

## Recommended Paper

# Bloom Filter Bootstrap: Privacy-Preserving Estimation of the Size of an Intersection

HIROAKI KIKUCHI<sup>1,a)</sup> JUN SAKUMA<sup>2,b)</sup>

Received: June 7, 2013, Accepted: December 4, 2013

**Abstract:** This paper proposes a new privacy-preserving scheme for estimating the size of the intersection of two given secret subsets. Given the inner product of two Bloom filters (BFs) of the given sets, the proposed scheme applies Bayesian estimation under an assumption of beta distribution for an a priori probability of the size to be estimated. The BF retains the communication complexity and the Bayesian estimation improves the estimation accuracy. A possible application of the proposed protocol is an epidemiological datasets regarding two attributes, *Helicobacter pylori* infection and stomach cancer. Assuming information related to *Helicobacter Pylori* infection and stomach cancer are separately collected, the protocol demonstrates that a  $\chi^2$ -test can be performed without disclosing the contents of the two confidential databases.

**Keywords:** privacy, privacy-preserving data mining, epidemiology, Bloom filter

## 1. Introduction

With the rapid development of database systems and online services, large amounts of information are being collected and accumulated from various data sources independently and simultaneously. Privacy-preserving data mining (PPDM) has been attracting significant attention as a technology that could enable us to perform data analysis over multiple databases containing sensitive information without violating subjects' privacy.

In this paper, we investigate the problem of set intersection cardinality. Given two private sets, the goal of this problem is to evaluate the cardinality of the intersection without disclosing the sets mutually. Set intersection cardinality has been extensively studied as a building block of PPDM, including association rule mining [19], model and attribute selection [18], and other aspects [4]. Our major application of this problem is epidemiological analysis, including privacy-preserving cohort studies. We wish to perform cohort studies over multiple independently collected medical databases, which are not allowed to disclose identifying information about patients.

Consider two databases developed independently by two organizations. One organization collects individual medical information, including patient ID, patient name, patient address, presence or absence of disease 1, disease 2, and so on. The other organization collects individual genome information from research participants; including participant ID, participant name, participant

address, presence or absence of genome type 1, genome type 2, and so on. The objective of a cohort study may be to investigate the association between the outbreak of a specific disease and genomes. For this analysis, the analyst makes use of four-cell contingency tables; each cell counts the number of patients who have (do not have) a specific disease and have (do not have) a specific genome type. If both tables are private, the set intersection cardinality may be used for evaluating of the count of each cell without sharing database content. In this study, we consider the following four requirements for practical situations.

**Requirement 1.** The time and communication complexity should be linear with respect to the number of records  $n$ . This is because statistical analysis, including cohort studies, usually treats databases with a large number of records.

**Requirement 2.** The time and communication complexity should be independent of the size of the ID space. In the use case described above, both organizations independently collect information from individuals. Thus, unique IDs are not given to records. Instead, the protocol must generate a unique ID for each record with the combination of individual attributes, such as the name and address. Because the space required for the combination of such user attributes is often much larger than the number of individuals, this requirement is important.

**Requirement 3.** The protocol should be designed considering the asymmetry of computational capabilities of organizations. Assume that a research institute that holds genome information provides epidemiological analysis services upon request to hospitals that hold medical information. In such

<sup>1</sup> Department of Frontier Media Science, School of Interdisciplinary Mathematical Sciences, Meiji University, Nakano, Tokyo 164–8525, Japan (The part of this work was done when author was with Tokai University.)

<sup>2</sup> Graduate School of Systems and Information Engineering, Computer Science Department, University of Tsukuba, Tsukuba, Ibaraki 305–8573, Japan

<sup>a)</sup> kikn@meiji.ac.jp

<sup>b)</sup> jun@cs.tsukuba.ac.jp

The initial version of this paper was presented at Computer Security Symposium 2012 (CSS 2012) in October 2012. This paper was recommended to be submitted to Journal of Information Processing (JIP) by Program Chair of CSS 2012.

a case, it is expected that the computational capabilities of the hospitals are poor. Therefore, a reasonable solution can be the outsourcing of computation; the research institute offers servers with high computational power and the hospital outsources most of the computation required for the analysis to the research institute. This example indicates that the protocol of set intersection cardinality should be designed considering the asymmetry of computational capabilities.

**Requirement 4.** The outputs of the protocol must be random shares. This requirement implicitly suggests that the set intersection cardinality may be used as a part of a larger-scale protocol. If the outputs of the protocol are random shares, these must be seamlessly used for inputs to other privacy-preserving protocols.

In this paper, we propose a set intersection cardinality protocol that satisfies these requirements.

### Related Work

Let  $S_A$  and  $S_B$  be private inputs of the set intersection cardinality. Let  $n_A$  and  $n_B$  be the cardinalities of  $S_A$  and  $S_B$ , respectively.

Agrawal et al. [1] presented a set intersection cardinality protocol using commutative encryption under DDH (Decisional Diffie-Hellman) assumption. The time complexity of this protocol is  $O(n_A + n_B)$ ; this is linear in the size of the databases and is independent of the size of the ID space. However, this protocol assumes that the two parties have nearly the same computation power. Furthermore, the protocol cannot output random shares. De Cristofaro and Tsudik [5] introduced an extension of Ref. [1]. It also requires  $O(n)$  computation by both parties.

Freedman et al. [7] proposed a set intersection protocol using oblivious polynomial evaluation. This protocol can be converted to the set intersection cardinality with a slight modification, and achieves  $O(n_B + \log \log n_A)$  time/communication complexity. Furthermore, the time complexity is independent of the ID space size and random shares can be output. This protocol also assumes that both parties have equal computational power.

All the above protocols guarantee exact outputs. Kantarcioglu et al. [11] approach the set intersection cardinality differently. Their protocol maps the input set onto a binary vector using a Bloom filter (BF) [2], and the set intersection cardinality is statistically estimated from the scalar product of the two binary vectors. With this approach, the results become approximations, although the computation cost is expected to be greatly reduced. The dimensionality of the vector used in this protocol is equal to the ID space size; this does not meet Requirement 2. In Ref. [11], a technique to shorten large IDs using hash functions was used with their protocol. As shown later by our theoretical analysis, given an error rate  $\epsilon$ , the optimal range of hash functions for  $n$  elements is  $O(n^2)$ . This indicates that such Naive ID generation can be too inefficient for practical use.

Camenisch and Zaverucha [3] has introduced the certified set intersection cardinality problem. This protocol considers asymmetry in the security assumptions of the parties, but does not consider asymmetry in their computational capability.

Ravikumar et al. used the TF-IDF measures to estimate the scalar product in Ref. [17]. As for epidemiological study, Lu et al.

studied the contingency tables in Ref. [13].

Thus, to our knowledge, no set intersection cardinality protocol satisfies the four requirements above, which should be met for practical privacy-preserving data analysis, especially for the outsourcing models.

### Our Contribution

In this manuscript, we present a protocol that satisfies the four requirements. Considering the first and second requirement, the sets are independently mapped onto BFs, and then the set intersection cardinality is statistically estimated from the scalar product of the two binary vectors representing the BFs.

As discussed later, the size of the BF must be  $O(n^2)$  to control the false positive rate in Ref. [11]; this does not meet Requirement 2. Our protocol therefore uses a number of BFs of size  $O(n)$ . The set intersection cardinality is obtained by iteratively applying Bayesian estimation to the scalar products of the BFs.

In the proposed protocol, the scalar product protocol is used as a building block. Modulo exponentiation is performed only by one party and this fits well with the outsourcing model (Requirement 3). In addition, the outputs can naturally be made random shares (Requirement 4).

We demonstrate our protocol with an epidemiological datasets regarding two attributes, *Helicobacter pylori* infection and stomach cancer. Assuming information related to *Helicobacter Pylori* infection and stomach cancer are separately collected, we demonstrate that a  $\chi^2$ -test can be performed without disclosing the contents of the two databases.

## 2. Preliminary

### 2.1 Bloom Filter

A BF is a simple space-efficient data structure for representing a set to support membership queries [2]. Recently, BFs have been used not only for database applications but also for network problems including detecting malicious addresses, packet routing, and the measurement of traffic statistics.

A BF for representing a set  $S = \{a_1, \dots, a_n\}$  of  $n$  elements is an array of  $m$  bits, initially all set to 0. The BF uses  $k$  independent hash functions  $H_1, \dots, H_k$  such that  $H_i : \{0, 1\}^* \rightarrow \{1, \dots, m\}$ . The hash functions map each element in the map to a random number uniformly chosen from  $\{1, \dots, m\}$ . Let  $B(S)$  be a set representing a BF defined by  $B(S) = \bigcup_{a \in S} B(a)$  such that  $B(a) = \{H_1(a), \dots, H_k(a)\}$ . Now let  $\mathbf{b}$  be an  $m$ -dimensional vector,  $(b_1, \dots, b_m)$ , which is an alternative representation of the BF, defined by  $b_i = \begin{cases} 1 & \text{if } i \in B(S), \\ 0 & \text{if } i \notin B(S), \end{cases}$  for  $i = 1, \dots, m$ . For example, the hash functions that map an element  $a$  as  $H_1(a) = 2$ ,  $H_2(a) = 7$  characterize a BF with  $m = 8$ ,  $B(a) = \{2, 7\}$ . Alternatively,  $\mathbf{b}(a) = (0, 1, 0, 0, 0, 0, 1, 0)$ . We can use either the set or vector representation of BF, depending on the cryptographic building blocks used. Note the following relationship between the set and vector representations,  $\mathbf{b}(S_1) \cdot \mathbf{b}(S_2) = |B(S_1) \cap B(S_2)|$ .

To test if  $a$  is an element of set  $S$ , we can verify that

$$\forall i = 1, \dots, k \ H_i(a) \in B(S), \quad (1)$$

which holds if  $a \in S$ . However, it also holds, with a small prob-

ability, even if  $a \notin S$ . That is, BFs suffer from false positives. According to Ref. [2], after all the elements of  $S$  are hashed into the BF, under an assumption that hash functions are perfectly random [2], the probability that element  $i$  does not belong to  $B(S)$ , i.e., that the  $i$ -th bit of  $b(S)$  is still 0, is  $p = \left(1 - \frac{1}{m}\right)^{kn} \approx e^{-kn/m}$ . We therefore have a probability of false positives given by  $p' = \left(1 - \left(1 - \frac{1}{m}\right)^{kn}\right)^k \approx \left(1 - e^{-kn/m}\right)^k$ . If  $k$  is sufficiently small for given  $m$  and  $n$ , Equation (1) is likely to hold only for the element of  $S$ . Conversely, with too large a value for  $k$ , the BF is mostly occupied by 1 values. In Ref. [2], [6], the optimal BF was found for  $k^* = \ln 2 \cdot (m/n)$ , which minimized the false-positive probability.

## 2.2 Cryptographic Primitives

### 2.2.1 Paillier Cryptosystem

Additively homomorphic public-key schemes – Paillier [16] or the modified ElGamal cryptosystems are both widely used. Both allow for key generation and decryption to be distributed amongst partially trusted authorities sharing private key. A cryptosystem  $E$  is said to satisfy the additively homomorphic property if: taking messages  $M_1$  and  $M_2$ ,

$$\begin{aligned} E[M_1]E[M_2] &= E[M_1 + M_2], \\ E[M_1]^{M_2} &= E[M_1 M_2]. \end{aligned}$$

The Paillier cryptosystem consists of three stages: key generation, encryption, and decryption.

- **Key generation:** Let  $n$  be  $pq$ , a product of two large prime numbers  $p$  and  $q$ , and  $g \in \mathbb{Z}_n^*$  be a generator whose order divides  $n$ . Compute  $\lambda = \text{LCM}(p-1, q-1)$  and  $\mu = (L(g^\lambda \pmod{n^2}))^{-1} \pmod{n}$ , where  $L$  is defined by  $L(u) = (u-1)/n$ . The public key is  $(n, g)$  and the private key is  $(\lambda, \mu)$ .
- **Encryption:** A ciphertext  $c$  of  $M$  is defined with randomly chosen  $r \in \mathbb{Z}_{n^2}^*$  as:

$$c = E(M) = g^M r^n \pmod{n^2}.$$

- **Decryption:** Given ciphertext  $c$ , plaintext  $M$  is computed as  $M = L(c^\lambda \pmod{n^2}) \cdot \mu$ .

Paillier is more efficient than ElGamal with respect to decryption overhead, as the latter requires a sort of brute force technique (in the limited domain) for decrypting candidates of messages. We implement the Paillier cryptosystem for performance evaluation since the single computational cost for encryption is more significant for our proposed protocol.

### 2.2.2 Secure Scalar Product.

The scalar product of two vectors is performed securely by using a public-key encryption scheme in Algorithm 1.

### 2.2.3 Secure Function Evaluation (SFE).

We use the generic two-party secure-function evaluation system, Fairplay [14] Fairplay is a compiler for a high-level procedural definition language, CFDL, producing a one-pass Boolean circuit in a language called SHDL. With Fairplay, we can perform secure functions without revealing their inputs.

### 2.2.4 Security Model

We assume that the parties are *honest-but-curious*, which

---

### Algorithm 1 Secure Scalar Product

---

Input: Alice has an  $n$ -dimensional vector  $\mathbf{x} = (x_1, \dots, x_n)$ . Bob has an  $n$ -dimensional vector  $\mathbf{y} = (y_1, \dots, y_n)$ .

Output: Alice has  $s_A$  and Bob has  $s_B$  such that  $s_A + s_B = \mathbf{x} \cdot \mathbf{y}$ .

- (1) Alice generates a key pair for a homomorphic public-key encryption scheme and sends the public key to Bob.
  - (2) Alice sends to Bob  $n$  ciphertexts  $E(x_1), \dots, E(x_n)$ , encrypted with her public key.
  - (3) Bob chooses  $s_B$  at random, computes  $c = E(x_1)^{y_1} \cdots E(x_n)^{y_n} / E(s_B)$  and sends  $c$  to Alice.
  - (4) Alice uses her secret key to decrypt  $c$  to obtain  $s_A = D(c) = x_1 y_1 + \cdots + x_n y_n - s_B$
- 

is known as *semi-honest* model, with parties that own private datasets following protocols properly but trying to learn additional information about the datasets from received messages.

The privacy of our proposed idea is defined in semi-honest model as follows.

**Definition 2.1** Let  $A$  and  $C$  be datasets (subsets) owned by two parties, Alice and Bob. A *secure* protocol tests whether the size of set intersection  $|A \cap C|$  over the two datasets is greater than a threshold without revealing  $A$ ,  $B$  and  $A \cap C$  in the semi-honest-model sense.

## 3. Difficulties in ID-less Datasets

### 3.1 Problem Definition

We are considering the problem of a two-party protocol that can evaluate the size of the intersection of two sets without revealing the sets themselves.

Let  $A$  and  $B$  be parties owning subsets  $S_A$  and  $S_B$ , respectively. For an agreed threshold  $t$ , they each wish to know if

$$X = |S_{A \cap B}| = |S_A \cap S_B| \geq t \quad (2)$$

is true, without revealing  $S_A$  or  $S_B$  to the other party. Here,  $X$  is a random variable describing the size of the intersection  $S_{A \cap B}$ .

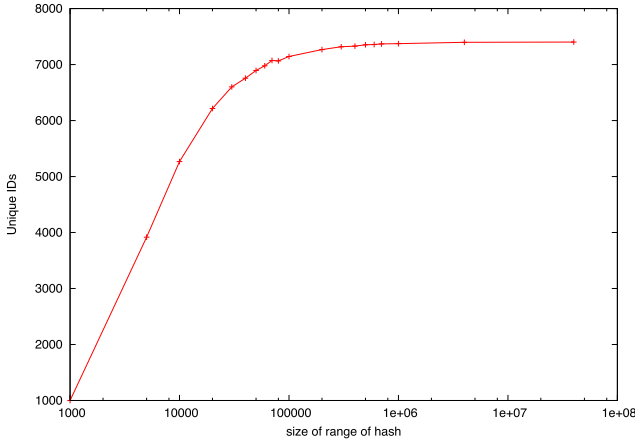
Note that we are not interested in learning about the intersection, itself but are only interested in evaluating its size because the size is often useful in many privacy-preserving applications. For example, an epidemic study might test if the difference between two subsets is statistically significant. The difference of  $|X_{A \cap B}|$  and  $t$  may even be confidential in some applications.

### 3.2 Naïve ID Generation

Consider a dataset of  $n$  elements with multiple attributes, such as name, sex, age and address, but with no unique identity being assigned. Instead, the elements are uniquely specified by attributes, e.g., name and birthday. Let  $A$  be a set of attributes  $A = \{a_1, \dots, a_n\}$ .

The simplest way to generate a pseudo identity is to use a hash function  $h : \{0, 1\}^* \rightarrow \{1, \dots, \ell\}$ . Using this hash function, we assign  $h(a_i)$  to the  $i$ -th element. For efficiency reasons, we assume the range is sufficiently large that we can neglect the occurrence of a collision such that  $h(a_i) = h(a_j)$  for some  $i \neq j$ . Letting  $h_A$  be the set of all pseudo identities, defined as  $h_A = \{h(a_i) \mid a_i \in A\}$ , we can see any collision of identities by testing whether  $|h_A| = n$ .

If the size  $\ell$  of the ID set increases, collisions can be avoided, but the computational cost will accordingly increase with  $\ell$ .



**Fig. 1** Unique hash values,  $|h_A|$  with respect to the range  $\ell$  (Experimental result using DBLP).

Clearly,  $\ell \geq n$ , but finding the optimal size is not trivial. To solve the tradeoff between accuracy and performance reduction, let us assume we have an optimal  $\ell$  that is sufficiently large to uniquely determine the given set of  $n$  elements.

This problem is equivalent to the problem known as “birthday paradox,” whereby, among a set of  $n$  randomly chosen people, there is a probability that some pair of them has the same birthday. When identities (birthdays) are chosen with a uniform probability of  $1/\ell$ , the probability that all  $n$  identities are unique is given by

$$\prod_{j=1}^{n-1} \left(1 - \frac{j}{\ell}\right) \approx \prod_{j=1}^{n-1} e^{-j/\ell} = e^{-n(n-1)/2\ell} \approx e^{-n^2/2\ell}.$$

Therefore, given the probability  $\epsilon$  with which  $n$  hash values are unique, we have

$$\frac{n^2}{2\ell} = \ln \epsilon^{-1}, \quad (3)$$

from which follows the solution of our problem. The optimal range of hash functions for  $n$  elements is given as  $\ell = n^2/2 \ln \epsilon^{-1}$ , for which  $n$  elements will have distinct identities with a probability of  $\epsilon$ .

For example, a dataset of  $n = 7,000$  users will be uniquely determined by pseudo identities generated by a hash function such that  $\ell = 4.7 \times 10^8$ , with a probability of 95%.

**Figure 1** shows the unique pseudo identities for  $n = 7,500$  names in the DBLP<sup>\*1</sup>, a public author dataset, with respect to the range  $\ell$  of the hash function used to generate the identities. It indicates that  $\ell = 4 \times 10^6$  satisfies to generate unique identities for  $h(S)$  with  $\epsilon < 1$ , and  $n = 7,500$ . Therefore, this naïve approach is nearly infeasible because of the large computational overhead occurred by the cryptographic protocols. For example, the secure scalar product [8] for evaluating the set intersection of the dataset requires  $n^2$  ciphertexts and  $n^2$  modular exponentiations. This clearly does not satisfy Requirement 2, shown in Section 1.

### 3.3 Kantarcioglu’s Scheme

In Ref. [11], Kantarcioglu, Nix and Vaidya proposed the following cryptographic protocol using BF in an approximate algorithm for the threshold scalar (dot) product.

<sup>\*1</sup> DBLP, A Citation Network Dataset, V1, (<http://arnetminer.org/citation>).

Let  $Y$  be a random variable representing the number of matching bits in the two BFs of  $S_A$  and  $S_B$ . That is,  $Y$  is defined by  $Y = |B(S_A) \cap B(S_B)|$ . There is a positive correlation between  $X$ , defined by true size of intersection  $S_A \cap B$ , and  $Y$ , which enables us to predict  $X$  from  $Y$  which can be obtained from BFs in a secure way.

Based on the properties of BFs [2], Eq. (2) is equivalent to

$$Z_A + Z_B + Z_{AB} \geq Z_A Z_B \frac{1}{m} \left(1 - \frac{1}{m}\right)^{-kt}, \quad (4)$$

where  $Z_A$  ( $Z_B$ ) is the number of 0s in  $B(S_A)$  ( $B(S_B)$ ), respectively.  $Z_{AB}$  is the number of matching 0s in the two BFs of  $S_A$  and  $S_B$ . That is,  $Z_{AB} = m - |B(S_A) \cap B(S_B)| = m - Y$ . To evaluate the inequality privately, Kantarcioglu et al. performs a secure protocol for the scalar product of two vectors [8] to obtain  $u_1$  and  $u_2$  such that  $b(S_A) \cdot b(S_B) = m - Z_{AB} = u_1 + u_2$  and a secure protocol for the multiplication of two integers  $Z_A$  and  $Z_B$  to obtain  $v_1$  and  $v_2$  such that  $v_1 + v_2 = (1 - 1/m)^{-kt}/m Z_A Z_B$ . Finally, they use SFE for the shared comparison of two integers to test if  $(Z_A + u_1 - m) + (Z_B + u_2) \geq (v_1 + v_2)$ .

According to their experimental results [11], their approximation algorithm using BFs with  $m = 3,000$ ,  $k = 2$ , and  $n = 20,000$  ran in 4 minutes, whereas an *exact* version required 27 minutes.

### 3.4 Difficulties in ID-less Datasets

In Ref. [11], Kantarcioglu et al. claim that as long as,  $m \ll n$ , their method would be much faster than the typical implementation of a secure scalar (dot) product protocol<sup>\*2</sup>. Their experimental results show that the accuracy of approximation increases as  $m$  increases<sup>\*3</sup>. We will show that these properties do not hold in our target, *ID-less datasets* model, where the two datasets have no consistent identities and hence  $n$  elements are specified with some unique attribute(s).

- (1) (**Accuracy**) The size of intersection is approximated in their scheme based on the expected value of probability of common bits in BFs. The accuracy is expected to be improved as  $m$  increases. However, this is not true in large  $m$  because that the vector becomes too sparse. To be adaptively dense vector, we must increase the number of hash functions,  $k$ . This is not trivial. In Ref. [11], the experimental behavior with some parameters were shown and no guarantee in accuracy.
- (2) (**Performance**) The size  $m$  of BF increases up to  $n^2$  in ID-less datasets. As we discussed in Section 3.2, the range of hash function should be as large as  $n^2$  in order to minimize the probability to fail to uniquely identify elements. This is too large to find the intersection since some schemes running in  $O(n)$  complexity in private set intersection are known, e.g., Refs. [1], [5].
- (3) (**Overhead**) Their scheme requires the secure multiplication as well as scalar product. It is not necessary in private set intersection.

In later section, we will present our scheme which overcomes

<sup>\*2</sup> In Section 2.2 (Computation and Communicational cost). In Section 3, they assume that the vector of 20,000 elements, whose density was 10%, that is, the vector contains 2,000 1’s ( $= n$ ), and it performs 20,000-dimensional vector’s scalar product for exact match and  $m = 3,000$  BF for their scheme.

<sup>\*3</sup> In Section 3.1, Fig. 1 (b).



**Table 1** Comparison between Ref. [11] and ours.

item	Ref. [11]	Proposed
approximation	Eq. (4)	Eqs. (8), (7)
priori distribution	–	Beta distribution
BF size ( $m$ )	large ( $n^2$ )	small ( $n/\ln 2$ )
accuracy	improved as increasing $m$	improved with Bayesian estimation from $s$ tests

the above limitations. **Table 1** gives a summary of comparison between the scheme in Ref. [11] and proposed scheme.

## 4. Proposed Scheme

### 4.1 Probability Distribution of Matching Bits in BFs

Suppose that given  $S_{A \cap B} = S_A \cap S_B$ , random variable  $X$  of the cardinality of  $S_{A \cap B}$ , and instance  $x = X$ , we wish to estimate the number of matching 1s bits in their two BFs, i.e.,  $y = |B(S_A) \cap B(S_B)|$ . The quantity  $y$  is equal to the number of 1s values in the conjunction of the two BF vectors. This subsection presents the mathematical properties of BFs, which will be used to estimate  $X$  in the subsequent subsection.

An element  $a$  in  $S_A \cup S_B$  belongs to either (1)  $S_{A \cap B}$  or (2)  $S_A \cup S_B - S_{A \cap B}$ . The case (1) always ensures that  $B(\{a\}) \subset B(S_A) \cap B(S_B)$ . Any element  $a$  in  $S_A$  yields 1s bits at the exactly same positions specified by  $b$  in  $S_B$ . While, in the case (2), 1s bits is set only if  $H_i(a) = H_j(b)$  arises for some  $i, j \leq m, a \in S_A$  and not in  $S_B$  and  $b \in S_B$  and not in  $S_A$  such that  $a \neq b$ . In other words, the case (2) happens by false positive. Since cases (1) and (2) are mutually exclusive events, we compute each conditional probability as follows.

Case (1): The probability that a certain bit in the conjunction of BFs is 0 after  $k$  random bits are set to 1 for all  $x$  element in  $S_A \cap S_B$  is  $q_X = (1 - \frac{1}{m})^{kx}$ .

Case (2): Suppose an element  $a$  that belongs to  $S_A$  and not to  $S_B$  can have the same hash value  $H_i(a) = H_j(b)$  as some element  $b \neq a$  in  $S_B$  and not in  $S_A$ . The probability that a certain bit is 0 in the BF for  $a$  in  $S_A - S_{A \cap B}$  is  $q_A = (1 - \frac{1}{m})^{k(n_A - x)}$ . Similarly, the BF of an element in  $S_B - S_{A \cap B}$  having a certain bit being 0 has a probability of  $q_B = (1 - \frac{1}{m})^{k(n_B - x)}$ . Therefore, the probability of a certain bit in the BF for  $S_A \cup S_B - S_{A \cap B}$  being 1 is given by the product of the complement of each event, namely  $(1 - q_A)(1 - q_B) = 1 - q_A - q_B + q_A q_B$ .

Because the conjunction of BF has 1 for a certain  $z$ th bit by being either an element of  $S_{A \cap B}$  or  $S_A \cup S_B - S_{A \cap B}$ , we have the probability  $\theta$  for a bit being 1 as the disjunction of the two events, namely,

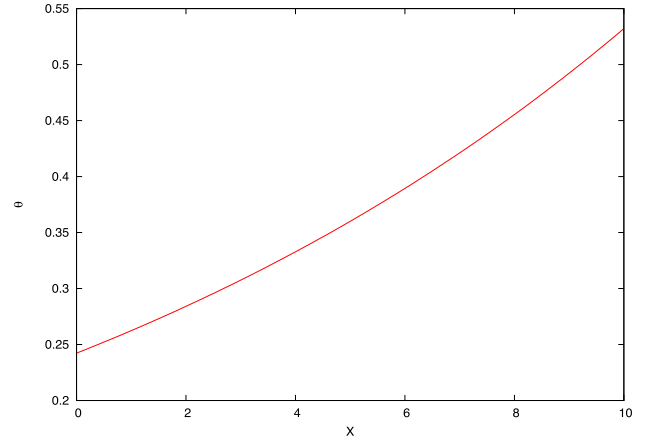
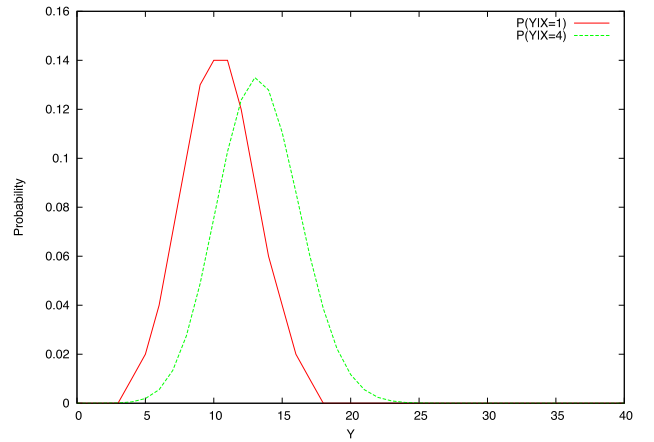
- (1)  $H_i(a) = z$  for some  $a \in S_A \cap S_B$  or,
- (2)  $H_i(a) = H_j(b) = z$  for some  $a \in S_A - S_{A \cap B}$ ,  $b \in S_B - S_{A \cap B}$ ,

equivalently,

- not not (1)  $H_i(a) \neq z$  for all  $a \in S_A \cap S_B$  and
- not (2)  $H_i(a) \neq z$  for all  $a \in S_A - S_{A \cap B}$ , and  $H_j(b) \neq z$  for all  $b \in S_B - S_{A \cap B}$ .

Therefore, we have the probability

$$\begin{aligned} \theta &= 1 - q_X(1 - (1 - q_A)(1 - q_B)) \\ &= 1 - \left(1 - \frac{1}{m}\right)^{kn_A} - \left(1 - \frac{1}{m}\right)^{kn_B} + \left(1 - \frac{1}{m}\right)^{k(n_A + n_B - x)}. \end{aligned} \quad (5)$$

**Fig. 2** Probability  $\theta$  of a certain bit being 1 in the conjunction of two BFs with respect to  $x = |S_A \cap S_B|$ .**Fig. 3** Probability distribution of  $Y$ , the number of 1s bit in the BF for  $Pr(Y|X=1)$  and  $Pr(Y|X=4)$ .

Consequently, the conditional probability of  $Y = |B(S_A) \cap B(S_B)|$  being  $y$ , given  $x = |S_A \cap S_B|$ , is given by the binomial distribution  $B(m, \theta)$ , of  $m$  independent binary events with success probability  $\theta$ . That is,

$$Pr(Y = y|X = x) = \binom{m}{y} \theta^y (1 - \theta)^{m-y}. \quad (6)$$

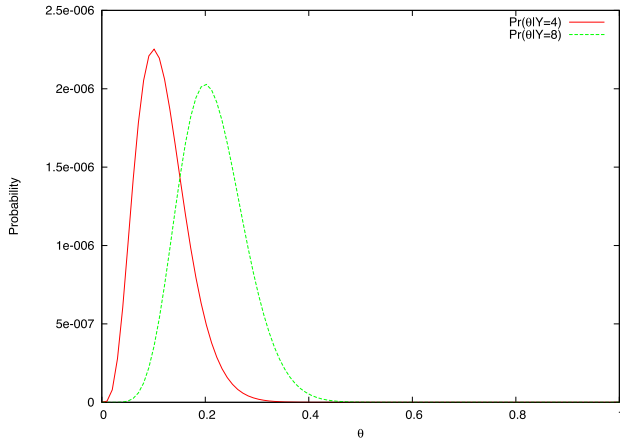
In a numerical example, consider two sets with  $n_A = 10$  and  $n_B = 8$ , whose BFs have  $k = 3, m = 40, q_A = 0.47$ , and  $q_B = 0.54$ . The conjunction of the BFs has 1s with a probability  $\theta = 0.33$  for  $x = 4$ . **Figure 2** shows the probability  $\theta$  with respect to  $x$ . Note that  $\theta$  is not 0 even for  $x = 0$  because a bit might be set to 1 by a false positive. Note also that  $\theta$  is monotone and onto mapping  $\{0, \dots, n\} \rightarrow [0, 1]$ , which makes the inverse mapping  $\theta^{-1}$  possible.

**Figure 3** shows the probability distribution of  $Pr(Y|X)$ , which is the conditional probability of the number of matching BF bits  $Y$  given the size of the intersection  $X$ . When  $X = 4$ , the number of matching bits in the BFs is distributed from 5 to 20 with a peak of 13.

### 4.2 Bayesian Estimation of $X$

Given known parameter values and  $Pr(X|Y)$ , we wish to identify the posterior distribution  $Pr(Y|X)$  using Bayes' rule.

One possible solution is an approximation based on a the likelihood value from a single observation, as described by



**Fig. 4** Posterior probability distribution  $Pr(\theta|Y)$  based on the beta distribution as a conjugate prior distribution.

Kantarcioglu et al. [11]. Their scheme suffers from the complexity of  $O(m)$ . That is, a secure scalar product will require  $m$  ciphertexts, which is greater than  $n$ . Moreover, the accuracy achieved is inadequate.

Instead, we will use recursive Bayesian estimation of several small BFs. That is more efficient because each individual BF used to perform the secure scalar product between two BFs will be smaller. Moreover, the iteration over multiple BFs improves the accuracy of the estimation. Given the properties of beta distribution, the iteration process can be performed with lightweight overheads.

Using the conjugate prior distribution of Eq. (6), we assume a beta distribution  $Be(\alpha, \beta)$ , which gives

$$Pr(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta}.$$

The initial prior distribution is given by  $Be(1, 1)$ , which yields a uniform distribution  $Pr(\theta) = 1$ . Using Bayes' theorem, we obtain the posterior probability of  $\theta$  given  $y$  as

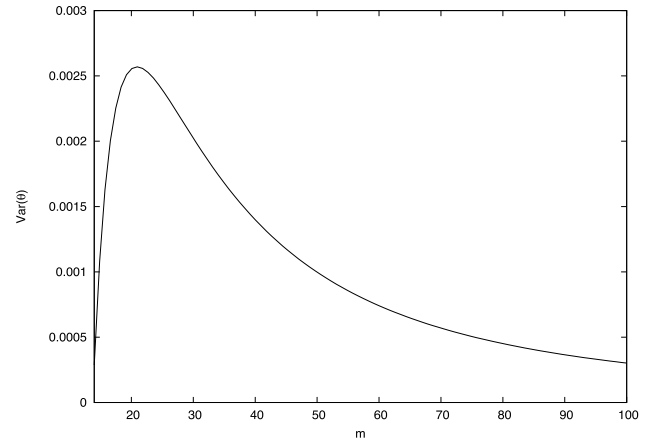
$$Pr(\theta|y) = \frac{Pr(\theta)Pr(y|\theta)}{\int Pr(\theta)Pr(y|\theta)d\theta} \propto Pr(\theta)Pr(x|\theta) \propto \theta^{\alpha-1+y}(1-\theta)^{\beta-1+m-y},$$

which results again in a beta distribution  $Be(\alpha', \beta')$  with new parameters as

$$\begin{aligned} \alpha' &= \alpha + y, \\ \beta' &= \beta + m - y \end{aligned}$$

For example, consider a posterior distribution  $Pr(\theta|Y)$  based on a BF with  $m = 40$ , for  $Y = 4$ , and 8, as shown in **Fig. 4**.

*Helicobacter Pylori* infection is considered to be an event that occurs to each individual independently. Modeling such a situation with the binomial distribution is considered to be reasonable; beta distribution, the natural conjugate prior distribution of the binomial distribution, is used as the prior distribution in our protocol mainly due to its mathematical convenience. The initial prior was set to the non-informative uniform distribution in the experiments. Nonetheless, it is difficult to exclude the subjectivity from the settings of the prior distributions, and the obtained experimental results need to be carefully examined.



**Fig. 5** Distribution of the variance of  $\hat{\theta}$ ,  $Var[\theta]$ , with respect to  $m$ , the size of the BF, for  $n = 10$ ,  $k = 3$ , and  $y = 14$ .

The mean of the beta distribution is denoted by  $E[\theta] = \alpha/(\alpha + \beta)$ . We can therefore estimate  $\hat{\theta}$  when the BFs of two sets have  $y$  matching bits as follows,  $\hat{\theta} = \frac{\alpha'}{\alpha' + \beta'} = \frac{1+y}{2+m}$ . After estimating  $\hat{\theta}$ , the size of the intersection is given by the inverse of Eq. (5), a mapping  $\theta^{-1}$ , as

$$\hat{x} = n_A + n_B - \frac{1}{k} \log_{1-\frac{1}{m}} \left( \hat{\theta} - 1 + \left(1 - \frac{1}{m}\right)^{kn_B} + \left(1 - \frac{1}{m}\right)^{kn_A} \right). \quad (7)$$

The inverse mapping can be evaluated locally in the final stage of privacy preservation (without encryption). We are not concerned that if Eq. (7) might appear complicated to evaluate.

#### 4.3 “Bootstrap” of BFs

To improve the accuracy, there are two approaches.

- (1) Enlarge the size of BF,  $m$ , and the estimate  $\hat{\theta}$ ,<sup>\*4</sup>
- (2) Estimate  $\hat{\theta}$  from multiple observations of  $Y_1, Y_2, \dots, Y_s$ .

Using a BF with more bits  $m$  could decrease the false positives in the membership test with the cost increasing as  $m$ . It is of interest that the value of  $m$  does not play a significant role in estimating of the intersection size, as we had expected. We will now show the mathematical properties that explain this observation.

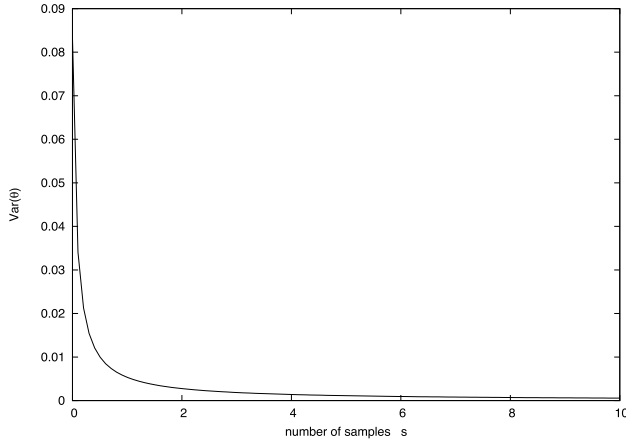
##### 4.3.1 (1) Variance of the beta Distribution for a Large BF

According to the known variance of the beta distribution  $Var[\theta] = \alpha\beta/((\alpha+\beta)^2(\alpha+\beta+1))$ , we illustrate the change of variance with respect to  $m$  in **Fig. 5**. Since the variance determines the standard deviation, which provides a confidence interval for the estimation, we can predict the accuracy via the reduction in variance. Figure 5 shows that the variance of  $\hat{\theta}$  decreases slightly as  $m$  increases. However, the reduction in variance is not significant, given the increased cost of the required ciphertexts. For example, a BF with  $m = 100$  requires 10 times more ciphertexts than that for an element in  $S$  with  $n = |S| = 10$ .

##### 4.3.2 (2) Variance from “Bootstrap” s Small BFs

Let  $y_1, y_2, \dots, y_s$  be the sequence of matching bits in  $s$  independent BFs for  $S_A$  and  $S_B$ . Recursive Bayesian estimation based on the sequence gives the posterior probability  $Pr(\theta|y_1, \dots, y_s)$  for the beta distribution  $Be(\alpha', \beta')$  defined by

<sup>\*4</sup> We do not consider the number of hash functions  $k$  because there are some constraints between  $m$  and  $k$ , such as  $kn < m$  and  $k = (\ln 2)m/n$  for minimizing false positives.



**Fig. 6** Distribution of the variance of  $\hat{\theta}$ ,  $\text{Var}[\hat{\theta}]$ , estimated from  $s$  independent BFs of the same size.

$$\alpha' = \alpha + \sum_{i=1}^s y_i, \quad \beta' = \beta - \sum_{i=1}^s y_i + sm.$$

The estimation of  $\hat{\theta}$  is provided from the mean of the beta distribution, namely

$$\hat{\theta} = \frac{\alpha + \sum_{i=1}^s y_i}{\alpha + \beta + sm} \quad (8)$$

**Figure 6** illustrates the reduction in the variance of  $\hat{\theta}$ . It implies that the bootstrapping reduces the confidence interval for the estimation of  $\theta$  significantly with increasing  $s$ .

#### 4.4 Proposed Scheme

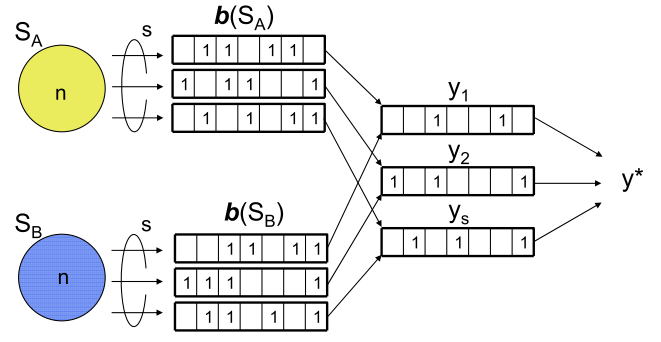
We give the procedure for estimating the size of the intersection without revealing each set in Algorithm 2. At Step 1, both parties  $A$  and  $B$  compute BFs for their  $n$ -element sets  $S_A$  and  $S_B$  with parameters, size of BF  $m$  and the number of hash function  $k$  such that  $k = (m/n) \ln 2$ . For tradeoff between efficiency and accuracy,  $k = 1$  and  $m = n / \ln 2$  can be used. Since this process can be performed locally and the hash function performs very efficiently, we consider the overhead is negligible. Both parties participate in the secure scalar product protocol (Algorithm 1), which is the most significant part in computation. The scalar product of two BFs,  $y$ , gives the number of common 1's bit in BFs, which can be divided into two integers, making the SFE possible to approximate  $\hat{\theta}$  in Eq. (8) without revealing any  $y_i$ . Note that the output of Step 2 are random shares,  $s_{i,A}$  and  $s_{i,B}$ , which satisfy Requirement 4. Step 5 is performed in public (or locally) after  $\hat{\theta}$  reaches at convergence.

The flow in improving accuracy through Bayesian estimation is illustrated in **Fig. 7**. Instead of extend the size of BF, we perform the secure scalar product protocols multiple times to get the sequence of  $y_1, y_2, \dots, y_s$ , which will be used to predict the  $\hat{\theta}$  in Bayesian estimation. Both parties iterate the test until the expected accuracy is given. The confidence interval is given by the standard deviation of estimated value.

#### 4.5 Security

The following theorem shows the security of Algorithm 2.

**Theorem 4.1** Suppose  $A$  and  $B$  behaves in the semi-honest model. Let  $S_A$  and  $S_B$  be inputs for Bloom Filter Bootstrap.



**Fig. 7** Bootstrap of Bloom Filters.

#### Algorithm 2 Bloom Filter Bootstrap $BFB(S_A, S_B)$

- Input: Alice has subset  $S_A$  of  $n$  elements. Bob also has  $S_B$ . Both know  $m$  (size of the BF),  $k$  (number of hash functions) and a threshold.
- Output:  $\hat{x}$  (estimate of the size of the intersection of  $S_A$  and  $S_B$ ).
- (1)  $A$  computes BF  $b(S_A)$  for  $S_A$  and  $B$  computes BF  $b(S_B)$ .
  - (2)  $A$  and  $B$  jointly perform Algorithm 1 to obtain  $s_{i,A}$  and  $s_{i,B}$ , respectively, such that  $y_i = s_{i,A} + s_{i,B}$  for  $i = 1, \dots, s$ .
  - (3)  $A$  sends  $s_{1,A}, \dots, s_{s,A}$  to SFE.  $B$  sends  $s_{1,B}, \dots, s_{s,B}$  to the SFE and make to evaluate if the right-hand side of Eq. (8) is greater than a given threshold. If it does not hold, stop (accept the null hypothesis).
  - (4)  $A$  and  $B$  reveal  $\sum_{i=1}^s s_{i,A}$  and  $\sum_{i=1}^s s_{i,B}$  and estimate  $\hat{\theta}$  using Eq. (8).
  - (5) Either  $A$  or  $B$  identifies  $\hat{x}$  using Eq. (7).

Then, the protocol Bloom Filter Bootstrap is secure in the sense of Definition 2.1.

*Sketch of the proof.* Since step 2 is multiple invocation of the scalar product protocol, the security is reduced to that of the scalar product protocol. Since step 3 is invocation of SFE, the security is reduced to that of SFE. By following the security proof in Refs. [8] and [14], the security of Bloom Filter Bootstrap is immediately proved. Note that computation in step 5 is performed by  $A$  without communication with  $B$ , the security is not compromised by execution of these steps.

#### 4.6 Complexity

We examine the complexities of our proposed scheme in terms of computation and communication costs. When these quantities are almost identical, we unify these by simply  $n$ . Protocols are compared in **Table 2**. In comparison with Ref. [11], we assume the ID-less model, where the size of BF can increase up to  $n^2$ .

Table 2 shows that the computational cost for  $A$  is linear to  $ms$ , while the cost for  $B$  is 0 (no modular exponentiation is required). Hence, it is preferable for outsourcing solution to our Requirement 3, where hospitals do not have powerful computational resources and become  $B$  in our protocol.

The protocols are classified into three groups. The first group is the scheme based on Oblivious Polynomial Evaluation. Scheme FNP [7] is designed to reveal not only the size of intersection but also the elements in the intersection. We show the performance for comparison purpose.

The second class, consisting of AES [1] and CT [5], is classified as Oblivious Pseudo-Random Functions (OPRF). AES depends on the commutative one-way function, while CT uses the RSA (Fig. 3 in Ref. [5]) and the blind RSA (Fig. 4 in Ref. [5]) encryptions. The privacy of scheme (Fig. 3 in Ref. [5]) is proved

**Table 2** Complexity Comparison of protocols.

	FNP [7]	AES [1]	CT [5]	KNV [11]	Proposed
primitives	OPE	commutative enc.	(blind) RSA	SSP w. BF	SSP w. BF
comp. at A	$n_A \log \log n_B$	$n_A + n_B$	$2n_A + 1$	$m$	$ms$
BF size	—	—	—	$n^2 \geq m > kn$	$m = n/\ln 2$
comp. at B	$n_B + n_A \log \log n_B$	$2n_A + n_B$	$n_A + n_B + 1$	1	1
complexity	$O(n_A \log \log n_B)$	$O(n)$	$O(n)$	$O(n^2)$	$O(n)$
comm. cost	$n_A + n_B$	$n_A + n_B$	$2n_A + n_B$	$m + 1$	$ms + 1$

OPE (Oblivious Polynomial Evaluation), SSP (Secure Scalar Product)

as the view of honest-but-curious party is indistinguishable under the One-More Gap Diffie-Hellman assumption in the random oracle model.

The last class is based on BF and Secure Scalar Product schemes. KNV [11] uses a single BF with large size, while ours iterates  $s$  independent BFs with small size. The sizes are shown in Table.

## 5. Accuracy Evaluation

### 5.1 Simulation with DBLP dataset

We evaluate the accuracy of the proposed scheme using a public dataset of author names, DBLP.

Four pairs of datasets  $S_A$  and  $S_B$  with  $n_A = n_B = 100$  were chosen from DBLP with the intersection sizes  $x = 20, 40, 60, 80$ . **Table 3** shows the experimental results for the estimation of  $x$ , for  $x = 20, 40, 60$ , and  $80$ , where we used a BF with of size  $m = 400$ , a number of hash functions  $k = 3$ , and iterated the estimation  $s$  times. The similar results for various BF sizes are given in **Table 4**. The results show that our scheme estimates the intersection within an error of  $\pm 1$ . The numbers of matching bits in the BFs,  $Y$ , are distributed according to the binominal distribution, as shown in **Fig. 13**. Note that all BFs estimate a size of the intersection close to the actual size of 40, but the differences are unstable.

### 5.2 Optimal BF design

The accuracy of estimation depends on the size of BF,  $m$ , and the number of hash function,  $k$ , and the iteration of testing,  $s$ . In order to clarify the strategy for optimal accuracy, we examine the Mean Absolute Error (MAE) with respect to  $m$  and  $k$ . **Figure 8** shows MAE in terms of  $m$  from 40 through 280, where  $n_A = n_B = 100$ ,  $x = 20$ ,  $k = 1$  and  $s = 20$ . **Figure 9** shows MAE with respect to  $k = 1, \dots, 6$  where  $m = 200$ . The MAE decreases as  $m$  increases, while the computational/communicational overhead increases accordingly. On the other hand, the increase of  $k$  does not reduce MAE.

A possible reason for the source of the error might be the restriction of  $m$  and  $k$ . As we discussed in Section 4.3, the optimal size for the BF is not trivial. We therefore suggest choosing  $k = 1$  first and then determining a near-optimal BF size by

$$m = kn / \ln 2 = 1 \cdot 100 / \ln 2 = 144.26.$$

Since large  $m$  increases the computational cost at secure scalar product, we conclude minimize  $k$ , i.e.,  $k = 1$  and optimize  $m = n / \ln 2$ .

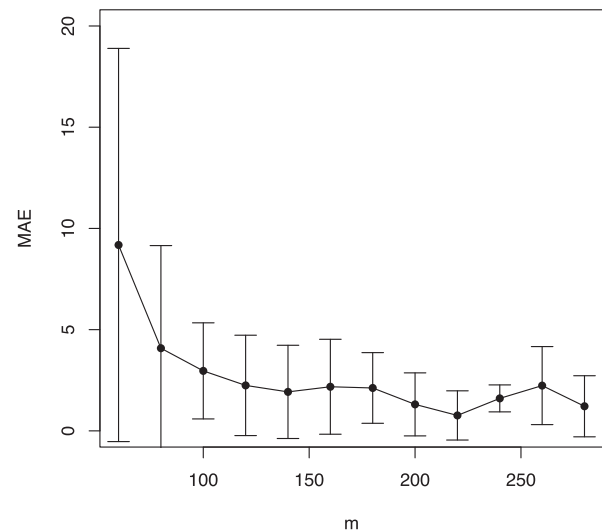
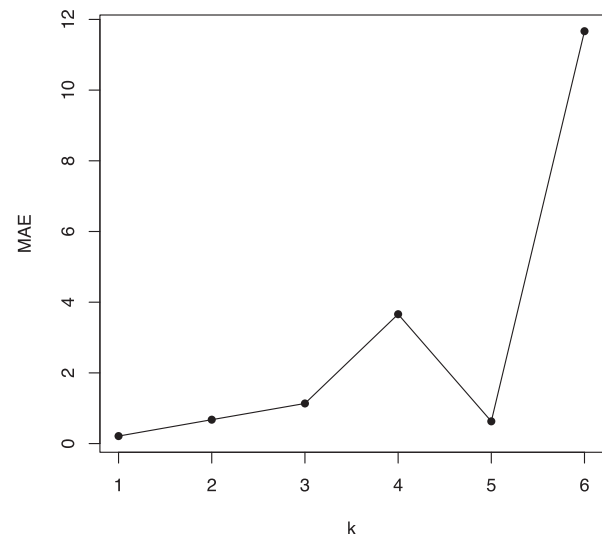
The distribution of the estimation for  $s = 10, 30$ , and  $100$  is shown in **Fig. 14**. As  $s$  increases, the distribution approaches a

**Table 3** Results of estimating  $X$  for various intersection sizes,  $x$ , for the dataset ( $n_A = n_B = 100$ ,  $m = 400$ ,  $k = 3$ ).

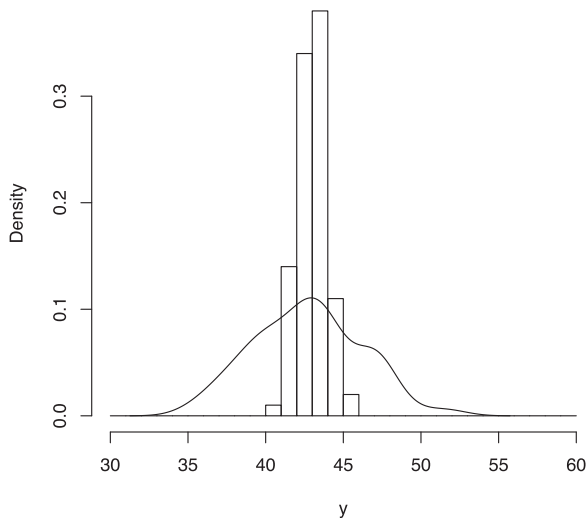
$x$	20	40	60	80
$E[Y]$	125.24	141.45	160.98	184.11
$\sigma(Y)$	6.78	5.92	5.34	5.15
$E(\theta)$	0.31	0.35	0.40	0.46
$\hat{x}$	19.523	38.869	58.969	79.411

**Table 4** Results of estimating  $X$  for various BF sizes,  $m$  for the dataset ( $n_A = n_B = 100$ ,  $x = 40$ ).

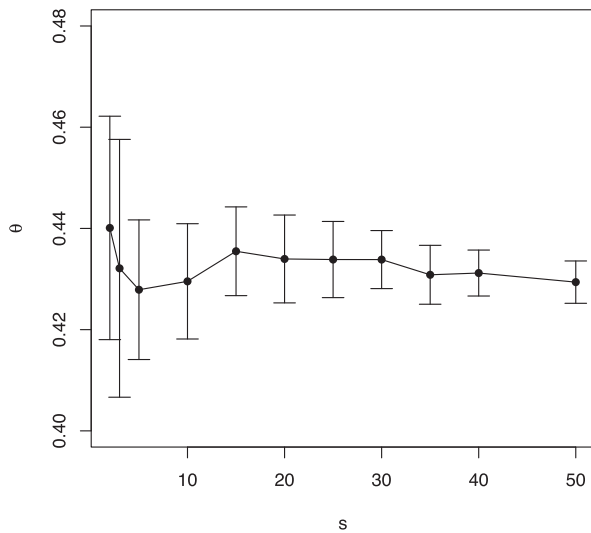
$m$	200	400	600	800
$k$	1	3	4	6
$E[Y]$	46.62	141.45	189.64	283.66
$\sigma(Y)$	3.146	5.923	6.436	7.488
$E(\theta)$	0.24	0.35	0.32	0.35
$\hat{x}$	39.490	38.869	39.604	39.227

**Fig. 8** Mean Absolute Error (MAE) with respect to the size of BF,  $m$ .**Fig. 9** Mean Absolute Error (MAE) with respect to the number of hash functions,  $k$ .





**Fig. 10** Distributions of matching bits in the BFs,  $y$ , shown in solid line, and distribution of the means,  $E[y]$ , in bars.

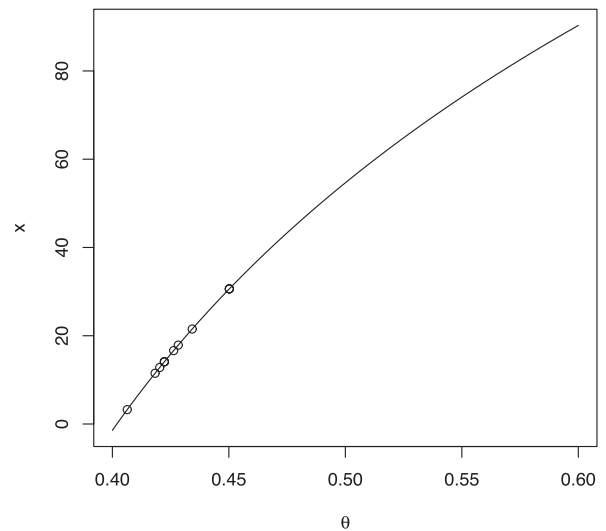


**Fig. 11** Conversion of probability of 1s bit in BF,  $\theta$ , with respect to iteration of BFs,  $s$ .

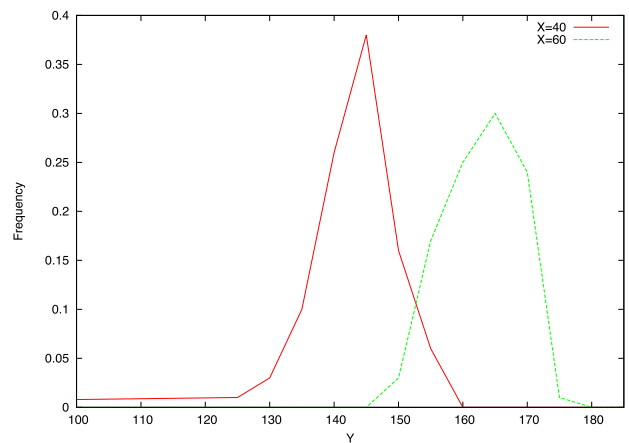
binominal distribution, with a mean equal to  $s\theta = 142.758$ . The accuracy improves as  $s$  increases as shown in **Fig. 15**. Note that the variance of the estimation  $\sigma(\theta)$  decreases as  $s$  increases, and the expected value  $E(\theta)$  is close to convergence.

The accuracy can be improved by iteration of small BF tests rather than increasing the size of BFs. In fact, **Fig. 10** demonstrates the reduction of variance of observation of  $E[Y]$ , indicated by bar plot, when  $s = 10$ . The solid line represents the distribution of  $Y$ , which is widely distributed than that of  $E[Y]$ . It is known as Central Limit Theorem [15], that as  $s$  increases, the amount of sampling variation decreases. **Figure 11** shows that the variance of estimated probability  $\hat{\theta}$  reduces as the iteration  $s$  increases. The experiment shows even small  $s = 10$  gives conversion of probability  $\theta$ . The selection of optimal  $s$  can be made based on the variance of the prediction of  $\theta$ . As we have showed in Section 4.3, the variance of beta distribution decreases with  $s$ , which determines the accuracy of approximation.

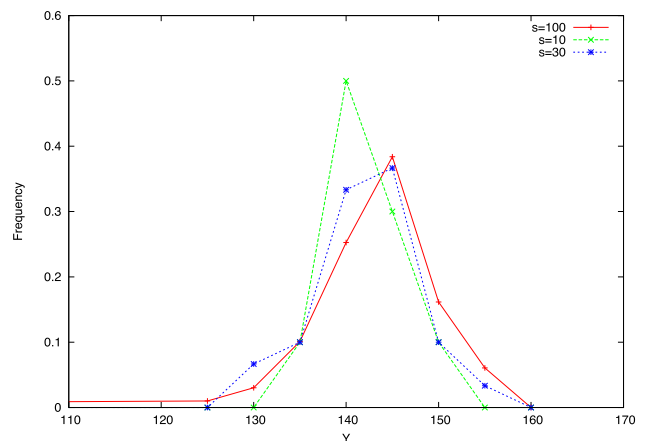
Finally, we obtain the estimate of intersection size,  $\hat{x}$ , by Eq. (7). We illustrate the distribution of  $\theta$  and the corresponding estimation of  $x$  in **Fig. 12**. The estimates  $\hat{x}$  are distributed



**Fig. 12** Relation between the size of intersection,  $x$ , and the probability of 1s bit,  $\theta$ .



**Fig. 13** Probability distribution of  $Y$ , the number of matching bits in BFs for  $X = 40$  and  $60$ .



**Fig. 14** Probability distribution of  $Y$ , the number of matching bits in BFs for  $s = 10, 30$ , and  $100$ .

normally at true size,  $x = 20$ .

### 5.2.1 Sufficient Number of Iterations

We have seen that increasing number  $s$  of iteration of Bayesian estimation improves the accuracy of estimation for particular case. In this section, we show that for any given problem with  $n, m, k$ , there exists sufficient number  $s$  of iteration to archive desired precision.

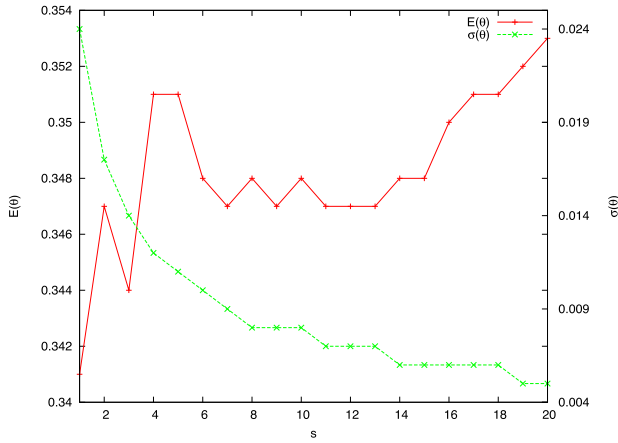


Fig. 15 Expected value  $E(\theta)$  and variance  $\sigma(\theta)$  of the Bayesian estimation of  $\theta$  with respect to the number of BF,  $s$ .

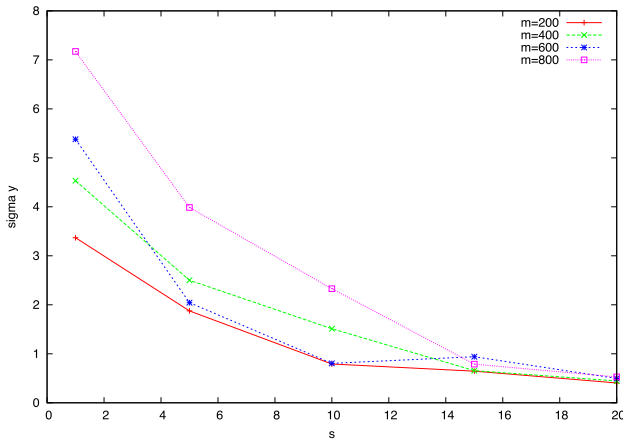


Fig. 16 Standard Deviation of  $\theta$  in terms of BF bit size  $m$  for 200, ..., 800, where  $n = 100$ ,  $x = 40$ ,  $k = 3$ .

First, let us remind that parameters  $n$ ,  $m$  and  $k$  are not independent in BF. As described in Section 2.1, parameters such that  $k = \ln 2 \cdot (m/n)$  minimize false-positive probability. So, we consider sufficient number  $s$  in terms of representative  $m$  only.

The estimation of  $x$  is equivalent to that of  $\theta$  since there is one-to-one correspondence. Hence, we estimate the accuracy of  $\theta$  instead of  $x$ . As studied in Section 4.3.2, the mean of the beta distribution gives an estimation of  $\theta$  as

$$\hat{\theta} = \frac{\alpha + \sum_{i=1}^s y_i}{\alpha + \beta + sm} = \frac{\alpha}{\alpha + \beta + sm} + \frac{\sum_{i=1}^s y_i}{\alpha + \beta + sm} \quad (9)$$

$$\leq \frac{\alpha}{\alpha + \beta + sm} + \frac{\sum_{i=1}^s y_i}{sm} = \frac{\alpha}{\alpha + \beta + sm} + E[Y] \frac{1}{m}. \quad (10)$$

Hence, the estimation of  $\theta$  is dominated by  $E[Y]$ , the average of  $s$  samples of the number of 1s bit in conjunction of two BFs. We regard the performing BF,  $y_i$ , as random sampling. According to the law of large numbers, the average of independent  $s$  samples should be close to the expected value with variance of  $\sigma^2/s$ . Hence, the confidence interval of estimation of  $\theta$  can be made small by increasing  $s$ .

Figure 16 shows the experimental results using DBLP dataset with  $n = 100$ ,  $x = 40$ ,  $k = 3$ . The standard deviation of sampled  $y_i$  is shown in terms of several BF size  $m = 200, 400, 600, 800$ . As shown in figure, the standard deviation converges around  $s = 15$  for all cases. Hence, we conclude that the proposed scheme finds sufficient number of  $s$  to estimate the set of intersection.

### 5.3 Performance

We implemented the proposed scheme in Java, JDK 1.6, with BigInteger class. As additive homomorphic public key algorithm, we use Paillier cryptosystem with 1024 bit key. With platform of commodity PC, Intel Core (TM) i7-663DQM, 2 GHz, 4 GB, running Windows 7 (64 bit), the encryption runs in  $t_e = 15.7$  [s], the decryption takes  $t_d = 21.5$  [s] in average. The secure scalar product of 64-bit vectors ( $n_A = n_B = 64$ ,  $x = 5$ ) is performed in 5.28 [s], i.e., 82.5 [ms/element]. With this platform, the processing time to deal with the problem in Ref. [11],  $n = 2,000$ ,  $k = 1$ , and  $m = n/\ln 2 = 2,885$ , is 4 minute and 125 second.

The naive pseudo identification in Section 3.2 suffers the complexity of  $n^2$ . Given the set with  $n = 100$ , Eq. (3) suggests the necessary range of hash function as  $\ell = n^2/2 \ln 1/\epsilon = 97,479$  with probability of 95%. The proposed scheme requires  $m = 200$ , which corresponds to  $s = 487$ .

## 6. Privacy-Preserving Risk Analysis of *H. pylori*

*Helicobacter pylori*, or *H. pylori*, is a bacterium that is found in the stomachs of two-thirds of the world's population. Epidemiology studies have shown that individuals infected with *H. pylori* have an increased risk of cancer of the stomach [10], [12].

Although *H. pylori* has been classified as a cancer-causing agent, it is not known how *H. pylori* infection increases the risk of cancer of the stomach. Some researchers have estimated that the risk of cancer the noncardiac region of the stomach is nearly six times higher for *H. pylori*-infected individuals than for uninfected people [9]. Some cohort studies revealed that the risk of gastric cardiac cancer among *H. pylori*-infected individuals was about one-third of that among uninfected individuals. The source of uncertainty is that the number of gastric cancers in the cohort study was too small to make a definitive statement. Cancer is a highly confidential matter and people will not reveal that they have it.

Our proposed methodology addresses the problem of epidemiology studies that preserve the privacy of the patients. The cryptographic protocol allows several small cohorts to be aggregated and analyzed for more certain evidence of increase or reduction of risk. Given two datasets of patients with cancer and *H. pylori*, the proposed protocol determines the size of the intersection of the two sets without revealing any entries in the datasets. With a secure hash function, the proposed scheme identifies a patient from their personal attributes.

### 6.1 Contingency Tables

The epidemiology study aims to determine whether an *H. pylori*-infected individual has increased the risk of gastric cancer. The evidence is shown by a measure of *relative risk* (RR), the probability of disease among exposed individuals divided by the probability of disease among the unexposed. Suppose that a sample of  $N$  individuals is arranged in the form of the  $2 \times 2$  contingency table in Table 5; the relative risk (RR) of *H. pylori* is estimated by

$$RR = \frac{\Pr(\text{cancer} | H. \text{pylori})}{\Pr(\text{cancer} | \text{unexposed})} = \frac{a}{a+b} \bigg/ \frac{c}{c+d} \approx \frac{ad}{bc},$$

**Table 5**  $2 \times 2$  Contingency table for *H. pylori* and stomach cancer.

<i>H. pylori</i>	Cancer	No cancer	total
Yes	$a$	$b$	$a + b$
No	$c$	$d$	$c + d$
total	$a + c$	$b + d$	$N$

where we assume  $a \ll b$  and hence  $a + b = b$ .

To examine whether *H. pylori*-infection increases the risk of cancer, i.e.,  $RR > 1$ , we test the null and the alternative hypotheses.

$H_0$ : The proportion of patients with cancer among individuals infected with *H. pylori* is equal to the proportion of patients with cancer among those uninfected.

$H_A$ : The proportions of patients with cancer are not identical in the two populations.

The chi-square test compares the observed frequencies in each category of the contingency table,  $O$ , with the expected frequencies given that the null hypothesis is true,  $E$ . To perform the test, we calculate the sum

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} = \frac{(N-1)((ad-bc) \pm N/2)^2}{(a+c)(b+d)(a+b)(c+d)},$$

where  $k$  is the number of cells in the table. The probability distribution of this sum is approximated by a  $\chi^2$  distribution with  $(2-1)(2-1) = 1$  degree of freedom. Alternatively, by taking its square root, we may assume that  $\chi$  is normally distributed with mean 0 and standard deviation 1.

## 6.2 Datasets

In our experiment, we have two datasets collected by independent agencies.

### (1) Patients with gastric cancer CAN.

The Chiba Cancer Center has performed an epidemiology study of causes and effects of cancer conditions since 1975 in Chiba Prefecture, Japan. **Table 6** shows the statistics for three years from 2003, used in this study. The dataset contains private attributes, including name, gender, birthday, mailing address, ZIP code, and medical treatments, e.g., patient ID, days of operations, day of death, type of cancers, and degree of tumor differentiation. The distribution of ages of patients is shown in **Fig. 17**.

### (2) Individuals infected with *H. pylori* PYL.

The Japanese Ministry of Health and Welfare (MHW) conducted a medical examination in 2001 in a small village in Chiba Prefecture. The dataset contains the number of *H. pylori*-infected individuals but their cancer status is not known.

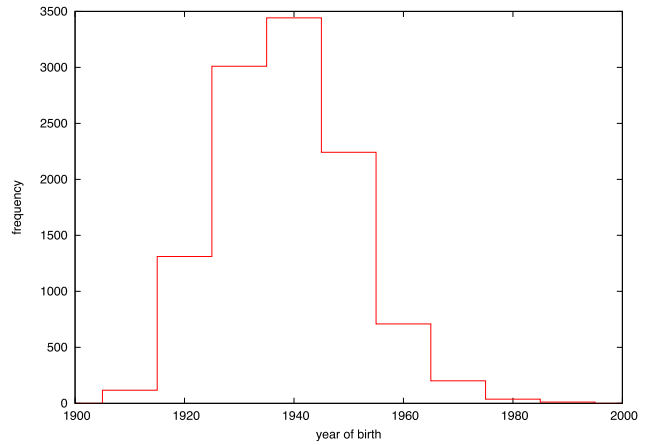
## 6.3 Hypothesis Testing

Our proposed algorithm estimates the size of the intersection of the two datasets, thus allowing the estimation of relative risk of *H. pylori*.

The statistics show that the size of the population in Chiba Prefecture in 2003 was 6,056,462 (3,029,486 male). The dataset in **Table 6** has  $n_A = 7,401$  recodes of patients with cancer. **Table 7** contains  $n_B = 2,629$  individuals infected with *H. pylori*. We apply a BF with size  $m = 14,000$ ,  $k = 1$  and  $s = 10$  to the two datasets and obtain the scalar product,  $y = b(\text{CAN}) \cdot b(\text{PYL})$  as

**Table 6** Chiba Cancer Center dataset CAN.

year	male	female	total
2003	2,330	1,134	3,464
2004	2,610	1,242	3,852
2005	2,559	1,205	3,763
total	7,500	3,581	11,081

**Fig. 17** Distribution of ages in CAN.**Table 7** MHW dataset of *H. pylori* infections PYL.

year	male	female	total
2001	2,671	5,206	7,877

**Table 8** Experimental results for CAN and PYL.

<i>H. pylori</i>	Cancer	No cancer	total
Yes	80	2,549	2,629
No	7,321	2,990,050	2,997,371
total	7,401	2,992,599	3,000,000 <sup>*5</sup>

$\mu(y) = 1,023.9$  on average. Based on Bayes' theorem, we estimate the probability  $\hat{\theta}$  in Eq. (8) as

$$\hat{\theta} = \frac{\alpha + \sum_i y_i}{\alpha + \beta + sm} = 0.073142.$$

From Eq. (7),  $\hat{x} = 81.1702$ , while the exact size of the intersection is 80. The number of individuals who are infected with *H. pylori* but do not have is therefore  $n_a - \hat{x} = 2,549$ . The other values can be obtained similarly. Finally, the numbers of individuals are summarized in **Table 8**.

An estimate of the relative risk of having cancer among *H. pylori*-infected individuals is therefore

$$RR = \frac{80 \cdot 222,964}{2,549 \cdot 7,321} = 12.81.$$

The chi-square test of the null hypothesis yields

$$\begin{aligned} \chi &= \frac{\sqrt{3,000,000 - 1}(80 \cdot 222,964 - 2,549 \cdot 7,321 - 3,000,000/2)}{\sqrt{7,401 \cdot 2,992,599 \cdot 2,629 \cdot 230,285}} \\ &= 28.71 > N(.05/2) = 1.960, \end{aligned}$$

which is too high to assume the null hypothesis. Therefore, we reject the null hypothesis at the 0.05 level of confidence.

In the experiment in Intel Xeon E5620 2.40 GHz, Memory 16 GB, the processing of the BF takes 17,030 second (= 4.7 hour), while the naive ID generation requires a scalar product of  $n^2 = 4.9 \times 10^7$ , which is estimated to be 223 hours.

<sup>\*5</sup> The number is referred from statistics in Chiba prefecture. There are potential individuals infected by *H. pylori* who was not counted in the table.

## 7. Conclusions

We have proposed an efficient algorithm for the estimation of the size of the intersection of two private sets. The proposed scheme gives a Bayesian estimation of the intersection size based on the mathematical properties of the number of matching bits in two BFs. A well-known secure scalar product protocol enables us to evaluate the number of matching bits in a privacy-preserving way and to test hypotheses that are useful in epidemiological studies. We have shown the properties of the accuracy of estimation for various parameters and the experimental results for the DBLP public dataset. One of our main results is that the bootstrap approach, iterating small BFs several times, is better than using a single large BF.

The extension of scalar product protocol to multiple parties can be done by replacing the Step 3 as that Bob forwards  $n$  ciphertexts computed with his secret vector as  $E(x_1)^{y_1}, \dots, E(x_n)^{y_n}$  to Carol who then perform the original Step 3 as  $c = E(x_1)^{y_1 z_1} \dots E(x_n)^{y_n z_n} / E(s_B)$ . The extension of Bloom filter to multiple parties is not trivial and one of our future work.

**Acknowledgments** Authors thank Dr. Haruo Mikami at Chiba Cancer Center for working with experiment using datasets. This work was supported by JSPS KAKENHI Grand-in-Aid Research (B), Grant Number 22300028 and was partly supported by FIRST program.

## References

- [1] Agrawal, R., Evfimievski, A. and Srikant, R.: Information Sharing Across Private Databases, *Proc. 2003 ACM SIGMOD International Conference on Management of Data*, pp.86–97, ACM Press (2003).
- [2] Broder, A. and Mitzenmacher, M.: Network Applications of Bloom Filters: A Survey, *Internet Mathematics*, pp.636–646 (2002).
- [3] Camenisch, J. and Zaverucha, G.: Private intersection of certified sets, *Financial Cryptography and Data Security*, pp.108–127 (2009).
- [4] Clifton, C., Kantarcioglu, M., Vaidya, J., Lin, X. and Zhu, M.: Tools for privacy preserving distributed data mining, *ACM SIGKDD Explorations Newsletter*, Vol.4, No.2, pp.28–34 (2002).
- [5] De Cristofarao, E. and Tsudik, G.: Practical private set intersection protocols with linear complexity, *Proc. 14th International Conference on Financial Cryptography and Data Security, FC'10*, pp.143–159, Springer-Verlag, Berlin, Heidelberg (online), DOI: 10.1007/978-3-642-14577-3\_13 (2010).
- [6] Fan, L., Cao, P., Almeida, J. and Broder, A.Z.: Summary cache: A scalable wide-area web cache sharing protocol, *IEEE/ACM Trans. Netw.*, Vol.8, No.3, pp.281–293 (online), DOI: 10.1109/90.851975 (2000).
- [7] Freedman, M.J., Nissim, K. and Pinkas, B.: Efficient private matching and set intersection, *Advances in Cryptology—EUROCRYPT*, pp.1–19, Springer-Verlag (2004).
- [8] Goethals, B., Laur, S., Lipmaa, H. and Mielikainen, T.: On private scalar product computation for privacy-preserving data mining, *Proc. 7th Annual International Conference in Information Security and Cryptology*, pp.104–120, Springer-Verlag (2004).
- [9] Helicobacter and Cancer Collaborative Group: Gastric cancer and Helicobacter pylori: A combined analysis of 12 case control studies nested within prospective cohorts, *Gut*, Vol.49, No.3, pp.347–353 (2001).
- [10] Atherton, J.C.: The pathogenesis of Helicobacter pylori-induced gastro-duodenal diseases, *Review of Pathology*, Vol.1, pp.63–96 (2006).
- [11] Kantarcioglu, M., Nix, R. and Vaidya, J.: An Efficient Approximate Protocol for Privacy-Preserving Association Rule Mining, *Proc. 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, PAKDD '09*, pp.515–524, Springer-Verlag (online), DOI: 10.1007/978-3-642-01307-2\_48 (2009).
- [12] Kusters J.G., van Vliet A.H. and Kuipers, E.: Pathogenesis of Helicobacter pylori infection, *Clinical Microbiology Reviews*, Vol.19, No.3, pp.449–490 (2006).
- [13] Lu, H., He, X., Vaidya, J. and Adam, N.: Secure Construction of Contingency Tables from Distributed Data, *Data and Applications Security XXII*, Atluri, V. (Ed.), Lecture Notes in Computer Science, Vol.5094, pp.144–157, Springer Berlin Heidelberg (online), DOI: 10.1007/978-3-540-70567-3\_11 (2008).
- [14] Malkhi, D., Nisan, N., Pinkas, B. and Sella, Y.: Fairplay – A secure two-party computation system, *USENIX Security Symposium*, pp.287–302 (2004).
- [15] Pagano, M., Gauvreau, K. and Pagano, M.: *Principles of biostatistics*, Duxbury Pacific Grove, CA (2000).
- [16] Paillier, P.: Public-key cryptosystems based on composite degree residuosity classes, *Advances in Cryptology - Eurocrypt 1999*, pp.223–238, Springer-Verlag (1999).
- [17] Ravikumar, P., Ravikumar, P., Fienberg, S.E. and Cohen, W.W.: A Secure Protocol for Computing String Distance Metrics, *PSDM* (2004).
- [18] Sakuma, J. and Wright, R.: Privacy-preserving evaluation of generalization error and its application to model and attribute selection, *Advances in Machine Learning*, pp.338–353 (2009).
- [19] Vaidya, J. and Clifton, C.: Privacy Preserving Association Rule Mining in Vertically Partitioned Data, *The 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.639–644 (2002).

## Editor's Recommendation

The authors propose a new efficient privacy-preserving scheme for estimating the size of the intersection of two given secret subsets. The proposed scheme successfully increases the efficiency of the estimation process compared to the previous schemes, by effectively combining Bloom filters and Bayesian estimation in its estimation algorithm. The proposed scheme is expected to increase the feasibility of the collaborative processing of big data with preserving citizens' privacy, which is one of the big challenges in the recent IT industries.

(Program Chair of Computer Security Symposium 2012,

Tsuyoshi Takagi)



**Hiroaki Kikuchi** was born in Japan. He received his B.E., M.E. and Ph.D. degrees from Meiji University in 1988, 1990 and 1994. After working in Fujitsu Laboratories Ltd. from 1990, in Tokai University from 1994, respectively, he joined Meiji University in 2013. He is currently a Professor at the Department of Frontier Media Science, School of Interdisciplinary Mathematical Sciences, Meiji University.

He was a visiting researcher at the School of Computer Science, Carnegie Mellon University in 1997. His main research interests are fuzzy logic, cryptographic protocol, network security, and privacy-preserving data mining. He is a member of IEICE, the Japan Society for Fuzzy Theory and Systems (SOFT), IEEE and ACM. He is a fellow of IPSJ.





**Jun Sakuma** was born in Japan. He received his B.E., M.E., and Ph.D. degrees from the Tokyo Institute of Technology, Tokyo Japan in 1997, 2000, and 2003. After working as a researcher at Tokyo Research Laboratory, IBM (2003–2004) and as an assistant professor at Tokyo Institute of Technology (2004–2009), he

joined University of Tsukuba, Tsukuba in 2009. He is currently an Associate Professor at the Department of Computer Science, School of System and Information Engineering, University of Tsukuba. He was a visiting researcher at DIMACS, Rutgers University in 2007. His main research interests include data mining, machine learning, data privacy, and cryptographic protocol. He is a member of the Japanese Society for Artificial Intelligence (JSAI) and IEICE.