Bloom Filterを用いた積集合サイズのベイズ推定とそのプライバシー保護疫学調査への応用

菊池 浩明† 佐久間 淳‡

† 東海大学情報通信学部通信ネットワーク工学科
106-8619 東京都港区高輪 2-3-23
kikn@tokai.ac.jp

‡ 筑波大学大学院システム情報工学研究科
305-8573 つくば市天王台 1-1-1 F934
jun@cs.tsukuaba.ac.jp

あらまし 本論文では，二つの部分集合の積集合の大きさを推定するプライバシー保護プロトコルを提案する．提案プロトコルでは，二つの集合のブルームフィルタが与えられたとき，推定するサイズの事前確率にベータ分布関数を仮定してベイズ推定を行う．ブルームフィルタは通信コストを下げ，ベイズ推定は推定の精度を向上する．ピロリ菌の癌に対するリスクの疫学調査への応用を議論する．

Application for Privacy-Preserving Epidemic analysis and Bays Estimation of Size of Intersection using Bloom Filter

Hiroaki Kikuchi† Jun Sakuma‡

†Dept. of Communication and Network Engineering,
School of Information and Telecommunication Engineering, Tokai University
2-3-23 Takanawa, Minato, Tokyo, 106-8619
‡Graduate School of SIE, Computer Science Department, University of Tsukuba
1-1-1 Tennodai, Tsukuba, 305-8573

Abstract This paper proposes a new privacy-preserving scheme for estimate the size of intersection of given two secret subsets. Given the inner product of two Bloom filters of given sets, the proposed scheme applies the Bayes estimation under assumption of beta distribution for a priori probability of the size to be estimate. The Bloom filter saves the communication complexity and the Bayes estimation improves the accuracy. The application of epidemic analysis of risk of cancer with H. pylori is discussed.

1 Introduction

二つのリストの要素を秘匿したままでその共通要素数，すなわち，積集合の大きさのみを評価する 2 者間のプロトコルを考える．ここで，互いの持つ集合の如何なる要素や積集合の要素は秘匿する．この問題は，セキュアな生体認証，プライバシー保護データマイニング，セキュア疫学調査などの多くの応用例に共通する基本的な要素技術の一つである．例えば，[9] では，がん罹患者のデータセットとピロリ菌の保有者のデータセットに秘匿内積プロトコルを適用して，ピロリ菌のがん罹患に対する相対危険度を算出し，疫学調査に応用している．

このような秘匿積集合評価を実行するいくつかの方式が知られている．準同型性を満たした公開鍵暗号によるセキュア内積プロトコル [7]，秘匿多項式評価 [3], 可換性を満たす一方向性関数を用いた秘匿積集合評価 [1]，などである．セキュア内積プロトコルはこれらの中で最も効率がよいが，二つの入力データは同一次元のベクトルで整合している要求条件がある．疫学調査を必要とする現実の多くのデータセットでは，異なる組織間で異種の ID 空間が用いられていることが多く，名前などを元に ID を近似しなくてはならない．

共通の ID がない問題に対するナイーブな解は，衝突困難性を満たした暗号学的ハッシュ関数を用いて

名前をハッシュ値に変換することである．しかし，十分な精度を保証するためには，比較する集合の大きさ n に対して $O(n^2)$ の値域のハッシュ値を用いる必要があり，疫学調査が対象としているビッグデータに適用できない．

そこで，集合の所属度を与える近似的なデータ構造である Bloom Filter(BF) を導入し，プライバシーを保護したまま安全に共通要素数を近似するアプローチを考える．プライバシー保護データマイニングへ BF を適用したのは，Kantarcioglu，Nix と Vaidya による相関ルール抽出がある [5]．BF の 1 のビット数と元の積集合の大きさとの間に成立する確率の近似式に基づいて，秘匿内積評価と秘密積計算を組み合わせている．これに対して，我々は事前確率を二項分布で近似してベイズの定理を用いて数値的に事後確率を推定する方法を提案した [8]．

しかし，二項分布は解析的には解けない欠点があった．そこで，本稿では，二項分布の自然な共役事前確率分布であるベータ分布 $Be(\alpha, \beta)$ を導入する．ベータ分布の性質に基づいて，BF のパラメーターを最適化し，精度の向上を試みる．

2 関連研究

2.1 ピロリ菌の相対危険度評価 [9]

[9] では，千葉がんセンター研究所予防疫学研究部によって収集されている，1975 年より千葉県内のがん登録から成るデータと，厚生省によって調査された 2001 年から 2002 年に，千葉県の一部の地域在住者の検診データからピロリ菌の保有者を抽出したデータに対して，秘匿内積プロトコルを用いて相対危険度を安全に算出した実験を報告している．

相対危険度 (relative risk) は，特定要因に暴露した群における比率の，暴露しなかった群での比率に対する比で定義される．例えば，ピロリ菌保有者という要因ががんに罹患する件数が表 1 の分割表で与えられたとき，相対危険度 RR は

$$RR = \frac{a}{a+b} \cdot \frac{c}{c+d} \approx \frac{ad}{bc}$$

で与えられる．ここで，一般に罹患率は小さいので，$a+b=b$ とみなしている．この相対危険度 $RR=1$ かどうかの有意性の検定は，$RR=1$ の仮説の元，統計量 χ が

$$\chi = \frac{\sqrt{N-1}\,((ad-bc) \pm N/2)}{\sqrt{(a+c)(b+d)(a+b)(c+d)}}$$

が標準正規分布 $N(0,1)$ に従うことで判定を行う．

それぞれのデータセットの大きさは既知なので，表の a が分かれば全て算出され，相対危険度の評価が可能である．結局のところ，疫学調査の問題は，がん患者のリストとピロリ菌保菌者のリストの照合を安全に行うことに帰着する．

表 1: 患者-対照調査によるデータの分布

要因	がん罹患	対照 (無)	罹患率
ピロリ菌	a	b	$a/(a+b)$
未感染	c	d	$c/(c+d)$

2.2 ナイーブな方式 – 名前 ID 変換

名前や住所などの属性情報の組みから成る集合 $A = \{a_1, \ldots, a_n\}$ がある．A の個数を n とする．A のデータには一意な ID がないので，ハッシュ関数 $h : \{0,1\}^* \to \{1, \ldots, \ell\}$ を適用して ID に変換する．このハッシュ関数の値域の大きさを ℓ と置くと，この名前・ID 変換は

$$h(A) = \{h(a_i) \bmod \ell \mid a_i \in A\}$$

と表される．ハッシュ関数のサイズ n によって衝突，すなわち，$h(a_i) = h(a_j)$ となる $a_i \neq a_j \in A$ が存在するので，変換された集合の大きさは必ずしも A の大きさにはならない．

データセットの大きさ n に対して最適なビット長 ℓ を考える．ℓ を挙げるほど，マッチングにおける精度は増加するが，比例して暗号処理速度と通信コストが増加する．明らかに，$\ell \geq n$ であるが，その大きさは自明ではない．

この問題は，$1/\ell$ の一様分布で生じる事象（ハッシュ値) を n 回繰り返し，全て異なっていることを意味しており，よく知られた誕生日パラドックスと同値である．大きさ ℓ のハッシュ関数で求めた n 個のハッシュ値が全て異なる確率は，

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

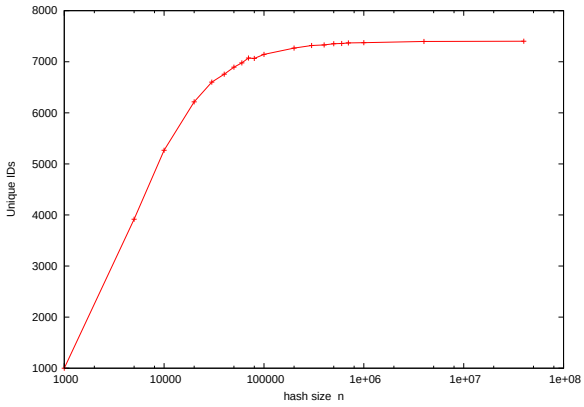

図 1: ハッシュ長 n に対するユニーク ID 数 [9]

で与えられる．従って，n 個のハッシュ値がユニークになる確率を ϵ とすると，

$$\frac{n^2}{2\ell} = \ln \epsilon^{-1} \tag{1}$$

の関係があり，n と精度 ϵ を与えた時のハッシュ関数のサイズ (値域) は $\ell = n^2/2 \ln \epsilon^{-1}$ で与えられる．例えば，がん患者サンプル数 $n = 7,000$ の時，ハッシュ値 $\ell = 4.7 \times 10^8$ は 95%の確率で一意になる．

図 1 は，患者データセットを対象として，ハッシュ長 ℓ を変化させた時に生成されるユニークな ID の総数の変化を表している．ハッシュ関数のサイズ ℓ に対して生成されるユニークな ID の個数は増加していき，$\ell = 4 \times 10^6$ の時にオリジナルのデータ数 $n = 7,500$ に達する．

従って，ナイーブな方式によって，秘匿内積プロトコルを実行するならば，約 n^2 個の暗号文を作って通信しなくてはならない．

2.3 Bloom Filter

S を n 個の要素からなる集合 $S = \{a_1, \ldots, a_n\}$ とする．Bloom filter (BF) は，k 個の独立したハッシュ関数 $H_i : \{0,1\}^* \to \{1, \ldots, m\}$, $(i = 1, \ldots, k)$ によって定められる S を表す m ビットのデータ構造である．S の BF を $B(S) = \bigcup_{a \in S} B(a)$ と定める．ここで，$B(a) = \{H_1(a), \ldots, H_k(a)\}$ とする．また，$\boldsymbol{b}$ を B から定まる m 次元ベクトル $(b_1, \ldots, b_m)$，すなわち，$i = 1, \ldots, m$ について，

$$b_i = \begin{cases} 1 & \text{if } i \in B(S), \\ 0 & \text{if } i \notin B(S), \end{cases}$$

と定める．例えば，$m = 8$ の時，$H_1(a) = 2, H_2(a) = 7$ ならば，$B(a) = \{2, 7\}$, $\boldsymbol{b}(a) = (0,1,0,0,0,0,1,0)$ である．ベクトルと集合の表現は一対一に対応しており，都合のよい表記を用いる．ここで，

$$\boldsymbol{b}(S_1) \cdot \boldsymbol{b}(S_2) = |B(S_1) \cap B(S_2)|$$

であることに注意せよ．

ある要素 a が集合 S に属するか否かは，

$$\forall i = 1, \ldots, k \; H_i(a) \in B(S) \tag{2}$$

で評価する．真に $a \in S$ ならば，式 (2) は常に成立するので，偽陰性 (false negative) はないが，$a \notin S$ が式 (2) を満たす，すなわち，false positive は生じる．$B(S)$ に要素 i がない (i ビット目が 0 である) 確率は，

$$p = \left(1 - \frac{1}{m}\right)^{kn} \approx e^{-kn/m} \tag{3}$$

で与えられることが知られている [4]. 従って，偽陽性が生じる確率は，

$$p' = \left(1 - (1 - \frac{1}{m})^{kn}\right)^k \approx \left(1 - e^{-kn/m}\right)^k \tag{4}$$

で与えられる．与えられた m と n に対して，k が小さすぎると式 (2) が容易に成立してしまい，逆に大き過ぎると B がほとんど 1 で埋まってしまう．[6] によると，$k^* = (\ln 2)m/n$ の時最適となることが知られている．

2.4 積集合の大きさを比較する秘匿プロトコル

Kantarcioglu，Nix と Vaidya らは，2 者間で相関ルールをデータマイニングする目的で，Bloom Filter を使った暗号プロトコルを提案している [5]. S_A を持つ A と S_B を持つ B が，互いの集合を秘匿したままで，共通の閾値 t について，

$$X = |S_A \cap S_B| \geq t \tag{5}$$

が成立するか否かだけを求めたい．ここで，X を共通集合の大きさを取る確率変数と定義する．

BF のベクトルにおける 1 のビット数は，登録する S の大きさに比例して増える．従って，集合を直接比較する代わりに，二つの BF の共通ビット数

$$Y = |B(S_A) \cap B(S_B)| \geq t'$$

を評価すれば，真の積集合の大きさが予測できる．Y は，X と同様に BF の積集合の確率変数である．

Broder らによる解析結果 [4] に基づくと，式 (5) の成立は，

$$Z_A + Z_B - Z_{AB} \geq Z_A Z_B \frac{1}{m}(1 - \frac{1}{m})^{-kt}$$

と同値である．ここで，Z_A, A_B は A, B における BF の 0 のビット数 ($Z_A = m - |B(S_A)|$)，$Z_{AB} = m - |B(S_A) \cap B(S_B)| = m - Y$ である．この不等式を秘匿したままで評価する為に，秘匿内積プロトコル [7] を用いて，

$$u_1 + u_2 = \boldsymbol{b}(S_A) \cdot \boldsymbol{b}(S_B) = m - Z_{AB}$$

となる u_1, u_2 の二つの値と，

$$v_1 + v_2 = (1 - 1/m)^{-kt}/mZ_A Z_B$$

となる v_1, v_2 を秘匿積プロトコルで計算する．最後に，分散比較プロトコルを用いて，

$$(Z_A + u_1 - m) + (Z_2 + u_2) \geq (v_1 + v_2)$$

で判定する．$n = 20,000$ のデータの例で，厳密に比較をすると 27 分かかるところが，BF で近似計算をすると 4 分に削減出来ることが報告されている [5]．

3 提案方式

3.1 アイデア

名前をハッシュして ID として，秘匿内積プロトコルを適用するナイーブな方式 [9] では，シンプルで実装も容易だが，$O(n^2)$ 個の暗号文を必要とするため，大規模なデータに適用できない．一方，Kantarcioglu らの方式 [5] では，BF を応用して小さなベクトル空間に変換することで $O(m)$ 個の暗号文のプロトコルを構成しているが，手順が複雑で比較的コストのかかる秘匿積評価プロトコルの実行を必要としている．そこで，二つの BF の共通ビット数 $Y = |B(S_A) \cap B(S_B)|$ にベイズの定理を適用し，共通要素数 $X = |S_A \cap S_B|$ を推定するアプローチを取る．Y を求めるには，m ビットのベクトルの秘匿内積プロトコル [7] を用いればよい．

[8] では，事前確率を二項分布で近似して数値的に事後確率を推定する方法を提案したが，二項分布は解析的には解けない．そこで，二項分布の自然な共役事前確率分布であるベータ分布 $Be(\alpha, \beta)$ を導入する．ベータ分布の性質に基づいて，BF のパラメーターを最適化し，精度の向上を試みる．

3.2 BF の共通ビット数 Y の確率分布

$X = S_A \cap S_B$ の要素数 $x = |X|$ が与えられたとき，二つの集合の BF の積，$\boldsymbol{b}(S_A) \wedge \boldsymbol{b}(S_B)$ のベクトルの 1 の要素数 y，すなわち，共通ビット数 $y = |B(S_A) \cap B(S_B)|$ を求めたい．

X と $S_A \cup S_B - X$ に分けて考える．まず，X の元 a については，必ず $a \in B(S_A) \cap B(S_B)$ である．よって，BF の積において，あるビットが 0 になる確率は，

$$q_X = (1 - \frac{1}{m})^{kx}$$

である．ここで，k はハッシュ関数 H_i の個数である．

一方，$S_A \cup S_B - X$ の要素についても，偶然に BF のビットが衝突して BF の積で 1 を生じさせることがあり得る．$S_A - X$ の BF のあるビットが 0 になる確率は，

$$q_A = (1 - \frac{1}{m})^{k(n_A - x)}.$$

同様に，$S_B - X$ の BF のビットが 0 になる確率も，$q_B = (1 - 1/m)^{(n_B - x)k}$ で与えられる．よって，$S_A \cup S_B - X$ の BF のあるビットが 1 になるのは，それぞれの余事象 (ビットが 1) の積，

$$(1 - q_A)(1 - q_B) = 1 - q_A - q_B + q_A q_B$$

で与えられる．

$\boldsymbol{b}(S_A) \wedge \boldsymbol{b}(S_B)$ のあるビットが 1 になるには，X の BF が 1 となるか，または，$S_A \cup S_B - X$ の BF が 1 となる時であるので，その確率 θ は，それぞれの和の確率により，

$$\begin{aligned} \theta &= 1 - q_X(1 - (1 - q_A)(1 - q_B)) \\ &= 1 - (1 - \frac{1}{m})^{kn_A} - (1 - \frac{1}{m})^{kn_B} \\ &\quad + (1 - \frac{1}{m})^{k(n_A + n_B - x)} \end{aligned} \tag{6}$$

である．以上より，$x = |S_A \cap S_B|$ の時に $y = |B(S_A) \wedge B(S_B)|$ となる条件付き確率は，成功確率 θ の m 回の独立試行の 2 項分布 $B(m, \theta)$ で，

$$P(Y = y | X = x) = \binom{m}{y} \theta^y (1 - \theta)^{m-y} \tag{7}$$

と与えられる．

例えば，$n_A = 10$, $n_B = 8$, $k = 3$, $m = 40$ の時，$q_A = 0.47$, $q_B = 0.54$ であり，$x = 4$ の時, $\theta = 0.33$ となる．x についての θ の関係を図 2 に示す．$x = 0$ の場合でも，偽陽性で BF のビットが 1 になるので，Y 切片は 0 ではない．θ は（明示していないが）x による単射 $\theta : \{0, \ldots, n\} \to [0, 1]$ であり，逆関数 θ^{-1} が定義できることに注意せよ．

この時，$P(Y|X)$ の確率分布を図 3 に示す．真の積集合の大きさ X に対して，BF の積集合の大きさ Y は 13 をピークとした分布を示している．

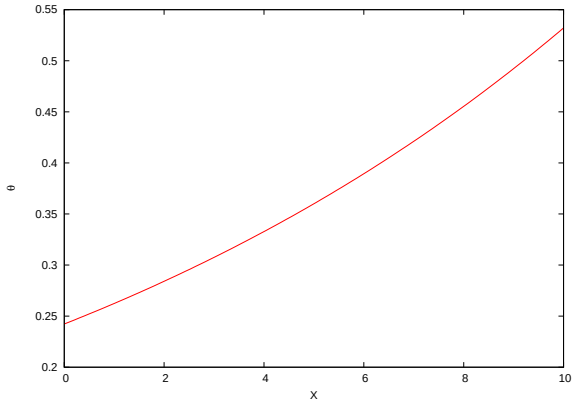

図 2: x についての θ の分布

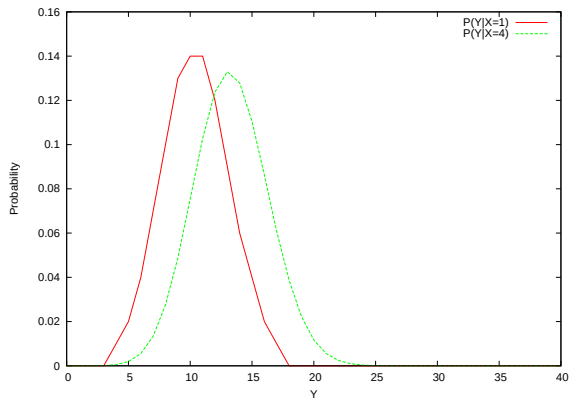

図 3: 確率分布 $Pr(Y|X = 1)$ と $Pr(Y|X = 4)$

3.3 共通要素数 X のベイズ推定

こうして定められた条件付き確率 $Pr(Y|X)$ が与えられたとき，ベイズの定理を用いて事後確率 $Pr(X|Y)$ を推定しよう．そのためには，事前確率 $P(X)$ を適切に決めなくてはならない．[8] では，事前確率を二項分布で近似して数値的に事後確率を推定する方法を提案したが，2 項分布は解析的には解けない．

そこで，式 (7) の二項分布の自然な共役事前確率分布である，ベータ分布 $Be(\alpha, \beta)$

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

に着目する．事前確率分布の初期値は $Be(1, 1)$，すなわち，一様分布 $P(\theta) = 1$ とする．

すると，事後分布はベイズの定理により，

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

と与えられ，やはりベータ分布となる．従って，ベータ分布のパラメータで，

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

の変換をするだけで事後分布が推定できる．例えば，$m = 40$, $Y = 4, 8$ の時の事後確率分布 $Pr(\theta|Y)$ を図 4 に示す．

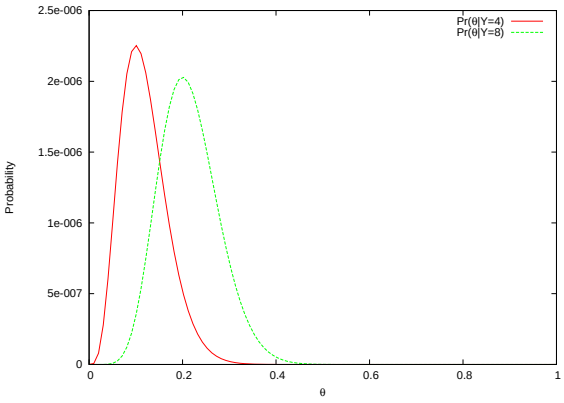

図 4: ベータ分布に基づく事後確率分布 $Pr(\theta|Y)$

ベータ分布の平均値は $E[\theta] = \alpha/(\alpha + \beta)$ であることを利用すると，y を観測した時の θ は，

$$\hat{\theta} = \frac{\alpha'}{\alpha' + \beta'} = \frac{1+y}{2+m}$$

で最尤推定できる．$\hat{\theta}$ が与えられれば，$\hat{x}$ は式 (6) の θ の逆関数を用いて

$$\begin{aligned} \hat{x} &= n_A + n_B \\ &\quad -\frac{1}{k}\log_{1-\frac{1}{m}}\left(\hat{\theta} - 1 + (1-\frac{1}{m})^{kn_B} + (1-\frac{1}{m})^{kn_A}\right) \end{aligned} \tag{8}$$

と推定できる．秘匿計算の際には，この処理は最後に局所的に行えばいいので，問題は $\hat{\theta}$ をいかに効率よく正確に求めるかにかかっている．

3.4 複数 BF による精度向上

推定精度を高めるには，次の方針が考えられる．

(1) BF のビット長 m を広げる[1]．

(2) ハッシュ関数を変えて，BF の共通ビット数 Y の観測を s 回繰り返す．

BF のビット数 m を増やせば精度は期待できるかも知れないが，秘匿内積の暗号文数が増えて性能は落ちる．一方，(2) は BF をある種のサンプリングとみなして，その試行を繰り返すことで精度を高める．

(1) ベータ分布の分散 $Var[\theta] = \alpha\beta/((\alpha+\beta)^2(\alpha+\beta+1))$ を用いて，m に対する分散を図 5 に示す．BF のビット数 m を決めると，偽陽性を最小化する様に $k = (m/n)\ln 2$ が決まるため，m が推定精度を支配するパラメータである．m を増やすと内積計算にかかる処理時間は比例して大きくなるが，精度は単調には下がらないことを示している．

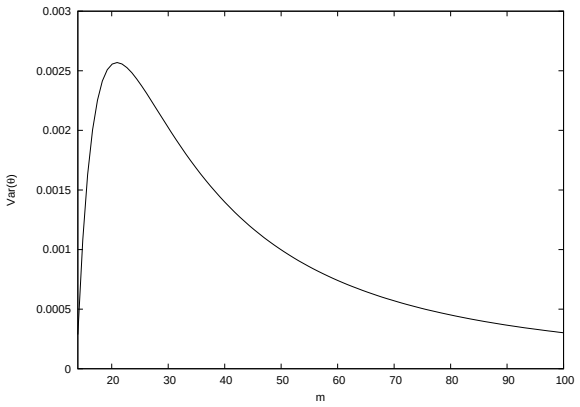

図 5: BF のサイズ m に対する $Var[\theta]$ の分布 ($n = 10, k = 3, y = 14$)

(2) BF を s 回繰り返して，観測された共通ビット数を $y_1, y_2, \ldots, y_s$ とする．この観測値について，ベイズ推定を逐次的に繰り返すと，事後確率のベータ関数が

$$\alpha' = \alpha + \sum_{i=1}^{s} y_i,$$

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

によって得られる．よって，s 回ベイズ推定をした時の θ の推定値は，ベータ関数の平均より，

$$E[\theta] = \frac{\alpha + \sum_{i=1}^{s} y_i}{\alpha + \beta + sm} \tag{9}$$

で与えられる．図 6 に s に対する $\hat{\theta}$ の分布を示す．s に対して，θ の分散が著しく減少している．

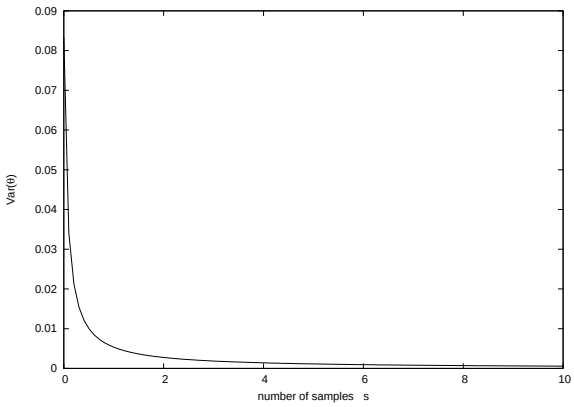

図 6: s 回の BF から推定される $Var[\theta]$ の分布

3.5 提案方式

1. A と B は，BF のパラメータ m, k を合意する．$i = 1$ とする．
2. A と B は，$B(S_A), B(S_B)$ をそれぞれ計算し，セキュア内積プロトコルを適用し，$y_i = \boldsymbol{b}(S_A) \cdot \boldsymbol{b}(S_B)$ を求める．
3. $y_1, \ldots, y_s$ について $\sigma(Y)$ を求め，決められた閾値に達するまで，Step 2 を繰り返す．
4. 式 (9) により，$\hat{\theta}$ を求め，式 (8) により，$\hat{x}$ を推定する．

4 DBLP を用いた評価

DBLP でインデックスされた論文データセット [10] から抽出した著者名でリスト S_A, S_B を用意し，提案方式に従って共通要素数 X を推定した．

[1]なお，(1) の変形として，BF を作るハッシュ関数の数 k を最適化する方法も考えられるが，$kn < m$ という自明な関係と，偽陽性を最小化する $k = (ln2)m/n$ の制約があるため，ここでは m のみを考える．

表 2: 共通要素数 X についての推定結果 ($n_A = n_B = 100, m = 400, k = 3$)

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

リストの大きさを $n_A = n_B = 100$ と一定にして，共通要素数 $x = 20, 40, 60, 80$ と変化させて，提案方式で推定した結果を表 2 に示す．BF のビット数 $m = 400$, ハッシュ関数の数は $k = 3$ としている．サンプル数は $s = 100$ と十分大きく取っている．± 1 の誤差の範囲内で推定している．$X = 40, 60$ の時の BF の共通ビット数 Y の分布を図 7 に示す様に，理論通りに 2 項分布に従っており，このばらつきが推定値の誤差の原因と考えられる．

次に，BF のビット数 m を変えて，$n_A = n_B = 100, x = 40$ のデータに適用した結果を表 3 に示す．k は $(m/n) \ln 2$ により最適値を定め，真の共通要素数 $x = 40$ に対して，推定結果 $\hat{x}$ は誤差 1 に収まっているが，その誤差の大きさは m には依存せずに変動している．m を増やしても，秘匿計算に時間がかかるだけで精度が向上しないのであれば，k を最小化，すなわち，$k = 1$ として，$m = kn/\ln 2 = 144.26$ で BF を構成し，それを複数回繰り返す方針 (2) が妥当と思われる．

そこで，試行回数 s についての推定値の変化を調べた．$s = 10, 30, 100$ に対する共通ビット数 Y の分布を図 8 に示す．$s = 10$ は分散が小さいが平均値は期待される値より離れている．$s = 30, 100$ と増加するに従って，二項分布の平均 $s\theta = 142.758$ に近づいている．更に，s に対する $\hat{\theta}$ の平均値 $E(\theta)$ と分散 $\sigma(\theta)$ を図 9 に示す．試行回数 s に対して，分散が小さくなり，平均が収束していることが観察される．

ナイーブに $n = 100$ 個のリストをハッシュした場合には，2.2 節で示した式 (1) により，$\ell = n^2/2 \ln 1/\epsilon = 97479$ で 95%の精度が得られることが分かっている．その時の秘匿内積計算の暗号文数と比較すると，$m = 200$ の BF を 487 回以上繰り返しては意味がない．

表 3: BF のビット数 m についての推定結果 ($n_A = n_B = 100, x = 40$)

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

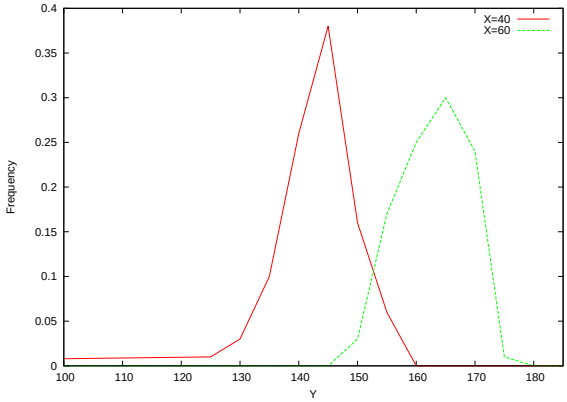

図 7: BF の交わり数 Y の分布

4.1 パフォーマンス

図 10 にハッシュ長 n を変化させた時のセキュア内積プロトコルの実行処理時間を示す．n に対して線形に増加している．

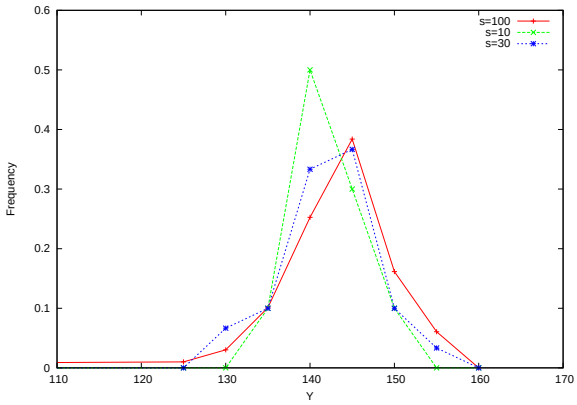

図 8: BF 試行回数 s に BF の共通ビット数 Y の分布

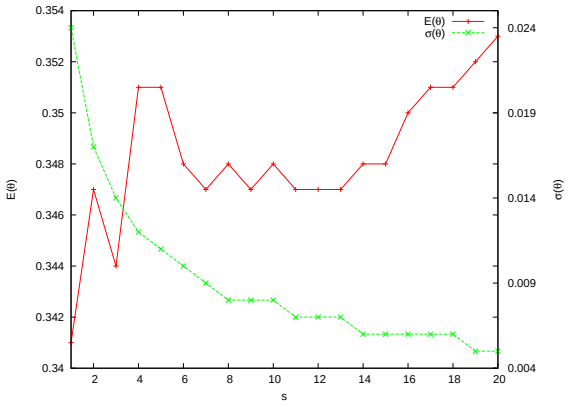

図 9: BF の試行回数 s についての推定値 $\hat{\theta}$ の期待値 $E(\theta)$ と分散 $\sigma(\theta)$ の変化

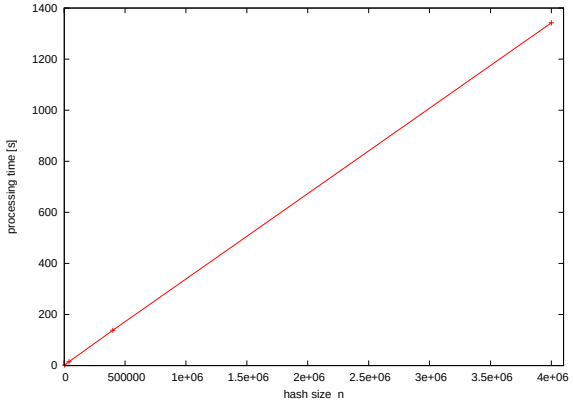

図 10: ハッシュ長 n に対する処理時間

5 結論

二つの集合を秘匿したままで共通要素数を推定する暗号プロトコルを提案した．提案方式は，小さなビット数の BF を繰返し求めて秘匿内積プロトコルを行うことで，計算効率と推定精度を向上している．BF のビット数 m を上げるよりも，小さな m を繰り返して用いる方が効率が良いことを示した．

参考文献

[1] Rakesh Agrawal, Alexandre Evfimievski, and Ramakrishnan Srikant, “Information shareing across private databases”, in proc. of ACM SIGMOD Intl. Conf. on Management of Data, 2003.

[2] Vaidya, J. and C. Clifton, “Privacy preserving association rule mining in vertically partitioned data”, The Eighth ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, SIGKDD, ACM Press, Edmonton, Canada, pp. 639-644, 2002.

[3] M. J. Freedman, K. Nissim, and B. Pinkas, “Efficient private matching and set intersection”, EUROCRYPT 2004, LNCS 3027, pp. 1?19, Springer-Verlag, 2004.

[4] A. Broder, M. Mitzenmacher, “Network Applications of Bloom Filters: A Survey”, Internet Math, Volume 1, Number 4 (2003), 485-509.

[5] Murat Kantarcioglu, Robert Nix and Jaideep Vaidya, “An Efficient Approximate Protocol for Privacy-Preserving Association Rule Mining”, 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD 2009), LNCS 5476, Springer, pp. 515-524, 2009.

[6] Li Fan, Pei Cao, Jussara Almeida, and Andrei Z. Broder, “Summary cache: a scalable wide-area web cache sharing protocol”, IEEE/ACM Trans. Netw. Vol. 8, No. 3, pp. 281-293, 2000.

[7] Bart Goethals, Sven Laur, Helger Lipmaa and Taneli Mielikainen, “On Private Scalar Product Computation for Privacy-Preserving Data Mining”, The 7th Annual International Conference in Information Security and Cryptology (ICISC 2004), Vol. 3506 of LNCS, pp. 104-120, 2004.

[8] 菊池，佐久間，“Bloom フィルタを用いたマッチング数の秘匿比較”, コンピュータセキュリティシンポジウム 2011， 情報処理学会 ,2C4-3, pp. 516-521, 2011.

[9] 菊池，佐久間，三上， “プライバシーを保護したピロリ菌疫学調査”, 第 26 回人工知能学会, 3I2-OS-20-9, pp. 1-4, 2012.

[10] A Citation Network Dataset, V1, (http://arnetminer.org/citation).