DEIM Forum 2013 F8-2

分散データベースにおける共有データの k-匿名化フレームワーク

櫻田　潤一† 　上土井陽子†† 　若林　真一††

† 広島市立大学大学院情報科学研究科　〒 731-3194 広島市安佐南区大塚東 3-4-1
†† 広島市立大学大学院情報科学研究科　〒 731-3194 広島市安佐南区大塚東 3-4-1
E-mail: †sakurada@lcs.info.hiroshima-cu.ac.jp, ††{yoko,wakaba}@hiroshima-cu.ac.jp

あらまし　異なる機関に保存された分散データベースを結合し，解析することで有益な情報を得ることができる．しかし，データ保有者のプライバシ保護を行うことができないのであればデータの共有・結合は実現されない．プライバシ保護の方法として k-匿名化が用いられている．従来研究では同じ ID の集合間での k-匿名化を満たす結合を行うためのフレームワークが提案されている．本稿では，異なる ID 要素集合の共通集合のみを安全に匿名化するためのフレームワークを提案する．

キーワード　分散データベース，k-匿名化

Jun'ichi SAKURADA†, Yoko KAMIDOI††, and Shin'ichi WAKABAYASHI††

† Graduate School of Information Sciences, Hiroshima City University
3-4-1 Otuka-higashi, Asaminami-ku, Hiroshima, 731-3194, Japan
†† Graduate School of Information Sciences, Hiroshima City University
3-4-1 Otuka-higashi, Asaminami-ku, Hiroshima, 731-3194, Japan
E-mail: †sakurada@lcs.info.hiroshima-cu.ac.jp, ††{yoko,wakaba}@hiroshima-cu.ac.jp

Key words　distributed databases, k-anonimization

1. はじめに

近年，様々な機関で蓄積された情報を共有，統合しデータマイニングに利用することの有効性が検討され始めた．データマイニングは大規模なデータベースを解析することで有益な情報を抽出する技術であるが，データ保有者のプライバシ保護を行うことができないのであればデータの共有は実現されない．プライバシ保護の方法としてデータセット内に同一データが k 個以上存在するまで一般化する k-匿名化が用いられている．本研究では複数の機関で分散管理されているデータベースをプライバシを保護した上で統合する分散 k-匿名化問題について考察する．

本研究で考察する分散 k-匿名化問題は，複数の機関が別々に保存している部分データを，他の機関に最終結果として出力するデータより読み取れる情報以外の情報を漏らすことなく，k-匿名性を満たす 1 つのデータに統合して出力する問題である．分散 k-匿名化問題に対する従来研究 [1] では各サイトで局所的に k-匿名部分データを作成し，2 つのサイトの局所的な k-匿名データの結合後にも k-匿名性を満たすデータとなるかを判定するため，2 つの集合間の積集合の要素数に関する判定問題を機密性を保ちながら解く k-匿名化の分散フレームワークが提案されている．従来研究では完全なデータセットが 2 つのサイト上に別々に保存されており，共通の識別子を用いた結合が可能であるとの仮定で，双方のプライバシ情報を保護しながら協調して k-匿名性を満たすフレームワークが提案されている．

本研究で扱うデータセットは表 1 のように表現されているとする．各行は個人情報を表し，各列は属性値を表すとする．属性 ID は各行異なる値をもつが個人を特定できない非識別子とする．属性勤務地，年齢，給与は準識別子と呼ばれる属性であり，組み合わせることにより個人を特定できる情報を含んでいる可能性がある．以降では表での各個人の情報に対応する行をレコードと呼ぶ．つまり，本研究で対象とするデータセットは従来研究とは異なり統合したときに部分的に属性値を欠いている行が存在してもよい．

本稿では，2 つの機関 P_1, P_2 がそれぞれ保持するデータセット T_1, T_2 双方において，識別子を表現する属性が存在するとする．このとき，データセット T_1, T_2 において共に有する識別子をもつタプルを結合し，かつ，k-匿名化する共有データ k-匿名化問題を考える．そのとき，分散 k-匿名化問題に対する従来分散フレームワークを拡張し，より広い分散データベースのクラスを対象とする共有データ k-匿名化問題に対する分散フレームワークを提案する．

表 1　対象とするデータセットの例

ID	勤務地	年齢
1	千葉	34
2	東京	33
3	岡山	46
4	広島	26
5	山口	49
6	鳥取	26
7	埼玉	34
8	広島	48
9	広島	26
10	岡山	45
11	大阪	32
12	奈良	34
13	京都	30

ID	給与
1	370,000
2	410,000
3	390,000
4	250,000
5	410,000
6	220,000
7	370,000
8	430,000
9	210,000
14	310,000
15	350,000
16	320,000

2. 準　　備

2.1 *k*-匿 名 化

本稿では，データセットはレコードの集合とし，各レコードは複数の属性の値の組合せで表現されているとする．k-匿名化とは元のデータセットを k-匿名性を満たすデータセットに変換することでプライバシ保護を行うものである．ここで k-匿名性を満たすとは，データセット内の各レコードと同じ値の組合せを持つレコードが少なくとも $(k-1)$ 個存在することとする．データを匿名化するため，特定の値を一般化する．一般化の規則は入力として各属性に対して値一般化階層の形式で与えられている．

例えば図 1 のように値一般化階層が設定されている場合，福岡，大分出身を九州出身と一般化することができる．

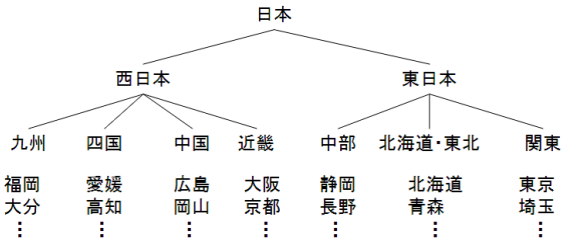

図 1　勤務地の値一般化階層

2.2　従来分散 *k*-匿名化フレームワーク

従来研究における分散 k-匿名化フレームワークを図 2 に示す．

例を用いて従来分散 k-匿名化フレームワークについて説明する．表 1 の 2 つのデータセットにおいて ID[1-9] を含むデータセットを 3-匿名性を保つよう結合する場合を考察する．まず属性〈勤務地，年齢〉を持つデータセットは図 1, 図 3 の値一般化階層より 1 段階一般化する．また属性〈給与〉を持つデータセットも図 4 から値を一般化する．しかし，給与を 1 段階あげただけでは結合後に 3-匿名性を保てない．よって判定で結合不

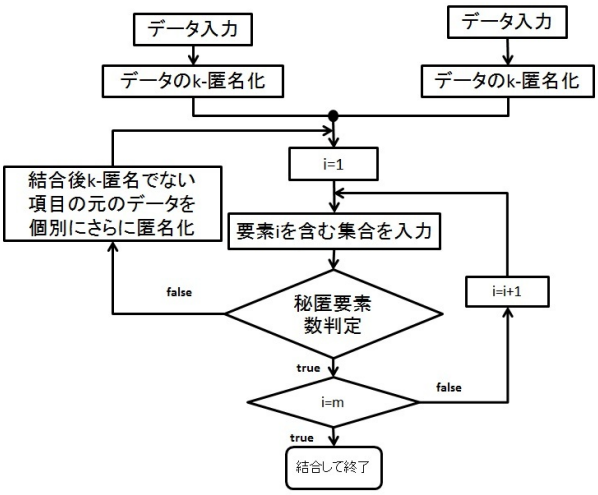

図 2　従来分散 k-匿名化フレームワーク

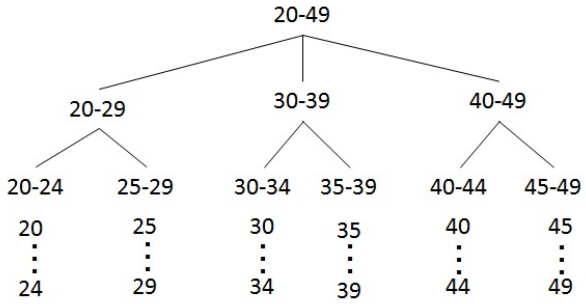

図 3　年齢の値一般化階層

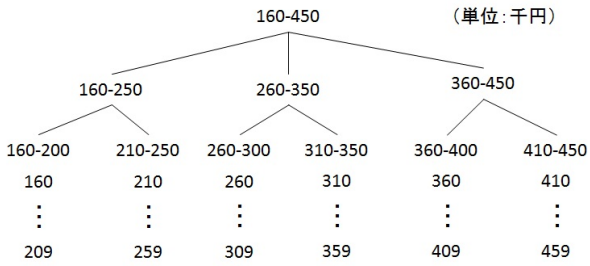

図 4　給与の値一般化階層

可となったデータセットの属性〈給与〉に関してさらに値を一般化する．それにより，結合後も 3-匿名性を持つデータセット表 2 を作成することができる．また，分散管理されているデータの局所的 k-匿名化は文献 [4] に記されている Datafly で行う．局所的な匿名化後の同じ値のレコードの集合において積集合の要素数を調べることで結合後に匿名性を満たすかどうか判定する．

例えば，勤務地が中国地方，年齢が 45-49 となる ID の集合を $D_1 = \{3, 5, 8\}$ とし，給与が 360-450 となる ID の集合を $D_2 = \{1, 2, 3, 5, 7, 8\}$ としたとき，$|D_1 \cap D_2| = 3 \geqq k$ を判定する．

2.3　要素数判定問題

ここでは別々に管理されている k-匿名化データセットの結合後も k-匿名性を満たすか判定するときに用いる要素数判定問題

を定義する．ここで整数集合 $\{1,...,m\}$ の部分集合 D_1, D_2 は２つの機関 P_1, P_2 それぞれがもつ情報とする．m は２つの集合の要素が存在する値域の最大値とする．

[要素数判定問題] 整数集合 $\{1,...,m\}$ の部分集合 D_1, D_2 において，$|D_1 \cap D_2| \geqq k$ を満たすか，否か． ◇

2.4 結合可能判定

結合可能判定を行う回数は秘匿でないならば最低 $|G_1| \times |G_2|$ 回の要素数判定を行えば結合可能と判定可能である．このとき，要素数判定は $|D_1 \cap D_2| = 0$ も $true$ である．従来分散 k-匿名化フレームワークにおいて秘匿要素数判定の入力に要素 i を含む集合を入力としている．これは秘匿要素数判定において互いの $|D_1|, |D_2|$ の情報を相手に漏らさないためである．このため，従来分散 k-匿名化フレームワークにおいて最低 $O(m)$ 回の秘匿要素数判定が必要となる．

3. 提案分散共有 k-匿名化フレームワーク

本稿では，２つの機関 P_1, P_2 がそれぞれ保持するデータセット T_1, T_2 双方において，識別子を表現する属性が存在するとする．このとき，データセット T_1, T_2 において共に有する識別子をもつタプルを結合し，かつ，k-匿名化する共有データ k-匿名化問題を考える．２つの機関に保存されたデータセットは表 1 のように各 ID に該当するデータを保持する場合と保持しない場合があるものとし，それぞれの機関を P_1, P_2 とし，それぞれのデータセットを T_1, T_2 とし，全体のデータセットの識別子の最大値を m とする．このとき，各機関が局所的に k-匿名化した k-匿名化データを表現する各データセットのグループ分けを表現するための用語を以下に定義する．ここで $i = 1, 2$ とする．

- データセット T_i の分割 G_i
- 分割 G_i 内の集合数 $n_i = |G_i|$
- 分割 G_i の要素 $D_i^j (1 \leqq j \leqq n_i)$.

つまり，$G_i = \{D_i^1, D_i^2, ..., D_i^{n_i}\}$ と表現できる．

従来研究において表 1 のように ID[10-16] のように一方のみが固有の ID を含む場合，秘匿要素数判定の入力として要素 i を含む集合を作成できない．また，空集合を入力として要素数判定を行うと false の判定のみが出力されてしまう．そこで本研究では秘匿要素数判定において $|D_1 \cap D_2| = 0$，$|D_1 \cap D_2| \geqq k$ のどちらかを満たすか否かを判定するものとする．つまり，$0 < |D_1 \cap D_2| < k$ を判定すればよい．この変形判定問題を要素数閉区間判定問題とする．このとき，要素数閉区間判定が全て $false$ だったときに，２つの匿名化データを結合可能と判定するようにフレームワークを変更する．

[要素数閉区間判定問題] 整数集合 $\{1,...,m\}$ の部分集合 D_1, D_2 において，$0 < |D_1 \cap D_2| < k$ を満たすか，否か． ◇

例えば，ID の値が 11 を含む集合を $D_1 = \{11, 12, 13\}, D_2 = \emptyset$ としたとき，$|D_1 \cap D_2| = 0$ なので $false$ と判定する．

従来フレームワークの秘匿要素数判定を秘匿要素数閉区間判定に変更することで一方のみしか属性値を保持していない識別子を結合データの対象外として両機関が提出した局所的 k-匿名化データを結合可能かどうか判定することができる．ここで，従来フレームワークが対象とする分散 k-匿名化問題では識別子順に並べた k-匿名化データを結合することでそのまま k-匿名化結合データを得ることができる．一方，本研究で対象とする共有 k-匿名化問題では，識別子順に並べた k-匿名化データを互いに公開することは相手に自分が保持している，または，保持していないデータの識別子を明かすことになってしまうので，安全ではない．よって，新たに結合データを作成するためのフェーズを必要とする．従来フレームワークに新しい秘匿要素数判定と結合データ作成フェーズを組み込んだ提案フレームワークを図 5 に示す．

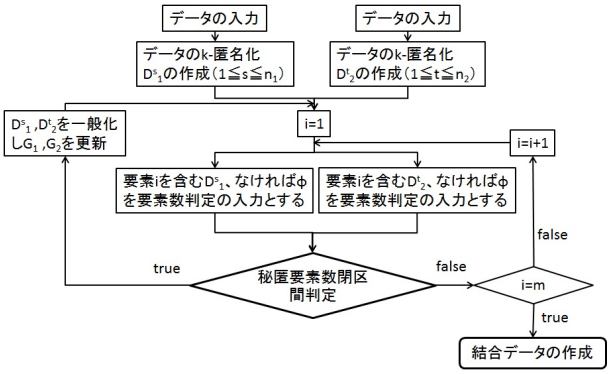

図 5　提案分散共有 k-匿名化フレームワーク

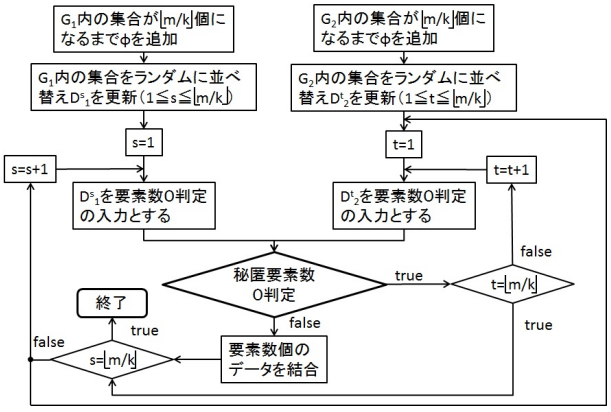

図 6　結合データ作成プロトコル

また，結合データ作成のフェーズのプロトコルの概要を図 6 に示す．結合可能判定後，G_1, G_2 内に空集合 $\emptyset$ を追加する．その後，G_1, G_2 内の集合を入力に $|D_1 \cap D_2| = 0$ を判定する秘匿要素数 0 判定を行う．結合後に要素数が 0 でない場合，その集合での結合後の要素数を判定し要素数分のデータセットを結合する．このとき，属性 ID はデータに含まない．結合データ作成の具体例を以下に示す．

3.1 両者の実際のやりとり

結合可能データセット T_1, T_2 の分割 $G_1 = \{\{1,2,7\}, \{3,5,8,10\}, \{4,6,9\}\}$,$G_2 = \{\{1,2,3,5,7,8\}, \{4,6,9\}, \{14,15,16\}\}$ を入力とする．

[フレームワーク]

・両者が要素 i をもつ場合

$$|\{1,2,7\} \cap \{1,2,3,5,7,8\}| \geqq 3 \quad (1)$$

$$|\{3,5,8,10\} \cap \{1,2,3,5,7,8\}| \geqq 3 \quad (2)$$

$$|\{4,6,9\} \cap \{4,6,9\}| \geqq 3 \quad (3)$$

・一方または両者が要素 i をもたない場合

$|\{11,12,13\} \cap \emptyset| = 0$

$|\emptyset \cap \{14,15,16\}| = 0$

以上の判定で結合可能と判定を得る．

[プロトコル]

・要素数 0 判定が true の場合

一方または両者が要素 i を含まない集合を入力としたときである．

$|\{1,2,7\} \cap \{4,6,9\}| = 0$

$|\{1,2,7\} \cap \{14,15,16\}| = 0$

$|\{1,2,7\} \cap \emptyset| = 0$

$|\emptyset \cap \emptyset| = 0$...etc

・要素数 0 判定が false の場合

両者が要素 i を含む集合を入力としたとき，つまり式 (1),(2),(3) のときである．このとき，それぞれ要素数を判定し P_1 側では属性 { 勤務地，年齢 } = { 関東, $30-34$} のデータ，P_2 側では属性 { 給与 } = $\{360-450\}$ のデータを 3 つ結合する．その後，{ 勤務地，年齢 } = { 中国, $25-29$}，{ 給与 } = $\{360-450\}$ を 3 つ結合する．最後に { 勤務地，年齢 } = { 中国, $45-49$}，{ 給与 } = $\{210-250\}$ を 3 つ結合し表 3 を得る．

表 2　結合可能データセット

ID	勤務地	年齢
1	関東	30-34
2	関東	30-34
3	中国	45-49
4	中国	25-29
5	中国	45-49
6	中国	25-29
7	関東	30-34
8	中国	45-49
9	中国	25-29
10	中国	45-49
11	関西	30-34
12	関西	30-34
13	関西	30-34

ID	給与（千円）
1	360-450
2	360-450
3	360-450
4	210-250
5	360-450
6	210-250
7	360-450
8	360-450
9	210-250
14	310-350
15	310-350
16	310-350

4. 結合データ作成プロトコルの安全性

本節では提案分散 k-匿名化フレームワークの安全性について考察する．提案フレームワーク全体としての安全性を示すには，結合可能判定までの安全性の検証と結合データ作成フェーズの安全性を検証する必要がある．結合可能判定までは，秘匿要素数閉区間判定を用いる以外，従来フレームワークとほぼ等しい．秘匿要素数閉区間判定は秘匿要素数判定のプロトコル

表 3　結合後 3-匿名データセット

勤務地	年齢	給与（千円）
関東	30-34	360-450
関東	30-34	360-450
関東	30-34	360-450
中国	25-29	360-450
中国	25-29	360-450
中国	25-29	360-450
中国	45-49	210-250
中国	45-49	210-250
中国	45-49	210-250

（文献 [1] の SSI プロトコル，または，文献 [2] のプロトコル）を拡張することで実現できる．したがって，本研究では従来フレームワークと大きく異なる結合データ作成フェーズの安全性に関して先ず考察する．

結合データ作成フェーズの安全性の検証では，結合可能と判定された結合データセットを入力としてプロトコルを実行する間に最終的に出力される結合データから得られる情報以外の相手機関の情報が漏えいしているか否かを判定する．

上記のような安全性の検証で用いられる方法の 1 つとして，シミュレーションがある [3]．本研究では提案するフレームワークにおいてそれぞれの機関で実際のやりとりをフレームワークを用いて分散データベースを結合後，出力される k-匿名化データセットを用いてシミュレート可能であるか否かを検証する．本稿では，まず，シミュレーションに用いる擬似データ作成方法を提案する．次に，結合可能判定後に表 2 を機関 P_1, P_2 はそれぞれもつものと提案擬似データ作成方法を用いて作成した擬似データを入力とした．シミュレーションの具体例を示す．

4.1 擬似データの作成

P_1, P_2 がそれぞれシミュレーションで使用する擬似データを結合後作成される匿名データセットと自分の保有する結合可能データセットから作成する方法を提案する．その方法を以下に示す．

(1) 自分がもっている結合可能 k-匿名化データセット $T_{initial}$ を値の種類ごとに分割し，それを $G_{initial} = \{D_1, D_2, ..., D_n\}$ とする．ここで各 D_i は同じデータ値を持っているレコードの識別子の集合とする．また，結合後のデータセット T_{final} を値の種類ごとに分割し，それを $G_{final} = \{A_1, A_2, ..., A_{n'}\}$ とする．ここで各 A_i は T_{final} で同じデータ値を持っているレコードの集合とする．また，$T_{initial}$ に含まれない識別子の集合を NID とし，作成中の擬似データを $G_{simulate}$ とし，初期値を $\emptyset$ とする．

(2)G_{final} に属するレコードの集合 A を選択する．

(3) レコードの集合 A と自分の保持している属性に関して，データセット T_{final} において同じ値の組合せをもつレコードの識別子の集合で，かつ，$G_{initial}$ に含まれる識別子の集合を見つけ，これを D とする．

(4) 集合 A の要素数 $\|A\|$ を p とし，集合 D の要素数 $\|D\|$ を q とする（$p \geqq q$）．$p = q$ ならばステップ (5.a) へ．$p > q$ ならば

ステップ (5.b) へ.
(5.a)($p \geqq q$ の場合)
相手機関は D 内に含まれる全ての識別子のデータを持っている．よって，$G_{simulate}$ に D をそのまま追加する．$G_{simulate} \leftarrow G_{simulate} \cup \{D\}$.
その後，$G_{initial}$ から D を削除する．$G_{initial} \leftarrow G_{initial} - \{D\}$.
(5.b)($p > q$ の場合)
相手機関は D に含まれる識別子のデータと同じ値の組合せのデータのうちの q 個を持っている．D からランダムに q 個の識別子を選び，識別子の集合 Q を作成し，$G_{simulate}$ に Q を追加する．$G_{simulate} \leftarrow G_{simulate} \cup \{Q\}$.
その後，$G_{initial}$ に含まれる集合 D を $D' = D - Q$ に更新する．$G_{initial} \leftarrow (G_{initial} - \{D\}) \cup \{D'\}$.
(6)G_{final} から集合 A を削除する．$G_{final} \leftarrow G_{final} - \{A\}$.
(7) もし，G_{final} が空集合でないならステップ (2) へ.
(8)$G_{simulate}$ の各要素 D に対応する相手機関のもつ属性の値組合せを結合後のデータセットより割り当てる．このとき，同じ値組合せに対応する複数の要素が存在するならそれらを $G_{simulate}$ から削除し，代わりに和集合を挿入する.
(9)$G_{simulate}$ の各要素 D に NID に含まれる識別子をランダムに追加，または，新しい集合を挿入し，擬似データとして $G_{simulate}$ を出力する.

4.2 P_1 のシミュレーション

結合後のデータセット表 3 と自身のデータセット T_1 からシミュレーションに使用する G'_2 を 4.1 節で示した擬似データ作成方法を用いて作成した．ここで作成した G'_2 は $G'_2 = \{\{1,2,7,a_1,a_2,a_3,d_1,...,d_n\},\{4,6,9,d_1,...d_n\},\{d_1,...,d_n\}\}$ である．(ここで，$a_i \in \{3,5,8,10\}$,$d_i \in \overline{T_1}$ とする)．このとき，3.1 節の実際の機関 P_2 とのやりとりを擬似データを使って模倣できることを示す.
[フレームワーク]
・お互いが要素 i をもつ場合
G_1, G'_2 から要素 i を含む D_1, D_2 を入力として要素数判定を行う.

$$|\{1,2,7\} \cap \{1,2,7,a_1,a_2,a_3,d_1,...,d_n\}| \geqq 3 \quad (4)$$

$$|\{3,5,8,10\} \cap \{1,2,7,a_1,a_2,a_3,d_1,...,d_n\}| \geqq 3 \quad (5)$$

$$|\{4,6,9\} \cap \{4,6,9,d_1,...d_n\}| \geqq 3 \quad (6)$$

・一方または両者が要素 i をもたない場合
空集合が入力となるので $|D_1 \cap D_2| = 0$ の判定で結合可能と判定を得る.
[プロトコル]
・要素数 0 判定が true の場合
G_1 の入力 D_1^s のどの要素も含まない $D_2'^t$ を入力とする.
$|\{1,2,7\} \cap \{4,6,9,d_1,...d_n\}| = 0$
$|\{1,2,7\} \cap \{d_1,...,d_n\}| = 0$
$|\{1,2,7\} \cap \emptyset| = 0$...etc
・要素数 0 判定が false の場合
G_1 の入力 D_1^s の要素を含む $D_2'^t$ を入力とする．つまり，式 (4),(5),(6) のときである．このとき，それぞれの要素数を判定したとき式 (4) から式 (1) を，(5) から式 (2) を，式 (6) から式 (3) をシミュレートし表 3 を得ることが可能である.

4.3 P_2 のシミュレーション

結合後のデータセット表 3 と自身のデータセット T_2 からシミュレーションに使用する G'_1 を作成した．ここで作成した G'_1 は $G'_1 = \{\{1,2,3,5,7,8,d_1,...,d_n\},\{4,6,9,d_1,...,d_n\},\{d_1,...,d_n\}\}$ である．(ここで，$\{d_i \in \overline{T_2}$ とする)．このとき，3.1 節の実際の機関 P_1 とのやりとりを擬似データを使って模倣できることを示す.
[フレームワーク]
・お互いが要素 i をもつ場合
G'_1, G_2 から要素 i を含む D_1, D_2 を入力として要素数判定を行う.

$$|\{1,2,3,5,7,8\} \cap \{1,2,3,5,7,8,d_1,...,d_n\}| \geqq 3 \quad (7)$$

$$|\{4,6,9\} \cap \{4,6,9,d_1,...d_n\}| \geqq 3 \quad (8)$$

・一方または両者が要素 i をもたない場合
空集合が入力となるので $|D_1 \cap D_2| = 0$ の判定で結合可能と判定を得る.
[プロトコル]
・要素数 0 判定が true の場合
G_2 の入力 D_2^t のどの要素も含まない $D_1'^s$ を入力とする.
$|\{1,2,3,5,7,8\} \cap \{4,6,9,d_1,...d_n\}| = 0$
$|\{1,2,3,5,7,8\} \cap \{d_1,...,d_n\}| = 0$
$|\{1,2,3,5,7,8\} \cap \emptyset| = 0$...etc
・要素数 0 判定が false の場合
G_2 の入力 D_2^t の要素を含む $D_1'^s$ を入力とする．つまり，式 (7),(8) のときである．このとき，それぞれの要素数を判定したとき式 (7) から式 (1),(2) を，式 (8) から式 (3) をシミュレートし表 3 を得ることが可能である.

5. ま と め

本稿では，異なる ID 要素集合の共通集合のみを安全に匿名化するためのフレームワークを提案した．今後の課題としては，作成擬似データを用いたシミュレーションを一般化し，安全性を検証することが挙げられる．また，提案結合データ作成プロトコルでは G_1, G_2 内の集合数をかく乱するために空集合を $\lfloor m/k \rfloor$ 個追加している．このため，結合データを作成するまでの秘匿要素数 0 判定の回数が増大している．よって，追加する空集合の適切な設定方法の導出もさらなる課題である.

文 献

[1] W. Jiang and C. Clifton:“A secure distributed framework for achieving k-anonymity,” The VLDB Journal, Vol.15, pp.316-333 (2006).
[2] 櫻田潤一, 上土井陽子, 若林真一:“分散データベースにおける安全で効率的なプロトコル,” DEIM2012 論文集, C5-5 (2012).
[3] B. Schneier:“Applied Cryptography, Second Edition,” John Wiley & Sons Inc.(1996).
[4] L. Sweeny; ”k-anonymity: a model for protecting privacy,” International Journal of Uncertainty, Fuzziness and Knowledge-based Systems, Vol.10, No.5, pp.571-588 (2002).