8**(5), 481–490 (1997)
10. Y. Desmedt, Y. Frankel, Threshold cryptosystems, in *CRYPTO 1989*. *LNCS*, vol. 1462 (Springer, Berlin, 1989), pp. 307–315
11. M.J. Freedman, K. Nissim, B. Pinkas, Efficient private matching and set intersection, in *EUROCRYPT 2004*. *LNCS*, vol. 3027 (Springer, Berlin, 2004), pp. 1–19
12. Y. Azar, A.Z. Broder, A.R. Karlin, E. Upfal, Balanced allocations. SIAM J. Comput. **29**(1), 180–200 (1999)
13. E. De Cristofaro, G. Tsudik, Practical private set intersection protocols with linear complexity, in *FC 2010*. *LNCS*, vol. 6052 (Springer, Berlin, 2010), pp. 143–159
14. E. De Cristofaro, J. Kim, G. Tsudik, Linear-complexity private set intersection protocols secure in malicious model, in *ASIACRYPT 2010*. LNCS, vol. 6477 (Springer, Berlin, 2010), pp. 213–231
15. F. Kerschbaum, Outsourced private set intersection using homomorphic encryption, in *ACMCCS 2012* (ACM, 2012), pp. 85–86
16. S. Goldwasser, S. Micali, Probabilistic encryption. J. Comput. Syst. Sci. **28**(2), 270–299 (1984)
17. Y. Ishai, J. Kilian, K. Nissim, E. Petrank, Extending oblivious transfers efficiently, in *CRYPTO 2003*. *LNCS*, vol. 2729 (Springer, Berlin, 2003), pp. 145–161
18. M.O. Rabin, How to exchange secrets with oblivious transfer. Technical Memo, TR-81 (1981)
19. C. Dong, L. Chen, Z. Wen, When private set intersection meets big data: an efficient and scalable protocol, in *ACMCCS 2013* (ACM, 2013), pp. 789–800
20. C. Aguilar, O. Blazy, J.-C. Deneuville, P. Gaborit, G. Zémor, Efficient encryption from random quasi-cyclic codes. IEEE Trans. Inf. Theory **64**(5), 3927–3943 (2018)

21. National Institute of Standards and Technology. Post-quantum cryptography, round 2 submissions (2019), https://csrc.nist.gov/projects/post-quantum-cryptography/round-2-submissions
22. A.C.-C. Yao, How to generate and exchange secrets, in *Proceedings of the 27th Annual IEEE Symposium on Foundations of Computer Science* (1986), pp. 162–167
23. D.J. Wu, T. Feng, M. Naehrig, K. Lauter, Privately evaluating decision trees and random forests, in *Proceeding on Privacy Enhancing Technologies*, vol. 4 (2016), pp. 1–21
24. I. Dangaard, M. Fitzi, E. Kiltz, J.B. Nielsen, T. Toft, Unconditionally secure constant-rounds multi-party computation for equality, comparison, bits and exponentiation, in *TCC2006: Theory of Cryptography* (2006), pp. 285–304
25. R. Cramer, I. Damgaard, J.B. Nielsen, Multiparty computation from threshold encryption, in *Eurocrypt* (2001), pp. 280–299

Chapter 4
Secure Data Management Technology

Tomoaki Mimoto, Shinsaku Kiyomoto, and Atsuko Miyaji

Abstract In this chapter, we introduce data anonymization techniques for several types of datasets. Data anonymity of anonymized datasets is an index for estimating the (maximum) reidentification risk from anonymized datasets and is generally defined as a quantitative index based on adversary models. The adversary models are implicitly defined according to the attributes in the datasets, use cases, and anonymization techniques. We first review existing anonymization techniques and the adversary models behind the data anonymity definitions for anonymization techniques; then, we propose a common anonymity definition and its adversary model, which is applicable to several types of anonymization techniques. Furthermore, some extensions of the definition, which is optimized for specific types of datasets, are presented in the chapter.

4.1 Introduction

Secure data management is a key issue in personal data distribution and analysis. Anonymization techniques have been used to harmonize the utility of data and their privacy risks. These techniques transform personal data into anonymized data to reduce the success probability of reidentification of data principals from the data. If the data are well anonymized, they cannot be connected to a person; thus, the privacy of the person is protected by anonymization techniques.

Secure computation is sometimes a realistic solution for commercial services due to its cost for data of very large size. Some anonymization techniques work

T. Mimoto (✉) · S. Kiyomoto
KDDI Research, Inc., 2-1-15 Ohara, 356-8502 Fujimino-shi, Saitama, Japan
e-mail: to-mimoto@kddi-research.jp

S. Kiyomoto
e-mail: kiyomoto@kddi-research.jp

A. Miyaji
Osaka University, Suita, Japan
e-mail: miyaji@comm.eng.osaka-u.ac.jp

A. Miyaji and T. Mimoto (eds.), *Security Infrastructure Technology for Integrated Utilization of Big Data*,
https://doi.org/10.1007/978-981-15-3654-0_4

on commercial services as a “practical” solution, even though the size of the data is very large. Thus, anonymization techniques have been applied for personal data distribution and data analysis. For example, k-anonymization was first proposed as a practical solution to reduce the reidentification risks of public data; since then, it has been considered to be able to be used for the secure management of personal data.

Quantitative measures for anonymity are required for estimating privacy risks and assessing the feasibility of privacy requirements. In several studies on anonymization, privacy notions providing quantitative measures for anonymity have been defined for each anonymization technique; however, no common notion for all anonymization techniques has been presented to date, which means that each privacy notion is not universal but is localized, and heuristic approaches are still used to harmonize the usability of data and privacy risks through whole processes or services. A common notion is required for consistent secure data management for the whole process.

In this chapter,[1] we discuss a new common privacy notion based on an adversary model, which is applicable to several anonymization techniques, and introduce a novel anonymization technique and implementation of the technique. In Sect. 4.2, we revisit adversary models on several anonymization techniques and review anonymization techniques. We propose a common adversary model and quantitative measures using the adversary model are presented in Sect. 4.3. An extension is discussed in Sect. 4.4. Our implementation of an anonymization tool is introduced in Sect. 4.5. We conclude this chapter in Sect. 4.6.

4.2 Anonymization Techniques and Adversary Models, Revisited

The related work presented below is grouped under k-anonymization and noise addition as anonymization methods.

4.2.1 k-Anonymization

k-anonymity [4–6] is a well-known privacy model. The property of k-anonymity is that each published record is such that every combination of values of quasi-identifiers can be matched to at least k respondents.

[1] This chapter is reprinted from [1–3].

4.2.1.1 Adversary Model

k-anonymized datasets are assumed to be in public domains. An adversary can obtain all the attribute values in a dataset and execute arbitrary operations on the attribute values.

There are few formal definitions or models for the adversary that aim to identify the attributes of a certain individual in a k-anonymized dataset. Kiyomoto and Martin modeled an adversary [7] for k-anonymized datasets based on two query functions as follows:

Let d be an index of the dth record, q_x be a set of m attribute values in T^{q*}, and s be a value for the sensitive attribute. The two query functions are defined as:

- **read.** For the input of an index value d, the function outputs the dth record. That is, $f(T^*, query = \{\textbf{read}, d\}) \rightarrow \{d, q_x^d, s^d\}$, where q_x^d and s^d are values of the quasi-identifier and the sensitive attribute in the dth record, respectively. If the dth record does not exist, then the function outputs $failed$.
- **search.** For input q_x and/or s, the function outputs the number u of records and index values that have a quasi-identifier q_x and/or sensitive attribute s. That is, $f(T^*, query = \{\textbf{search}, q_x, s\}) \rightarrow u, D$, where u and D are the number of records and a sequence of index values that have the same quasi-identifier and/or sensitive attribute, respectively. If s or q_x do not exist, then the function outputs $failed$.

4.2.1.2 k-Anonymization Algorithm

This idea is easy to understand, and many types of k-anonymization algorithms have been proposed. The Incognito algorithm [8] generalizes the attributes using taxonomy trees, and the Mondrian algorithm [9] averages or replaces the original data with representative values and achieves k-anonymization. In this paper, we use a k-anonymization algorithm based on clustering and denote $A_k(D)$ as k-anonymization for dataset D. The algorithm finds close records and creates clusters such that each partition contains at least k records. For details of the algorithm, see [10].

4.2.2 Noise Addition

Noise addition works by adding or multiplying stochastic or randomized numbers to confidential data [11]. The idea is simple and is also well known to be an anonymization technique.

4.2.2.1 Adversary Model

One objective of an adversary against noise-added datasets is to remove the noise or estimate the original values from the noise-added attribute values. One potential scenario is a probabilistic approach in which an adversary estimates the distribution of noise and chooses an attribute value with high probability. There is no formal adversary model on static noise-added datasets, but *Differential Privacy* settings assume data include dynamically added noise, and their adversary simulations are defined as query-based.

4.2.2.2 Anonymization Algorithm by Noise Addition

The first work on noise addition was proposed by Kim [12], and the idea was to add noise ϵ with a distribution $\epsilon \sim N(0, \sigma^2)$ to the original data. Additive noise is uncorrelated noise and preserves the mean and covariance of the original data, but the correlation coefficients and variance are not retained. Another variation of additive noise is correlated additive noise, which keeps the mean and allows the correlation coefficients in the original data to be retained [13]. Differential privacy is a state-of-the-art privacy model that is based on the statistical distance between two database tables differing by at most one record. The basic idea is that, regardless of background knowledge, an adversary with access to the dataset draws the same conclusions, irrespective of whether a person's data are included in the dataset. Differential privacy is mainly studied in relation to perturbation methods in an interactive setting, although it is applicable to certain generalization methods.

In this paper, we use Laplace noise as a noise addition and add noise $\epsilon \sim Lap(0, 2\phi^2)$ to each attribute. We denote $A_\phi(D)$ as noise addition for dataset D.

4.2.3 K-Anonymization for Combined Datasets

We introduce an adversary model for a combined dataset from datasets produced by two service providers and anonymization methods [14].

4.2.3.1 Adversary Model

If we consider the existing adversary model and assume that the anonymization tables produced by the service providers satisfy k-anonymity, the combined table also satisfies k-anonymity. However, we have to consider another type of adversary in our new service model. In our service model, the combined table includes many sensitive attributes; thus, the adversary can distinguish a data owner using background knowledge of combinations of sensitive attribute values of the data owner. If the adversary finds a combination of known sensitive attributes on only one record, the adversary can obtain information; the record is a data owner that the adversary knows, and the adversary also knows the remaining sensitive attributes of the data owner. We model the above type of new adversary as follows:

π-knowledge Adversary Model. An adversary knows certain π sensitive attributes $\{s_1^i, ..., s_j^i, ..., s_\pi^i\}$ of a victim i. Thus, the adversary can distinguish the victim with an anonymization table in which only one record has any combinations (maximum π-tuple) of the attributes $\{s_1^i, ..., s_j^i, ..., s_\pi^i\}$.

4.2.3.2 Modification of Quasi-identifiers

The first strategy is to modify the quasi-identifiers of the combined table. The data user generates a merged table from two anonymization tables as follows: First, the data user simply merges the records in the two tables as $|q_C^g|s_{AB}^h|s_A^i|s_B^j|$. Then, the data user modifies q_C^q to satisfy the following condition, where θ is the total number of sensitive attributes in the merged table.

4.2.3.3 Modification of Sensitive Attributes

The second approach is to modify the sensitive attributes in the combined table for the condition. If a subtable $|s_{AB}^h|s_A^i|s_B^j|$ that consists of sensitive attributes is required to satisfy k-anonymity, some sensitive attribute values are removed from the table and are changed to $*$ to satisfy k-anonymity. Note that we do not accept that all sensitive attributes are $*$ due to having no information record.

4.2.3.4 Algorithm for Modification

One algorithm that finds a k-anonymized combined dataset is executed as follows:

1. The algorithm generalizes quasi-identifiers to satisfy the condition that each group of the same quasi-identifiers has at least $\pi \times k$ records.
2. The algorithm generates all the tuples of π sensitive attributes in the table.
3. For each tuple, the algorithm finds all the records that have the same sensitive attributes as the tuple or has $*$ for sensitive attributes and makes them a group. We define the number of sensitive attributes in the group which is θ. The algorithm generates a partial table that consists of $\theta - \pi$ sensitive attributes and checks whether the partial table has at least k different combinations of sensitive attributes.
4. If the partial table does not satisfy the above condition, the algorithm chooses a record from other groups that have different tuples of π sensitive attributes and changes the π sensitive attributes to $*$. The algorithm executes this step until the partial table has up to π different combinations of sensitive attributes.

5. The algorithm executes step 3 and step 4 for all the tuples of π sensitive attributes in the table.

4.2.4 Matrix Factorization for Time-Sequence Data

Some studies have used matrices for time-sequence datasets. Zheng et al. [15, 16] proposed predicting a user's interests in an unvisited location. They assumed users' GPS trajectory as a user-location matrix where each value of the matrix indicates the number of visits of a user to a location. The matrix is very sparse because each user visits only a handful of locations, so a collaborative filtering model is applied to the prediction. Zheng et al. [17] built a location-activity matrix, M, which has missing values. M is decomposed into the two low-rank matrices U and V. The missing values can be filled by $X = UV^{\mathsf{T}} \simeq M$, and locations can be recommended when some activities are given. Chawla et al. [18] constructed a graph from the trajectories of taxis and transformed the graph into matrices. The authors of [19] proposed a method of identifying traffic flows that cause an anomaly between two regions.

4.2.5 Anonymization Techniques for User History Graphs

In this subsection, we introduce two anonymization techniques for user history graphs, which are proposed in [1].

4.2.5.1 Adversary Model

Privacy leakage from a merged history graph is the disclosure of the actions of a particular person from the graph. Attacks against user history graphs are intended to obtain the private information of a particular user from the graph. We assume that the merging process is executed on a trusted domain and that only the merged history graph is published; thus, the adversary can only obtain the merged graph. Furthermore, we assume that the adversary has the following knowledge about the user: The history of the user is included in the merged graph and the user performs an action t. The adversary tries to discover other actions of the user to be able to guess which edges connecting to node t can be assigned to the user.

We summarize the adversary model as follows:

Adversary against a Merged History Graph. It is assumed that an adversary knows that a victim A executed an action t. The objective of the adversary is to obtain the actions that A executed before or after the action t. Thus, the adversary searches the merged history graph, which includes actions of other people and finds the actions of A using the knowledge that action t was executed.

We define privacy notions to use with the above adversary model in a later subsection.

4.2.5.2 Notions for the Untraceability of a Graph

We consider two levels of privacy notions: partial k-untraceability and complete k-untraceability. Partial k-untraceability accepts the leakage of some partial actions of a user but prevents all the actions of the user from being revealed. The definition of complete k-untraceability involves meeting the requirement that no action of the user is leaked. The symbol $Act^{A}_{N_{x\to y}}$ for user A denotes the sequence of all the actions of user A from action x to action y. For example, the sequence of actions from the first action to action x and the sequence of actions from action x to the final action are denoted as $Act^{A}_{N_{start\to x}}$ and $Act^{A}_{N_{x\to end}}$, respectively.

Definition 4.1 (*Partial k-untraceability*) We assume that an adversary knows an action t of a user A, and we consider all the possible adversaries defined for any action t of the user in the merged graph. If at least k sequences of actions are potentially associated with user A and $k-1$, other users exist as candidates for all actions $Act^{A}_{N_{start\to t}}$ and $Act^{A}_{N_{t\to end}}$, the digraph satisfies k-untraceability for A. If the digraph satisfies the above condition for all users, then the digraph is said to satisfy partial k-untraceability.

Definition 4.2 (*Complete k-untraceability*) We assume that an adversary knows an action t of a user A and we consider all the possible adversaries defined for any action t of the user in the merged graph. If at least k actions are potentially associated with user A and $k-1$ other users exist as candidates for each action in $Act^{A}_{N_{start\to t}}$ and $Act^{A}_{N_{t\to end}}$, the digraph satisfies k-untraceability for A. If the digraph satisfies the above condition for all users, the digraph satisfies complete k-untraceability.

Generally, many trivial actions are performed by many users. It is not important for privacy purposes where we keep the information about such actions. Thus, we relax the above definitions to produce an anonymized graph that includes much of the information needed to analyze a user's history. Let v be the threshold value for the number of performing users that establishes that an action is trivial; that is, we judge the actions $x \to y$ to be trivial if the label $L(x \to y) \geq v$. Both definitions are modified as follows:

Definition 4.3 (*Partial (k, v)-untraceability*) We assume that an adversary knows an action t of a user A, and we consider all the possible adversaries defined for any t in the merged graph. If at least k sequences of actions are potentially associated with user A and $k-1$ other users exist as candidates for all actions $Act^{A}_{N_{start\to t}}$ and $Act^{A}_{N_{t\to end}}$ except trivial actions $x \to y$ that have a label $L(x \to y) \geq v$, then the digraph satisfies partial (k, v)-untraceability for A. If the digraph satisfies the above condition for all users, then the digraph satisfies partial (k, v)-untraceability.

Definition 4.4 (*Complete (k, v)-untraceability*) We assume that an adversary knows an action t of a user A, and we consider all the possible adversaries defined for any t in the merged graph. If at least k actions are potentially associated with user A and $k - 1$ other users exist as candidates for each action in $Act^A_{N_{start \to t}}$ and $Act^A_{N_{t \to end}}$ except trivial actions $x \to y$ that have a label $L(x \to y) \geq v$, then the digraph satisfies complete (k, v)-untraceability for A. If the digraph satisfies the above condition for all users, then the digraph satisfies complete (k, v)-untraceability.

In a complete (k, v)-untraceable graph, each action t except trivial actions has k outgoing edges and incoming edges; thus, an action of user A that connects to action t cannot be identified from k candidates. Thus, the graph satisfies untraceability for an adversary who knows action t of the user. It is trivial that a complete (k, v)-untraceable graph satisfies partial (k, v)-untraceability; all actions except trivial actions are connected to k potential actions in a complete (k, v)-untraceable graph. A graph that satisfies partial (k, v)-untraceability generally produces much more information than a complete (k, v)-untraceable graph, where the partial (k, v)-untraceable graph and the complete (k, v)-untraceable graph are generated from a user history graph. However, the (k, v)-untraceable graph may reveal partial actions of users due to the relaxed definition of the privacy notion; an attack is successful when an adversary obtains all the actions of a user. To trace all the actions of the user, the adversary has to select a sequence of actions from k sequences of actions; thus, all the actions of the user are untraceable, even though some actions are traceable by the adversary. The parameter k means that an action (or a sequence of actions) is potentially associated with a user and $k - 1$ other users in the untraceable graph, and the parameter v means that v users perform the same action in the graph. Generally, we should select the parameter $v = k$ with regard to the privacy requirement for a merged graph. The actions of a user are hidden in the actions of a group that consists of k members including the user. A privacy notion for the graph should be selected from the above two notions according to a use case of the graph and its privacy requirements.

4.2.5.3 Algorithm Generating a Partial (k, V)-Untraceable History Graph

The details of the algorithm are denoted as **Algorithm 4.1**, where oe_t and ie_t are defined as the number of outgoing edges and incoming edges of a node t, respectively. The algorithm for generating a partial (k, v)-untraceable history graph is as follows:

1. This step consists of a part of the detailed algorithm, from line 1 to line 3. For the input of a user history graph **G**, the algorithm adds a virtual incoming edge $(s_r \to r)$ to each node $r \in start$ until the number of incoming edges is the same as the number of outgoing edges. Then, the algorithm adds a virtual outgoing edge $(q \to u_q)$ to each node $q \in end$ until the number of outgoing edges is the same as the number of incoming edges. A label of a virtual incoming edge $L(s_x \to x)$ denotes the number of users who first perform the action, and a label of a virtual

outgoing edge $L(y \to u_y)$ denotes the number of users who perform the action at the end.

2. This step consists of a part of the detailed algorithm, from line 4 to line 12. The algorithm searches for a node t that has fewer outgoing edges than k and for which all its lower nodes $\mathcal{N}_{t\to end\backslash t}$ have fewer outgoing edges than k. Then, the algorithm removes all the outgoing edges $(t \to *)$ that satisfy $L(t \to *) < v$. Next, the algorithm searches for a node t' that receives incoming edges numbering less than k and all upper nodes $\mathcal{N}_{start\to t'\backslash t'}$ that receive fewer incoming edges than k. Then, the algorithm removes all the incoming edges $(* \to t')$ that satisfy $L(* \to t') < v$. The algorithm repeats this step until no node that meets the conditions is found.
3. This step is the same as line 13, line 14 and line 15 in the detailed algorithm. The algorithm removes virtual incoming and outgoing edges, removes nodes that have no edges, and outputs the modified graph.

Algorithm 4.1 Generation of a Partial (k, v)-Untraceable History Graph

```
Input: User History Graph G, parameters $k$ and $v$
Output: Anonymized Graph $G^{\alpha}(G, k, v)$
1: $G^{\alpha}(G, k, v) \leftarrow G$
2: Add virtual incoming edges to start nodes
3: Add virtual outgoing edges to end nodes.
4: $T \leftarrow$ all nodes $t$, where $oe_{\mathcal{N}_{t\to end}} < k$ and all of its edges do not have $L(t_i \to *) \geq v$
5: $T' \leftarrow$ all nodes $t'$, where $ie_{\mathcal{N}_{start\to t'}} < k$ and all of its edges do not have $L(* \to t'_j) \geq v$
6: while $T \neq \emptyset$ or $T' \neq \emptyset$ do
7:     Choose $t_i$ from $T$
8:     Remove all outgoing edges of $t_i$ where $L(t_i \to *) < v$ from $G^{\alpha}(G, k, v)$
9:     Choose $t'_j$ from $T'$
10:    Remove all incoming edges of $t'_j$ where $L(* \to t'_j) < v$ from $G^{\alpha}(G, k, v)$
11:    Update $T$ and $T'$
12: end while
13: Remove virtual edges
14: Remove all nodes $t''$ where $oe_{t''} = 0$ and $ie_{t''} = 0$ from $G^{\alpha}(G, k, v)$
15: return $G^{\alpha}(G, k, v)$
```

4.2.5.4 Algorithm Generating a Complete (k, V)-Untraceable History Graph

The details of the algorithm are denoted as **Algorithm 4.2**. The algorithm for generating a complete (k, v)-untraceable history graph is as follows:

1. The algorithm first executes **Algorithm 4.1** except line 13 and line 15.
2. This step consists of a part of the detailed algorithm, from line 3 to line 11. The algorithm searches for a node t that has fewer outgoing edges than k and removes

all the outgoing edges ($t \rightarrow *$) that satisfy $L(t \rightarrow *) < v$, until no node is found. Then, the algorithm searches for a node t' that receives fewer incoming edges than k and removes all the edges ($* \rightarrow t'$) that satisfy $L(* \rightarrow t') < v$. The algorithm repeats this step until no node that meets the conditions is found.

3. This step consists of line 12, line 13, and line 14 in the detailed program. The algorithm removes virtual edges, removes nodes to which no edge is connected, and outputs the modified graph.

4.2.6 Other Notions

Differential Privacy [20, 21] is a notion of privacy for perturbative methods based on the statistical distance between two database tables differing by, at most, one element. The basic idea is that, regardless of background knowledge, an adversary with access to the dataset draws the same conclusions whether a person's data are included in the dataset. That is, a person's data have an insignificant effect on the processing of a query. Differential privacy is mainly studied in relation to perturbation methods [22–24] in an interactive setting. Attempts to apply differential privacy to search queries have been discussed in [25]. Li et al. proposed a matrix mechanism [26] applicable to predicate counting queries under a differential privacy setting. Computational relaxations of differential privacy were discussed in [27–29]. Another approach for quantifying privacy leakage is an information-theoretic definition proposed by Clarkson and Schneider [30]. They modeled an anonymizer as a program that receives two inputs: a user's query and a database response to the query. The program acted as a noisy communication channel and produced an anonymized response as the output. Hsu et al. provides a generalized notion [31] in decision theory for making a model of the value of personal information. An alternative model for the quantification of personal information is proposed in [32]. In the model, the value of personal information is estimated by the expected cost that the user has to pay for obtaining perfect knowledge from given privacy information. Furthermore, the sensitivity of different attribute values is taken into account in the average benefit and cost models proposed by Chiang et al. [33]. Krause and Horvitz presented utility-privacy tradeoffs in online services [34, 35].

4.2.7 Combination of Anonymization Techniques

A combination of anonymization methods leads to the construction of datasets that are useful and that preserve privacy. Some countries publish census data, and they combine several anonymization methods, such as generalization, noise addition, and sampling [36, 37]. However, some problems remain. One problem is that it is difficult to evaluate the privacy risks of anonymized datasets when anonymization methods are combined. Some research is available about the relationships among anonymization methods. Chaudhuri et al. proposed (c, ϵ, δ)-privacy [38] and studied the relationship among sampling and differential privacy [39]. Li et al. proposed $(\beta, \epsilon, \delta)$-differential privacy and studied the relationship among sampling, differential privacy, and k-anonymity. Soria-Comas et al. proposed a k-anonymized algorithm for differential privacy using an insensitive algorithm [40].

4.3 (p, N)-Identifiability

4.3.1 Common Adversary Model

Existing privacy measures are supposed to protect against idealized attackers, and it is difficult to maintain their utility and assess their reidentification risk. We designed adversary models to describe more realistic attackers by structuring a real setting for the attackers. In the case of exchanging anonymized datasets between companies, for instance, a data-providing company first anonymizes and encrypts datasets for transmission to a receiver company via a secure channel. The receiver company locates the dataset in a secure room and allows only authorized employees to access the anonymized dataset. This process can reduce the reidentification risk in the anonymized dataset, and it specifies the attacker and limits the ability to access datasets so that the attacker must know the quasi-identifiers of the neighbors or acquaintances. For example, it seems to be quite rare for an attacker to know all the quasi-identifiers of a target because the target is a neighbor of the attacker. Thus, a more stringent analysis of the reidentification risk can be achieved when we assume a more realistic situation, such as that the attacker has only limited knowledge of the victim.

Access rights to an anonymized dataset may be given to attackers, and attackers may acquire some information about the original dataset or obtain the anonymization algorithm used to generate the anonymized dataset. Information about the original dataset is categorized into three parts as follows: information on a specified record such as a neighbor; the original dataset; and any other information except the target information that the attacker is seeking. The case of William Weld, who was governor of Massachusetts [41], is a typical example of reidentification, and an attack on the Netflix Prize dataset was carried out by a strong attacker who gained access to the Internet Movie Database [42].

We can consider the abilities of an attacker in two areas: knowledge about the dataset and the ability to simulate anonymization algorithms. Many previous studies such as [43, 44] assumed that an attacker has all the information required except knowledge of the target of the attack. In this paper, we consider an attacker who has knowledge of only the target record and can simulate anonymization algorithms to obtain anonymized records that may correspond to the target record.

4.3.1.1 Definitions of Actual Attackers

Generally, when an anonymized dataset is published on the Web, anyone who can access the dataset is a potential attacker; thus, the adversary model should be ideal because we cannot assume there is only a limited-knowledge adversary, and we have to assume all possible adversaries are present. On the other hand, when the dataset is managed under strict controls, the model adversary is not considered to be an unlimited-knowledge adversary. We design two realistic adversary models under the assumption that the dataset is managed in a restricted area (not public) and only a limited set of attackers can access the dataset; and then, we propose a privacy metric for privacy risk analysis.

Definition 4.5 (*Anonymization Simulator f_{sim}*) Let D_0 with n_0 records, D_1 with n_1 records, $r_i^x[QI]$, and $r_i^x[SI]$ be an original dataset, an anonymized dataset generated from the original dataset, the quasi-identifiers of a record $r_i^x \in D_x$, and sensitive information from the record $r_i^x \in D_x$, respectively. An anonymization simulator f_{sim} simulates an anonymization algorithm used to generate an anonymized dataset as an oracle and outputs $r_i^1[QI] \in D_1$ for the input $r_i^0[QI] \in D_0$. That is, $f_{sim} : r_j^0[QI] \to \{\mathbf{r}^1[QI], \perp\}$, where $\mathbf{r}^1[QI]$ is a set of $r_i^1[QI]$ and no output is produced in the case of $\perp$.

The simulator is a deterministic process for deterministic anonymization, such as top-coding and bottom-coding, and a probabilistic process for probabilistic anonymization, such as random sampling. The simulator can provide access to D_0 to simulate the anonymization algorithm, even though no adversary can access D_0. Next, we define two adversary models.

Definition 4.6 (*Deanonymizer for Anonymized Datasets, $\mathcal{DA}$*) When $\exists_1 r_j^0[QI] \in D_0$, $\forall r_i^1[QI||SI] \in D_1$ and f_{sim} are given, a deanonymizer $\mathcal{DA}$ lines up potential candidates r_i^1 corresponding to r_j^0 by executing the simulator f_{sim}; then, the deanonymizer $\mathcal{DA}$ outputs a list of candidates $r_i^1[QI||SI]$ for r_j^0, where the number of records in the list is n_q, the number of sensitive information items in the list is n_s and $0 \leq n_s \leq n_q \leq n_0$.

If an attacker knows the actual anonymization function f, the attacker can use f as f_{sim}, and the evaluation result should be more credible.

Definition 4.7 (*Reidentifying Adversary versus Anonymized Datasets*) When $\exists_1 r_j^0[QI] \in D_0$, $\forall r_i^1[QI||SI] \in D_1$ and f_{sim} are given, a reidentifying adversary executes the deanonymizer $\mathcal{DA}$ and can identify r_i^1, which is a record of the same person in the record r_j^0, from the records in a dataset D_0, where $r_j^0 \in D_0$ is given. The success probability of the attack is calculated as $1/n_q$ when r_j^1 is included in the output by $\mathcal{DA}$; otherwise, it is 0.

Assuming an attacker who has $\exists_1 r_j^0[QI] \in D_0$ is the same as assuming $|D_0|$ attackers who have $r_j^0 (j = 1, ..., |D_0|) \in D_0$.

Definition 4.8 (*Revealing Adversary versus Anonymized Datasets*) When $\exists_1 r_j^0[QI] \in D_0, \forall r_i^1[QI||SI] \in D_1$ and f_{sim} are given, a revealing adversary executes the deanonymizer $\mathcal{DA}$ and finds a $r_j^0[SI]$ from $r_i^1[SI]$ such that r_i^1 is a record of the same person as the record r_j^0. The success probability of the attack is calculated as $1/n_s$ when r_j^1 is included in the output of $\mathcal{DA}$; otherwise, it is zero.

A revealing adversary does not try to identify the record but tries to access sensitive information. In other words, the attacker seeks only to obtain sensitive information from the record in question. More precisely, the success probability of the *revealing adversary* can be calculated as $[n_s]/n_q$, where the correct number of sensitive items in the list is $[n_s]$, but the probability itself may be uncertain. Assume that when the probability is 0.99, some attackers are convinced that the target should be the majority. Furthermore, in the case that the deanonymizer $\mathcal{DA}$ is leaked and the f_{sim} used in the deanonymizer is a deterministic process, an attacker can infer the sensitive information of r_j^0. On the other hand, when the f_{sim} used in the deanonymizer is a probabilistic process, even if $\mathcal{DA}$ is leaked, outputting the result should not involve uncertainty.

4.3.1.2 (p, N)-Identifiability

Here, we assume that anonymized datasets are strictly controlled and that the attacker has knowledge of a specific record and the anonymization algorithms. We assume that the attacker is the strongest type of attacker and has knowledge of the most characteristic record. Nevertheless, it is difficult to quantify this characteristic, so we assume that each attacker has an original record. In other words, we assume there are as many attackers as there are original records.

Definition 4.9 (*(p, N)-identifiability*) Let p be the success probability for an adversary who has $\exists_1 r^0[QI] \in D_0, \forall r_i^1[QI||SI] \in D_1$ and f_{sim}, and N be the number of adversaries whose attack success probability is p.

The probability p is the conditional probability that the adversary can select the correct record from the list produced by the deanonymizer $\mathcal{DA}$ when the collected record is included in the list. The probability that the deanonymizer successfully produces the list, including the correct record, depends on the anonymization algorithms.

Our model can extend to an adversary who has knowledge of two or more records. For simplicity, we use an adversary model that knows a single record and consider N single knowledge adversaries in our risk analysis. The idea of (p, N)-identifiability is studied in [2].

4.3.2 Success Probability Analysis Based on the Common Adversary Model

In this section, we assume the attackers described in the previous section and explain the calculation to obtain the success probability of attacks on representative anonymization methods: generalization, noise addition, and sampling. We consider that f_{sim} is constructed as a typical combined algorithm selected from three anonymization algorithms, $f_{generalization}$, $f_{sampling}$ and f_{noise}. We explain the above three anonymization algorithms and show combined anonymization using an example dataset.

4.3.2.1 Generalization

We include deletion of records or cells and top- or bottom-coding as steps in generalization. One step of $f_{generalization}$ is similar to k-anonymity in checking the number of identical combinations of quasi-identifiers. When an anonymized dataset has k-anonymity, p equals $1/k$. k-anonymity is an intuitive privacy metric, but the greater the number of attributes, the more difficult it is for the datasets to achieve k-anonymity. If an attacker has generalization trees for each attribute, the attacker adds records which satisfy the requirements of the trees of the list of candidates. When there is a record whose address attribute is Tokyo, for instance, an attacker who has the generalization tree adds records whose addresses are in the Kanto region as well as records whose addresses are in Eastern Japan to the list of candidates. It is appropriate that an attacker can infer the generalization tree and in our experiment, f_{sim} can be considered capable of accessing the generalization trees of each attribute.

4.3.2.2 Random Sampling

When an attacker who has one original record is assumed, the privacy risk differs greatly among the original datasets. Consider an original dataset with many unique records, and assume that random sampling is implemented. Let M be the number of unique records and α be the sampling rate. The probability that unique records will not appear is $(1-\alpha)^M$. Even when $\alpha = 0.1$ and $M = 44$, the probability is less than 0.1%. When a large dataset is anonymized, it is possible that there will be more than 44 unique records, which shows that if sampling is implemented, a characteristic record may be identified or suspected.

We evaluate sampling as follows: For simplicity, we consider the case where the anonymization method is only random sampling. When a unique record is sampled, an attacker who knows the person is certain that the record is for that person. Thus, the probability p does not change. On the other hand, sampling reduces the number of unique records, and N decreases accordingly. When unique records are very few and do not appear in an anonymized dataset, p decreases. We apply this approach to the case of combining different anonymization methods.

The approaches to sampling vary, and we can also consider $f_{sampling}$ in various ways. For instance, the probability of disclosing the identity of any individual is evaluated by using the posterior probability of population uniqueness [45].

4.3.2.3 Noise Addition

There are two cases of noise addition: One is adding noise to the numerical data itself, and the other is adding noise to its quantity. In the former case, the data consist of original numerical data or data anonymized by a process, such as microaggregation, and in the latter case, the data are original quantity data or anonymized data, such as 11–20 in the age attribute.

In the former case, we can consider f_{noise} as follows. Noise is added based on a probability distribution, such as normal, Laplace, and exponential distributions. In particular, it has been mathematically proven that adding Laplace noise to the output of some queries achieves differential privacy [39], so this type of noise is widely used. Therefore, when an anonymized record is included in the 90 or 95% confidence interval, the record is added to the list of candidates. More simply, when original data and anonymized data have small differences such as 10 or 20% for each attribute, the attacker may consider the possibility that they are the same.

In the latter case, we cannot use the same method. When a record has 72 and is anonymized to 95, for instance, the attacker whose target is a specific person may not regard the target to be that person. However, the attacker can link them after the top-coding is executed and change the value to 70-. On the other hand, when a record is 19, is anonymized to 20 and is generalized to 20–29, the attacker may not link them. One of the ideas of f_{noise} is that a group with each attribute can be changed to next group and such records are output as candidates. As in the generalization step, an attacker can infer the next group for each group and f_{noise} can be thought of as defining the distance of each classification.

The description above shows that when the order of anonymization is changed, f_{sim} will also be changed.

4.3.2.4 Combination of Anonymization Methods

The principles of each anonymization can be combined by evaluating each anonymization step by step. Stated differently, an attacker has $f_{generalization}$, $f_{sampling}$, and f_{noise} as f_{sim}. We show examples of combined cases by using a sample dataset (Fig. 4.1). An attacker should change his or her approach when the order of anonymization is changed if he or she knows this fact. We assume five attacker models, A_1 to A_5, in the following example, and the candidates of each attacker model are represented as C_1 to C_5. We denote C_i of r_j in the following figures as the candidates of an attacker A_i who has r_j as a target. The adversary model for A_1 to

Fig. 4.1 Sample dataset

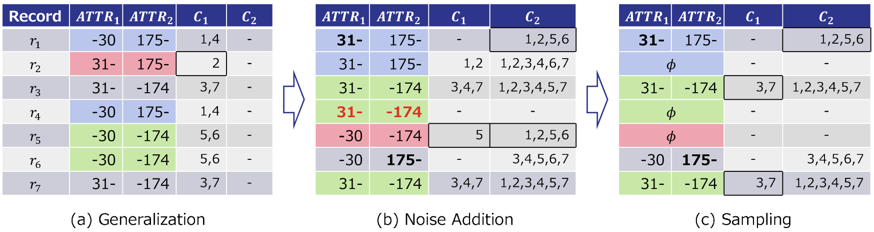
record	$ATTR_1$	$ATTR_2$	$ATTR_S$
r_1	28	178	Hospital
r_2	31	179	Office
r_3	38	165	Office
r_4	30	180	Shop
r_5	27	167	Hospital
r_6	29	171	Shop
r_7	33	173	Hospital

A_4 is the *reidentifying adversary* defined in Definition 4.3, and the adversary model in Fig. 4.4 is the *revealing adversary* defined in Definition 4.4.

Let the conditions of attackers be as follows: A_1 and A_3 do not consider noise-adding and generalization but simply compare $r_i^1 \in D_1$ with $r_j^0 \in D_0$. This is one approach to f_{noise} and $f_{generalization}$. On the other hand, A_2, A_4, and A_5 do consider the added noise and generalization. We define the noise addition shown in Fig. 4.2 as follows: the classifications of each attribute change to the next classification with a certain probability. We assume A_2 knows the rule of noise addition and that f_{noise} of A_2 outputs candidates that have a different classification in one attribute from an original record. On the other hand, let a small amount of noise be added in step (a) of Figs. 4.3 and 4.4. We assume the attackers A_4 and A_5 know the rule and that f_{noise} of A_4 and A_5 outputs candidates whose values of $ATTR_1$ are different but within 2 from the original record and whose values of $ATTR_2$ are different but within 4 from the original record. In the figures, the boldface sections show that the classifications are not correct but are within the permissible range for f_{noise} of A_2, A_4, and A_5: The red boldface sections show that there are substantial distances from the original values and that attackers who have the record cannot link them.

4.3.2.5 Examples of Analyses

The Case of A_1



Record	$ATTR_1$	$ATTR_2$	C_1	C_2
r_1	-30	175-	1,4	-
r_2	31-	175-	2	-
r_3	31-	-174	3,7	-
r_4	-30	175-	1,4	-
r_5	-30	-174	5,6	-
r_6	-30	-174	5,6	-
r_7	31-	-174	3,7	-

(a) Generalization

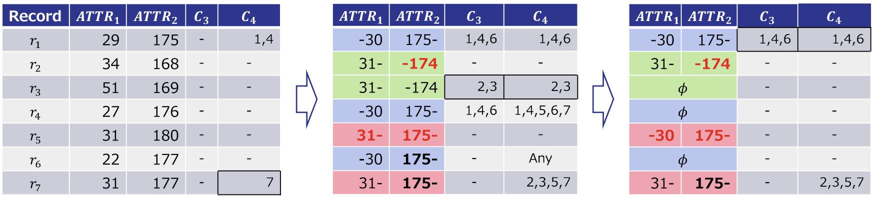
$ATTR_1$	$ATTR_2$	C_1	C_2
31-	175-	-	1,2,5,6
31-	175-	1,2	1,2,3,4,6,7
31-	-174	3,4,7	1,2,3,4,5,7
31-	**-174**	-	-
-30	-174	5	1,2,5,6
-30	**175-**	-	3,4,5,6,7
31-	-174	3,4,7	1,2,3,4,5,7

(b) Noise Addition

$ATTR_1$	$ATTR_2$	C_1	C_2
31-	175-	-	1,2,5,6
ϕ		-	-
31-	-174	3,7	1,2,3,4,5,7
ϕ		-	-
ϕ		-	-
-30	**175-**	-	3,4,5,6,7
31-	-174	3,7	1,2,3,4,5,7

(c) Sampling

Fig. 4.2 Sample anonymization and the result of simulation attack 1



Record	$ATTR_1$	$ATTR_2$	C_3	C_4
r_1	29	175	-	1,4
r_2	34	168	-	-
r_3	51	169	-	-
r_4	27	176	-	-
r_5	31	180	-	-
r_6	22	177	-	-
r_7	31	177	-	7

(a) Noise Addition

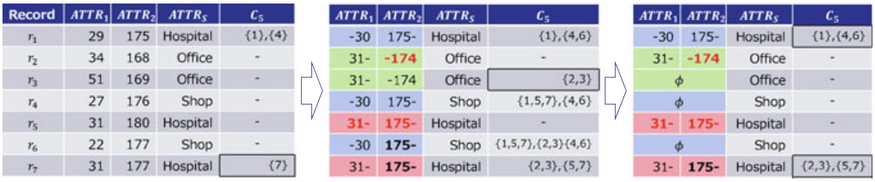
$ATTR_1$	$ATTR_2$	C_3	C_4
-30	175-	1,4,6	1,4,6
31-	-174	-	-
31-	-174	2,3	2,3
-30	175-	1,4,6	1,4,5,6,7
31-	175-	-	-
-30	175-	-	Any
31-	175-	-	2,3,5,7

(b) Generalization

$ATTR_1$	$ATTR_2$	C_3	C_4
-30	175-	1,4,6	1,4,6
31-	-174	-	-
ϕ		-	-
ϕ		-	-
-30	175-	-	-
ϕ		-	-
31-	175-	-	2,3,5,7

(c) Sampling

Fig. 4.3 Sample anonymization and the result of simulation attack 2



Record	$ATTR_1$	$ATTR_2$	$ATTR_S$	C_5
r_1	29	175	Hospital	{1},{4}
r_2	34	168	Office	-
r_3	51	169	Office	-
r_4	27	176	Shop	-
r_5	31	180	Hospital	-
r_6	22	177	Shop	-
r_7	31	177	Hospital	{7}

(a) Noise Addition

$ATTR_1$	$ATTR_2$	$ATTR_S$	C_5
-30	175-	Hospital	{1},{4,6}
31-	-174	Office	-
31-	-174	Office	{2,3}
-30	175-	Shop	{1,5,7},{4,6}
31-	175-	Hospital	-
-30	175-	Shop	{1,5,7},{2,3}{4,6}
31-	175-	Hospital	{2,3},{5,7}

(b) Generalization

$ATTR_1$	$ATTR_2$	$ATTR_S$	C_5
-30	175-	Hospital	{1},{4,6}
31-	-174	Office	-
ϕ		Office	-
ϕ		Shop	-
31-	175-	Hospital	-
ϕ		Shop	-
31-	175-	Hospital	{2,3},{5,7}

(c) sampling

Fig. 4.4 Sample anonymization and the result of simulation attack 3

Generalization, noise addition, and sampling are executed as anonymizing methods in Fig. 4.2. In the generalization step (a), all records are generalized to be divisible into equal parts. As a result, only r_2 is unique, and this dataset has (1, 1)-identifiability.

In step (b), r_1, r_4, and r_6 are changed by the addition of noise. As a result, r_1 and r_2 are indistinguishable. r_3, r_4, and r_7 are also indistinguishable, but r_5 and r_6 become unique. We define A_1 as not considering the addition of noise, so that an attacker who has r_6 cannot link the original record but an attacker who has r_5 can. Therefore, identifiability becomes (1, 1)-identifiability.

After sampling, in step (c), r_2, r_4, and r_5 do not appear. Then, r_3 and r_7 become the focus are focused and identifiability becomes (1/2, 2)-identifiability. This attacker simply checks how many of the same records there are in the dataset. Even if various anonymization methods are implemented, some records may not be affected. Therefore, it is important to assume such attackers. When we can say that a dataset has a certain level of privacy from such attackers, it means that an attacker cannot link the target with the original record by accident.

The Case of A_2

We omit the explanation of step (a) because noise is not added. In step (b), the attacker with r_1, for example, chooses r_1, r_2, r_5, and r_6 as candidates because one or more of their attributes match r_1 = {-30, 175-}. On the other hand, an attacker with r_4 cannot output candidates because both attributes of r_4 are changed. Hence, identifiability is (1/4, 2)-identifiability. In step (c), r_5 does not appear, and identifiability becomes (1/4, 1)-identifiability.

The Case of A_3

In Fig. 4.3, the dataset is anonymized by the addition of noise, generalization, and sampling.

In the case of A_3, the dataset with added noise is safe enough from attackers who do not consider the added noise and we omit this case; however, this does not mean that noise addition is safe, and when another attacker, such as A_4, is considered, the result should be different. In step (b), we focus on the attacker with r_3. This is the strongest attacker, and this attacker suspects that r_2 and r_3 are the candidates. More specifically, the scope is $r_3 = \{38, 165\} = \{31\text{-}, \text{-}174\}$ and r_2, r_3 meet the requirement. The attacker with r_2 seems to have the same risk but cannot identify the actual target r_2 is a possible candidate because the noise of $ATTR_2$ is great enough. Hence, the identifiability becomes $(1/2, 1)$-identifiability. In step (c), r_3 does not appear, and the privacy risk is $(1/3, 1)$-identifiability.

The Case of A_4

Next, we show the case of A_4. In step (a), every record but r_1 and r_7 has enough added noise, and attackers cannot infer which is the correct record. The attacker with r_7 regards the records within $\{33 \pm 2, 173 \pm 4\}$ as candidates. Only r_7 satisfies the condition, and the privacy risk is $(1, 1)$-identifiability. In step (b), the effect of noise addition becomes weak, and the number of attackers who should be considered increases. The attacker with r_6, for instance, regards the records within $\{29 \pm 2, 171 \pm 4\} = \{(\text{-}30, 31\text{-}), (\text{-}174, 175\text{-})\}$, namely, all records, as candidates. The privacy risk becomes $(1/2, 1)$-identifiability after generalization is finished. In step (c), similar to the previous steps, the privacy risk becomes $(1/3, 1)$-identifiability.

The Case of A_5

Finally, we show an example of a *revealing adversary*.

An attacker can claim to succeed when the sensitive information $ATTR_S$ of the target can be correctly identified. Step (a) is similar to that of the case of A_4. In step (b), the attacker with r_3 suspects r_2 and r_3 are the candidates. Their $ATTR_S$ are, however, "Office" and the attacker claims to identify the person. Thus, the privacy risk is $(2/2 = 1, 1)$-identifiability, which is similar to l-diversity. In step (c), the attacker with r_1 suspects r_1, r_4 and r_6 are the candidates; the $ATTR_S$ of r_1 is "Hospital," and that of the others is "Shop." Therefore, the probability of reidentification is 1/2. More precisely, the probability is 1/3 because there are three candidates and one is correct, but the probability may be important information for the attacker with r_1. The same can be said of the attacker with r_7; therefore, the risk according to our definition is $(1/2, 2)$-identifiability.

As described above, when the adversary model is different, the result of the risk is also different. Assuming attackers who disregard noise, we consider the risk to the records whose fluctuations are due to anonymization to be small. On the other hand, assuming attackers who do consider the actual added noise, we consider the risk to the dataset as a whole. Moreover, strong attackers can be assumed to use the inverse function of the actual noise or anonymization method. In the case that noise based on a normal distribution is added, for instance, an optimal distance-based record linkage can be performed [46].

It is important to consider the various types of attackers in this way, because the most important factor of privacy is the inability to definitely link an anonymized record X' and original record X. Our metrics ensure that the attackers considered can neither identify a record nor make an identification by chance, by considering many attackers.

4.3.2.6 Implementation of the Analysis Algorithm

Processing time is a problem when our metric is applied to a large dataset. In this section, we discuss this problem.

First, we have to evaluate the risk from attackers with each record, and when sampling is implemented, the candidates in each record need to be preserved across the sampling. However, we do not need to store the candidates for every record or the records that have certain risks because the metric does not consider attackers who have knowledge of a record that does not have the highest risk. Moreover, when anonymization and evaluation are performed repeatedly, it takes a long time to evaluate the risk because the same number of attackers as the number of records are assumed. Thus, a threshold risk can be introduced to resolve the problem. When the risk of an attack does not exceed the threshold, attackers do not need to be evaluated. It is possible, however, that the risk may increase depending on the situation (see r_5, r_6 in Fig. 4.2). Therefore, when a threshold is introduced, the accuracy of the privacy risk may worsen. We describe the pseudocode of risk analysis as follows:

Algorithm 4.5 (D_0, D_1, A, f_{sim}) : Risk analysis.

```
Input: Original dataset $D_0$, Anonymized dataset $D_1$, Adversary model $A$, and attack simulator $f_{sim}$
while $\forall r_i^0 \in D_0$ do
    $p_i \leftarrow$ simulation attack($r_i^0, D_1, A, f_{sim}$)
end while
$p \leftarrow \max(p_i)$
$N \leftarrow$ count(max($p_i$))
return $p, N$
```

Second, the attackers do not have to compare their records with every record because the method of evaluation is similar to that of k-anonymity, and the attackers only need to compare a representative of each group. The attackers need to compare their records with {-30, 175-}, {31-,-174}, and {31-, 175-} in (b) of Fig. 4.3, for instance. However, when the levels of generalization are different, such methods cannot be applied, and every record should be checked. To solve the problem, we first count the number of values of each attribute and then compare each attribute of r_j^0 with that of each record of D_1 in accordance with the large number of varieties.

Finally, when the procedure for anonymization is known in advance, it is possible to perform the evaluation more quickly by considering the effect of the initial part of the anonymization. For instance, in Fig. 4.3a, we only have to consider cells whose values do not exceed 30 in $ATTR_1$ or fall short of 174 in $ATTR_2$.

4.3.3 *Experiment*

4.3.3.1 Experimental Environments

We conducted experiments to evaluate the validity of the proposed metrics. We measured the time to output the risk and confirmed that the privacy metric was appropriate. We used three parameters, k, β, ϵ, for comparison and verified the relationships among k-anonymity, sampling, and noise addition. We implemented our risk analysis method on a PC with an Intel Core i7-4790 3.6-GHz CPU and a 16.0-GB memory.

4.3.3.2 Dataset and Adversary Model

We used a pseudomedical dataset based on an actual medical dataset. The dataset had 10,000 records and two attributes, total cholesterol (TC) and HbA1c, and the

Fig. 4.5 Distribution of TC

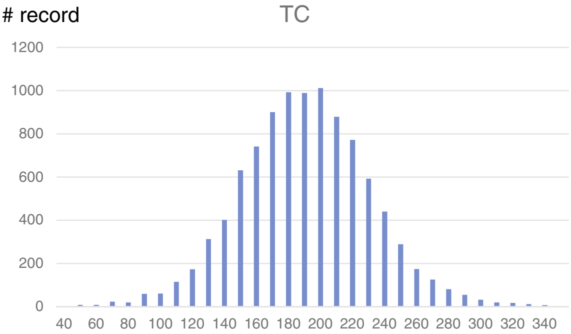
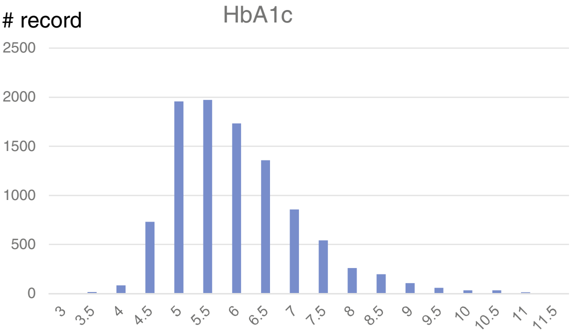


Fig. 4.6 Distribution of HbA1c



distribution of each attribute is shown in Figs. 4.5 and 4.6. We first measured the computation time while changing the number of records and then evaluated the validity of our metrics while changing the parameters of each anonymization method. Noise addition, generalization, and sampling were used as representative anonymization methods, and we adopted the Mondrian algorithm [9] for k-anonymization, Laplace noise for noise addition, and random sampling for sampling. We assumed *reidentifying adversary* A_1 to A_4. The conditions of the attacker models are the same as those of Sect. 4.3.2.4 except for noise addition. We define the f_{noise} of the A_2 and A_4 output records, whose value for each attribute differed by 5% from the original value, to be candidates.

4.3.4 Results

4.3.4.1 Computational Complexity

Our proposed privacy metrics are intended to be able to applied to large datasets. We measured the execution time by changing the number of records (Table 4.1) and parameters (Table 4.2, 4.3 and 4.4).

It takes little time to evaluate the risk when simple attackers, such as A_1 and A_3, are considered. On the other hand, when reflective attackers are assumed, the number of calculations increases and more time is required for evaluation. However, some of the processing described above reduces the time. For instance, the number of combinations of attributes increases with increasing numbers of records, and once an attacker has checked the risk of a record, that attacker does not have to calculate the risk of other records that have the same values. Therefore, the analysis algorithm is appropriate for large datasets.

Table 4.1 Execution time

# of records	A_1 (ms)	A_2 (ms)	A_3 (ms)	A_4 (ms)
1000	1.8	699.6	131.8	569.0
5000	2.6	17,005.6	751.2	8,920.8
10000	4.7	32,764.2	1,361.6	12,925.5

Table 4.2 The case of $\epsilon = 0.5, k = 2$

β	A_1 (ms)	A_2 (ms)	A_3 (ms)	A_4 (ms)
0.05	2.6	17,005.6	751.2	8,920.8
0.10	1.2	18,950.8	512.8	5,084.8
0.30	2.0	26,715.4	139.2	8,285.4

Table 4.3 The case of $\beta = 0.05, k = 2$

ϵ	A_1 (ms)	A_2 (ms)	A_3 (ms)	A_4 (ms)
0.5	2.6	17,005.6	751.2	8,920.8
1.0	1.4	17,002.4	628.6	9,256.4
3.0	1.6	16,894.8	945.0	8,968.2

Table 4.4 The case of $\beta = 0.05, \epsilon = 0.5$

k	A_1 (ms)	A_2 (ms)	A_3 (ms)	A_4 (ms)
2	2.6	17,005.6	751.2	8,920.8
3	2.9	16,828.6	744.2	8,788.4
4	2.8	17,211.9	755.8	9,013.1

Table 4.5 Relationship among parameters and our metrics (p, N)

$k = 2$		β		
		0.05	0.1	0.3
ϵ	0.1	(0.0196, 1)	(0.0303, 2)	(0.0909, 1)
	0.5	(0.0204, 1)	(0.0250, 1)	(0.1000, 1)
	1.0	(0.0208, 1)	(0.0278, 1)	(0.1000, 1)

When the sampling rate is changed, the computation time differs depending on the attacker. This is because there are two loop processes, one for sampled records and one for nonsampled records, and the calculation methods of each process differ depending on the attacker.

The effect of noise addition on computation time is not different in this experiment, but when a very large amount of noise is added, the distribution of the records is uniform and the different kinds of records increase; as a result, the computation time may increase.

The effect of k-anonymity also seems minimal, but when k is large the number of different types of records decreases and the computation time may decrease.

Validation

We observed p and N by changing the sampling rate β and the noise parameter ϵ to verify the validity of our metrics. We evaluated the attacker model A_4 while changing the parameters k, β, and ϵ. The evaluation result is shown below (Table 4.5, 4.6).

The risk to privacy decreases as k increases and as β and ϵ decrease, and the risk is a valid privacy metric. Sampling rates are the key factor that reduces the risk in this experiment. There are some outliers in the datasets, and they are the cause of the risk. In fact, if such records are not sampled, the privacy risk decreases. We conducted this experiment multiple times, and the result was different each time. Table 4.7 presents a sample of the evaluation results. Some outliers were included in

Table 4.6 Relationship among parameters and our metrics (p, N)

$k = 4$		β		
		0.05	0.1	0.3
ϵ	0.1	(0.0154, 1)	(0.0270, 2)	(0.0667, 2)
	0.5	(0.0192, 1)	(0.0227, 2)	(0.0625, 3)
	1.0	(0.0200, 1)	(0.0238, 2)	(0.0625, 1)

Table 4.7 Case of $\beta = 0.05, \epsilon = 1.0$

Times	A_1	A_2	A_3	A_4
1	(1.0000,3)	(0.0035,1)	(0.0083,1)	(0.0049,1)
2	(1.0000,2)	(0.0013,4)	(0.0108,1)	(0.0035,1)
3	(1.0000,4)	(0.0217,1)	(0.1667,1)	(0.0204,1)
4	(0.5000,5)	(0.0030,1)	(0.0667,1)	(0.0050,1)
5	(1.0000,5)	(0.0032,1)	(0.0294,1)	(0.0051,1)

the third operation, and the risk was higher than that of other operations. Therefore, the key factor may change when outliers are removed in advance.

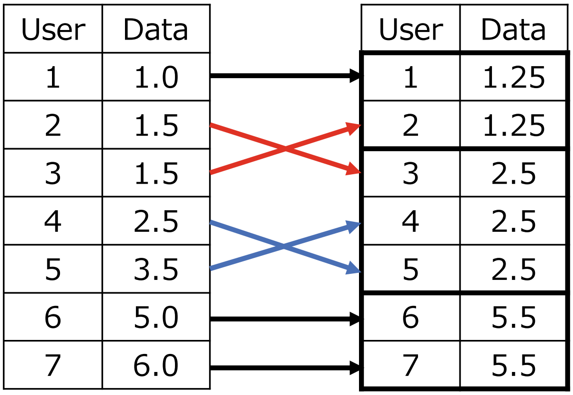
4.4 Extension to Time-Sequence Data

4.4.1 Privacy Definition

We define two types of attack models for time-sequence datasets. The first, a reidentification attack, is a general attack model where an attacker has information on the original dataset M and tries to reidentify it in an anonymized dataset $A(M)$. This model assumes that an attacker has maximal information about the original dataset. This model is the same as that of k-anonymization, where even if an attacker has an original dataset, the probability of the reidentification of a k-anonymized dataset is $1/k$.

Definition 4.10 (*Reidentification attack*) Let an attacker have a matrix $M_{t_1} \in \mathbb{R}^{n \times m}$ and an anonymized matrix $A(M_{t_1}) \in \mathbb{R}^{n \times m}$. A reidentification attack against a record r_i succeeds if record $r_i \in M_{t_1}$ is linked to record $r'_j \in A(M_{t_1})$, where r_i and r'_j are the same user.

A linkage attack, which is an attack on a valid user, is one in which an attacker tries to obtain information from the given datasets $A(M_{t_1})$ and $A(M_{t_2})$. $A(M_{t_1})$ and $A(M_{t_2})$ are assumed to include the same users, but the primary keys are different. An attacker in this model has only anonymized datasets, so a valid user is assumed



User	Data
1	1.0
2	1.5
3	1.5
4	2.5
5	3.5
6	5.0
7	6.0

User	Data
1	1.25
2	1.25
3	2.5
4	2.5
5	2.5
6	5.5
7	5.5

Fig. 4.7 Example of a risk evaluation

to be an attacker in this model. There are few studies concerning this problem, and we evaluate the risk using actual datasets in this paper.

Definition 4.11 (*Linkage attack*) Let an attacker have two anonymized matrices, $A(M_{t_1}) \in \mathbb{R}^{n\times m}$ and $A(M_{t_2}) \in \mathbb{R}^{n\times m}$. M_{t_1} and M_{t_2} include the same users and items, where each user and item of M_{t_2} are the same as those of M_{t_1}. A linkage attack against a record r_i succeeds if record $r'_i \in A(M_{t_1})$ is linked to record $r''_j \in A(M_{t_2})$, where r'_i and r''_j are the same user.

We next define the privacy metric as follows:

Definition 4.129 (*Privacy metric*) Let n be the total number of users of a dataset M and n' be the number of users that are successfully attacked. The privacy risk of M is defined as $\frac{n'}{n}$.

We consider the attacks to be the same as the previous ones to solve an assignment problem. An assignment problem is to find an appropriate task assignment when there are n users and tasks, and the Hungarian algorithm [47] solves the assignment problem in such a way that the entire cost is minimal.

We apply the same algorithm as used for reidentification and linkage attacks and assume that when an attacker assigns a record to the correct user, the attack succeeds. When a dataset is k-anonymized, there are at least $k-1$ of the same records. Hence, when a record is assigned to the cluster to which the correct record belongs to, we regard the record as being assigned correctly even if the assigned record is not actually correct. Furthermore, we define the privacy metric as the result obtained by multiplying the probability, and we define $1/k$ because the probability is the ratio of correctly assigned clusters (Fig. 4.7).

Figure 4.1 shows an example of a risk evaluation. The dataset on the left is the original dataset and that on the right is the anonymized dataset. The arrows indicate the assignment result. User 2 of the original dataset, for instance, is assigned to user 3 of the anonymized dataset, so the attack on user 2 fails. When noise addition is used as the anonymization method, users 2, 3, 4, and 5 are assigned to the wrong users and the privacy risk is $3/7$. On the other hand, when k-anonymization is used, in this case, $k = 2$, users 4 and 5 are assigned to the wrong users (blue arrows) but are assigned to the clusters that are the same as those of the correct users. Therefore, we consider the attacks on users 4 and 5 to be successful. The failed attacks are only for users 2 and 3 (red arrows), and the privacy risk is $5/7 \times 1/2 = 5/14$.

4.4.2 Utility Definition

We define the utility metric here. In previous research, most utility metrics are based on either the distance between the original dataset and the anonymized dataset, or the amount of information loss [48, 49]. However, the utility depends on the situation (i.e., context and use case), and these metrics do not necessarily match the actual utility. Therefore, we consider a use case scenario and present a utility definition that matches the scenario. Specifically, we consider a use case in which an anonymized dataset is used as training data for a machine learning algorithm. In the case of a Web access log dataset, for example, a client, who is a developer of an anti-virus software, may generate a machine learning model from an anonymized dataset and predict whether their user will access a phishing Web site.

Definition 4.13 (*Utility metric*) Let $F(M, E)$ be the F-measure of a machine learning model, where the training data are M and the test data are E. The utility metric is defined as follows:

$$Uti(A(M)) = \frac{F(A(M), E)}{F(M, E)}. \tag{4.1}$$

Figure 4.8 gives an overview of the utility evaluation. We first generate two machine learning models: One is from an original dataset, and the other is from its anonymized dataset. An item is randomly chosen as an objective variable, and the remaining items are explanation variables. Then, we use these models and predict an attribute of each record of an evaluation dataset that has the same attributes as those of the original dataset. This operation is performed several times while an objective variable is changed. The utility is defined as the average of the ratio of the F-measure of a model of the anonymized dataset to that of a model of the corresponding original dataset. In this paper, we apply logistic regression as the machine learning algorithm and predict fifty attributes.

4.4.3 Matrix Factorization

Matrix factorization is a fundamental task in data analysis, and the technique is used in various scenarios, such as text data mining, acoustic analysis, and product recom-

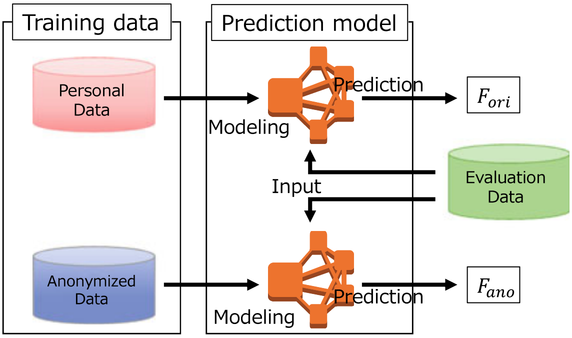


Fig. 4.8 Overview of utility evaluation

mendation by collaborative filtering. We use matrix factorization as an anonymization technique, so we present an overview of matrix factorization in this section.

4.4.3.1 SGD Matrix Factorization

We consider an unknown rank-r matrix $M \in \mathbb{R}^{n \times m}$ and assume that we know a set of elements $\Omega \subset [n] \times [m]$. $P_\Omega(M) \in \mathbb{R}^{n \times m}$ is defined as:

$$P_\Omega(M) = \begin{cases} M_{ij} & \text{if}(i, j) \in \Omega, \\ 0 & \text{otherwise}. \end{cases} \tag{4.2}$$

The goal of matrix factorization is to find two matrices $U \in \mathbb{R}^{r \times n}$ and $V \in \mathbb{R}^{r \times m}$ which approximate the original matrix $M_{ij} \approx X_{ij}$ s.t. $\forall M_{ij} \in \Omega(M)$ with lower dimensionality $r << min(n, m)$. Here, $X = U^{\mathsf{T}} V$.

This problem is defined to solve the following optimization problem:

$$\min_{u^*, v^*} \sum_{(i,j) \in P_\Omega(M)} (M_{ij} - u_i^{\mathsf{T}} v_j)^2 + \lambda(||u_i||^2 + ||v_j||^2), \tag{4.3}$$

where u_i is a vector of user factors and v_j is a vector of item factors. When u_i and v_j are variables, this function is not a convex set, so the problem described above cannot be solved. Some techniques are proposed to solve the problem, and gradient descent [50], for example, is a fundamental technique to find a local minimum value. However, gradient descent needs to update vectors iteratively to obtain an optimal solution and using gradient descent is computationally expensive, so stochastic gradient descent (SGD) is widely used, for example, in the KDD Cup 2011 [51] and the Netflix Prize [52].

There has been some research to speed up SGD-based matrix factorization, such as [53–56], and each algorithm updates the matrices in parallel or in a distributed manner.

In this paper, we apply a simple SGD technique to optimize formula (2) and denote $Update(A)$ as the update of a matrix A using the SGD technique.

4.4.4 Anonymization Using Matrix Factorization

We consider matrix factorization to be an anonymization method, and rank r contributes to the accuracy of the matrix approximation. Moreover, we propose combining matrix factorization with another anonymization method ano, such as k-anonymization or noise addition. We denote p as a parameter of the anonymization method, and p is k or ϕ in this paper. A basis matrix U and weighting matrix V can be assumed to be the characteristics of the rows and columns, respectively, and U is a characteristic matrix of users in our dataset. Therefore, we propose to anonymize U and maintain V so that the characteristics of the domain are preserved. In our algorithm, we first divide the dataset M into U and V, and anonymize U. Then, we optimize V once and recombine it with the anonymized U. The algorithm is described below.

We indicate that $A_r(D)$ applies matrix factorization to matrix D and that $A_{(ano,r)}(D)$ combines matrix factorization and the anonymization method ano by:

$$A_{(ano,r)}(D) = (A_{(ano)}(U))^{\mathsf{T}} V, \text{where } U \in \mathbb{R}^{r\times n},\ V \in \mathbb{R}^{r\times m}. \tag{4.4}$$

Algorithm 4.6 (M, r, I, ano, p) : Anonymization using Matrix Factorization

```
Input: Original dataset $M$, rank $r$, and the number of iterations $I$.
$t = 0$
Construct $U_t \in [0, 1]^{n\times r}$ and $V_t \in [0, 1]^{m\times r}$ randomly
while $t < I$ do
    $U_{t+1} = Update(U_t)$
    $V_{t+1} = Update(V_t)$
    $t = t + 1$
end while
$U'_{t+1} = A_{(ano)}(U_{t+1})$
return $X = U'^{\mathsf{T}}_{t+1} V_{t+1}$
```

Table 4.8 Dataset format

ID (= i)	Date	URL (= j)
x_{t_1} (= 1)	2016-12-01 16:13:48	www.google.com (= 1)
y_{t_1} (= 2)	2016-12-01 16:15:14	www.mail.google.com (= 2)
x_{t_1}	2016-12-01 16:17:13	www.youtube.com (= 3)
z_{t_1} (= 3)	2016-12-01 16:19:01	www.facebook.com (= 4)
x_{t_2} (= 1)	2016-12-01 16:21:15	www.youtube.com
x_{t_2}	2016-12-01 16:22:42	www.google.com
z_{t_2} (= 3)	2016-12-01 16:25:01	www.youtube.com

4.4.5 Experiment

4.4.5.1 Dataset

We use an actual Web access log dataset as a time-sequence dataset. The dataset consists of an ID, a time stamp, and the access domain, as shown in Table 4.8. We convert the dataset into a matrix as follows:

$$M_T = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nm} \end{bmatrix} \tag{4.5}$$

Here, T is the observation time.

We say $r_{ij} = 1$ if a user whose ID is i accesses domain j during time T, and otherwise, $r_{ij} = 0$. For example, we construct the datasets in Table 4.8 as follows:

$$M_{t_1} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{4.6}$$

$$M_{t_2} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \tag{4.7}$$

Here, t_1 is the 10-min span between 2016-12-01 16:10:00 and 2016-12-01 16:19:59, and t_2 is the similar 10-min span between 2016-12-01 16:20:00 and 2016-12-01 16:29:59. The IDs are different between t_1 and t_2, but x_{t_1} and x_{t_2}, and z_{t_1} and z_{t_2} represent the same users.

Table 4.9 Linkage attack against a non-anonymized dataset

Observation time (h)	Linkage attack probability
2	0.51
4	0.64
8	0.80

In the following experiments, we chose randomly 200 users and 1,000 domains from an actual Web access log and let the pseudonymous ID be changed at each designated time T.

4.4.5.2 The Privacy Risk Against a Linkage Attack

First, we evaluate whether a linkage attack is possible. We set the observation time t_1 as 2, 4, and 8 h from 16:00 on a weekday and the observation time t_2 as the same time on another weekday. The probability of a linkage attack between M_{t_1} and M_{t_2} is shown in Table 4.9.

The matrix only includes information on whether a domain has been accessed, and even if the observation time is 2 h, the linkage attack probability, i.e., risk, is very high (over 50%). Moreover, the risk increases as the observation time increases because when the observation time increases, the trend of a user becomes noticeable. The result shows that the pattern of Web access for people has consistent characteristics. Hence, we need to consider not only reidentification attacks but also linkage attacks to avoid privacy leakages.

4.4.5.3 Effects of Matrix Factorization

Observation times t_1 and t_2 are fixed as 8 h from 16:00 h on a weekday in the following experiments. The inputs of matrix factorization are the original dataset M, the number of iterations I, and the rank r. Furthermore, λ and γ and are the hyperparameters. We fix $I = 100$, which is enough to converge, $\gamma = 0.05$, and $\lambda = 0.01$. The convergence result is shown in Fig. 4.9. The rank r can be treated as the parameter of anonymization by matrix factorization because the accuracy of dataset $X = UV^{\mathsf{T}}$ depends on the rank r, so r is the parameter of our algorithm; we set $r = 10, 20, 30, 40$. We set larger values in the experiments in [3], but the results of the case $r > 40$ are saturated. The probabilities of reidentification and linkage attacks are shown in Table 4.10.

The results show that matrix factorization itself does not have much effect on reidentification attacks. Note that matrix factorization can preserve the relative positional relationship among the records so that the privacy risk of the reidentification attack does not decrease much by using a matching algorithm. When the rank is

Fig. 4.9 Convergence result

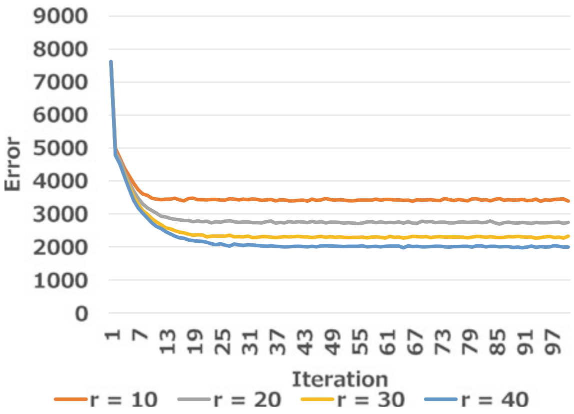
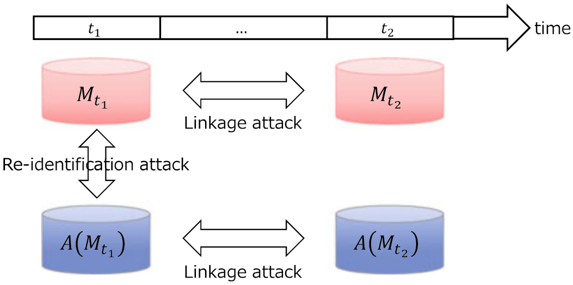


Table 4.10 Attacks against matrix factorization

Rank	Reidentification attack	Linkage attack
10	0.98	0.31
20	1.00	0.45
30	1.00	0.54
40	1.00	0.58

Fig. 4.10 Overview of the experiment



small enough, $r = 10$, the positional relationship is broken, and the privacy risk is lowered.

On the other hand, compared with the reidentification attack presented in Table 4.9, the linkage attack probability between $A_r(M_{t_1})$ and $A_r(M_{t_2})$ is better. This is because the relationship between the records of M_{t_1} and M_{t_2} is weaker than that between M_{t_1} and $A_r(M_{t_1})$. In our experiment, the dataset of the observation time is 8 h and $r = 30$ has almost the same privacy level as when the observation time is 2 h (Fig. 4.10).

Table 4.11 Experiment 1

k	Reidentification attack	Linkage attack
2	0.500	0.185
4	0.250	0.050
6	0.167	0.038
8	0.125	0.027
10	0.098	0.023

4.4.6 Results

4.4.6.1 Risk Evaluation

We evaluate our anonymization method, Algorithm 4.1, in the following experiments. We apply the method described in [10] as k-anonymization and Laplace noise as the noise addition. When noise addition is applied, noise $\epsilon \sim Lap(0, 2\phi^2)$ is added to each element, and the parameter is ϕ.

1. Evaluate the privacy risk of a reidentification attack between $A_k(M_{t_1})$ and M_{t_1} and a linkage attack between $A_k(M_{t_1})$ and $A_k(M_{t_2})$.
2. Evaluate the privacy risk of a reidentification attack between $A_\phi(M_{t_1})$ and M_{t_1} and a linkage attack between $A_\phi(M_{t_1})$ and $A_\phi(M_{t_2})$.
3. Evaluate the privacy risk of reidentification attacks between $A_k(U_{t_1})^\mathsf{T} V$ and M_{t_1} and linkage attacks between $A_k(U_{t_1})^\mathsf{T} V$ and $A_k(U_{t_2})^\mathsf{T} V$.
4. Evaluate the privacy risk of reidentification attacks between $A_\phi(U_{t_1})^\mathsf{T} V$ and M_{t_1} and linkage attacks between $A_\phi(U_{t_1})^\mathsf{T} V$ and $A_\phi(U_{t_2})^\mathsf{T} V$.

The evaluations of the reidentification attacks in experiments 1 and 2 are almost the same as those conducted in many previous studies. The difference is the privacy metric (see 4.4.1), and these results are used for comparison with experiments 3 and 4, which are evaluations of our algorithm. There are few studies on linkage attacks, and evaluations of this type of attack are one of our contributions.

The evaluation of the reidentification attack in experiment 1 (Table 4.11) is simple, and the result is almost the same as for k-anonymization. However, our privacy metric is slightly different from that for k-anonymity, so the result is also slightly different from $1/k$. The result of the linkage attack also shows that k-anonymization can greatly improve the privacy of linkage attacks and that 2-anonymization can reduce the privacy risk by $77\%(0.8 \rightarrow 0.185)$.

The evaluations of experiment 2 are shown in Table 4.12. The privacy of the reidentification attack is improved from $\phi \geq 0.9$, and when ϕ is large, for example, $\phi = 1.5$, the score appears to be good. However, almost half of the records are changed by more than 1 by the added noise, and each original value of M is 0 or 1, namely, $M_{ij} \in \{0, 1\}$, so that the noise is too large to preserve utility. Therefore, we conclude that simple noise addition is not good, in terms of utility preservation, as an

Table 4.12 Experiment 2

ϕ	Reidentification attack	Linkage attack
0.3	1.00	0.33
0.6	1.00	0.10
0.9	0.95	0.01
1.2	0.81	0.03
1.5	0.62	0.00

Table 4.13 Experiment 3: reidentification attack

k	$r = 10$	$r = 20$	$r = 30$	$r = 40$
2	0.44	0.50	0.50	0.50
4	0.21	0.24	0.25	0.25
6	0.12	0.14	0.15	0.16
8	0.10	0.11	0.11	0.12
10	0.08	0.08	0.08	0.08

Table 4.14 Experiment 3: linkage attack

k	$r = 10$	$r = 20$	$r = 30$	$r = 40$
2	0.11	0.15	0.15	0.15
4	0.05	0.07	0.08	0.07
6	0.04	0.03	0.03	0.04
8	0.03	0.03	0.03	0.03
10	0.02	0.02	0.02	0.02

anonymization method. On the other hand, we obtain an interesting result for linkage attacks. The privacy for linkage attacks is improved even if the noise is very small and adding even a small amount of noise is an effective countermeasure against a linkage attack.

In experiment 3, we evaluate the effect of our proposed algorithm, which is a combination of matrix factorization and k-anonymization. Table 4.13 presents the result of the reidentification attack. In the experiment, we cannot find the effect of the matrix factorization very well, but the privacy slightly improves as r increases. This is because k-anonymization has a large effect on the reidentification risk, and the effect of the matrix factorization does not appear.

The results of the linkage attack in experiment 3 are shown in Table 4.14. In the experiment, we cannot obtain new knowledge about the effect of matrix factorization. When the datasets, which are observed at different time periods, are sufficiently anonymized by k-anonymization, there is no relationship among the same users of each dataset and only outliers can be linked.

Table 4.15 Experiment 4: reidentification attack

ϕ	$r = 10$	$r = 20$	$r = 30$	$r = 40$
0.05	0.75	0.95	0.97	1.00
0.10	0.42	0.72	0.85	0.86
0.15	0.25	0.50	0.61	0.70
0.20	0.18	0.28	0.40	0.49

Table 4.16 Experiment 4: linkage attack

ϕ	$r = 10$	$r = 20$	$r = 30$	$r = 40$
0.05	0.21	0.34	0.34	0.50
0.10	0.12	0.15	0.14	0.20
0.15	0.07	0.11	0.09	0.10
0.20	0.03	0.03	0.03	0.02

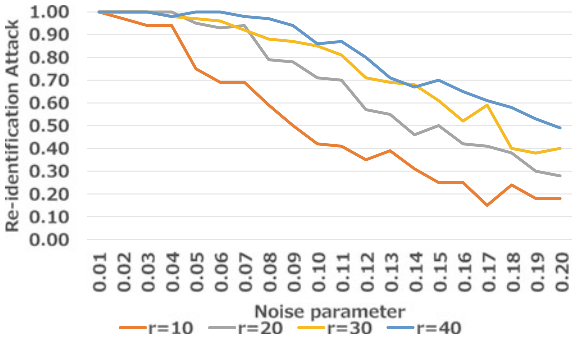


Fig. 4.11 Reidentification risk of the combination of matrix factorization and noise addition

In experiment 4, we evaluate the impact of our method, which is a combination of matrix factorization and noise addition. The evaluation results of the reidentification attack are presented in Table 4.15. Noise is added to U, which is the user's characteristics, and then, $U^{\top}$ is multiplied by V. Therefore, we cannot simply compare the results with those of experiment 2, but the impact of the matrix factorization is high. This result shows that using matrix factorization can help to construct anonymized datasets flexibly from the viewpoint of privacy. For example, the privacy risk of $A_{(\phi=0.15,r=20)}(M_{t_1})$ and $A_{(\phi=0.20,r=40)}(M_{t_1})$ is almost the same as that of $A_{(k=2)}(M_{t_1})$ and $A_{(\phi=1.5)}(M_{t_1})$.

The results of the linkage attack in experiment 4 are presented in Table 4.16. The trend is the same as that of the reidentification attack, and the matrix factorization is compatible with noise addition. We present the details of the results of the reidentification attack and the linkage attack in Figs. 4.11 and 4.12.

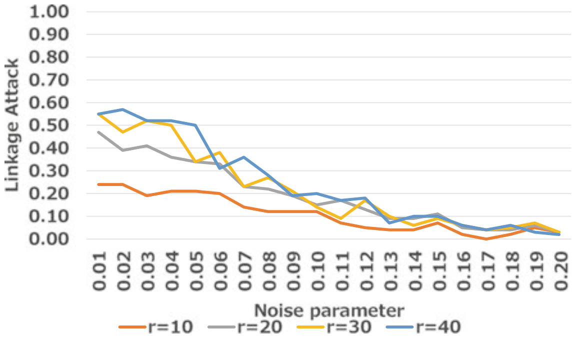


Fig. 4.12 Linkage risk of the combination of matrix factorization and noise addition

Table 4.17 Utility evaluation 1

Dataset D	Precision	Recall	F-measure	$Uti(D)$
$A_{(k=2)}(M_{t_1})$	0.780	0.720	0.749	0.981
$A_{(k=4)}(M_{t_1})$	0.741	0.688	0.714	0.936
$A_{(k=6)}(M_{t_1})$	0.755	0.691	0.721	0.946
$A_{(k=8)}(M_{t_1})$	0.737	0.659	0.696	0.913
$A_{(k=10)}(M_{t_1})$	0.748	0.677	0.711	0.932

4.4.6.2 Utility Evaluation

We next evaluate the utility of anonymized datasets. We evaluate the utility of datasets by applying a machine learning algorithm. Logistic regression (https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html) is applied in the following experiment, and the parameters are those of the default setting. One of the applications of an access log dataset is to predict a malicious site and inform the web browser's users. Therefore, we use a machine learning algorithm and predict whether each user will access a malicious site. We generate learning models using the original (non-anonymized) dataset and the anonymized datasets and input the test dataset to these models. The utility score is defined in Definition 4.13, and the F-measure of the model of the original dataset is 0.763. Each result of the evaluation is shown in Tables. 4.17, 4.18, 4.19, and 4.20.

1. Evaluate the utility of $A_{(k)}(M_{t_1})$ for $k = 2, 4, 6, 8$, and 10.
2. Evaluate the utility of $A_{(\phi)}(M_{t_1})$ for $\phi = 0.3, 0.6, 0.9, 1.2$, and 1.5.
3. Evaluate the utility of $A_{(k=2,r)}(M_{t_1})$ for $r = 10, 20, 30$, and 40.
4. Evaluate the utility of $A_{(\phi,r)}(M_{t_1})$ for $\phi = 0.1$ and 0.15 and $r = 10, 20, 30$, and 40.

In experiment 1, each element is $M_{ij} \in \{0, 1\}$ and the matrix is sparse, even when k-anonymization is effective. However, when the dataset is more complex, the utility of k-anonymization will decrease; this is widely known as the curse of dimensionality.

Table 4.18 Utility evaluation 2

Dataset D	Precision	Recall	F-measure	$Uti(D)$
$A_{(\phi=0.3)}(M_{t_1})$	0.780	0.664	0.717	0.941
$A_{(\phi=0.6)}(M_{t_1})$	0.738	0.610	0.668	0.876
$A_{(\phi=0.9)}(M_{t_1})$	0.719	0.541	0.618	0.810
$A_{(\phi=1.2)}(M_{t_1})$	0.652	0.507	0.571	0.748
$A_{(\phi=1.5)}(M_{t_1})$	0.625	0.520	0.567	0.744

Table 4.19 Utility evaluation 3

Dataset D	Precision	Recall	F-measure	$Uti(D)$
$A_{(k=2,r=10)}(M_{t_1})$	0.686	0.735	0.710	0.930
$A_{(k=2,r=20)}(M_{t_1})$	0.699	0.767	0.731	0.959
$A_{(k=2,r=30)}(M_{t_1})$	0.695	0.773	0.732	0.960
$A_{(k=2,r=40)}(M_{t_1})$	0.712	0.786	0.747	0.980

Table 4.20 Utility evaluation 4

Dataset D	Precision	Recall	F-measure	$Uti(D)$
$A_{(\phi=0.10,r=10)}(M_{t_1})$	0.742	0.650	0.693	0.909
$A_{(\phi=0.10,r=20)}(M_{t_1})$	0.752	0.688	0.719	0.943
$A_{(\phi=0.10,r=30)}(M_{t_1})$	0.736	0.703	0.719	0.943
$A_{(\phi=0.10,r=40)}(M_{t_1})$	0.737	0.735	0.736	0.965

Table 4.21 Utility evaluation 5

Dataset D	Precision	Recall	F-measure	$Uti(D)$
$A_{(\phi=0.15,r=10)}(M_{t_1})$	0.718	0.614	0.662	0.868
$A_{(\phi=0.15,r=20)}(M_{t_1})$	0.748	0.655	0.698	0.915
$A_{(\phi=0.15,r=30)}(M_{t_1})$	0.704	0.680	0.692	0.907
$A_{(\phi=0.15,r=40)}(M_{t_1})$	0.716	0.711	0.713	0.935

The results of experiment 2 show that the utility of the dataset decreases as noise increases. As stated in the risk evaluation section, each element of the original dataset is 0 or 1, and the utility drastically worsens when the noise parameter is large, such as $\phi = 1.5$.

When k-anonymization and matrix factorization are combined, the effect of matrix factorization is small, as is the case for the privacy risk. In this experiment, the effect of k-anonymization is large, and the effect of matrix factorization is relatively small.

The evaluation results of the combination of noise addition and matrix factorization show a good performance (Tables 4.20 and 4.21). A dataset generated by combining matrix factorization and noise addition has more utility than a dataset generated by noise addition when each dataset has the same privacy level.

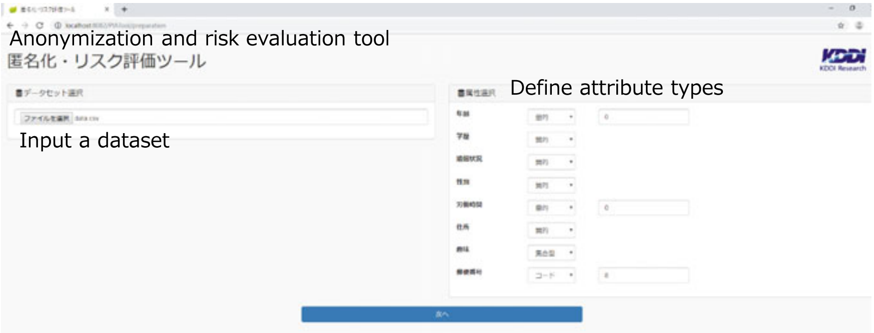


Fig. 4.13 Anonymization and privacy risk evaluation tool 1

4.5 Anonymization and Privacy Risk Evaluation Tool

In this section, we introduce an anonymization and privacy risk evaluation tool. So far, we have shown how to evaluate the privacy and utility of several datasets. We focus on static datasets and apply the theory we have described in the tool. First, we explain the outline of the tool. The tool requires a dataset that is the target of anonymization and privacy risk evaluation. At this time, the data type is defined for each attribute (see Fig. 4.13). Numerical, qualitative, set, code, and sensitive types can be defined. Age, height, and weight are defined as numerical types, and a user can assign a range of values. For instance, a user may want to divide age into groups of two years or five years depending on the situation. Qualitative-type records have nonnumerical value, such as gender and occupation. The set type is an extended numerical or qualitative type, and attributes that include multiple data correspond to this type. The code type is defined when every value is the same digit, such as a postcode. The sensitive type corresponds to sensitive information. The privacy risk is evaluated using quasi-identifiers in our tool, and the attributes that are sensitive do not effect the privacy risk. However, it is known that sensitive information may cause privacy leakages, and the tool can cover the risk for sensitive information such as l-diversity.

After the type of each attribute is decided, a user defines the noise and sampling parameters. Our tool can evaluate datasets that are anonymized by the combined method. Then, the user generates a hierarchical tree for each attribute, and the tool anonymizes the values in accordance with the tree. The user can generate and change the construction of hierarchical trees by using a UI (see Fig. 4.14.).

After these preparations are finished, the user can define the conditions and generate a dataset flexibly. A sample operation screen is shown in Fig. 4.15. Let us introduce a method commonly used as an example. First, a user searches records that do not achieve k-anonymity. Namely, the user searches records that do not include more than k copies of the same record, and then the user changes the level of an attribute of the records. The records that are secure enough are not processed, so the

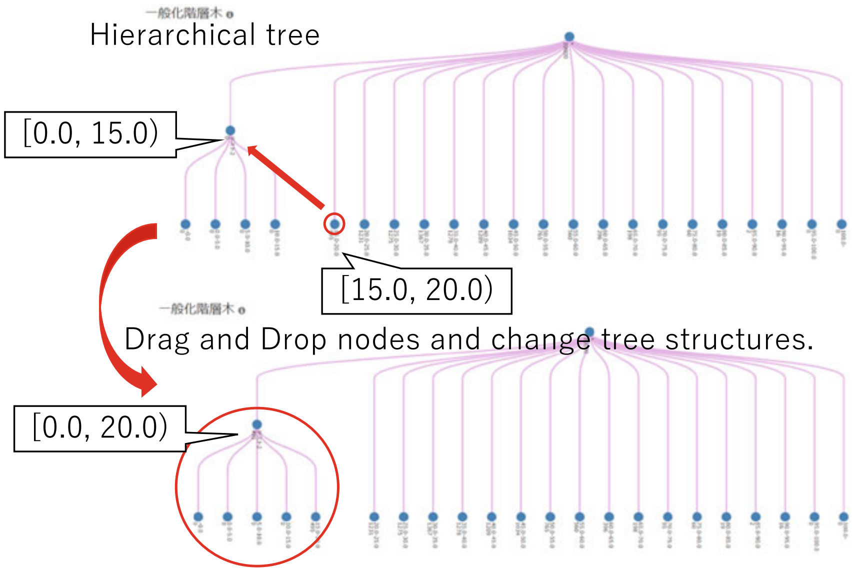
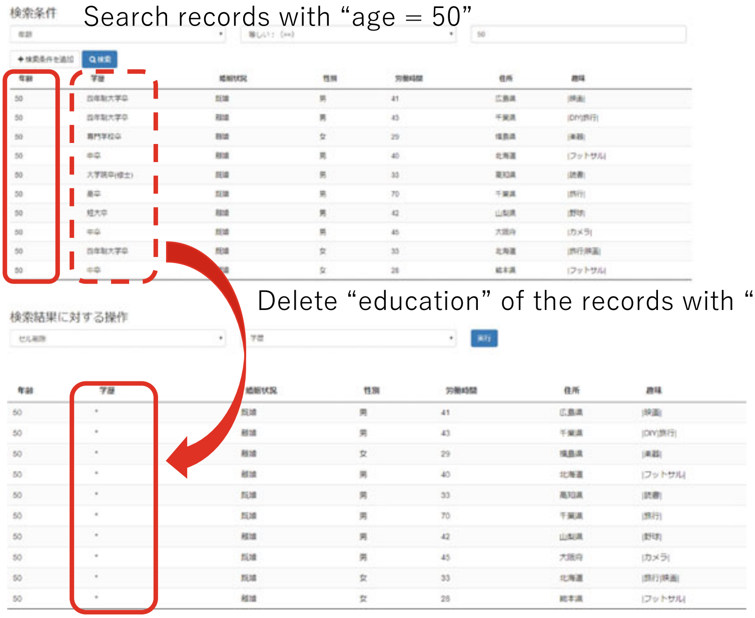


Fig. 4.14 Anonymization and privacy risk evaluation tool 2



年齢	学歴	婚姻状況	性別	労働時間	住所	趣味
50	四年制大学卒	既婚	男	41	広島県	|映画|
50	四年制大学卒	離婚	男	43	千葉県	|DIY|旅行|
50	専門学校卒	離婚	女	29	福島県	|楽器|
50	中卒	離婚	男	40	北海道	|フットサル|
50	大学院卒(修士)	既婚	男	33	高知県	|読書|
50	高卒	既婚	男	70	千葉県	|旅行|
50	短大卒	離婚	男	42	山梨県	|野球|
50	中卒	既婚	男	45	大阪府	|カメラ|
50	四年制大学卒	既婚	女	33	北海道	|旅行|映画|
50	中卒	離婚	女	28	熊本県	|フットサル|

年齢	学歴	婚姻状況	性別	労働時間	住所	趣味
50	*	既婚	男	41	広島県	|映画|
50	*	離婚	男	43	千葉県	|DIY|旅行|
50	*	離婚	女	29	福島県	|楽器|
50	*	離婚	男	40	北海道	|フットサル|
50	*	既婚	男	33	高知県	|読書|
50	*	既婚	男	70	千葉県	|旅行|
50	*	離婚	男	42	山梨県	|野球|
50	*	既婚	男	45	大阪府	|カメラ|
50	*	既婚	女	33	北海道	|旅行|映画|
50	*	離婚	女	28	熊本県	|フットサル|

Fig. 4.15 Anonymization and privacy risk evaluation tool 3

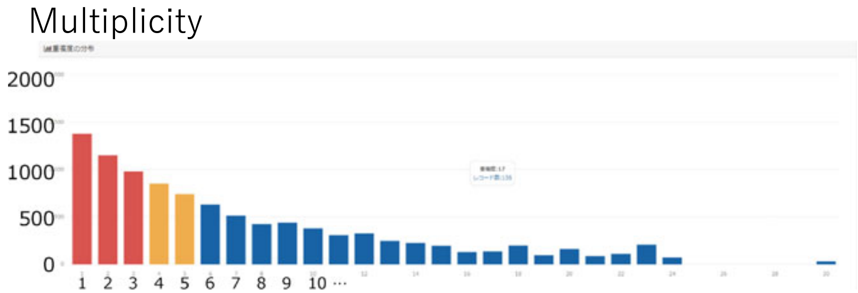


Fig. 4.16 Anonymization and privacy risk evaluation tool 4

utility of the dataset can be maintained. The conditions can be more complex. For example, the records that have a value of "age" over 80 and a value of "occupation" that is not "self-employed" are identified and anonymized. The ranks of the records are "balanced" according to the hierarchical tree. The privacy risk can be seen in real time (in Fig. 4.16), and the user can anonymize a dataset by trial and error. The operation procedure can be output as a setting file, and once the operation is decided, the procedure can be performed automatically, such as in batch processing.

4.6 Conclusion

In this chapter, we considered the importance of data and privacy. Several anonymization techniques, including k-anonymization, are introduced in Sect. 4.2, and the privacy and adversary model for static data are shown in Sect. 4.3. We focused on static data and time-sequence data in this project, and we discuss time-sequence data in Sect. 4.4. Finally, in Sect. 4.5, we introduce an anonymization and privacy risk evaluation tool. The tool is partly developed in this project, and we are proactive in using it commercially.

References

1. S. Kiyomoto, K. Fukushima, Y. Miyake, Privacy preservation of user history graph, in *Information Security Theory and Practice. Security, Privacy and Trust in Computing Systems and Ambient Intelligent Ecosystems*, ed. by I. Askoxylakis, H.C. Pöhls, J. Posegga (Springer, Berlin, 2012), pp. 87–96
2. T. Mimoto, S. Kiyomoto, K. Tanaka, A. Miyaji, (p, n)-identifiability: anonymity under practical adversaries, in *2017 IEEE Trustcom/BigDataSE/ICESS* (IEEE, 2017), pp. 996–1003
3. T. Mimoto, S. Kiyomoto, S. Hidano, A. Basu, A. Miyaji, The possibility of matrix decomposition as anonymization and evaluation for time-sequence data, in *2018 16th Annual Conference on Privacy, Security and Trust (PST)* (IEEE, 2018), pp. 1–7

4. P. Samarati, L. Sweeney, Generalizing data to provide anonymity when disclosing information, in *Proceedings of PODS 1998* (1998), p. 188
5. P. Samarati, Protecting respondents' identities in microdata release. IEEE Trans. Knowl. Data Eng. **13**(6), 1010–1027 (2001)
6. L. Sweeney, Achieving k-anonymity privacy protection using generalization and suppression. J. Uncert. Fuzziness Knowl.-Base Syst. **10**(5), 571–588 (2002)
7. S. Kiyomoto, K.M. Martin, Towards a common notion of privacy leakage on public database, in *2010 International Conference on Broadband, Wireless Computing, Communication and Applications* (2010), pp. 186–191
8. K. LeFevre, D.J. DeWitt, R. Ramakrishnan, Incognito: efficient full-domain k-anonymity. Proceedings of SIGMOD **2005**, 49–60 (2005)
9. K. LeFevre, D.J. DeWitt, R. Ramakrishnan, Mondrian multidimensional k-anonymity, in *Proceedings of the 22nd International Conference on Data Engineering (ICDE '06)* (IEEE, 2006), pp. 25–35
10. J.-W. Byun, A. Kamra, E. Bertino, N. Li, Efficient k-anonymization using clustering techniques, in *International Conference on Database Systems for Advanced Applications* (Springer, 2007), pp. 188–200
11. K. Mivule, Utilizing noise addition for data privacy, an overview (2013). arXiv:1309.3958
12. J.J. Kim, A method for limiting disclosure in microdata based on random noise and transformation, in *Proceedings of the Section on Survey Research Methods* (American Statistical Association, 1986), pp. 303–308
13. T. Yu, S. Jajodia, *Secure Data Management in Decentralized Systems*, vol. 33 (Springer Science & Business Media, Berlin, 2007)
14. S. Kiyomoto, K. Fukushima, Y. Miyake, Data anonymity in multi-party service model, in *Security Technology*, ed. by T.-H. Kim, H. Adeli, W.-C. Fang, J.G. Villalba, K.P. Arnett, M.K. Khan (Springer, Berlin, 2011), pp. 21–30
15. Y. Zheng, L. Zhang, Z. Ma, X. Xie, W.-Y. Ma, Recommending friends and locations based on individual location history. ACM Trans. Web (TWEB) **5**(1), 5 (2011)
16. Y. Zheng, L. Zhang, X. Xie, W.-Y. Ma, Mining interesting locations and travel sequences from gps trajectories, in *Proceedings of the 18th International Conference on World Wide Web* (ACM, 2009), pp. 791–800
17. V.W. Zheng, Y. Zheng, X. Xie, Q. Yang, Collaborative location and activity recommendations with gps history data, in *Proceedings of the 19th International Conference on World Wide Web* (ACM, 2010), pp. 1029–1038
18. S. Chawla, Y. Zheng, J. Hu, Inferring the root cause in road traffic anomalies, in *2012 IEEE 12th International Conference on Data Mining (ICDM)* (IEEE, 2012), pp. 141–150
19. Y. Zheng, Trajectory data mining: an overview. ACM Trans. Intell. Syst. Technol. (TIST) **6**(3), 29 (2015)
20. C. Dwork, Differential privacy, in *Proceedings of ICALP 2006*. LNCS, vol. 4052 (2006), pp. 1–12
21. C. Dwork, Differential privacy: a survey of results, in *Proceedings of TAMC 2008*. LNCS, vol. 4978 (2008), pp. 1–19
22. C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, M. Naor, Our data, ourselves: privacy via distributed noise generation, in *Proceedings of Eurocrypt 2006*. LNCS, vol. 4004 (2006), pp. 486–503
23. C. Dwork, F. McSherry, K. Nissim, A. Smith, Calibrating noise to sensitivity in private data analysis, in *Proceedings of TCC 2006*. LNCS, vol. 3876 (2006), pp. 265–284
24. C. Dwork, G.N. Rothblum, S. Vadhan, Boosting and differential privacy. Proceedings of IEEE FOCS **2010**, 51–60 (2010)
25. P. Kodeswaran, E. Viegas, Applying differential privacy to search queries in a policy based interactive framework, in *Proceedings of PAVLAD '09* (ACM, 2009), pp. 25–32
26. C. Li, M. Hay, V. Rastogi, G. Miklau, A. McGregor, Optimizing linear counting queries under differential privacy, in *Proceedings of PODS '10* (ACM, 2010), pp. 123–134

27. I. Mironov, O. Pandey, O. Reingold, S. Vadhan, Computational differential privacy, in *Proceedings of CRYPTO 2009, LNCS*, vol. 5677 (2009), pp. 126–142
28. A. McGregor, I. Mironov, T. Pitassi, O. Reingold, K. Talwar, S. Vadhan, The limits of two-party differential privacy. Proceedings of IEEE FOCS **2010**, 81–90 (2010)
29. A. Groce, J. Katz, A. Yerukhimovich, Limits of computational differential privacy in the client/server setting, in *Proceedings of TCC 2011*, to appear. LNCS (2011)
30. M.R. Clarkson, F.B. Schneider, Quantification of integrity, in *Proceedings of 23rd IEEE Computer Security Foundations Symposium* (IEEE, 2010), pp. 28–43
31. T.-S. Hsu, C.-J. Liau, D.-W. Wang, J.K.-P. Chen, Quantifying privacy leakage through answering database queries, in *Proceedings of ISC '02*. LNCS, vol. 2433 (2002), pp. 162–176
32. Y.C. Chiang, T.-S. Hsu, S. Kuo, D.-W. Wang, Preserving confidentially when sharing medical data, in *Proceedings of Asia Pacific Medical Information Conference* (2000)
33. Y.T. Chiang, Y.C. Chiang, T.-S. Hsu, C.-J. Liau, D.-W. Wang, How much privacy? - a system to safe guard personal privacy while releasing database, in *Proceedings of 3rd International Conference on Rough Sets and Current Trends in Computing, LNCS*, vol. 2475 (2002), pp. 226–233
34. A. Krause, E. Horvitz, A utility-theoretic approach to privacy and personalization, in *Proceedings of AAAI'08*, vol. 2 (2008), pp. 1181–1188
35. A. Krause, E. Horvitz, A utility-theoretic approach to privacy in online services. J. Artif. Intell. Res. **39**, 633–662 (2010)
36. L. Zayatz, Disclosure avoidance practices and research at the us census bureau: an update. J. Offic. Stat. **23**(2), 253 (2007)
37. M. Freiman, J. Lucero, L. Singh, J. You, M. DePersio, L. Zayatz, The microdata analysis system at the us census bureau, in *Proceedings of the Joint Statistical Meetings, American Statistical Association Section on Survey Research Methods* (2011)
38. K. Chaudhuri, N. Mishra, When random sampling preserves privacy, in *Annual International Cryptology Conference* (Springer, 2006), pp. 198–213
39. C. Dwork, A. Roth et al., The algorithmic foundations of differential privacy. Found. Trends Theor. Comput. Sci. **9**(3–4), 211–407 (2014)
40. J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, S. Martínez, Enhancing data utility in differential privacy via microaggregation-based k-anonymity. The VLDB J. **23**(5), 771–794 (2014)
41. L. Sweeney, k-anonymity: a model for protecting privacy. Int. J. Uncert. Fuzziness Knowl.-Based Syst. **10**(5), 557–570 (2002)
42. A. Narayanan, V. Shmatikov, How to break anonymity of the netflix prize dataset (2006), arXiv:cs/0610105
43. A. Machanavajjhala, J. Gehrke, D. Kifer, t-closeness: privacy beyond k-anonymity and l-diversity, in *Proceedings of ICDE'07* (2007), pp. 106–115
44. C. Dwork, Differential privacy: a survey of results. Proceedings of TAMC **4978**(2008), 1–19 (2008)
45. Y. Omori, Posterior probability of population uniqueness in microdata. Proc. Inst. Stat. Math. **51**(2), 223–239 (2003)
46. R.J.A. Little, Statistical analysis of masked data. J. Offic. Stat. **9**(2), 407 (1993)
47. H.W. Kuhn, The hungarian method for the assignment problem. Naval Res. Logist. (NRL) **2**(1–2), 83–97 (1955)
48. J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, A.W.-C. Fu, Utility-based anonymization for privacy preservation with less information loss. SIGKDD Explor. Newsl. **8**(2), 21–30 (2006)
49. J.-W. Byun, A. Kamra, E. Bertino, N. Li, Efficient k-anonymity using clustering technique, in *Proceedings of the International Conference on Database Systems for Advanced Applications* (2007), pp. 188–200
50. J. Nocedal, S. Wright, *Numerical Optimization* (Springer Science & Business Media, Berlin, 2006)
51. G. Dror, N. Koenigstein, Y. Koren, M. Weimer, The yahoo! music dataset and kdd-cup'11, in *Proceedings of the 2011 International Conference on KDD Cup 2011*, vol. 18 (JMLR.org, 2011), pp. 3–18

52. R.M. Bell, Y. Koren, Lessons from the netflix prize challenge. SiGKDD Explor. **9**(2), 75–79 (2007)
53. B. Recht, C. Re, S. Wright, F. Niu, Hogwild: a lock-free approach to parallelizing stochastic gradient descent, in *Advances in Neural Information Processing Systems* (2011), pp. 693–701
54. R. Gemulla, E. Nijkamp, P.J. Haas, Y. Sismanis, Large-scale matrix factorization with distributed stochastic gradient descent, in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2011), pp. 69–77
55. Y. Zhuang, W.-S. Chin, Y.-C. Juan, C.-J. Lin, A fast parallel sgd for matrix factorization in shared memory systems, in *Proceedings of the 7th ACM Conference on Recommender Systems* (ACM, 2013), pp. 249–256
56. J. Oh, W.-S. Han, H. Yu, X. Jiang, Fast and robust parallel sgd matrix factorization, in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2015), pp. 865–874

Chapter 5
Living Safety Testbed Group

Koji Kitamura and Yoshifumi Nishida

Abstract Safety technology for everyday activities is strongly needed for children, the elderly, and persons with disabilities. However, it is difficult to understand problems related to everyday life from injury data, medical data, and so on because such data are distributed over multiple organizations and cannot easily be shared or integrated due to privacy protection concerns. To address this issue, our project is developing technologies for integrating and utilizing multi-organizational distributed big data based on security technology. The authors research school safety based on the developed technologies. In this chapter, the authors describe a trend analysis technology for time series injury data, a cliff analysis technology for extracting serious injury situation, and child behavior prediction technology as the necessary functions for finding and predicting serious injuries and evaluating the effectiveness of an intervention. We also present some analysis examples using the developed function. Furthermore, we describe some social implementation projects for injury prevention for the serious injuries found by analyzing injury data using our developed system.

5.1 Necessity of Living Safety

Community safety is highly desirable for children, the elderly, persons with disabilities, and others with special needs for functional support in daily life. People with variances in the functions of daily life experience insufficiencies in bodily or cognitive function under conditions or environments that had previously been problem-free. Risk arises at certain times, and maintenance of their safety through their own care or the care of people around them is thereafter difficult. It is accordingly important

K. Kitamura (✉)
National Institute of Advanced Industrial Science and Technology, 2-4-7, Aomi, Koto, Tokyo 135-0064, Japan
e-mail: k.kitamura@aist.go.jp

Y. Nishida
Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro-ku, Tokyo 152-8550, Japan
e-mail: nishida.y.af@m.titech.ac.jp

A. Miyaji and T. Mimoto (eds.), *Security Infrastructure Technology for Integrated Utilization of Big Data*,
https://doi.org/10.1007/978-981-15-3654-0_5

to seek out data that will serve as a basis for identification of states of risk and related conditions, implement effective corrective measures, and verify the results.

In the realm of community safety, historical data on the past accidents and therapies commonly exist in a state of dispersion among many different organizations, and it is therefore difficult to determine the total number of accidents that have occurred and gain an overall perspective extending from cause of accident to resulting injury. If relevant data held at many different organizations can be integrated and utilized, this may then lead to problem identification and effective solution based on the data.

In actuality, sharing and integration of data across institutions is difficult because of the need to protect information on individuals, maintain privacy, prevent information leakage, and other needs. So long as non-engagement in active sharing and integration of such data remains blameless, it will tend to discourage advancement of community safety. In this light, we are now engaged in advancing the development of technology for utilization and application of multi-organizational dispersed data using security-based technology, in a Japan Science and Technology Agency (JST) CREST (systemization of the security base technology for expediting/accelerating of/for big data integration and utilization) project. The research group of the authors is working in collaboration with data-holding medical/therapy organizations and with product design and other data-user sites to develop technology for effective utilization of organizationally dispersed data. To date, in collaboration with Fire and Disaster Management Agency, Japan Sport Council, multiple medical institutions, nursery, elementary, and junior high schools, and other entities, we have advanced the development of technology for integration and utilization of dispersed injury-related data.

With school safety as a specific field of application, we are engaged in proof of concept and system by demonstration. So far, we have compiled medical cost and other KPI-bearing big data from accident data dispersed in multiple elementary schools, performed presumed integration without specifying the schools, and conceived and developed a serious injury accident analysis system using the multiparty private set intersection (PSI) protocol privacy-preserving information-sharing technique and severity cliff analysis technology, for analysis of the main accidents causing severe injury, and verified the system effectiveness by applying it to actual data. With this system, the analyst identifies the task to be performed at the school site and presumably has it applied as a preventive measure.

In the present report, we describe the on-site use of the proposed system for task identification focused on temporal changes that becomes necessary and function expansion and application to actual data in intervention results evaluation. We also report on actual utilization of the system and on identified tasks as we engaged in acquisition and analysis of fine data necessary for injury prevention.

5.2 Overview of Test Bed System for Living Safety

For problem identification and solution, a system of privacy preservation is necessary to permit sharing and integration of data held by multiple organizations. It also requires an analytical method of obtaining useful information from the shared and integrated data. One method for this purpose is embodied in the JST CREST (systemization of the security base technology for expediting/accelerating of/for big data integration and utilization) project in which the authors participate, have developed the dataset (PSI: private set intersection) computation technology that preserves privacy, and have proposed a system including the severity cliff analysis technology.

PSI technology enables extraction of intersections in relation to specified data items left uncoded and held by multiple organizations. With its utilization, accident information meeting conditions specified by the user can be provided to the user in an integrated state while leaving concealed the identity of the school where the accident occurred.

The severity cliff analysis technology provides a means of analyzing the cause of severe injury accidents by seeing medical cost as severity. It enables analysis of the severity of accidents occurring in similar circumstances, location of the point of departure between cases of high and low severity, and differences between accidents with severe and slight injury, thus enabling causal analysis of accidents involving severe injury.

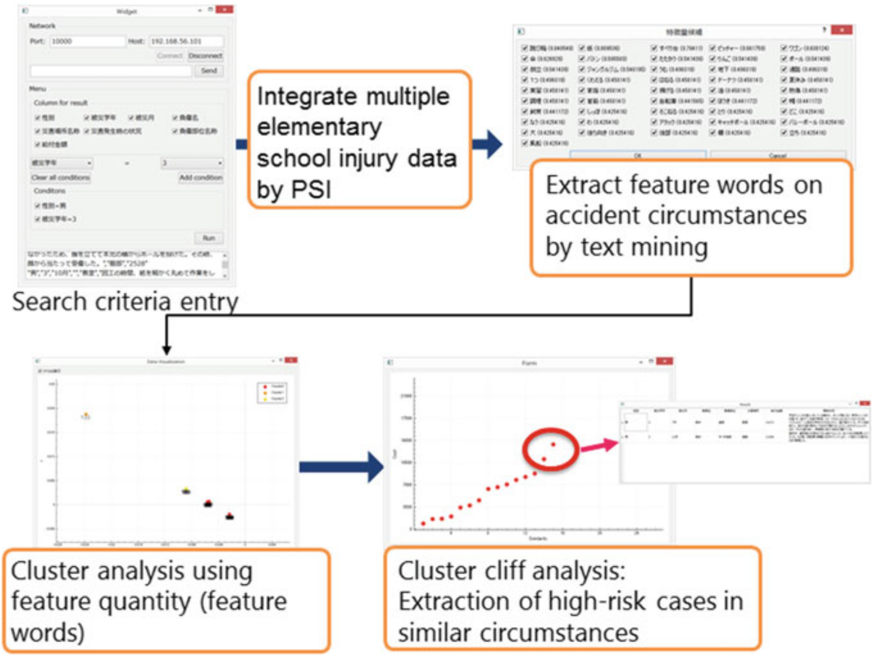

Widget
Network
Port: 10000
Host: 192.168.56.101
Connect
Disconnect
Send
Menu
Column for result
Clear all conditions
Add condition
Conditions
Run
Search criteria entry
Integrate multiple elementary school injury data by PSI
Extract feature words on accident circumstances by text mining
OK
Cancel
Data Visualization
Form
Cluster analysis using feature quantity (feature words)
Cluster cliff analysis: Extraction of high-risk cases in similar circumstances

Fig. 5.1 System for sharing and analyzing life-safety-related data with secure function

In combination, these two technologies can be used to integrate information on accidents in multiple school environments while preserving privacy, identify severe injury accidents from the integrated accident data, and analyze their causes. More specifically, we have conceived and developed a system as shown in Fig. 5.1. Accident-related information (e.g., grade, sex, and accident and injury categories) desired by the user is entered and criteria-meeting injury data from multiple schools are acquired and integrated. Severity cliff analysis is then applied to the accident circumstances described by textual data accompanying the acquired injury data, thus enabling determination of the severe injury accidents for the specified accident circumstances and analysis of the cause.

5.3 Severity Cliff Analysis of School Injury

5.3.1 Development of Severity Cliff Analysis System

5.3.1.1 System Overview

As shown schematically in Fig. 5.2, the developed severity cliff analysis system comprises four functions: accident circumstance registration, similar accident circumstance search, severe injury accident search, and severity cliff analysis. These functions are described in detail in the following corresponding subsections.

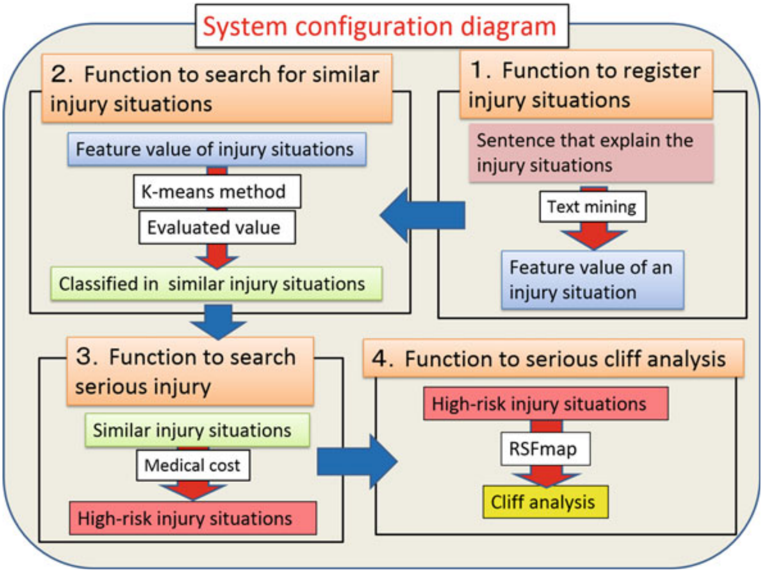

System configuration diagram
2. Function to search for similar injury situations
Feature value of injury situations
K-means method
Evaluated value
Classified in similar injury situations
1. Function to register injury situations
Sentence that explain the injury situations
Text mining
Feature value of an injury situation
3. Function to search serious injury
Similar injury situations
Medical cost
High-risk injury situations
4. Function to serious cliff analysis
High-risk injury situations
RSFmap
Cliff analysis

Fig. 5.2 System configuration for cliff analysis

5.3.1.2 Accident Circumstance Registration

The accident circumstance registration function assigns the accident circumstance feature values to the accident circumstances present in the accident database. The accident database is first subjected to morphological analysis of text representing accident circumstances in order to extract the nouns and verbs. In this analysis, the Japanese concept dictionary (Japanese WordNet) is used to consolidate the noun and verb orthographic variants. Important words are next extracted with TF-IDF weighting of each. In the present study, words with high TF-IDF values were selected as representing accident circumstance feature values. These accident circumstance feature values are assigned to the accident samples in order to construct the accident database with assigned feature values.

5.3.1.3 Similar Accident Circumstance Search

With this second function, the accident circumstances registered by the first function for their assigned feature values are sorted into similar accident circumstance groups. Clustering is performed using the Euclidean distance of the accident circumstance feature value vectors assigned in the accident database. The optimum cluster number is determined with the gap statistic value resulting from the cluster number assessment. Figure 5.3 shows the results of sorting the accident database into similar accident circumstances.

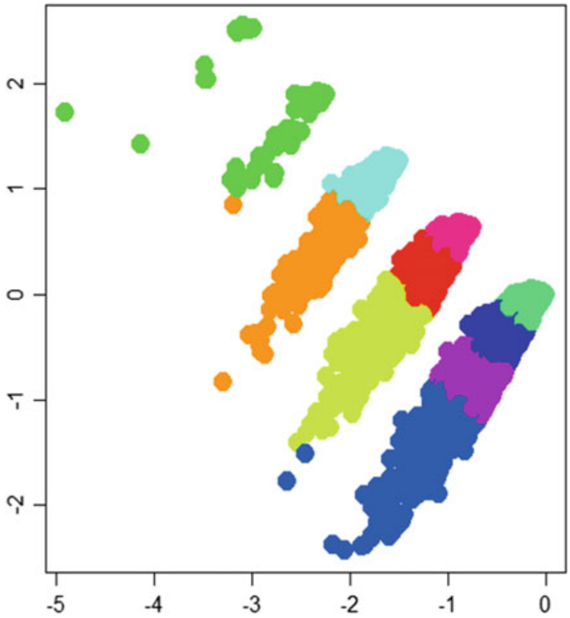


Fig. 5.3 Clustering of injury cases

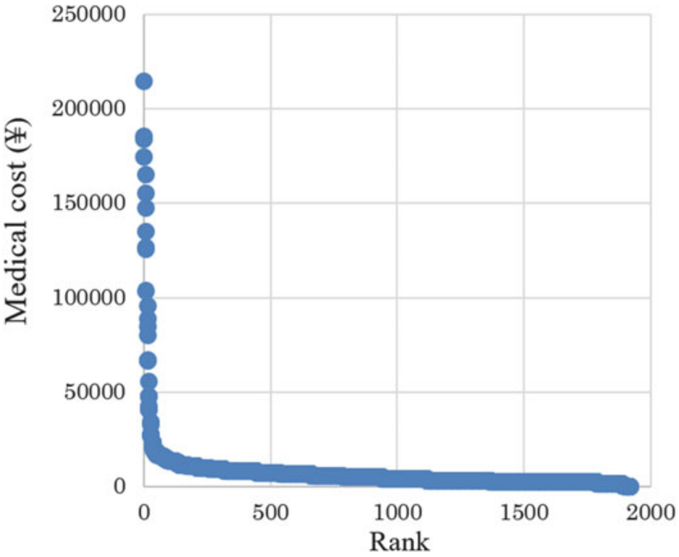


Fig. 5.4 Example of severe injury analysis of injuries occurring under similar situations

5.3.1.4 Severe Injury Accident Search

The medical costs included in the accident database were used to identify severe injury accidents, with medical cost presumed high for severe injury accidents. Figure 5.4 shows medical cost in decreasing order for injuries occurring under similar circumstances. As shown, medical cost may differ substantially even for accidents occurring in similar circumstances, and cliffs marked by specific changes may exist. This indicates that severe injury accidents can be identified by focusing on specific differences in medical cost.

5.3.1.5 Severity Cliff Analysis

Figure 5.5 shows the relation between degree of circumstance similarity and medical cost in similar states of accident, where the degree of circumstance similarity is the degree of cosine similarity in comparison with the highest medical cost accident cases (severe injury accident cases). Figure 5.6 shows the three-dimensional graph obtained on addition of frequency to the graph. Similarity 1.0 denotes the highest similarity. With these graphs, comparison of severe injury and slight injury accidents under similar circumstances enables performance of severity cliff analysis focused on the difference between severe injury and slight injury accidents.

Fig. 5.5 Relationship between similarity and cost

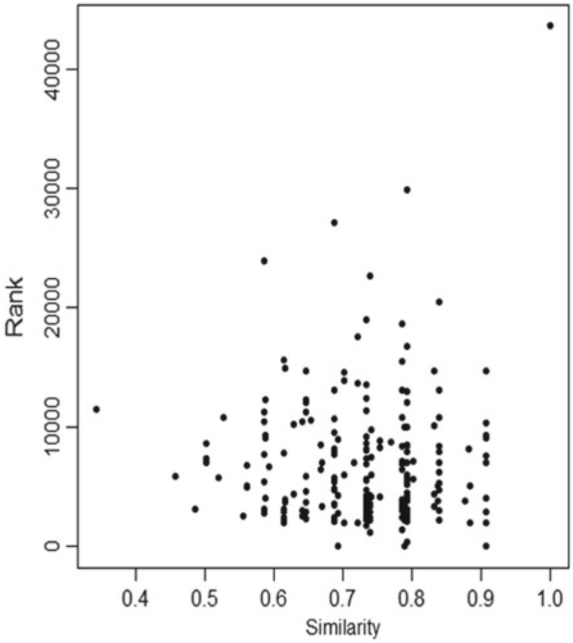
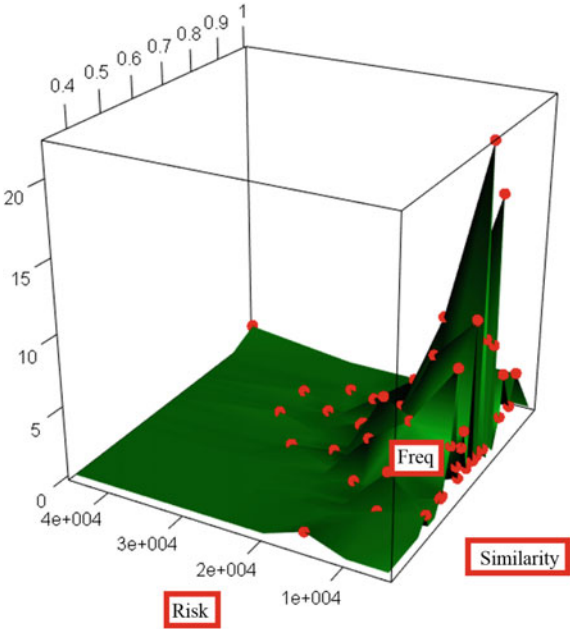


Fig. 5.6 Relationships among similarity, risk, and frequency



5.3.2 Severity Cliff Analysis

To test the effectiveness of the developed method when applied to investigating the causes of actual severe injury accidents, we used the accident data of 19,948 cases from the Injury and Accident Mutual Aid Benefit System for multiple junior high schools gathered by the Japan Sport Council.

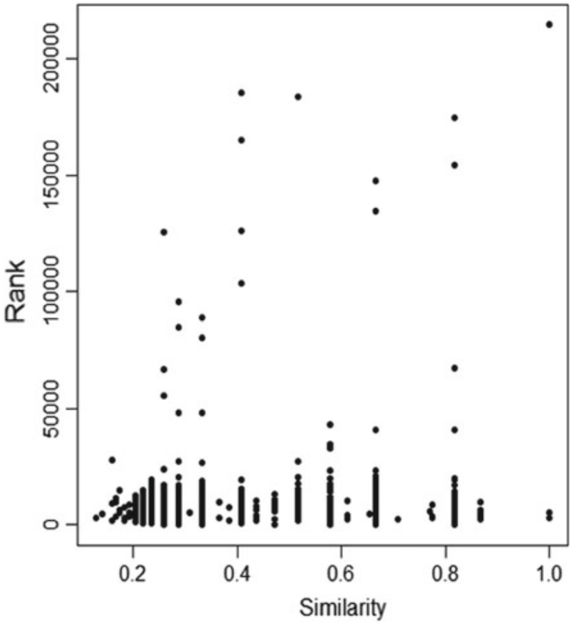
We performed the cliff analysis for similar accident circumstances with the relation shown between similarity degree and medical cost as shown in Fig. 5.7. Figure 5.8 shows the graph of Fig. 5.7 with frequency added.

The severe injury accidents in the similarity range of 1.0–0.6 in Fig. 5.8 were as follows:

- Strongly impacted and injured right shoulder in fall on contact with opponent during soccer match. (first-year junior high school, bone fracture, ¥174,504)
- In competing for ball with opponent, on contact with that opponent fell from the left side, impacting with the ground and injuring the left clavicle. (third-year junior high school, bone fracture, ¥154,475)
- In competing for the ball with an opponent, encountered strong contact and fell over from the right shoulder, thereby strongly impacting the right shoulder on the ground and fracturing the right clavicle. (3rd year junior high school, ¥147,297)

In the same similarity range, the slight injury accidents were as follows:

Fig. 5.7 Relationship between similarity and cost



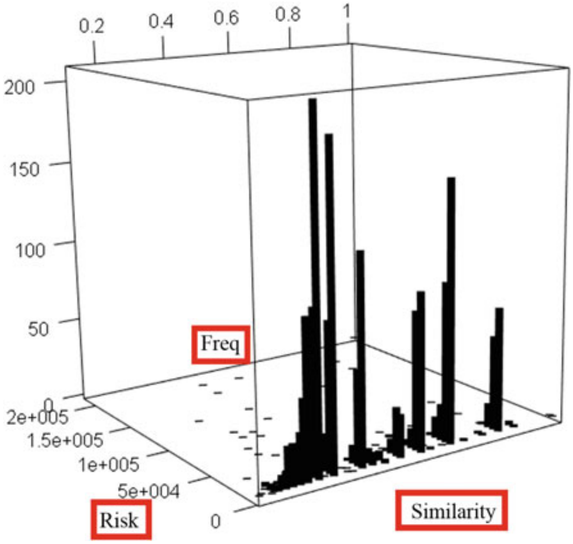


Fig. 5.8 Relationships among similarity, risk, and frequency

- At afternoon homeroom starting time, in carrying a bag from a locker and returning to a seat, the student tripped over the extended leg of a nearby student and fell, impacting his/her jaw on the leg of a desk and injuring a finger on the left hand. (second-year junior high school, contusion/bruise, ¥3,452)
- In recess from third class hour, while walking and conversing with a friend, tripped and fell at entrance to classroom with hands in pockets and therefore impacting jaw on floor. (second-year junior high school, dislocation, ¥3,152)
- In noon recess, while walking in a corridor, tripped on a friend's leg and fell, impacting right eye on wall. (first-year junior high school, contusion/bruise, ¥2,984)
- In classroom before start of class, collided with a friend and fell, impacting face on floor. (first-year junior high school, bone fracture, ¥2,476)
- While cleaning, engaged in shoving match with friend and fell with left elbow impacting floor. (third-year junior high school, contusion/bruise, ¥2,256)

In summary, it was found that severe injury accidents occurred in a soccer match in contacting an opponent and falling, in competing with an opponent for the ball and contacting the opponent and falling, and in competing for the ball with an opponent and encountering strong contact and falling over and thus, all during soccer matches in contact with an opponent and falling, whereas slight injuries occurred in tripping over someone's leg and falling, tripping and falling at an entrance, tripping on a friend's leg and falling, colliding with a friend and falling, and engaging in a shoving match and falling and thus were all in tripping on or colliding with something or someone and falling. Taken together, the results show that among similar instances in a circumstance of tripping and falling, severe injuries more readily occur in colliding with an opponent and falling in a soccer match.

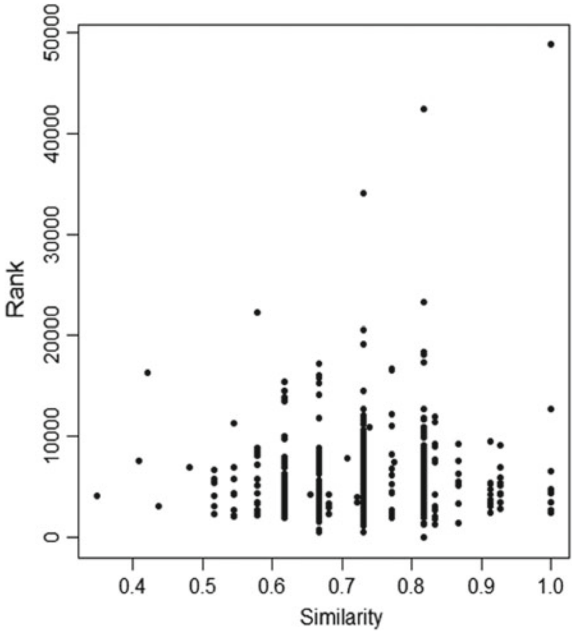


Fig. 5.9 Relationship between similarity and cost

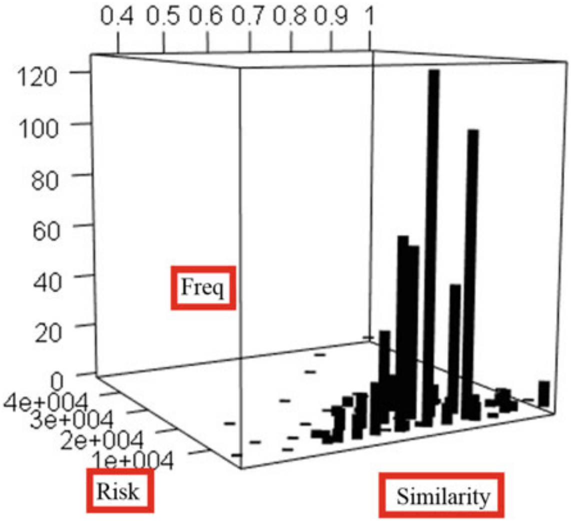


Fig. 5.10 Relationships among similarity, risk, and frequency

Let us next consider the severe injury accidents in the similarity range of 1.0–0.7 in the clusters shown in Figs. 5.9 and 5.10, which were as follows.

- In a “soft tennis” club morning practice session, a ball came flying unseen and unevaded, directly striking a student in the right eye. (first-year junior high school, contusion/bruise, ¥48,740)

- In a “soft tennis” club morning practice session, while throwing ball on the tennis court for a two-step hit, a student was struck in the left eye by a ball hit by an opponent, suffering a bruised left eyeball, left retinal tear, and left eye conjunctivitis. (first-year junior high school, contusion/bruise, ¥42,416)
- In a softball club activity, a student playing catch with a third-year student on the school ground was struck in the face by the ball after losing sight of it and having it hit his/her own glove. (second-year junior high school, bone fracture, ¥18,112)

In the same similarity range, the slight injuries were as follows:

- While a student was playing handball in a physical education class on the schoolyard in the fourth school period, during the match, the ball came flying toward the student, who tried to catch it but mistakenly was struck by the ball on the ring finger of the left hand. (second-year junior high school, bone fracture, ¥6,304)
- In a volleyball activity, when boys and girls were practicing hitting serves over the net, the ball hit by a boy struck a student on the left thumb, breaking a bone. (second-year junior high school, bone fracture, ¥4,484)
- During a volleyball club tournament in a seaside region, in a practice serve at a gymnasium before a match, a hit ball from the opposite side of the court struck and injured the right hand of the student. (third-year junior high school, contusion/bruise, ¥3,492)
- During dribbling practice in a club activity at a gymnasium, a ball bounced off the leg of a club member and struck the right hand and sprained the thumb of the student. (second-year junior high school, sprain, ¥2,164)
- During bunting practice of the baseball club after class, in a bunting attempt, the bat was mispositioned and the ball struck the thumb of the right hand. (third-year junior high school, sprain, ¥2,032)

Concerning these severe and slight injury accidents, in summary, it was found that the severe injury accidents involved a tennis ball that came flying and struck the right eye, a tennis ball hit by an opponent that struck the eye, and softball a ball striking the eye, and thus, all involving an eye being struck by a ball, whereas that the slight injury accidents involved a handball striking a finger, a hit volleyball striking a thumb, a basketball striking the right hand, and a baseball striking the right thumb, and thus, all involving a ball striking a hand or leg. These findings clearly show that, for accident circumstances in which a ball similarly strikes the body, those in which the ball strikes an eye tend to result in severe injury. This in turn indicates the existence of certain parts of the body and types of sports for which injuries tend to be serious and for which a preventive measure such as an eye protector is seldom implemented but necessary.

5.4 Trend Analysis of School Injury

5.4.1 Trend Analysis for Evaluating Intervention

Annual trends can provide an effective perspective in the search for problems that need to be solved. Examples include accidents that have sharply increased in recent years and cases that have been large in number with no change over many years, which may represent problems requiring consideration of preventive measures. It is also important to focus on annual trends when assessing the effects of measures or interventions. In this light, we have developed a trend analysis function that can be integrated and applied in combination with the previously developed severe injury accident analysis system. It has thus become possible to analyze changes in trends focused on circumstances and on verbal words characteristic of accident occurrence.

5.4.2 Analysis of Judo Accident

We have applied this trend analysis function to analyze data on 60,300 senior high school cases among 152,695 cases of judo-related injury included in the Injury and

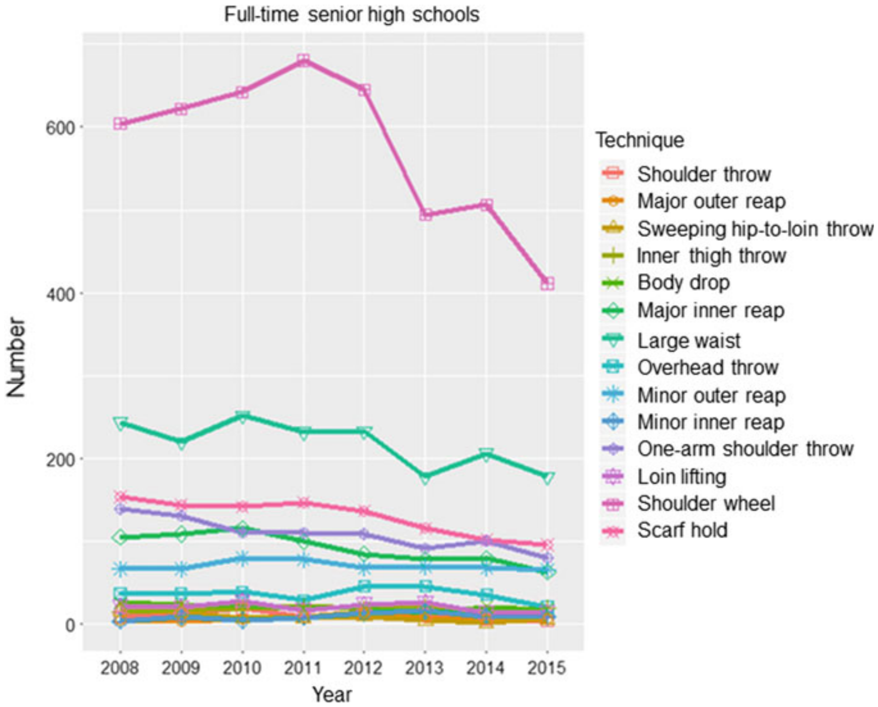


Fig. 5.11 Analysis of judo accident trends relative to judo techniques

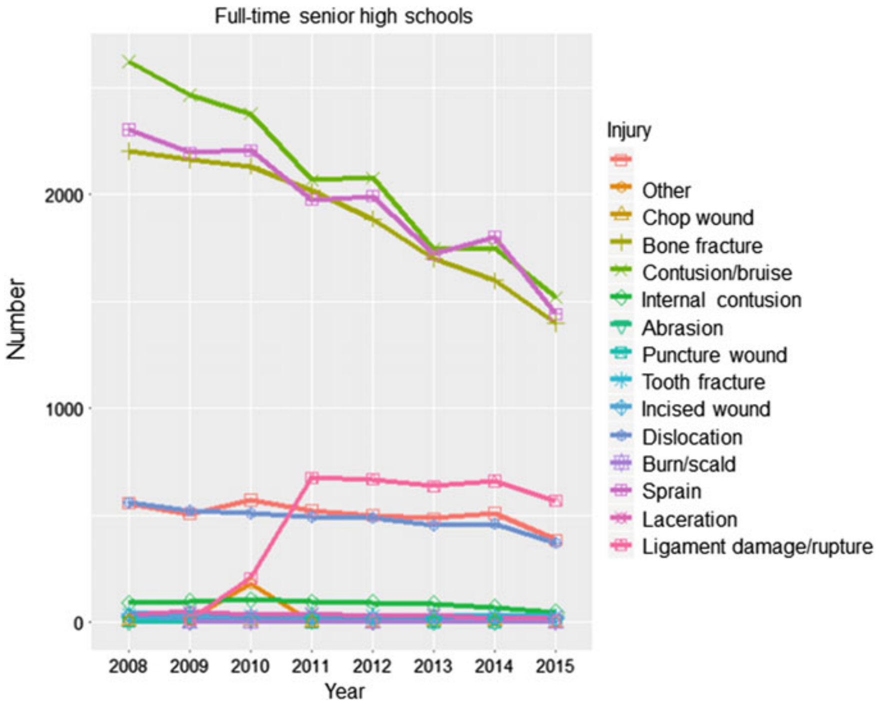


Fig. 5.12 Analysis of trends in injuries in judo accidents

Accident Mutual Aid Benefit System data of the Japan Sport Council from 2008 to 2015.

Figure 5.11 shows the results of an analysis of trends in judo techniques as related to accidents, and Fig. 5.12 shows the results of an analysis of trends in injuries due to judo accidents. A publication on judo accidents was issued in 2013, leading to their recognition as a social problem, issuance of a related alert, and notification of the risks of shoulder throwing and major outer reaping in particular. A manual on safe teaching methods was also produced and on-site initiatives were implemented. All of these apparently had considerable effect.

A marked decrease from 2013 in instances of accident-related shoulder throwing was confirmed by the authors, but they also found that no clear reduction occurred in major outer reap accidents. With this trend analysis, it is thus possible to assess the effects of intervention important for injury prevention. Application of the analysis to moderate injuries showed sharp reductions in contusion and bruise, sprain, and bone fracture, but sharp increases in ligament injury and rupture occurred in 2011 and high levels of their occurrence continued thereafter. It has thus been found possible to sharply reduce the occurrence of some injuries for which sharp increases had preceded, by investigating their cause followed by actions such as intervention for their prevention.

5.5 Childhood Home-Injury Simulation

5.5.1 Background of Simulation

Most accidents involving children below the age of five occur within their homes. Since it is important to maintain a safe home environment for children, it is imperative to be able to predict what kinds of accidents may occur in a particular environment and then to find ways to improve that environment. However, the various and scattered statistical data sources and scientific knowledge related to accident prediction have not been structured for integrative utilization. In this section, the authors report on the development of a new simulation technology that can be used to predict the kinds of accidents that may occur in a particular environment by means of a hybrid memory- and model-based approach. The system consists of a graph-structuralized accident database created from large-scale accident data (which enables the memory-based approach) and a development behavior model which describes the statistical relationship between a body interaction abilities and the age of children.

5.5.2 Home-Injury-Situation Simulation System

In this study, in order to predict child-related accident situations which may occur in an individual environment, we propose a home-injury-situation simulation system which consists of three functions: a development-related behavior prediction function, an accident situation search function, and a function for classifying products involving similar risks. The configuration of the proposed system is shown in Fig. 5.13.

The development-related behavior prediction function is used to estimate the area that can be reached by a child's hands and then visualize that area in 3D space on a computer.

The accident situation search function is used to look for specific accident situations that involve a product extracted from accident situation structure data. These accident situation structure data reports describe time series changes of the accident situation in a graph-structuralized form by utilizing text mining technique.

The similar-risk-product classification function uses a clustering method to identify products that involve similar risks. In the clustering, shape features and the accident types are used as feature vectors.

With these functions, when a user inputs target environment and child age information, the system calculates possible interactions such as "grasping object" using the developmental behavior model by considering the range of products which exist in the target environment. The system also locates accident data related to such products using the graph-structuralized accident database and then outputs possible accidents corresponding to the target child's development stage. In addition, the system attempts to determine the potential product risks using the third function even

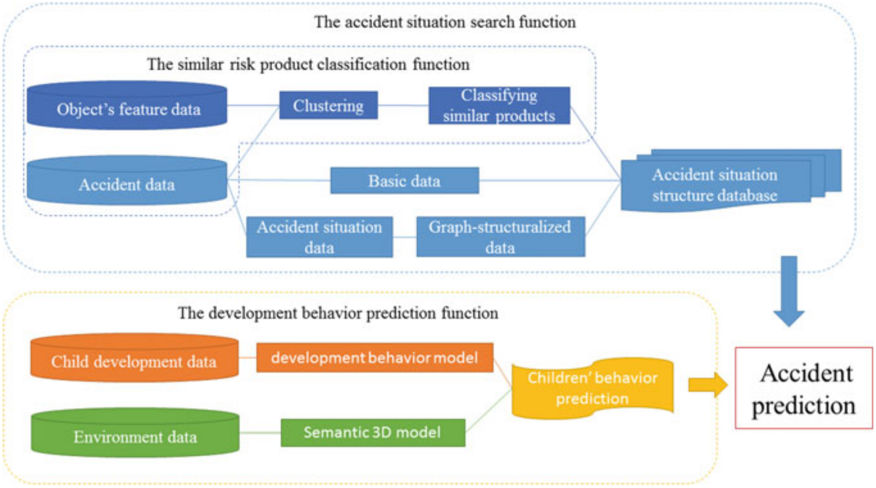

The accident situation search function
The similar risk product classification function
Object's feature data
Clustering
Classifying similar products
Accident data
Basic data
Accident situation structure database
Accident situation data
Graph-structuralized data
The development behavior prediction function
Child development data
development behavior model
Children' behavior prediction
Environment data
Semantic 3D model
Accident prediction

Fig. 5.13 System configuration of Home-injury-situation simulation

if there are very few or no past reports of accidents involving the products in the target environment. This case-based prediction facilitates accident forecasting even if product and children interaction knowledge is insufficient.

In this study, we select accidental ingestion and burn/scald injury as concrete example injuries in order to confirm the effectiveness of the system.

5.5.3 Development Behavior Prediction Function

Since child behavior changes significantly as development progresses, it is necessary to consider developmental stages when predicting child-related accidents. The development-related behavior prediction function visualizes the behavior of children in a virtually constructed environment using the development behavior and semantic 3D models described below.

Touch and climbing behaviors are among the primary causes of accidental ingestion and burn/scald injuries. One example reads, "When an electric cooking plate was being used on the table, a boy climbed onto a chair and touched the edge of the plate, thus burning his finger." This example shows that even if an object is not placed on a floor, it can burn a child if he or she is capable of climbing. Therefore, in the current system, we implemented a function for predicting climbing and reaching behaviors based on body measurement and behavior characteristics collected from more than 2,000 Japanese children.

The statistical data using this database were published as a book for a product designer in 2013. Using this database, we created a behavior model that describes the probabilistic relation between the height of a pedestal that a child can climb to

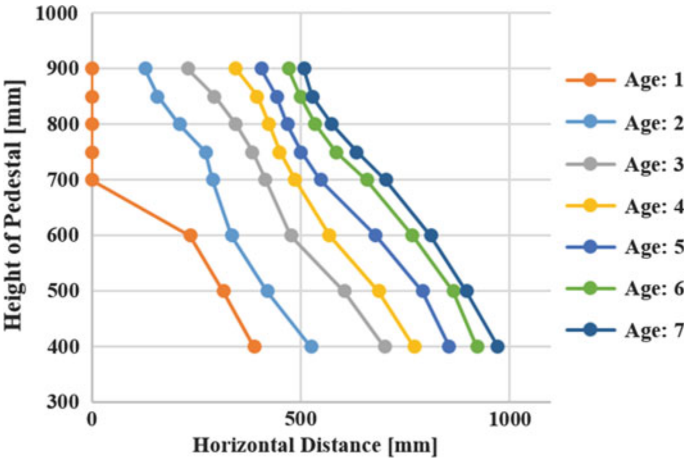


Fig. 5.14 Statistical data on reachable area

and the reachable horizontal distance from the edge of a pedestal. This model allows the system to calculate the probability that the child might touch an object placed at one of a variety of heights. Figure 5.14 shows the relationship between the reachable horizontal distance from the edge of a pedestal and the pedestal height.

When a user inputs information on a target environment, such as a furniture arrangement, as shown in Fig. 5.15, the system can predict the range of child behavior that can occur within the target environment. The user inputs environmental information by constructing and arranging 3D object models in a virtual environment. The system utilizes the 3D game engine Unity to achieve a function suitable for constructing a target 3D environment on a computer. Each 3D object model has semantic information such as the object name and child-related interaction behavior. Figures 5.15 and 5.16 show visualization examples.

Figure 5.15 shows that the child can touch yellow objects and that whether the object is touchable depends on the pedestal height, the horizontal distance from the edge, and child's age. For example, although two-year-old children cannot touch the object put at a height of 800 mm, four-year-old children can touch the object put at a height of 800 mm and a distance of 100 mm from a edge. Figure 5.16 shows that, depending on age, the child can climb to the red top faces.

5.5.4 Accident Situation Search Function

Conventional accident data contain detailed information in a free descriptive sentence format. However, it is difficult to utilize free descriptive data for situation predictions. Recently, our research group has been developing a graph-structuralization-based data mining technique [1] to provide a useful tool for obtaining knowledge on causal

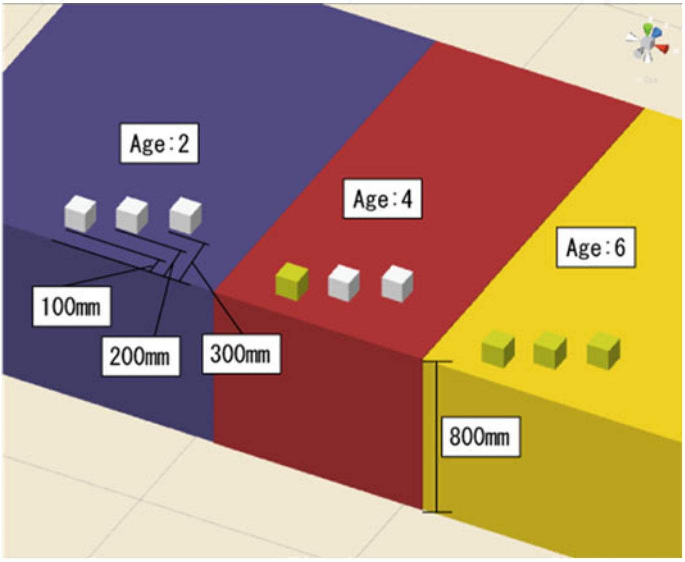


Fig. 5.15 Visualization of reachable objects

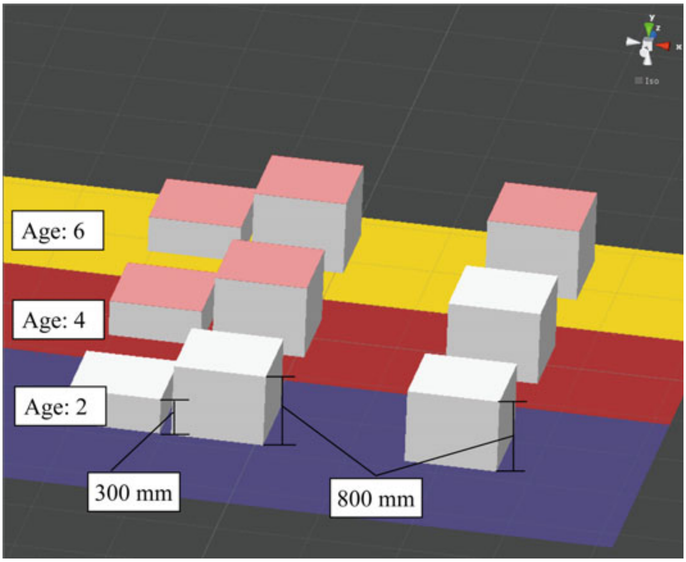


Fig. 5.16 Visualization of climbable places

relationships arising from interactions between objects and human beings. The graph-structuralization-based technique allows data mining by first converting the free descriptive sentence into graph-structured data and then applying a graph analyzing method to the data. Using our software, a user can transform free descriptive data into graph-structured data that express time series relationship changes between agents such as a child and a parent, a product, and interaction behavior with the product. We have also collected over 30,000 childhood injury case data reports in cooperation with hospitals, with which we created an accident situation structure database which consists of the data on 681 burn/scald accidents and 1,221 accidental ingestion incidents. The accident situation search function can be used to find possible accidents from the accident situation structure database by taking into consideration both the child's behavior development stage and the past accident data.

5.5.5 Similar-Risk-Product Classification Function

Objects that cause similar accidents often show shape and characteristic resemblances. For example, objects related to hot water, such as electric kettles and electric pots, can cause burn injuries. Therefore, classification of products from the viewpoint of product characteristics is important for predicting potential risks from products. Such risk predictions allow us to find potential risks even if a new product has not been responsible for any previous injuries. To implement the similar-risk-product classification function, the authors conduct hierarchical clustering using the features of the objects.

5.5.6 Simulation Example of the Accident Situation

Figure 5.17 shows examples of behavior visualization in a target 3D environment. Each simulation was performed using the functions stated above. By visualizing a child's behavior by age, it is possible to check changes in child behavior on the input environment. For example, in Fig. 5.17, although neither a desk nor a chair can be reached at when a child is less than 1 year old, they can both be reached when the child is more than 2 years old.

Figure 5.18 shows an example of similar objects found when an accident situation is input. In this example, the system simulated not only accidents related to tobacco and soup, which exists in the environment, but also those resulting from objects similar to soup, such as boiling water, tea, and heated baby food.

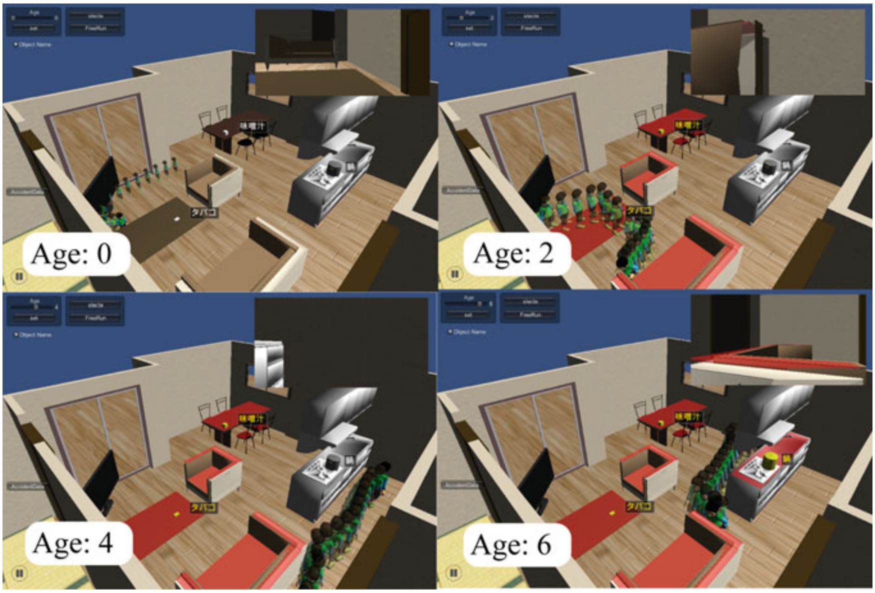


Fig. 5.17 Comparison of accident situation simulations by child age

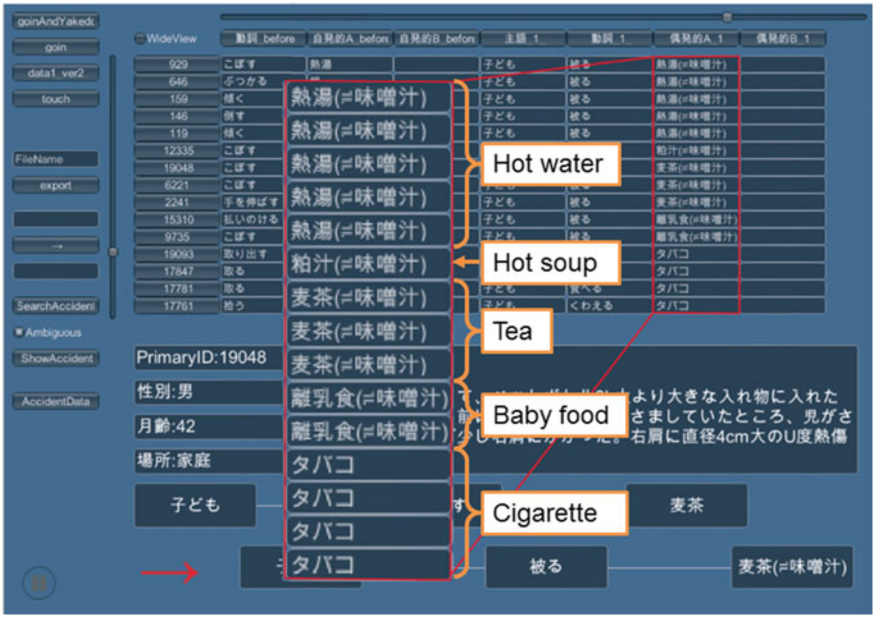


Fig. 5.18 Search for potential risks from objects having similar features

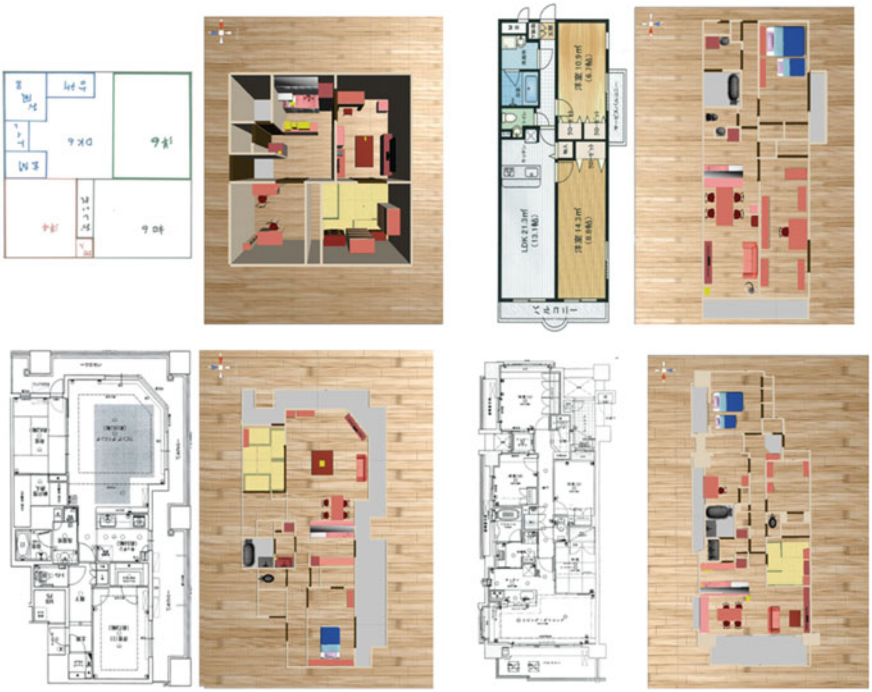


Fig. 5.19 Search for potential risks from objects having similar features

5.5.7 System Verification

To demonstrate the validity of the developed simulation, we reproduced actual ordinary home environments in which accidents had occurred and compared the incident reports with the simulation results predicted by the system. Actual injury data and environmental information were collected during home visit investigations. To date, we have collected such data from 21 ordinary homes where children were injured. At this stage of the evaluation, we selected four environments where burn/scald injuries occurred and one where an accidental ingestion occurred.

The evaluation process proceeded as follows: First, we input environmental information such as the house layout, furniture placement, and the accident situation and conduct a simulation of injury prediction. Figure 5.19 shows the simulated home floor plans and the 3D environmental models created using the information provided in the investigation.

Table 5.1 compares actual data with simulated results. In Table 5.1, the “Product” column indicates a the type of product related to an accident. “Age in accident data” indicates the age of the children when the accident occurred. “Minimum age” indicates the minimum age set in the simulation that children could touch the products that could cause burn/scald and/or accidental ingestion. “Number of accident cases”

Table 5.1 Comparison between actual data and simulation result

Product	Age in accident data (months old)	Minimum age in simulated results (months old)	Number of accident cases
Internal medicine	17	12	10
Pot	57	36	25
Detergent	17	12	21
Stove	50	12	5
Fan heater	11	0	0(8)

indicates the number of accident cases, and the number in the parenthesis indicates the number of accidents due to similar products found by the similar risk product classification function. The minimum age in the simulation is always less than the ages given in the accident data. This suggests that the minimum age set by the simulation was appropriate.

It should also be noted that the simulation succeeded in finding 13 out of 14 accident cases that actually occurred in the environment used for verification in this study. This confirms that the developed simulation works for finding various accident types. The single incident that the simulation failed to identify involved a parent holding a child who grasped an electrical pot located at a high level. Since this incident relates more to the parent's behavior than to the child's, we believe that the simulation is capable of replicating all incidents that a child might cause by his or herself.

5.6 Social Impact Engagement Based on Big Data Analysis in Cooperation with Multiple Stakeholders

5.6.1 Engagement for Preventing Soccer Goal Turnover

The developed system was applied to the Injury and Accident Mutual Aid Benefit System data compiled by the Japan Sport Council, to analyze 1,921 cases of injury involving soccer goals that occurred at elementary and junior and senior high schools in AY2014. Accident circumstances included injury suffered from colliding with a soccer goal, tripping on a soccer goal or net and falling, or transporting, installing, cleaning, hanging from, or jumping into a soccer goal, by a soccer goal overturning by wind, from falling while climbing or sitting on a soccer goal, or by tools or weights used to secure a soccer goal. Some of these accidents were fatal, and in analysis for accidents involving soccer goal overturn, we found 29 [2]. More specifically, the circumstances were as follows:

- A student acting as goalkeeper on the school grounds in a soccer match during a physical education class was overjoyed when a shot flew wide of the goal frame and then hung from the goal, fell, and became pinned under the goal and had one or more teeth knocked out by the goal.
- While playing soccer in a tournament, the goalkeeper was struck in the neck by a goal tipped over by strong wind.
- In the lunch-hour break, a student was playing tag at an outlying area of the sports ground when several other children pulled on the net of a mini-soccer goal, which fell over and happened to hit the student, who was passing by, in the right side of his/her face, bruising the student in the head.

In the analysis, it was possible to roughly identify the circumstances of accidental overturning of the soccer goal, but quantitative determination of the size of the risk in analysis with these data alone was difficult, and it was therefore difficult to quantitatively assess the importance and specific method of preventive measures. In our attempt to determine means of prevention, we therefore measured the impact of the overturning soccer goal and the force required to overturn it. Because a soccer goal overturning accident had occurred when someone hung from the crossbar, we also measured the force on the soccer goal when an individual hung and swung from it.

For two aluminum goals and one steel goal, we overturned each by ropes attached to the crossbar and measured the resulting impact with an impact force gauge holding a load cell sensor mounted on the crossbar where it hit the ground.

In each case, the ropes were pulled gently to avoid imparting a shock load and the pulling was stopped when the soccer goal began to tip over, and the goal was thereafter left to turn over under its own weight. The pulling force was simultaneously measured by a small load cell sensor attached between the ropes.

As shown in Fig. 5.20, for the measured impacts when each goal overturned, the maximum value was 9,521 N for one aluminum goal, 18,980 N for the other, and 29,283 N for the steel goal. The impact of the steel goal was thus found to be 1.5–3 times those of the aluminum goals. Consideration of the relation between impact and injury indicates that the human skull will fracture under an impact of 3,000–5,000 N [3], and the results thus showed that impact by any one of these goals would be sufficient to pose a risk of skull fracture.

As noted above, we measured the force required to overturn a goal in the experiment with a small load cell sensor mounted between the ropes used to pull on the goal. The measurement was performed for an aluminum goal alone and with one of the various weights (from 20 to 80 kg in 20 kg increments) attached to its lower rear bar, with the results shown in Fig. 5.21. With no weight attached, the goal was found to be overturned by the small minimum force of 242.2 N (24.7 kgf), and the pulling force required to overturn the goal was found to increase in an approximately linear correlation with the increase in the attached weight, with a slope of 0.94 when the pulling force was expressed in kilograms. This was approximately equal to the 0.89 ratio of the 223 cm length of the rearward-directed bar relative to the goal post height of 250 cm, thus indicating that the goal post lower end functioned as the fulcrum in the principle of the lever.

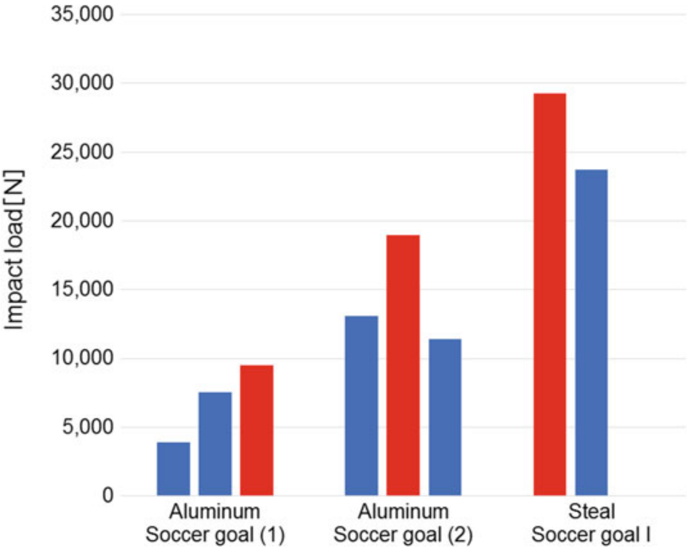


Fig. 5.20 Impact of overturning soccer goal

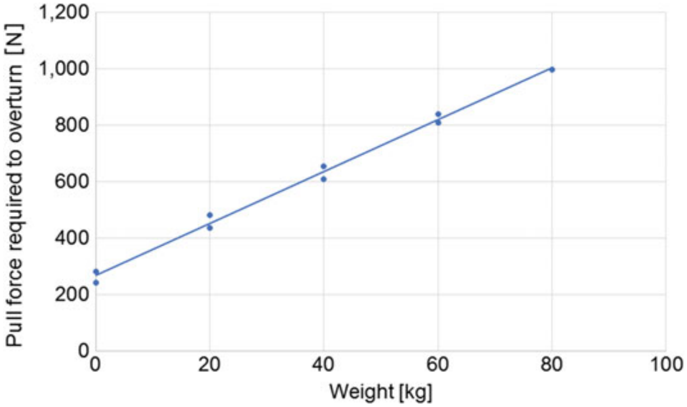


Fig. 5.21 Force required to overturn soccer goal

The most common circumstance of soccer goal accidents at schools is that of a child hanging and swinging forward and rearward from the goal crossbar. The horizontal load required to overturn the goal in such circumstances was simulated and measured in an experiment with a constructed steel-post assembly in which a biaxial load sensor was attached to each of the two ends of the horizontal bar and the horizontal and vertical loads were measured. The experiment was performed as one of 10 cooperating junior high school students hung and swung. Figure 5.22 shows the maximum horizontal loads found in the trials. Overall, the maximum applied force found for any of the forward and rearward swinging was 405.4 N (41.4 kgf).

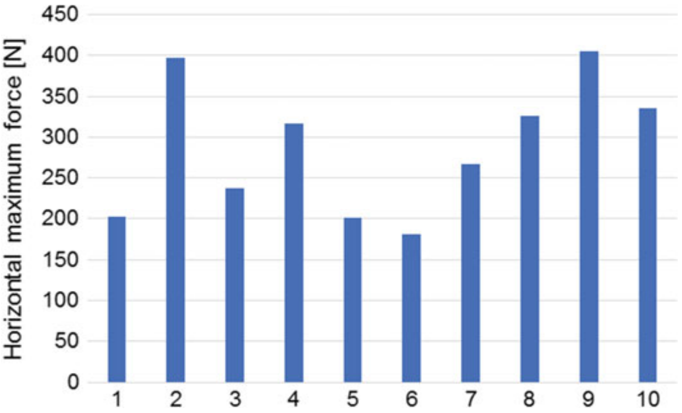


Fig. 5.22 Horizontal load with forward and rearward swinging

Taken together, the results indicated that the force of crossbar impacts near ground level when the goal overturned ranged from a minimum of 3,887 N to a maximum of 29,283 N and thus posed a high risk of causing skull fracture. It was found that an aluminum goal was overturned by a small force of 242.2 N (24.7 kgf) and that a child hanging and swinging forward and rearward imparted a horizontal force of 405.4 N (41.4 kgf) on the crossbar, and thus, it was found that a soccer goal will be readily overturned by the swinging action of just one student if not securely fastened down or having movement curtailed by a mounted weight.

These results have been presented at symposia, and the specific data have been shown and led to consciousness-raising activities.

5.6.2 Engagement for Preventing Vaulting Box Accidents

Analysis of 97,716 accidents relating to elementary school exercise activities recorded in the Injury and Accident Mutual Aid Benefit System data of the Japan Sport Council in AY2014 showed that vaulting box exercise accidents were most numerous [4]. They numbered 14,715 and thus accounted for approximately 15% of the total accident number. Among injuries suffered in vaulting box accidents, bone fractures were most numerous and accounted for approximately 37% of all injuries. The circumstances of vaulting box accident occurrence include run-up, takeoff, time from start to end of hand contact, landing, and forward somersault on platform, with accidents occurring in the largest number during the time from start to end of hand contact. Data analysis showed that many bone fractures occurred in the vaulting box exercise, again with most occurring during the time from start to end of hand contact. Further details are lacking, however, and in the present state of data on accident circumstances or child movements, application to injury prevention would be difficult.

We therefore performed observation and pattern classification of the relationship between vaulting box vaulting, and the risks involved in actual classes, in collaboration with Toshima Ward Fujimidai Elementary School and physical therapists. The patterns found included low momentum in takeoff, incorrect arm support, and insufficient center of gravity movement resulting in contact of buttocks with hand on vaulting box and leading to wrist sprain or failure to vault from vaulting box and impact of buttocks on vaulting box, and concentration on forward movement alone leading to loss of balance and falling on landing. Based on this analysis, we have developed a system that shows vaulting with risk of accident, vaulting action checkpoints, and practice methods for correction of ineffective moves (Fig. 5.23) and will proceed with its evaluation and modification through actual utilization at elementary schools.

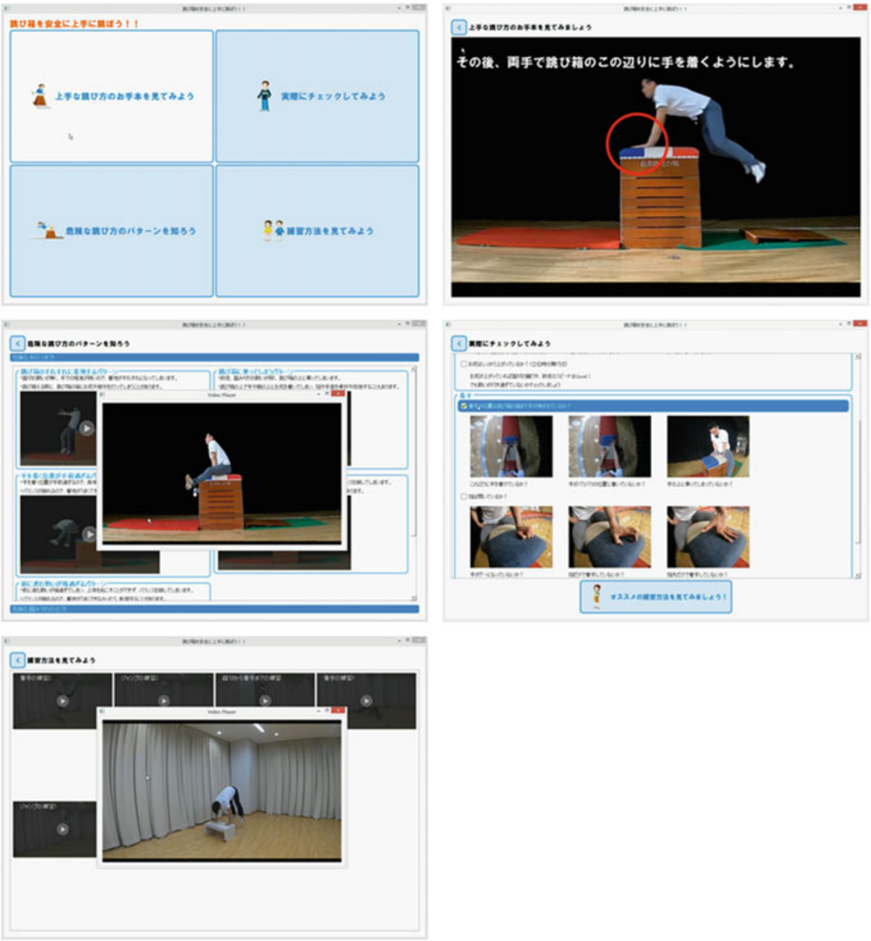

跳び箱を安全に上手に跳ぼう！！
上手な跳び方のお手本を見てみよう
実際にチェックしてみよう
危険な跳び方のパターンを知ろう
練習方法を見てみよう
上手な跳び方のお手本を見てみましょう
その後、両手で跳び箱のこの辺りに手を着くようにします。
危険な跳び方のパターンを知ろう
Video Player
実際にチェックしてみよう
オススメの練習方法を見てみましょう！
練習方法を見てみよう
Video Player

Fig. 5.23 Software supporting guidance on vaulting box safety

5.7 Conclusion

In this report, we have described trend analysis functions important for advancement of school safety in application to multi-organizational dispersed data utilizing basic security technology and performance analysis of actual judo accidents at schools. In problems elucidated through use of the system under development in this study, we have engaged in acquisition and analysis of detailed data necessary for injury prevention and described our engagement in studies on accidents in soccer goal overturning and vaulting box activities.

We will further apply our system currently under development to actual sites of activity while further advancing verification and investigate ecosystems for performance of injury prevention in actual on-site utilization of the system.

References

1. A. Hirata, K. Kitamura, Y. Nishida, Y. Motomura, H. Mizoguchi, Accident-data-aided design: visualizing typical and potential risks of consumer products by data mining an accident database, in *Proceedings of 2013 IEEE/SICE International Symposium on System Integration* (2013), pp. 376–381
2. T. Yamanaka, K. Kitamura, M. Oono, Y. Nishida, Importance of anchoring soccer goal posts based on scientific evidence, in *Injury Prevention*, vol. 24, supp. 2 (2018), pp. A12
3. N. Yoganandan, F.A. Pintar, Biomechanics of temporo-parietal skull fracture. Clin. Biomech. **19**(3), 225–239 (2004)
4. M. Oono, Y. Nishida, K. Kitamura, M. Nonoyama, H. Saito, H. Itakura, Prevention of vaulting box-related injuries in PE class: developing an educational tool by collaboration between physical therapists and artificial intelligence, in *The 9th Asian Regional Conference on Safe Communities in Atsugi* (2018)

Chapter 6
Health Test Bed Group

Katsuya Tanaka and Ryuichi Yamamoto

Abstract Under the new law for the secondary use of medical information, which was activated in May 2018, the future expected secondary use with information anonymization may contribute to research and development in the medical field of integrated medical research and public health. On the other hand, under the revised Personal Information Protection Law and the revised ethical guidelines in medical research, privacy protection and patient consent management is a crucial issue for the management of researches. Our JST CREST project, which started in March 2014, has issued the development of technological elements and synthesized the developed methods for real-world system for the secondary use and privacy protection of big data on cloud infrastructure, including safe clinical information management, commercial cloud utilization, and privacy risk evaluation. In this paper, assuming the utilization of the Standardized Structured Medical Record Information Exchange version 2 storage, the following target issues are described: (1) effective utilization of existing standardized storage, (2) secure data collection across medical institutions, (3) privacy risk evaluation in analysis, and (4) traceability while secondary use.

K. Tanaka (✉)
National Cancer Center Japan, 5-1-1 Tsukiji, Chuo-ku, Tokyo 104-0045, Japan
e-mail: katstana@ncc.go.jp

R. Yamamoto
Medical Information System Development Center, Kagurazaka 1-1,
Shinjuku-ku, Tokyo 162-0825, Japan
e-mail: yamamoto@medis.or.jp

A. Miyaji and T. Mimoto (eds.), *Security Infrastructure Technology for Integrated Utilization of Big Data*,
https://doi.org/10.1007/978-981-15-3654-0_6

6.1 Overview of Legislation and Standardization for the Secondary Use of Electronic Medical Records

6.1.1 Personal Information Protection Act and Next-Generation Medical Infrastructure Act

The Personal Information Protection Act [1] was revised in September 2015 and was fully enforced in May 2017. Prior to the revision, the Personal Information Protection Act was established in 2003 and fully enforced in 2005. At the time that the previous law was established, both Houses of Councilors recognized that it was insufficient, with the establishment of separate laws being required in multiple fields, including medicine, and it was not actually reviewed. This situation remained for a decade or more. It is now high time for it to be revised. Consequently, several problems in the previous law were improved; however, a few problems still exist, and some new concerns have emerged. These include fears that secondary use, which is essential in the medical treatment field, will become problematic.

To avoid a negative impact on innovation, including drug discovery and medical equipment development, the Next-Generation Medical Infrastructure Act (official name: Act on Anonymously Processed Medical Information to Contribute to Medical Research and Development) [2], specializing in the secondary use of medical data, was enacted in April 2017 and enforced on May 11, 2018. In this paper, we investigated the issues related to the Personal Information Protection Act and the predicted effects of the revision to the law, provided an overview, and considered the impact of the Next-Generation Medical Infrastructure Act.

6.1.1.1 The Personal Information Protection System in Japan and Related Issues

The main objective of reviewing the previous law was to respond to the EU directive concerning cross-border personal information in 1995 and the concerns about privacy infringement as a result of the resident registration network brought about by revisions to the Basic Resident Registration Act. Furthermore, there were concerns regarding the prospect of eavesdropping being made possible with court approval through revisions to the Criminal Procedure Code. Also, as previously stated, this was fully enforced in April 2005. Basically, this conformed to the OECD personal information cross-border guidelines [3]; however, allusions to several issues have been identified. The problems in the medical area include the following facts: the law is a comprehensive one that does not specify the field; the characteristics of medicine frequently provided to third parties, an essential purpose, are not being considered; the definition of personal information is ambiguous, which inevitably makes anonymization difficult; and focus is placed on protection, and so, the promotion of reuse that does not violate the right of the individual, which was the original purpose of the law, is largely ignored. There is also the fact that, in a private sense, it is aimed at the operator, and where there is an individual causing the infringement, it is an indirect regulation concerning supervision by the operator. Moreover, the penalties are light and indirect and thus lacking in effectiveness. The different systems for acquiring personal information are enforced for governments, independent administrative agencies, local governments, and private enterprises, and this is considered to obstruct the utilization of personal information across these frameworks.

In the revised Personal Information Protection Act, the concept of important information has been introduced, and as the vast majority of medical data is designated as important information, this is a step forward from a comprehensive law that does not specify fields. When acquiring important information, explicit consent is required, and third-party provision based on opting-out, which can occur when providing such information to third parties while the intentions of the person concerned are still unclear, is prohibited. This is clearly a step forward and promises to suppress provision to third parties where this is not intended by the person concerned. On the other hand, in the case of third-party provision, which is essential in collaborative medicine, while there were concerns about the explanation of symptoms to family members, consultations with specialists, etc., this is largely covered by the clear definition of opt-out consent as "implicit consent" in guidance concerning the appropriate handling of personal information by healthcare and nursing care providers (hereafter, "healthcare and nursing care guidance"), implementation guidelines issued jointly by the Personal Information Protection Commission (hereafter, "PPC") and Ministry of Health, Labor and Welfare (hereafter, "MHLW"). However, the fact that, under law, clear consent is required for provision to third parties for the purpose of drug discoveries for the development of medicine or medical equipment remains unchanged. Gaining clear consent places a considerable burden on medical sites, and even where there may be no intention to violate rights, it must be considered that this is significantly more problematic than with the previous law. In the revised law, the concept of anonymized information has been introduced, and by anonymizing data in accordance with the standards of the PPC, this may be provided without consent under certain conditions.

However, it is necessary to impose the conditions of prohibition on reidentification and safety management on the recipient of the data, and it is procedurally complex to make third-party provision with anonymized information a public duty. Additionally, to meet the anonymization standards of the PPC, a certain amount of information processing capability is required, which is not simple. While this is not a legal item, in regard to important information, another feature of the revised law is that traceability must be secured. Moreover, although a significant impact is feared in healthcare and nursing care fields where there is frequent provision to third parties, in the healthcare and nursing care guidance, this is virtually all considered as essential for healthcare and nursing care, thus avoiding a major increase in the workload of the healthcare and nursing care institutions. On the other hand, in regard to provision to third parties involved in secondary use that is not essential for healthcare and nursing care, the creation of records and their confirmation at the time of receipt are required. For genetic information, as well, having a personal identifier code specified from which the individual can be identified, provided certain conditions are met, has immense medical significance.

The points alluded to earlier are all features of the previous and revised law as seen from the perspective of the healthcare and nursing care field. Furthermore, as the responsibility for enforcing the revised law has been centralized in the PPC and penalties have been significantly increased, it has become more effective. Major changes, such as the conditions for distributing personal information overseas being clarified, have been determined, but these will only be listed in this chapter.

The revised law promises to improve several issues in the previous law. The strengthening of punitive measures increases its effectiveness, and the introduction of the concept of important information reduces discrimination based on the illegal use of special personal information, preventing its use through provision to a third party not intended by the person concerned. However, several issues remain unresolved. The first of these is that, as operations are based on different regulations from the government, independent administrative bodies, local governments, and private enterprises, there are about 1,800 autonomous bodies and close to 2,000 statutes. Certainly, there are not any major differences in their basic thinking, but the executing body varies depending on the statute, and subtle decisions are made by each executing body. In the case of healthcare and nursing care, the body is a private company, but the local government and institutions at its rank often contribute, as do national institutions and independent administrative bodies.

For example, if one prefectural, two city, and two town hospitals and five private medical institutions collaborate to share organic patient data, it will be necessary for at least four autonomous bodies to review whether this is possible. For healthcare providers looking to move forward, this can become a significant burden. Currently, the fact that the statute may be different depending on the acquiring body has not been improved at all. For hereditary information, it is expected that genetic information will be specified with a personal identification code and handled prudently; however, under the Personal Information Protection Act, consent gives an absolute pardon. On the other hand, in the case of hereditary information, even if the person providing the information provides consent, the impact of such may extend to blood relatives such as parents and offspring. If, as a result of a parent's consent, a child became the victim of discrimination, this could not be handled under the Personal Information Protection Act. At the current time, improving this point seems to be not possible, and several people indicate that this is an issue. This should be reviewed in the near future, and it is to be hoped that it will be resolved quickly.

6.1.1.2 Review and Establishment of the Next-Generation Medical Infrastructure Act

As previously described, with the revisions to the Personal Information Protection Act, although several of the issues in the previous law were improved, secondary use, where there is no intention to violate personal rights and the aim is to use personal information for the public good, was previously possible via opt-out. However, with the revisions, this is no longer possible. Healthcare and nursing care must be performed based on medicine, but this cannot develop without the use of patient and user data. Immediately utilizing research results obtained using the laboratory or animals in medicine and healthcare is not possible, and human knowledge is essential. In other words, if this type of usage is suppressed, the acquisition of medical knowledge itself will likely be suppressed, and this may obstruct the development of healthcare and nursing care itself. If medical institutions and nursing care providers are able to anonymize, they will be able to provide data for secondary use without consent. However, in the case of regional comprehensive care and collaborative medicine, information is distributed between multiple operators. Therefore, unless information can be concentrated in a single institution through a joint use declaration, linking and anonymizing the disparate information will not be possible. Anonymization makes reidentification impossible, and so anonymized information cannot necessarily be linked. The simple solution would entail making a joint declaration of use; however, in this case, the perimeter of information for joint use and other information must be clarified. In Japan, the healthcare and nursing care services can be freely chosen by patients and users, so setting the perimeter is essentially difficult. Additionally, it is necessary to announce the fact that anonymization is taking place, and this cannot be provided without restriction. A prohibition on reidentification is sought from the recipient, and although this is effort based at best, safety management is also required. The provider has no duty to supervise the destination, but if an incident or illegal use occurs, the complaint from the individual embodying the information will be directed at the providing medical or nursing care institution, which may result in a civil lawsuit. Supervision may be considered to be mandatory. Although it is not impossible, some preparedness and effort are required. However, it is not desirable that this situation impacts the development of medicine/medical equipment or drug discovery. Regarding academic research, in Chap. 4 of the revised law, it states that although various duties are not placed on the operators acquiring the personal information, this is limited to academic research by academic research institutions, and although there are calls to draw up and execute guidelines for those not covered by Chap. 4 of the revised law, such guidelines are difficult to implement on a statutory basis.

Faced with this situation and the awareness of the need to promote use for the public good that does not violate the rights of the individual, the Cabinet Secretariat and Office of Healthcare Policy primarily reviewed the measures, whereupon the Next-Generation Medical Infrastructure Act was submitted to the Diet as Cabinet legislation. The basis of this held that if operators with the ability to perform reliable and safe anonymization, who were able to provide safe information for the public good in a broad sense, were accredited and medical information was provided from the accredited operators to the medical institutions, consent could be provided via an opt-out system. This was established at the end of April 2017 and delivered in May.

6.1.1.3 Content of the Next-Generation Medical Infrastructure Act

This law focuses on accredited anonymizer medical data creation operators, and as previously described, operators who can perform anonymization reliably and handle and provide information safely are accredited by the government. The law intends, "through the safe and appropriate utilization of anonymized medical data, to promote cutting-edge R&D related to health and healthcare, and new industries, and contribute to the development of a society where people live healthy and long lives." The aim is not simply commercial use but use for the public good in a broad sense. Although its scope is narrow, it is positioned as an individual law from the Personal Information Protection Act, and this overwrites such Act.

Definition of Wording

This law is not aimed at general personal information but "medical data." The main target is the information related to healthcare, which is a type of important information, and the definition has been slightly expanded. The Personal Information Protection Law covers the information of living individuals, but the Next-Generation Medical Infrastructure Act includes medical information on deceased people as well. In healthcare, life and death exist consecutively, and so, this can be considered to be a reasonable extension. Additionally, in the revised Personal Information Protection Act, the guidelines for anonymization are indicated by the PPC, whereas the guidelines for the anonymizer medical data in the Next-Generation Medical Infrastructure Act are provided by the minister in charge. However, the wording in the definition is the same, and the law clarifies that this should be determined after consultation with PPC. Anonymizer information and anonymizer medical information are basically the same, but with the latter, it is possible to provide detailed guidelines depending on the case in which it is used.

Accredited Anonymizer Medical Data Creation Operators

The core of this law is the stipulation of accredited anonymizer medical data creation operators. This is limited to companies who possess appropriate anonymizer capabilities and can provide information to operators who can handle the safe anonymizer information in accordance with the law. The anonymizer work of such operators does not apply to stipulations regarding the creation of anonymizer information in Article 36 of the Personal Information Protection Act. Additionally, the safe management of information and an appropriate response to this are required. This also does contravene the concept that this is provided to contribute to R&D in the medical field, and use that exceeds the scope of achieving the objectives of the accredited operator is not recognized.

With this, no particular restriction is noted on the operator other than the accredited work. Additionally, provided that the information before the anonymization was for the operator to create the anonymizer medical data, it may be provided to other accredited anonymizer medical data creation operators within the scope of that purpose. In this way, if, for example, accredited operator A is mainly accumulating hospital information and operator B is mainly collecting clinic information, it is possible for A to provide to B and B to provide to A and create anonymizer medical data after linking the medical information of the clinics and hospitals.

Operators Handling Medical Data
This refers to medical institutions, and broadly speaking, two types of regulations when providing medical data to accredited anonymizer medical data creation operators are described. The first is notice to the patients and notification to the minister in charge. In the notice, it must be clarified that provision shall be stopped if there is a request for such from the patient or a bereaved family member. A point to note here is that this is just described as "notice" to the patient. Simply presenting it is not enough, and the content of the notice must be actually notified to the patient, etc. The second point is that if provision is stopped due to the request of the person concerned or the bereaved family, there is a duty to issue evidence in writing that there has been a request to stop provision, and a copy of this must be stored. In case there is a request to stop the provision of medical data owned by the accredited anonymizer medical data creation operator, this information may not be received.

Operators Handling Anonymizer Medical Data
Recipients provided with anonymizer medical data from the accredited anonymizer medical data creation operators are exempt from the stipulations of Articles 37 (provision of anonymizer information), 38 (prohibition on identification action), and 39 (safety management measures, etc.) of the Personal Information Protection Act. On the other hand, in the Next-Generation Medical Infrastructure Act, reidentification itself is prohibited. This should not just be a penalty stipulation for "operators handling anonymizer medical data," and the restriction of agreements with accredited anonymizer medical data creation operators is also necessary. If an actual breach occurs or the agreement conditions lack effectiveness, the application of the Unfair Competition Prevention Act should also be considered.

Accredited Medical Data Handling Contractors
Operators undertaking the work of accredited anonymizer medical data creation operators need to be accredited by the government.

6.1.1.4 Opting-Out Under the Next-Generation Medical Infrastructure Act

Provision to third parties is specified with opt-out under Article 23, paragraph 2, of the Personal Information Protection Act. When providing to a third party after notifying the party concerned or in a situation where the person concerned could easily learn of the fact, provided that there is no motion of refusal from the person concerned, it may be provided to a third party. Originally, this was prohibited in cases involving sensitive information, and so, medical data cannot be provided to a third party in this way. In contrast, in the Next-Generation Medical Infrastructure Act, as long as the third-party provision is to an accredited anonymizer healthcare information creation body, an exception shall be granted, and this may be provided in an opt-out form. However, it is only permitted to be provided to a third party after notifying the person concerned if there is no motion of refusal. In other words, it just being a situation where they could easily learn of the fact is not enough.

6.1.1.5 Safety Management Measures

The safety management measures section stipulates the safety management measures to be taken by accredited anonymizer healthcare information creation bodies, and the contents are as follows:

1 Purpose and target of application
2 Concrete measures
2-1 Organizational safety management measures
2-2 Human safety management measures
2-3 Physical safety management measures
2-4 Technical safety management measures
2-5 Other measures

These can be considered to be typical chapter headings, and the majority of these are not particularly different from the MHLW "Security Guidelines for Medical Information Systems." However, the network is limited to dedicated lines and IP-VPNs within accredited operators. In terms of availability, while the superiority of dedicated lines is unquestionable, as they are not clearly superior in regard to completeness or anonymity, implementation may be difficult when cost is considered.

6.1.1.6 The Future of the Next-Generation Medical Infrastructure Act and Issues

If the Next-Generation Medical Infrastructure Act functions as intended, the regulations strengthened in the revised Personal Information Protection Act can be introduced in a form without the risk of violating the rights of individuals. Moreover, regarding the purpose restricted to R&D in the medical field, there are expectations that it can be promoted in a safe and significant manner. However, two main issues are identified. The first is the establishment of the system itself. Although the law has been established, we are still waiting for the establishment of a basic policy as well as the government and ministerial ordinances delegating the main part of this work. At present, only the outline has been fixed. We do not yet have a system with "meat on the bones" to be used in actual operation, and efforts by all related parties are required. Additionally, it is expected that there will be public comments once a draft of the government and ministerial ordinance or guidelines is determined, and hopefully, several people will have constructive comments from many people. The second issue is that although the accreditation of anonymizer medical data creation operators is a public work contributing, in a meaningful way, to R&D in the medical field, the law presumes the accreditation of private operators. In other words, the accredited operators must both maintain their own survival and continue work with significant public work elements. This is certainly not simple for a private operator. Unless the accredited operator can gain trust, the medical institutions, etc., will lose enthusiasm for the provision, and the system itself may fail. If we consider the world aimed for by this law to be significant, the support of not only the administration and operators aiming for accreditation but also a wide range of people, including medical-related parties and patients, is required.

Even if the aforementioned issues can be safely overcome and operations begin, problems will remain. We have repeatedly indicated that this framework is based on accredited anonymizer medical data creation operators who are private companies. However, at present, the government, local governments, and insurers are systematically accumulating information, and much of the useful medical data is owned by the government. For example, information on life and death is the ultimate outcome of treatment, and to determine this outcome with certainty, basic resident registration information and death certificates, etc., must be accessed. Although according to this law, cooperation on consent with accredited anonymizer medical data creation operators is possible, there is no consideration at all regarding collaboration on the information owned by the government, local governments, and insurers under this system. Despite the fact that R&D in the medical field is urgent in terms of maintaining social security, it must be indicated that there is a problem in terms of efficiency, based on the Next-Generation Medical Infrastructure Act alone. It is considered necessary to establish an external system to promote a comprehensive system for using information for the public good for government and private enterprise. Additionally, the security and anonymizer standards are somewhat abstract. In the case of technology that uses individual data with Privacy Preserving Data Mining and multiparty protocols in its anonymous form for calculation purposes only, despite the fact that it has been demonstrated that a technical solution is possible, as no consideration has been given from a statutory or system viewpoint, it is difficult to judge whether this can be used under the Personal Information Protection Law. While promoting technical initiatives, it is also necessary to clarify positioning in a statutory and system sense.

6.1.2 Ethical Guidelines and Anonymization of Medical Information

6.1.2.1 Ethical Guidelines

The Ethical Guidelines for medical and health research involving human subjects [4] apply to medical research for human beings and basically requires researchers to respond to the request sought by the Act on the Protection of Personal Information. However, Chap. 4 of the Personal Information Protection Law shown in the following is exempt from clinical research:

> Chapter IV Obligations, etc., of a Personal Information Handling Business Operator
> Section 1 Obligations of a Personal Handling Business Operator
> Section 2 Obligations of an Anonymously Processed Information Handling Business Operator, etc.
> Section 3 Supervision
> Section 4 Private Sector Body's Promotion for the Protection of Personal Information

It also applies to the information infrastructure for collecting and analyzing medical information that this project aims to build. In large-scale data collection research, specific responses required by the Ethics Guidelines are mainly described in the informed consent. The description in the Ethics Guidelines is as follows:

> Researchers do not necessarily need to receive informed consent however, if you do not receive informed consent, the subject of the study appropriate consent of the However, in cases where it is difficult to obtain appropriate consent, information used in other studies to conduct research. from 4(1) to (6) for the implementation of the research, if there is a particular reason for to be notified or published to the subject of the research, and to ensure that the research is carried out or continued. opportunities for research subjects, etc., to be denied. personal information may be used.

Also, the Ethical Guidelines need to notify or publish the following matters to the patient, etc:

> (1) The purpose of use and use of samples and information (including methods when provided to other organizations)
> (2) Items of samples and information used or provided
> (3) Scope of use
> (4) Name or name of the person responsible for the management of samples and information
> (5) To use samples and information to identify the subject of the research or to other research institutions at the request of the research subject or its agent
> (6) (5) How to accept the request of the subject or its agent

This is also true for information systems that deal with large-scale data. Furthermore, if the target personal information is the anonymized one, it is not necessary to notify the patients. At this time, the opportunity of the consent withdrawal is not guaranteed to the patient.

In fact, regardless of the presence or absence of anonymization processing for electronic medical record (EMR) items, in most cases, the content of research that uses medical information is made public on the homepage of each medical or research institution and patients. It is difficult for patients to understand how their own EMR items are used and provided.

6.1.2.2 Anonymizer Medical Data

Anonymizer medical data is an extension of the anonymizer information under the Personal Information Protection Act. Under this Act, the target information is limited to personal information surviving, while under the Next-Generation Medical Infrastructure Act, the information of deceased individuals may also be covered, depending on the situation. Additionally, anonymizer information is information from which the

Table 6.1 Contents of guidelines for anonymizer information

1 Positioning of these guidelines
2 Definition
2-1 Medical information
2-2 Anonymizer medical data (related to Article 2, paragraph 3)
2-3 Anonymizer medical data creation business
3 Duties of accredited anonymizer healthcare information creation bodies and bodies handling anonymizer healthcare information
3-1 Thinking behind duties regarding the handling of anonymizer medical data
4 Processing required when creating anonymizer medical data
4-1 Processing standards for anonymizer medical data
4-1-1 Deletion of descriptions, etc., from which specific individuals may be identified
4-1-2 Deletion of individual identification codes
4-1-3 Deletion of codes that interconnect information
4-1-4 Deletion of peculiar descriptions, etc.
4-1-5 Other measures based on the nature of medical information databases
4-2 Items requested for investigation when creating anonymizer medical data
4-2-1 Format for using anonymizer data
4-2-2 Possibility of identification by referring to other information
4-3 Anonymizer medical data creation process
4-4 Method of anonymization based on medical data categories
4-5 Medical data-specific anonymization
4-5-1 Medical images
4-5-2 Genome data
5 Safety management measures, such as anonymizer medical data
6 Prohibition on identification actions
7 Provision registration

individual cannot be identified by ordinary people, whereas anonymizer medical data is information from which the individual cannot be recognized by general healthcare-related people. As this fulfills the stipulations of the guidelines on anonymizer information determined under the Personal Information Protection Act, processing based on additional risk analysis is required. Furthermore, another characteristic can be considered to be the fact that even after the provision of the information, there is a duty to follow up, including confirming how it is used. The content of these guidelines is shown in Table 6.1.

Another feature is that medical information is categorized from a risk perspective, which is shown in Tables 6.2 and 6.3.

Table 6.2 Categorization of the risk of individual identification in medical data

Category	Overview
Identifier	Information directly linked to an individual (name, number of insured, etc.)
Quasi-identifier	Information that, when multiple types are combined, can lead to the identification of the individual (date of birth, organization, etc.)
	*The medical institution code is considered a quasi-identifier
Static attributes	Highly invariant information (height, blood type, allergies, dates [such as consultation dates], etc.)
	Information related to external characteristics such as disabilities
	*Handling of information on chronic illnesses with a high level of invariance needs to be reviewed
Semi-static attributes	Data with universality for a fixed period (weight, etc.)
	It is assumed that this relates to information on diseases, procedures, administered medicines, etc.
Dynamic attributes	Information that is constantly changing (data on inspection values, food, other treatment, etc.)

Table 6.3 Anonymizer examples through the categorization of medical data

Category	Example of anonymization method
Identifier	Deleted or irreversible pseudonymization
Quasi-identifier	Generalization (date of birth -> year born, address -> prefecture) or micro application that satisfies k-anonymity
	Delete data items
	Add attributes (geographical, scale, etc.), such as medical institution codes, and convert codes into an unidentifiable form
Static attributes	Numerals are top-to-bottom coding
	Generalization or micro application
	Generalization or offset based on treatment date, etc.
Semi-static attributes	Numerals are top-to-bottom coding
	Delete sensitive diseases when not necessary
Dynamic attributes	Anonymization not required, but where necessary, numbers are top-to-bottom coding
	In consideration of the significance of abnormal values, look at the distribution of values and carry out processing, such as rounding of upper and lower % values

6.1.3 *Standardization of EMRs*

The Standardized Structured Medical Record Information Exchange (SS-MIX) [5, 6] aims to promote/develop the results of the standardized electronic medical chart information exchange system development commission project conducted by the Health Policy Bureau of the MHLW in FY 2006 in Japan.

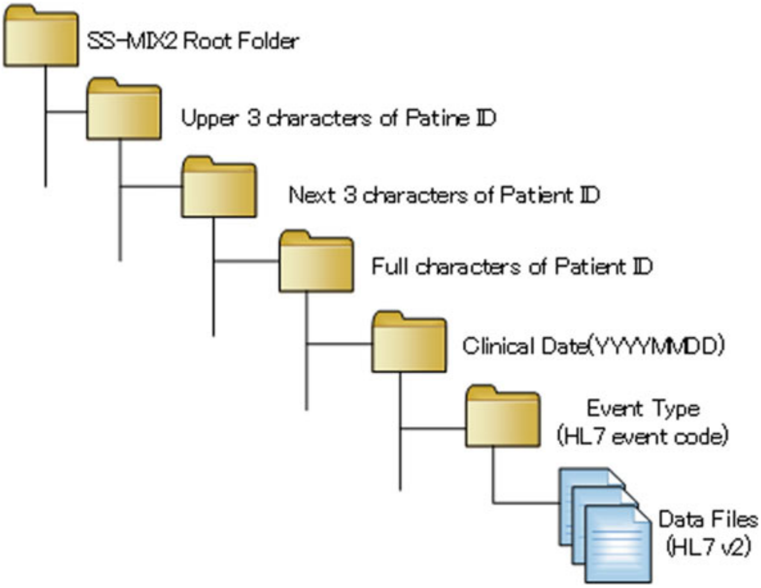

SS-MIX2 Root Folder
Upper 3 characters of Patine ID
Next 3 characters of Patient ID
Full characters of Patient ID
Clinical Date(YYYYMMDD)
Event Type
(HL7 event code)
Data Files
(HL7 v2)

Fig. 6.1 Overview of SS-MIX2 directory layout

SS-MIX includes the following:

(1) Hospital information system information gateway telegraphic message specification
(2) "Standardized Storage Specification" directory structure
(3) Electronic medical information CD and patient referral document CD specification

Furthermore, the scenes where the utilization of this standardized storage is expected are as follows:

Ensuring continuation of medical information Repository in community healthcare coordination Information sharing among multiple vendors Utilization as backup information

Figure 6.1 shows the file system layout of SS-MIX2 storage. Directories are sorted by patient ID, clinical date, and event type.

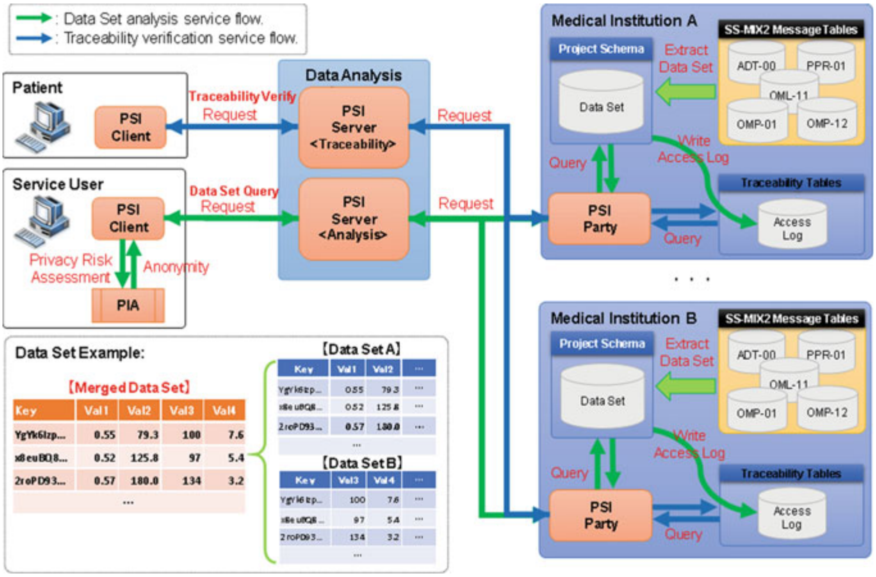
Table 6.4 shows the clinical event types covered by SS-MIX2 storage represented by HL7 v2. We can represent 30+ clinical events using this storage [6].

6.2 Medical Test Bed Concepts and Requirements

Considering the current situation surrounding medical information as described earlier and the development of future utilization, the public cloud is used for the secondary use of medical data scattered through medical institutions across the organization. We are developing a secure information utilization base test bed in the medical information field, assuming the utilization promotion by adopting.

Table 6.4 SS-MIX2 data types

No	Data type	Name	HL7 message type
1	ADT-00	Update of patient’s basic information	ADTˆ08
2	ADT-00	Deletion of patient’s basic information	ADTˆ23
3	ADT-01	Change of investigator	ADTˆ54
4	ADT-01	Cancellation of investigator	ADTˆ55
5	ADT-12	Reception of outpatient physical examination	ADTˆ04
6	ADT-21	Hospitalization plan	ADTˆ14
7	ADT-21	Cancellation of hospitalization plan	ADTˆ27
8	ADT-22	Conduct of hospitalization	ADTˆ01
9	ADT-22	Cancellation of conduct of hospitalization	ADTˆ11
10	ADT-31	Conduct of staying outside	ADTˆ21
11	ADT-31	Cancellation of conduct of staying outside	ADTˆ52
12	ADT-32	Conduct of return from staying outside	ADTˆ22
13	ADT-32	Cancellation of conduct of return from staying outside	ADTˆ53
14	ADT-41	Plan of change of department/building (change of room/bed)	ADTˆ15
15	ADT-41	Cancellation of plan of change of department/building (change of room/bed)	ADTˆ26
16	ADT-42	Conduct of change of department/building (change of room/bed)	ADTˆ02
17	ADT-42	Cancellation of conduct of change of department/building (change of room/bed)	ADTˆ12
18	ADT-51	Plan of discharge	ADTˆ16
19	ADT-51	Cancellation of plan of discharge	ADTˆ25
20	ADT-52	Conduct of discharge	ADTˆ03
21	ADT-52	Cancellation of conduct of discharge	ADTˆ13
22	ADT-61	Registration/update of allergy information	ADTˆ60
23	PPR-01	Registration/update of disease name (history) information	PPRˆD1
24	OMD	Food order	OMDˆ03
25	OMP-01	Prescription order	RDEˆ11
26	OMP-11	Prescription conduct notice	RASˆ17
27	OMP-02	Injection order	RDEˆ11
28	OMP-12	Injection conduct notice	RASˆ17
29	OML-01	Specimen examination order	OMLˆ33
30	OML-11	Specimen examination result notice	OULˆ22
31	OMG-01	Radiological examination order	OMGˆ19
32	OMG-11	Notice of radiological examination conduct	OMIˆ23
33	OMG-02	Endoscopy order	OMGˆ19
34	OMG-12	Notice of endoscopy conduct	OMIˆ23
35	OMG-03	Physiological examination order	OMGˆ19
36	OMG-13	Notice of physiological examination result	ORUˆ01



Data Set Example:

【Merged Data Set】

Key	Val1	Val2	Val3	Val4
YgYk6Izp...	0.55	79.3	100	7.6
x8euBQ8...	0.52	125.8	97	5.4
2roPD93...	0.57	180.0	134	3.2
...				

【Data Set A】

Key	Val1	Val2	...
YgYk6Izp...	0.55	79.3	...
x8euBQ8...	0.52	125.8	...
2roPD93...	0.57	180.0	...
...			

【Data Set B】

Key	Val3	Val4	...
YgYk6Izp...	100	7.6	...
x8euBQ8...	97	5.4	...
2roPD93...	134	3.2	...
...			

Fig. 6.2 Overview of the system developed for secure data collection and analysis

The main points in the development of a medical test bed are as follows:

- Unnecessary sensitive information not used for research should not be leaked outside the medical institution.
- Securely extract and combine medical information across organizations.
- Information extraction control linked with patient consent is possible.
- Privacy risk can be evaluated for the extracted dataset.
- Patients can verify the history of information utilization on the platform.

Figure 6.2 presents an overview of the developed system [7]. The key concepts are the following:

1. Each medical institution has EMR data in SS-MIX2 storage, including billions of HL7 v2 messages.
2. HL7 v2 messages are periodically parsed and stored to relational database management system (RDBMS) tables, maintaining synchronization with the billions of message files in SS-MIX2.
3. Analysis requests from researchers and data collection are managed by the private set intersection (PSI) service on the cloud, which communicates with a client agent located at a client terminal and PSI agents located at each medical institution.
4. Target data criteria, such as diseases, age, and gender, must be defined before the PSI executes data collection. The PSI party agent deploys the target dataset in advance from the local RDBMS to memory.

5. Data collection is achieved using PSI software, which is based on Bloom filter technology for record verification across institutions. The application of bloom filter technology is aimed at realizing data matching in which personal information does not leak outside each hospital during the data collection process.
6. The collected dataset can be verified considering the possibility of patient identification using the extracted attributes.
7. Patients can trace the use of their medical records during data collection.
8. If they choose, patients can withdraw consent for the secondary use of their data. Consent withdrawal information is assumed to be an input to existing the EMR system and exported SS-MIX2 storage in each hospital.

6.3 Features and Implementations of Secondary Use Infrastructure Development

This section describes key features and implementation details of our developed test bed for medical field.

6.3.1 SS-MIX2 Standardized Storage

6.3.1.1 Objective

In Japan, SS-MIX2, which is the domestic standard of exporting whole EHR data as HL7 v2 message files to the external storage for the purposes of backup, regional collaboration, disease repository, and others, is common. In this standard, EHR data is exported to a storage in a directory structure using patient id and clinical date and event type. Therefore, the use of the exported storage for cross-patient analysis such as epidemiological studies is challenging. We are applying an RDB-based virtual file system technology to the storage to achieve cross-patient/cross-institution analysis without collecting data files.

6.3.1.2 Methods

The overview of the system is shown in Fig. 6.3. The storage is developed based on Filesystem in Userspace (FUSE), a virtual file system technology. We adopted pgfuse and PostgreSQL as the FUSE and RDBMS, respectively. The recorded HL7 messages are stored to the DB tables as BLOB data, and the RDBMS traces the transaction in real time. The HL7 messages are parsed by PL/SQL, and parsed medical records (HL7 segments, fields) are recorded to user-defined tables in the RDBMS. Parsing tasks are intended to be executed periodically. Once the records have been stored to

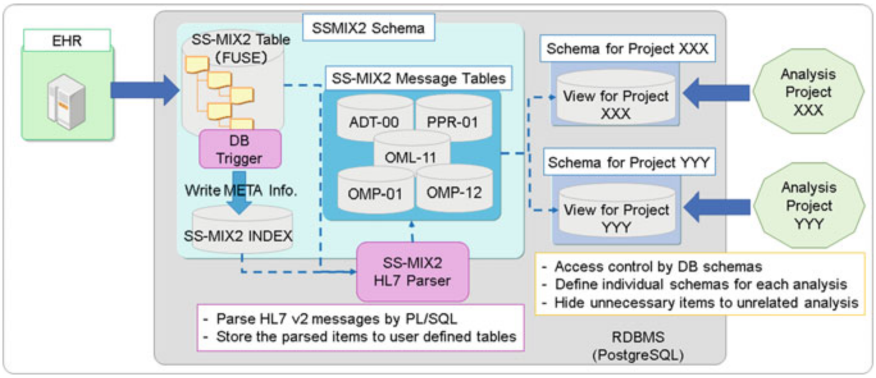


Fig. 6.3 Overview of the developed storage with a virtual file system

Table 6.5 Evaluation results (s) for various numbers of medical records

COPY (VFS to HDD)	COPY (HDD to VFS)	Delete (VFS)	Parse files (VFS)
3,261 s	1,466 s	775 s	1,989 s
5.03 Mbps	2.26 Mbps	9.52 Mbps	54.9 files/s

the tables, the minimum required items can be queried through individually applied view schemas according to the purpose of each analysis project. Performance tests are executed with dummy messages of 109,174 files (922 MB in total, 1,689 patients) including 27 clinical events defined by SS-MIX2 standard, such as ADT-00, OMP-01, and OML-11.

6.3.1.3 Results

Performance test results are shown in Table 6.5. All types of messages could be parsed by PL/SQL. Based on the performance, this storage can process the daily generated medical records of our hospital in less than 2 h.

6.3.1.4 Discussion

The developed storage enables the rapid cycle for the secondary use of medical records analysis among institutions and also prevents the disclosure of unnecessary patient information to each analysis by the regulations of applying view schemas for queries. Moreover, using the developed storage, exported medical records and parsed result tables can be more easily backed up to a remote place in real time using DB replication technology, compared with synchronizing the enormous number of files.

6.3.1.5 Summary

This section describes the development of a standardized storage for the purpose of cross-patient/cross-institution analysis based on the domestic EHR data exporting standard. We will try to develop a secure data collection infrastructure assuming the distributed environment of the developed storages.

6.3.2 Secure Collection of Distributed Medical Information

In this section,[1] we propose an alternative method of collecting and storing EMR data, wherein only necessary items are included in collected data, eliminating the need for individual identifiable information to spread outside the medical institution. The system facilitates EMR data distribution within each medical institution, enabling cross-patient or cross-facility data collection and analysis. The PSI library developed by Miyaji [8] is used for the data integration and encryption of the extracted EMR data. This paper aims to provide an overview of the system and its major technical elements and evaluate the transaction performance of data extraction and collection from the distributed SS-MIX2 storage.

6.3.2.1 Methods

Experimental Environment
The transaction performance of data extraction and collection from the distributed SS-MIX2 storage was evaluated using an experimental environment comprising a server (PSI Server), three data stores (PSI Party), and a client (PSI Client). The Server and Party machines were deployed as VMware ESXi virtual machines. The PSI Client can be deployed on any machine that can run Java.

Experimental data were virtually produced by anonymizing laboratory test result data in the SS-MIX2 storage exported from the EMR system of The University of Tokyo Hospital (Tokyo, Japan). Storage assumed to have 10% overlap between each node was arranged and used for the evaluation tests. The hash value of the character string combining the patient's name, date of birth, and sex was used as the key attribute of each record for the bloom filter.

[1] This section is reprinted from "Studies in Health Technology and Informatics, Vol 255, Katsuya Tanaka, Ryuichi Yamamoto, Kazuhisa Nakasho, Atsuko Miyaji, Development of a Secure Cross-Institutional Data Collection System Based on Distributed Standardized EMR Storage, pp. 35–39," Copyright (2018), with permission from IOS Press. The publication is available at IOS Press through http://dx.doi.org/10.3233/978-1-61499-921-8-35.

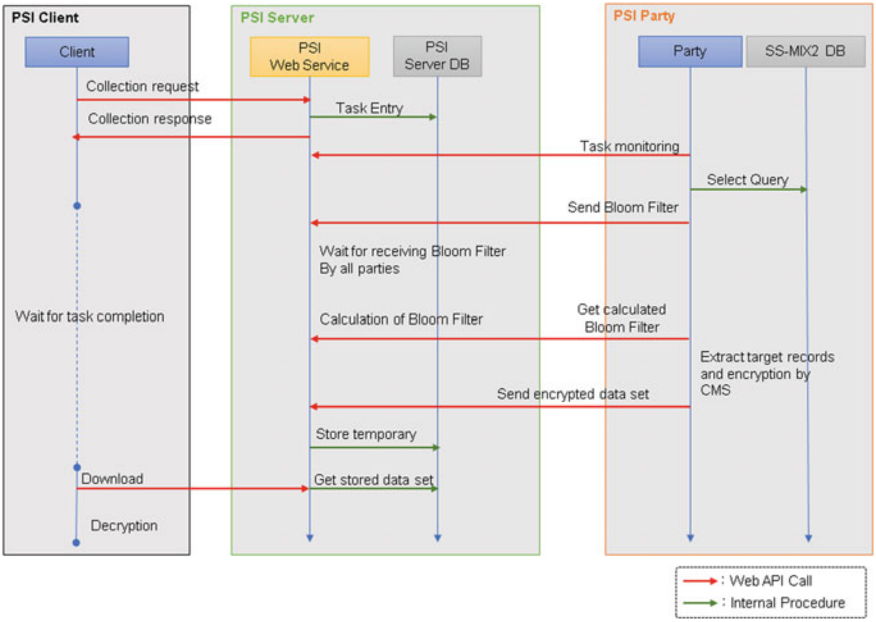


Fig. 6.4 Overview of the transaction flow during data collection

Data Collection with PSI

Figure 6.4 presents an overview of the transaction flow during a secure data collection using the system. The entire system was designed as a Web service so that in the future the service could be available via a commercial cloud. The PSI application programming interface was developed in Java using SOAP Web services and deployed on an Apache Tomcat. All Web communications were implemented with client authentication under TLS 1.2. Extracted EMR data are encrypted by Cryptographic Message Syntax and can be decrypted only by the user requesting the collection.

6.3.2.2 Results

Table 6.6 summarizes the evaluation test results for data queries for calculations, bloom filter calculations, and result data extraction for increasing numbers of EMRs. The processing time linearly increased with the number of records.

Table 6.6 Evaluation results (s) for various numbers of medical records

Records	20,000	40,000	80,000	160,000	320,000	640,000	960,000	1,280,000
Query data	0.1	0.2	0.2	0.3	0.6	1.0	1.5	2.0
Bloom filter processing	0.5	0.5	1.0	2.0	2.9	7.4	13.5	20.7
Data extraction	3.3	3.0	3.1	3.4	4.0	10.1	15.4	23.5
Total	3.8	3.7	4.3	5.7	7.4	18.5	30.4	46.2

6.3.2.3 Discussion

Significance of the System

The system was completely achieved using Web service architecture with the encryption of the extracted EMR data, indicating that medical institutions participating in research would not need to maintain a secure connection to the specific service provider if the developed PSI services are operated on the commercial cloud. The encryption of EMR data avoids any disclosure of the extracted information to the cloud service providers. Furthermore, because the infrastructure makes it unnecessary to connect an EMR storage to the Internet, this eliminates the possibility of experiencing network attacks to the data storage. To meet the requirements of a given analysis, the PSI can execute not only intersection operations but also union operations on distributed datasets.

Performance

The experimental results showed that an intersection operation involving approximately 1 million records was completed within a minute. With this level of processing performance, there should not be any problems with actual operations. We now intend to verify this with larger datasets.

Future Work

The remaining issues for development include (1) the management of consent information, (2) risk assessment for the extracted dataset, and (3) traceability management against data collection. The first issue can be addressed by scanning paper-based consent information related to patients opting-out of the secondary use of their data and storing the scanned data files to the SS-MIX2 storage. We intend to represent consent information as XML files, such as HL7 CDA Privacy Consent Directives, Release 1 [9]. The other two issues are under discussion.

6.3.2.4 Summary

This section describes the underlying concepts and implementation of a secure data collection infrastructure with distributed standardized EMR storage. Using the PSI data collection technology, the experimental results demonstrated high performance. A few issues remain for future implementation.

6.3.3 *Privacy Risk Assessment of Extracted Datasets*

6.3.3.1 Overview

This section describes a prototype of a Web service that enables a series of operations to perform privacy risk evaluation against a dataset extracted from multiple storages by the PSI service developed.

6.3.3.2 Method

The PSI and privacy impact assessment (PIA) libraries are applied using SS-MIX2 standardized storage that adopts FUSE, one of the virtual file systems developed so far. As a FUSE, pgfuse corresponding to PostgreSQL was adopted. Assuming that a service for finally collecting data safely will be operated in the public cloud, the server and client, which will be the nodes of the data collection infrastructure based on the PSI and PIA libraries, are configured as Web services using SOAP. The configuration of the experimental system is shown in Fig. 6.5.

The data for verification was constructed by virtually distributing the HL7 v2 format data obtained by anonymously processing the SS-MIX2 standardized storage data held by The University of Tokyo Hospital to three storages and constructing a virtual multi-facility environment. Storing 1 million specimen test result messages for each storage, creating a dataset using the PSI library, and developing a user interface that can apply the extracted dataset to the risk assessment function in a one-stop manner. In addition, patients between SS-MIX2 standardized storages were artificially adjusted with 10% duplication as a count of the patients.

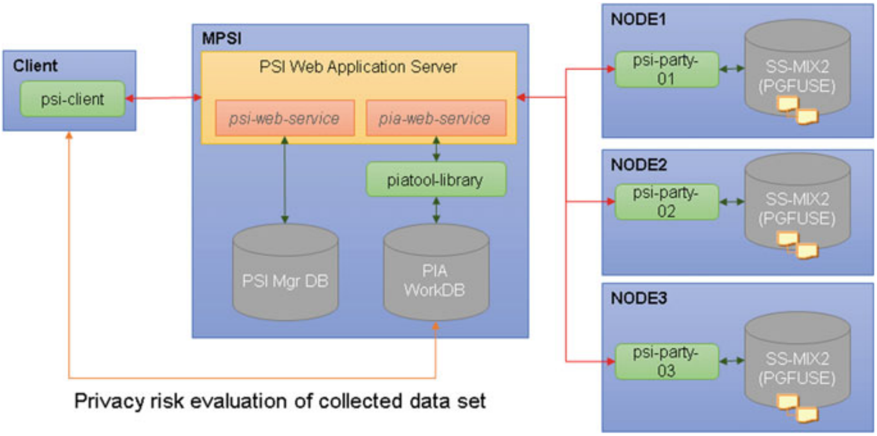


Fig. 6.5 Overview of experimental settings for privacy risk assessment

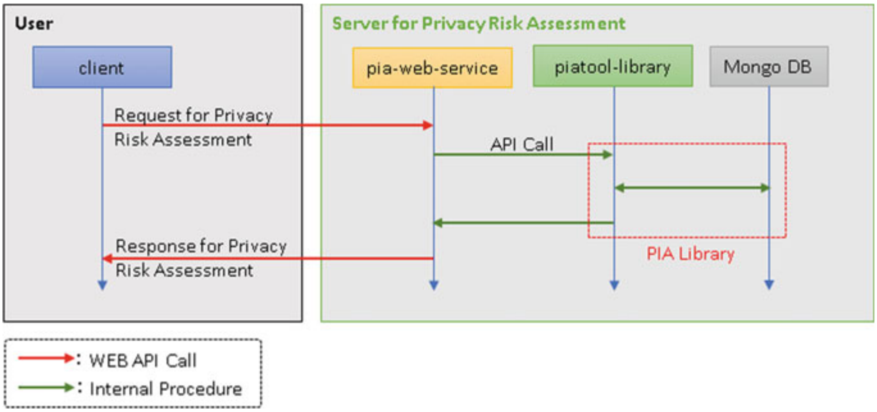
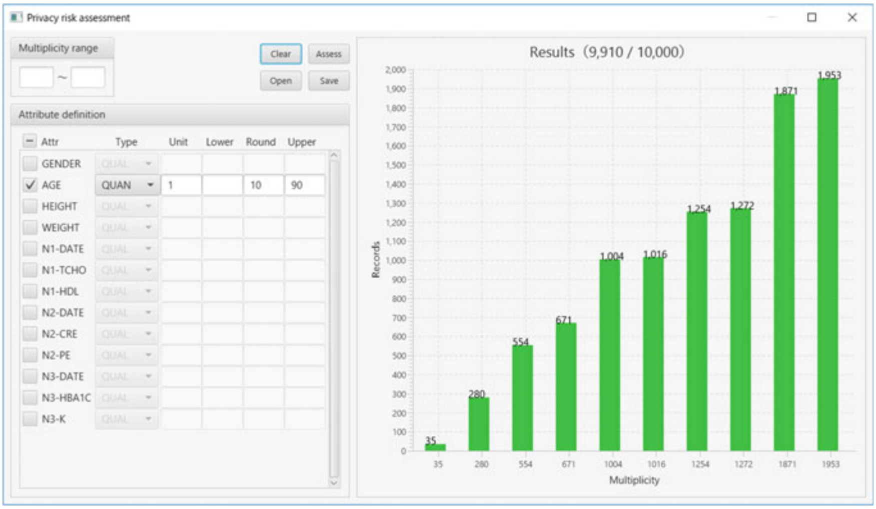

User
client
Request for Privacy Risk Assessment
Response for Privacy Risk Assessment
Server for Privacy Risk Assessment
pia-web-service
piatool-library
Mongo DB
API Call
PIA Library
: WEB API Call
: Internal Procedure

Fig. 6.6 Operation flow of the developed privacy risk assessment service



Attr	Type	Unit	Lower	Round	Upper
□ GENDER	QUAL				
☑ AGE	QUAN	1		10	90
□ HEIGHT	QUAL				
□ WEIGHT	QUAL				
□ N1-DATE	QUAL				
□ N1-TCHO	QUAL				
□ N1-HDL	QUAL				
□ N2-DATE	QUAL				
□ N2-CRE	QUAL				
□ N2-PE	QUAL				
□ N3-DATE	QUAL				
□ N3-HBA1C	QUAL				
□ N3-K	QUAL				

Fig. 6.7 Overview of the developed GUI for privacy risk assessment

The privacy risk evaluation function is configured as a separate Web service and positioned so that it can be operated as data extraction processing by PSI and data processing after acquisition. The system operation flow of the risk evaluation function is shown in Fig. 6.6. While checking the maximum and minimum values and the number of data in each data item of the extracted dataset on the screen, top and bottom coding and generalization processing (processing of numerical data with the specified division accuracy) were constructed.

Figure 6.7 shows the user interface for evaluating privacy risk developed in our project. The dataset extracted by the PSI service can be read, and the maximum and minimum values of each data item can be confirmed. For numerical data, processing can be performed by specifying the upper and lower limits and division unit. In addition, it is possible to calculate the degree of overlap after processing the target dataset using the attribute value group specified on the screen and have an interface for confirming an index for privacy risk evaluation.

6.3.3.3 Results and Discussion

The privacy risk evaluation service operates with a response of up to several tens of seconds when numerous attribute values are specified for a post-extraction dataset with a scale of 100,000. Although there is no problem in performance, it is a configuration in which functions are centrally arranged on the service side regarding the processing of numerical data and evaluation of redundancy, and it does not function unless the original data is exposed to the service side. From the viewpoint of data concealment, it remains a problem, and the functional layout needs to be reconsidered.

6.3.4 Secondary Use and Traceability

This section describes how to implement the capability of traceability in the developed system for secure data collection and analysis.[2]

6.3.4.1 Objective

This section describes how to implement the capability of traceability in the developed system for secure data collection and analysis. Blockchain technology has been recently applied in healthcare fields, including primary patient care, data aggregation for research purposes, and connecting healthcare providers [10–12]. The system that we are developing has a second purpose: to secure the traceability of EMR data, methods to disclose the logs of secondary use are needed. In the present situation, where patients do not have any common ID, it is difficult for a patient to audit all the secondary use logs across the distributed hospital storages that he/she visited. By blockchain technology, we expect to provide patients a common search infrastructure with immutable secondary use logs. Thus, we plan to apply blockchain technology to the aggregation of data extraction log records. This method has several possible implementations, and they must be evaluated assuming operations in real use.

[2] This section is reprinted from "Studies in Health Technology and Informatics, Vol 264, Katsuya Tanaka, Ryuichi Yamamoto, Assessment of Traceability Implementation of a Cross-Institutional Secure Data Collection System Based on Distributed Standardized EMR Storage, pp. 1373–1377," Copyright (2019), with permission from IOS Press. The publication is available at IOS Press through http://dx.doi.org/10.3233/shti190452.

The following experimental results mainly concern data structure and transaction performance compared with traditional implementation for achieving the aggregation of distributed log records of EMR data extraction.

6.3.4.2 Methods

Traceability for Patients

EMR storage for the developed secure data collection system is supposed to process queries from clinical researchers using the standard interface implemented by the PostgreSQL database. EMR data are extracted by data extraction requests handled by the PSI service. Thus, the selected records are identifiable based on each query result, and the records represent the disclosure history of EMR data during data collection through the use of the developed PSI service. By making the log record of extraction searchable by patients, we suppose that traceability in the secure data collection system will be achieved. However, because storage is supposed to be distributed at each hospital, log records must be aggregated by some secure method to be made auditable.

Log data is assumed to be represented by a combination of the following attributes:

1. Identifier of target patient (patient identifier)
2. Storage source (medical institution identifier)
3. Disclosed destination (extracting user identifier)
4. Purpose of use
5. Type of extracted EMR data
6. Extraction timestamp

Attribute 1 (patient identifier) is mandatory for patient identification. In Japan, at present, universal patient identifiers are not available. We assume that insurance numbers may be desirable for searching log records across medical institutions because the patient ID at one medical institution is only applicable for searching log records at that medical institution.

Attribute 2 (medical institution identifier) is used to distinguish the institution storing the extracted EMR data.

Attribute 5 (type of extracted EMR data) is represented by HL7 v2 message types such as “ADT-00,” “OMP-01,” and “OML-11.”

Attributes 3, 4, 5, and 6 are used to distinguish the secondary use of target EMR data by patients. By verifying these attributes, patients can determine whether actual secondary uses meet their consent.

Data Structure for Query

A query for EMR storage may extract the records of several patients at one time. For disclosing extracted history to patients, the extracted history should be sorted by patient, and each history should include the aforementioned attributes.

Table 6.7 Sample data representing extraction history

```
{
 "patientID":"781e5e245d69b566979b86e28d23f2c7",
 "insuti-tionID":"aabd258c8894b996e8d8561fa868364d",
 "disclosedDestination":"AnalysisUser001",
 "purposeofUse":"DrugDevelopment",
 "typeofRecords":"OMP-01",
 "extractionTime":"2018/11/12 01:23:45"
}
```

By focusing on one patient, the extracted history grows as queries hit the target patient EMR record. Moreover, this extracted history is distributed at each EMR storage site across the participating medical institutions.

For achieving desirable response, the aggregation of extracted history should be obtained in a realistic time. This is closely related to the data structure and size of each log record. Future studies should focus on the data size of stored log records.

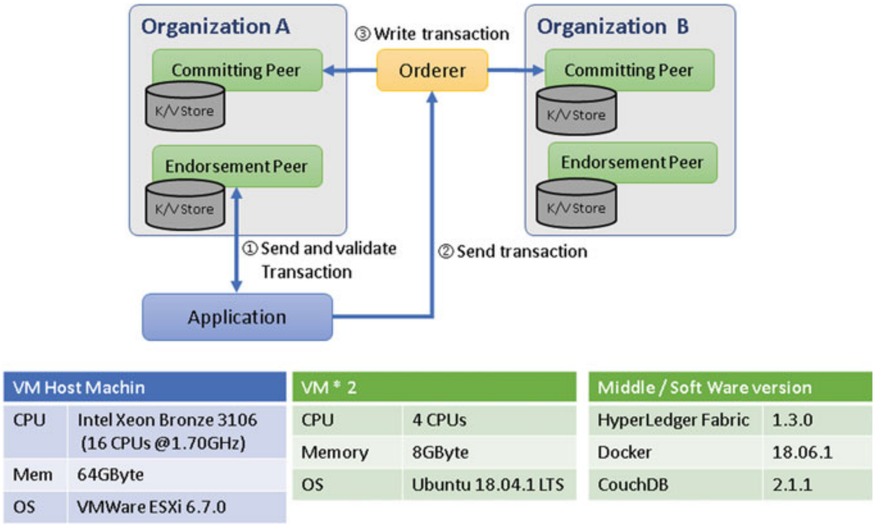
In the performance test, a simple message structure is defined as a JSON (shown in Table 6.7). The identifiers of patient and institution are represented as hash values. Each log record can be stored separately in the blockchain (separate style) or aggregated in a block by a patient appending records to the corresponding block (appending style). In the former method, the pieces of the records related to the patient of interest must be gathered. In the latter method, the block size grows as the system is used. We examined performance differences when the data size of a record to be written is changed.

Experimental Setups

We evaluated the following three approaches to implement traceability function. Of these, two are based on blockchain technology. The last approach uses the same method of secure data collection as PSI against log records stored in distributed PostgreSQL databases.

- Hyperledger Fabric [13]
- BigchainDB [14]
- PSI (Bloom filter)

The experimental settings for each approach are described as follows. Between Hyperledger Fabric and BigchainDB, key/value store implementation for search use differs from each other.



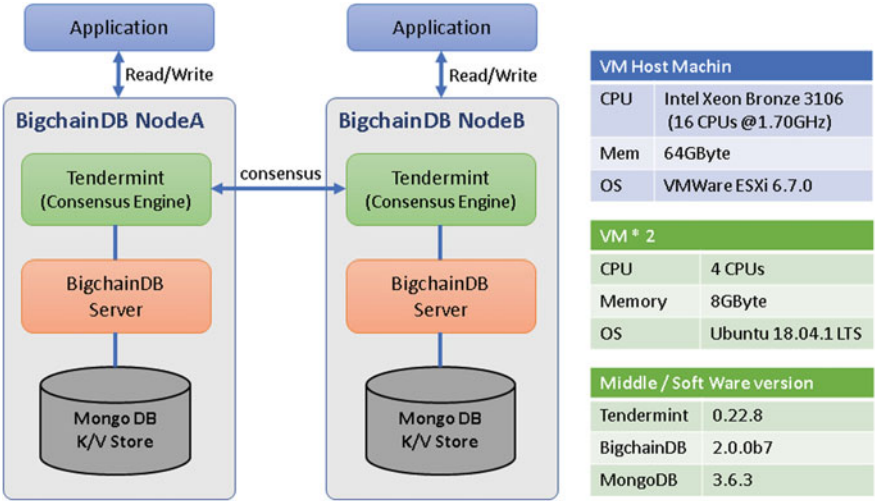
VM Host Machin	
CPU	Intel Xeon Bronze 3106 (16 CPUs @1.70GHz)
Mem	64GByte
OS	VMWare ESXi 6.7.0

VM * 2	
CPU	4 CPUs
Memory	8GByte
OS	Ubuntu 18.04.1 LTS

Middle / Soft Ware version	
HyperLedger Fabric	1.3.0
Docker	18.06.1
CouchDB	2.1.1

Fig. 6.8 Experimental settings (hyperledger fabric)

1. Hyperledger Fabric
   Figure 6.8 shows the experimental setup using Hyperledger Fabric to store query log records during data collection. Assuming two participating institutions, two nodes were set for the performance test. Native implementation only offers key-value storage and is applicable to a separate style. Furthermore, we evaluated Hyperledger implementation with CouchDB [15], which enables query against the value of the JSON message described earlier. Thus, both separate and appending styles can be implemented.
2. BigchainDB
   Figure 6.9 shows the experimental settings for using BigchainDB to store log data. As mentioned earlier, two nodes were prepared for evaluation. MongoDB [16] was selected as the backend database. In this case, both separate and aggregated structures are possible on the same implementation.
   Query key candidate is the transaction ID of the stored block or stored JSON value.
3. PSI (Bloom filter)
   Figure 6.10 shows the experimental settings in the case of PSI implementation. The log records of data extraction are recorded at the time of extraction. Using the same method of EMR data collection, we can gather the log records against distributed storages under encryption. Particularly, although the search is performed by specifying the insurance number, date of birth, and gender by patients, since the matching is performed using the bloom filter, these values are not directly disclosed on the infrastructure.
   In this test, three nodes were prepared for evaluation, but the performance test measurement was executed on only one node.

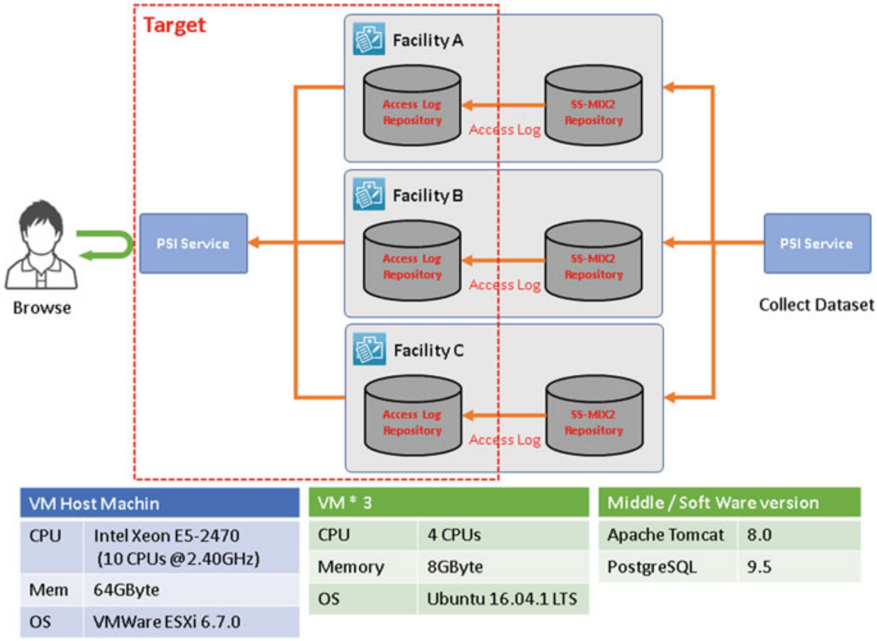


VM Host Machin	
CPU	Intel Xeon Bronze 3106 (16 CPUs @1.70GHz)
Mem	64GByte
OS	VMWare ESXi 6.7.0

VM * 2	
CPU	4 CPUs
Memory	8GByte
OS	Ubuntu 18.04.1 LTS

Middle / Soft Ware version	
Tendermint	0.22.8
BigchainDB	2.0.0b7
MongoDB	3.6.3

Fig. 6.9 Experimental settings (BigchainDB)



VM Host Machin	
CPU	Intel Xeon E5-2470 (10 CPUs @2.40GHz)
Mem	64GByte
OS	VMWare ESXi 6.7.0

VM * 3	
CPU	4 CPUs
Memory	8GByte
OS	Ubuntu 16.04.1 LTS

Middle / Soft Ware version	
Apache Tomcat	8.0
PostgreSQL	9.5

Fig. 6.10 Experimental settings (PSI)

6.3.4.3 Results

Performance by Data Size

Figure 6.11 shows the performance of writing records to the blockchain storage by record size for Hyperledger. In the experimental environment, it worked normally for records with a size of 7 MB or smaller. As the record size grew, the response became unstable.

Figure 6.12 shows the same test for BigchainDB. The maximum record size was 0.6 MB, which was much lower than that for Hyperledger. However, the transaction time to commit was larger than that for Hyperledger.

By contrast, the data size for PSI can be as large as allowed by the database system.

Transaction Performance

Figure 6.13 shows the performance results of writing records to the blockchain storage for Hyperledger with/without CouchDB and BigchainDB under one or five thread processings. In all cases, processing by threads contributed to storage performance, but the throughput did not increase linearly with the number of threads.

Comparing the three implementations, BigchainDB was slightly faster than Hyperledger. Hyperledger with CouchDB had the worst performance; this is likely caused by the cost of indexing within CouchDB. In the best case, 1 million records were written to the blockchain storage in 3–4 h. This performance is equivalent to writing 10 million records or less in one day.

Comparing these implementations using blockchain technology, the performance of PSI was equivalent to the "insert" performance of the PostgreSQL database used. The necessary time for inserting 1 million records to the database was below 10 min. This performance is about 1,000 times faster than the blockchain implementations.

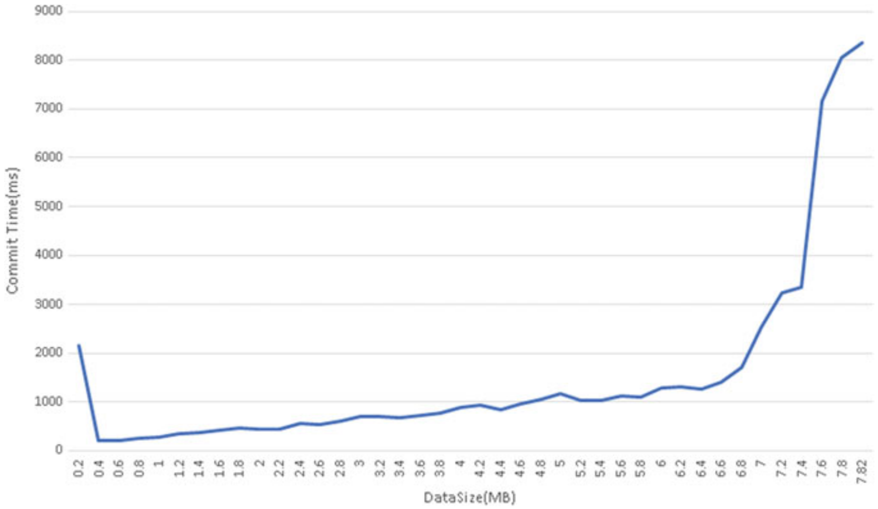


Fig. 6.11 Performance results by record size (Hyperledger)

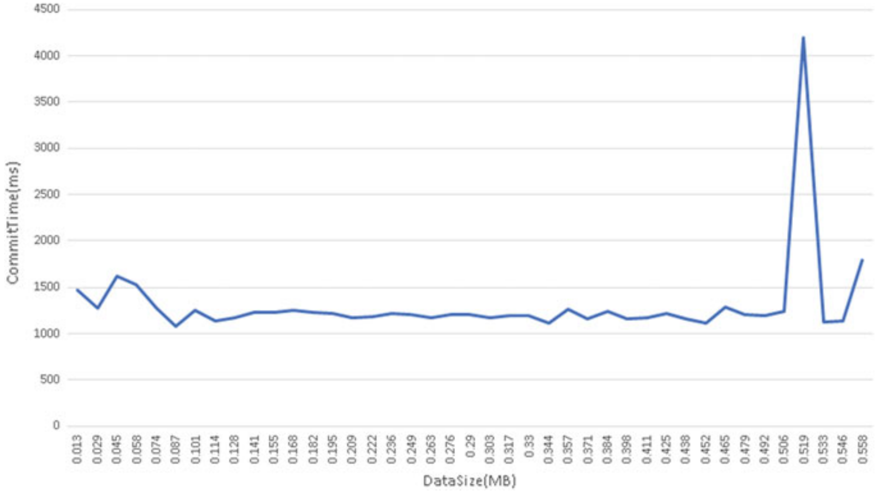
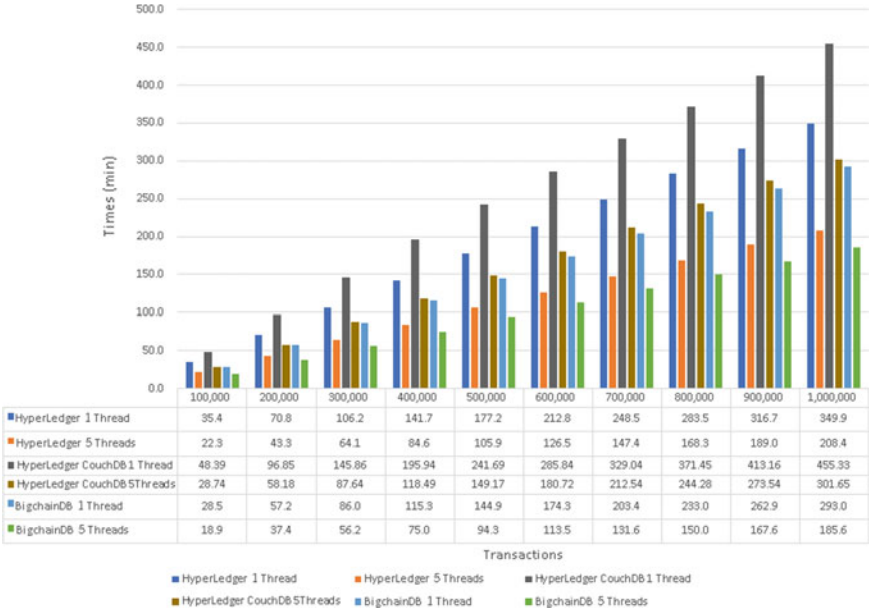
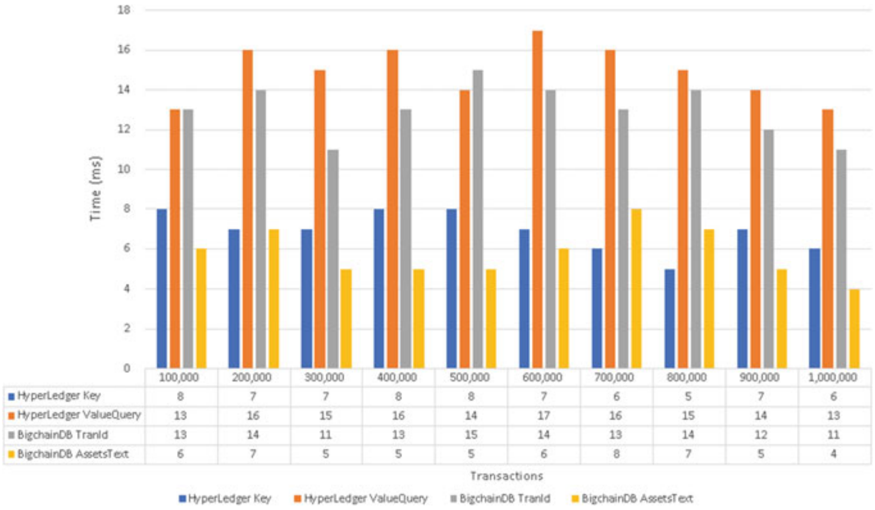


Fig. 6.12 Performance results by record size (BigchainDB)



	100,000	200,000	300,000	400,000	500,000	600,000	700,000	800,000	900,000	1,000,000
■ HyperLedger 1 Thread	35.4	70.8	106.2	141.7	177.2	212.8	248.5	283.5	316.7	349.9
■ HyperLedger 5 Threads	22.3	43.3	64.1	84.6	105.9	126.5	147.4	168.3	189.0	208.4
■ HyperLedger CouchDB 1 Thread	48.39	96.85	145.86	195.94	241.69	285.84	329.04	371.45	413.16	455.33
■ HyperLedger CouchDB 5Threads	28.74	58.18	87.64	118.49	149.17	180.72	212.54	244.28	273.54	301.65
■ BigchainDB 1 Thread	28.5	57.2	86.0	115.3	144.9	174.3	203.4	233.0	262.9	293.0
■ BigchainDB 5 Threads	18.9	37.4	56.2	75.0	94.3	113.5	131.6	150.0	167.6	185.6

Fig. 6.13 Performance result of writing records



	100,000	200,000	300,000	400,000	500,000	600,000	700,000	800,000	900,000	1,000,000
■ HyperLedger Key	8	7	7	8	8	7	6	5	7	6
■ HyperLedger ValueQuery	13	16	15	16	14	17	16	15	14	13
■ BigchainDB TranId	13	14	11	13	15	14	13	14	12	11
■ BigchainDB AssetsText	6	7	5	5	5	6	8	7	5	4

Fig. 6.14 Performance result of querying records

Query Performance

Figure 6.14 shows the performance test results of retrieving one record from the blockchain storage using four types of implementation. No significant differences were noted in the query response times between Hyperledger and BigchainDB. “Hyperledger Key” and “BigchanDB transid” represent the separate style of storage, whereas “Hyperledger Value” and “BigchainDB AssetsText” represent the aggregated style.

Query response is fast enough for actual use in the case of 1 million records in the storage. This result shows hitting 1 record, and the response time linearly increases as hit records increase.

On the other hand, PSI implementation needs 1 min or less to aggregate the extracted results across the distributed databases.

6.3.4.4 Discussion

Based on the initial evaluations, the following recommendations are made.

Transaction Performance

The transaction performance of a blockchain network was quite low for storing massive numbers of log records generated by queries in the developed system. In the case of blockchain, at most 100 transactions per second is best for a node to register to storage. Compared with implementation with PostgreSQL, the total transactions per day will be 1,000 times smaller. If we do not implement any aggregation of log records, it will be impossible to process the enormous numbers of log records generated for each EMR item. Some patient-based aggregation of log records should be considered to overcome performance limitation.

Data Size

The results by data size show the upper limit for storing log records to the blockchain storage. As writing large records to storage makes the system unstable, writing in the appending style is not suitable because of the long operation time of the system. Considering the transaction performance test results mentioned earlier, the total number of transactions to the blockchain network per day should be limited.

Query Response

As the amount of storage increases, the search function must query all storage in the network. The whole log records thus require some possible indexes for searching by patient. The query performance test results show a good response for searching for a log record in the blockchain network despite the increase in the number of log records.

Proposed System for Future Implementation

Based on the performance evaluation results, we decided to implement the following policy as the basis for making the search log history visible to patients when using the developed secure data collection system:

- Aggregate log data by patient in each facility.
- All log records are stored at each facility.
- Record the minimum amount of data, such as the log record identifier key and facility identifiable key, for retrieving index data in the blockchain.
- For query log data, use personal identification information, such as insurance number, date of birth, and gender.

By following these policies, a patient can search the blockchain and find the storage facility. Moreover, the number of records that must be recorded per period can be reduced to the number of related patients. Figure 6.15 shows an overview of the proposed log search system. The log records should include the following:

- Facility identifiable key
- Log record identifiable key
- Digest to audit each log record
- Key to identify each patient (this could be generated by encrypting a patient identifier such as insurance number, date of birth, and gender)

We plan to develop a log search system with the described structure.

6.3.4.5 Limitations

Because we did not have sufficient time to set up larger records, performance tests were executed for 1 million records or less. As the number of records increases, the test results and system stability may change. Performance tests with more records are required in the future work. Similarly, performance should be estimated for larger numbers of nodes.

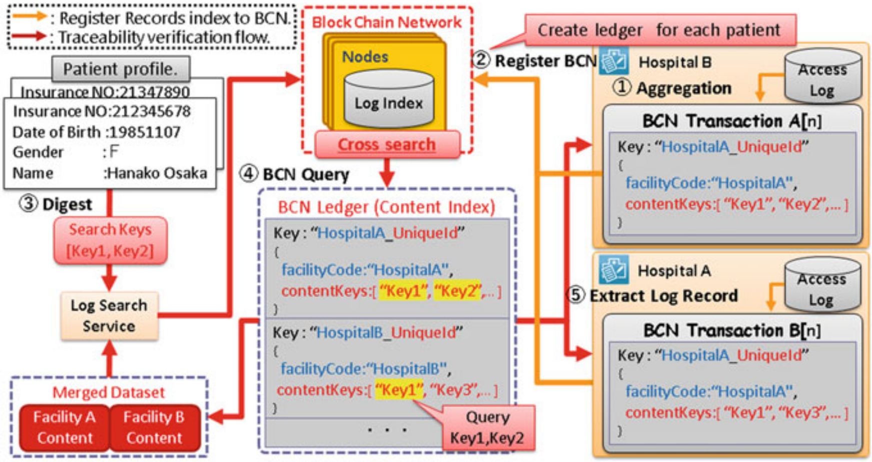

: Register Records index to BCN.
: Traceability verification flow.
Patient profile.
Insurance NO:21347890
Insurance NO:212345678
Date of Birth :19851107
Gender :F
Name :Hanako Osaka
③ Digest
Search Keys [Key1,Key2]
Log Search Service
Merged Dataset
Facility A Content
Facility B Content
Block Chain Network
Nodes
Log Index
Cross search
④ BCN Query
BCN Ledger (Content Index)
Key: "HospitalA_UniqueId"
{
facilityCode:"HospitalA",
contentKeys:["Key1", "Key2",...]
}
Key: "HospitalB_UniqueId"
{
facilityCode:"HospitalB",
contentKeys:["Key1", "Key3",...]
}
. . .
Query Key1,Key2
Create ledger for each patient
② Register BCN
Hospital B
① Aggregation
Access Log
BCN Transaction A[n]
Key: "HospitalA_UniqueId"
{
facilityCode:"HospitalA",
contentKeys:["Key1", "Key2",...]
}
Hospital A
⑤ Extract Log Record
Access Log
BCN Transaction B[n]
Key: "HospitalA_UniqueId"
{
facilityCode:"HospitalA",
contentKeys:["Key1", "Key3",...]
}

Fig. 6.15 Overview of the proposed log search service using a blockchain network

6.3.4.6 Summary

This section reports the initial performance results related to traceability for a secure data collection system under development. The desired data structure and system infrastructure were examined. Although blockchain implementation is a strong candidate for establishing an audit infrastructure to verify the use of EMR data for clinical research, there are some challenges for maintaining long-term operation as the amount of data increases. Thus, we proposed a data structure and querying implementation to overcome the implementation performance.

6.4 Integration and Prospects

As described earlier, the implementation and verification of the following element function have been carried out for a secure secondary use of medical data with the capability of access control by consent information and secondary use status confirmation by traceability function. The key features of our medical test bed are the following:

1. Improvement of the searchability of medical data in SS-MIX2 standardized storage
2. Safe medical data extraction function from SS-MIX2 standardized storage using PSI
3. Electronic description of consent information and mechanism for checking consent information when extracting data
4. Privacy risk assessment function for the extracted dataset
5. Traceability function that can be verified by patients.

Currently, the development of the aforementioned functions is being integrated and developed with the in mind that it can be used as a Web service applicable to public cloud.

If our developed system is ready on public cloud, it would help clinical researchers to conduct cross-institutional data collection and analysis with a certain level of security guaranteed.

References

1. Japan Personal Information Protection Commission. Amended act on the protection of personal information (2017), https://www.ppc.go.jp/files/pdf/Act_on_the_Protection_of_Personal_Information.pdf
2. Japan Ministry of Justice. Act on anonymously processed medical information to contribute to medical research and development (2017), http://www.japaneselawtranslation.go.jp/law/detail/?re=01&dn=1&x=33&y=11&co=1&ia=03&yo=&gn=&sy=&ht=&no=&bu=&ta=&ky=%E5%8C%BB%E7%99%82%E5%88%86%E9%87%8E&page=1
3. Organisation for Economic. OECD privacy guidelines (2013), https://www.oecd.org/internet/ieconomy/privacy-guidelines.htm
4. Ethical guidelines for medical and health research involving human subjects (2015), https://www.mhlw.go.jp/file/06-Seisakujouhou-10600000-Daijinkanboukouseikagakuka/0000080278.pdf
5. M. Kimura, K. Nakayasu, Y. Ohshima, N. Fujita, N. Nakashima, H. Jozaki, T. Numano, T. Shimizu, M. Shimomura, F. Sasaki, T. Fujiki, T. Nakashima, K. Toyoda, H. Hoshi, T. Sakusabe, Y. Naito, K. Kawaguchi, H. Watanabe, S. Tani, SS-mix: a ministry project to promote standardized healthcare information exchange. Methods Inf. Med. **50**(2), 131–139 (2011)
6. Japan Association for Medical Informatics. "SS-mix2 standardized storage" explanation of the structure and guidelines for implemenation ver. 1.2 (2014), https://www.jami.jp/jamistd/docs/SS-MIX2/descript-implemglonSS-MIX2_V1.2.pdf
7. K. Tanaka, R. Yamamoto, K. Nakasho, A. Miyaji, Development of a secure cross-institutional data collection system based on distributed standardized EMR storage. Stud. Health Technol. Inform. **255**, 35–39 (2018)
8. A. Miyaji, K. Nakasho, S. Nishida, Privacy-preserving integration of medical data. J. Med. Syst. **41**(3), 37 (2017)
9. Health Level Seven International. HL7 standards product brief - HL7 CDAR R2 implementation guide: privacy consent directives, release 1 (2017), http://www.hl7.org/implement/standards/product_brief.cfm?product_id=280
10. A. Dubovitskaya, Z. Xu, S. Ryu, M. Schumacher, F. Wang, Secure and trustable electronic medical records sharing using blockchain. AMIA Annu. Symp. Proc. **2017**, 650–659 (2017)
11. M.N. Kamel Boulos, J.T. Wilson, K.A. Clauson, Geospatial blockchain: promises, challenges, and scenarios in health and healthcare. Int. J. Health Geogr. **17**(1), 25 (2018)
12. J.M. Roman-Belmonte, H. De la Corte-Rodriguez, E.C. Rodriguez-Merchan, How blockchain technology can change medicine. Postgrad. Med. **130**(4), 420–427 (2018)
13. Hyperledger fabric - hyperledger (2018), https://www.hyperledger.org/projects/fabric
14. @bigchaindb. Bigchaindb - the blockchain database (2018), https://www.bigchaindb.com/
15. Apache couchdb (2018), http://couchdb.apache.org/
16. Open source document database (2018), https://www.mongodb.com/index