Vol.2018-DBS-167 No.4
Vol.2018-IFAT-132 No.4
2018/9/12

線形計画法を利用したプライバシ保護再公開の安全性評価

上土井 陽子[1]　村上 頼太[2]　若林 真一[1]

概要：本稿では個人の情報を表すタプルの挿入や削除により変化する動的データセットの再公開における公開表の系列のプライバシ保護に関する安全性評価について考察する．まず，公開表の系列において，各公開表に属する各個人に対して，攻撃者がもつ背景知識と個人の情報の一貫性と矛盾ない対応付けの候補（機密な値）の数の整数下限の最大値 c を評価値とする c-可能性という安全性評価基準を新しく提案する．次に，公開表の系列と攻撃者が知りうる情報が背景知識として与えられたとき，c-可能性に基づく評価値 c を線形計画法により算出できることを示し，提案基準に基づいた安全性評価の実用性を実験的に検証する．さらに，従来の確率的なリスク評価基準で 100%の確率でプライバシ漏洩があると判定される任意の公開表の系列に対しては，提案安全性評価基準でも $c = 1$ しか許さないことを理論的に示す．

1. はじめに

近年, 個人の情報を表すタプルの挿入や削除が行われる動的データセットに対して時系列で公開表の系列を公開する再公開，または，連続的公開を対象とする研究 [1][2][4][6][8] が行われている．それらの研究で扱われているデータセットの動的な変化には主にタプルの挿入のみとするもの，タプルの挿入と削除とするもの，タプルの挿入と準識別子などの属性の追加とする 3 タイプがある．本研究では，それらの中でもプライバシ漏洩のリスクを評価することが特に複雑なタプルの挿入と削除により動的に変化するデータセットの再公開に注目する．

タプルの挿入や削除が行われる動的データセットの公開表を時系列で公開する再公開においてプライバシを保護する方法として，先行研究 [8] では m-不変性という一般化の基準が考案され，その安全性を評価するための計算方法として全射関数を利用した方法が提案されている．m-不変性は動的に変化するデータの再公開においても個人の機密情報が漏洩する確率（リスク）を $1/m$ 以下に抑制することを目標に設定された一般化基準である．しかし，文献 [8] で m-不変性を保つことで $1/m$ 以下に抑制できることが証明されたリスクは攻撃者に公開表が m-不変性の基準に従って作成されたことを考慮することを強制することを仮定しており，厳密な意味での確率とは異なっている．一方，文献 [8] のリスク評価を参考にして定義できる厳密な確率によるリスク評価では公開表から攻撃者がもつ知識への矛盾のない対応を全て数え上げることが必要となり，背景知識のサイズに関して指数関数的に増加する計算時間を必要とする．

本稿では動的データセットの公開表の系列と攻撃者の背景知識が与えられたときに，各タプルに結びつけることが可能な個人 (候補者という)，および，各個人に結び付けられる機密な値の候補の数の整数下限の最大値を安全性の評価値 c とする新しい安全性評価基準 c-可能性 (c-*possibility*) を提案する．そして，何らかの一般化基準に従って作成されているという仮定がなくても，公開表の列と背景知識を入力として，評価値 c を計算する問題が線形計画問題として定式化できることを示す．さらに，c-可能性による安全性評価の実用性を検証するため，評価値 c を算出する方法を線形計画ソルバを用い実現し行ったシミュレーション実験の結果を示す．最後に，従来の厳密な確率によるリスク評価基準で 100%のプライバシ漏洩があると判定される入力に対しては，提案安全性評価基準でも $c = 1$ しか許さず常に 1-可能性となる，すなわち 100%のプライバシ漏洩が判定できることを理論的に示す．

2. 準備

2.1 用語の定義

本研究において必要な定義を述べる．元データを次の (1)∼(3) に分類する．以下，T は公開者が公開表作成時に用いた元データを示すとし，列と行で構成される表とする．さらに T の各行をタプルという．

(1) T は識別属性 A^{id} をもつ．例えば，A^{id} は名前の集合を表す．

(2) T は d 個の準識別子（QI）属性 A_i^{qi} をもつ．例えば，

[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三丁目 4-1
[2] 広島市立大学 情報科学部
〒 731-3194 広島市安佐南区大塚東三丁目 4-1

A_1^{qi} は年齢の集合，A_2^{qi} はジップコードの集合である．

(3) T は機密属性 A^S をもつ．例えば，A^S は病名の集合である．

また，各タプル t に対し $t[A]$ は属性 A における t の値を示す．初期表として与えられた元データ T に対し，順次，タプルの挿入と削除が行われるものとし，時刻 j における表を $T(j)$ で表し，時刻 j における T のスナップショットという．時刻 j は初期値を 1 とし，1 ずつ増加するとする．$T = T(1)$ とする．$T^*(j)$ は $T(j)$ を一般化した表を示し，$T^*(j)$ の情報歪曲に関する情報を提供する補助表を $L(j)$ とする．公開者は $T^*(j), L(j)$ の組を公開し，その際にデータは後で述べる偽造を含む一般化条件を満す状態で公開される場合があるとする．以上を踏まえ，文献 [8] を参考に基礎定義を述べる．

表 1 元データ $T(1)$
Table 1 Microdata $T(1)$

Name	Age	Zip.	Desease
Bob	21	12k	dyspepsia
Alice	22	14k	bronchitis
Andy	24	18k	flu
David	23	25k	gastritis
Gray	41	20k	flu
Helen	36	27k	gastritis
Jane	37	33k	dyspepsia
Ken	40	35k	flu
Linda	43	26k	gastrotis
Paul	52	33k	dyspepsia
Steve	56	34k	gastritis

表 2 表 1 を一般化した表 $T^*(1)$

T.ID	Age	Zip.	Desease
t1	[21,22]	[12k,14k]	dyspepsia
t2	[21,22]	[12k,14k]	bronchitis
t3	[23,24]	[18k,25k]	flu
t4	[23,24]	[18k,25k]	gastritis
t5	[36,41]	[20k,27k]	flu
t6	[36,41]	[20k,27k]	gastritis
t7	[37,43]	[26k,35k]	dyspepsia
t8	[37,43]	[26k,35k]	flu
t9	[37,43]	[26k,35k]	gastrotis
t10	[52,56]	[33k,34k]	dyspepsia
t11	[52,56]	[33k,34k]	gastritis

定義 1 偽造を含む一般化

公開表 $T^*(j)$ は $T(j)$ に基づいて作成され，以下の特性を持つ．

(1) $T^*(j)$ はタプル ID(T.ID) と呼ばれる A^t と，A^{id} を除いた $T(j)$ の全ての属性をもつ．

(2) 一般化されたタプルは元のタプルと等しい機密属性値をもち，一般化された QI 属性値の範囲は元のタプルの属性値を含む範囲である．

(3) $T(j)$ に対して，$T^*(j)$ は偽造されたタプル t_c^* を任意の個数含み，その機密属性値 $t_c^*[A^S]$ は A^S の領域の任意の値であり，任意の QI 属性値の範囲をもち，偽造タプルである可能性のあるタプル ID を別の情報として指定することでその範囲を限定する． □

定義 2 補助表

補助表 $L(j)$ は “タプル ID” と “Count” と呼ばれる 2 つの列を持つ．少なくとも 1 つの偽造を含む公開表 $T^*(j)$ での $L(j)$ において行 $\langle g, c\rangle$ が存在する．ここで g は偽造タプルの可能性があるタプル ID のグループを，c はこのグループにある偽造タプルの数を示す．$T^*(j)$ において偽造タプルが存在しない場合，$L(j)$ は空である．$L(j)$ により偽造の情報を公開することは，公開表を用いた統計的な解析において偽造タプル挿入による情報歪曲の影響を勘案するための助けとなる． □

表 3 元データ $T(2)$

Name	Age	Zip.	Desease
Bob	21	12k	dyspepsia
David	23	25k	gastritis
Emily	25	21k	flu
Jane	37	33k	dyapepsia
Linda	43	26k	gastritis
Gray	41	20k	flu
Mary	46	30k	gastritis
Ray	54	31k	dyspepsia
Steve	56	34k	gastritis
Tom	60	44k	gastritis
Vince	65	36k	flu

表 4 $T(2)$ を一般化した表 $T^*(2)$

T.ID	Age	Zip.	Desease
t12	[21,23]	[12k,25k]	dyspepsia
t13	[21,23]	[12k,25k]	gastritis
t14	[25,43]	[21k,33k]	flu
t15	[25,43]	[21k,33k]	dyapepsia
t16	[25,43]	[20k,30k]	gastritis
t17	[41,46]	[20k,30k]	flu
t18	[41,46]	[20k,30k]	gastritis
t19	[54,56]	[31k,34k]	dyspepsia
t20	[54,56]	[31k,34k]	gastritis
t21	[60,65]	[36k,44k]	gastritis
t22	[60,65]	[36k,44k]	flu

表 5 偽造を含む一般化した表 $T^*(2)$

Name	T.ID	Age	Zip.	Desease
Bob	t12	[21.22]	[12k,14k]	dyspepsia
c_1	t13	[21.22]	[12k,14k]	bronchitis
David	t14	[23,25]	[21k,25k]	gastritis
Emily	t15	[23,25]	[21k,25k]	flu
Jane	t16	[37,43]	[26k,33k]	dyspepsia
c_2	t17	[37,43]	[26k,33k]	flu
Linda	t18	[37,43]	[26k,33k]	gastritis
Gray	t19	[41,46]	[20k,30k]	flu
Mary	t20	[41,46]	[20k,30k]	gastritis
Ray	t21	[54,56]	[31k,34k]	dyspepsia
Steve	t22	[54,56]	[31k,34k]	gastritis
Tom	t23	[60.,65]	[36k,44k]	gastritis
Vince	t24	[60.,65]	[36k,44k]	flu

表 6 補助表

T.ID	Count
t12,t13	1
t16,t17,t18	1

以降，攻撃者がもつ背景知識を定義する．

定義 3 背景知識表の列 $B(1), B(2), \ldots, B(n)$

時刻 n において，攻撃者は予め元データセットの系列 $T(1), T(2), \ldots, T(n)$ の各表から機密属性 A^S を除いた全ての属性を持つ表により構成される背景知識表の系列 $B(1), \ldots, B(n)$ を持つ．各背景知識表 $B(i)$ に属するタプル b は個人の情報を示しているので，以降では背景知識表のタプルを個人と呼ぶ．このとき，$B(1), B(2), \ldots, B(n)$ には全ての偽造タプルに関する公開された詳細情報が組み込まれている．よって，以降では補助表 $L(1), L(2), \ldots, L(n)$ についての記載を省くものとする．

また，$UB(n)$ を $B(1) \cup B(2) \cup \ldots \cup B(n)$ を表す背景知識の統合表とする．$UB(n)$ は公開表に含まれる全ての個人の情報を統合したものであり，$b \in UB(n)$ なる各個人 b は $Lifespan$ という項をもつことにより $Lifespan[b] = (x, y)$ にて存在する時間を明記し，時刻以外の情報の重複をまとめて，集約した形で表現される．ここで，x は個人 b が元データ $T(j)$ に追加された時の時刻 j，y は b が元データ $T(j)$ から削除される 1 つ前の時刻 $j-1$ を示すとする． □

2.2 従来のリスク評価基準

2.2.1 元データの再構築

先行研究 [8] では再公開表列のプライバシ漏洩のリスクを確率的に評価する基準が定義されている．この定義を説明するにあたり元データの再構築の概念が必要になる．元データの再構築とは，相手が公開表の列 $T^*(1), ..., T^*(n)$ と背景知識表 $B(1), \ldots, B(n)$ を組み合わせて，個人の機密情報を推測することとする．以降では，公開表の列の和集合 $T^*(1), ..., T^*(n)$ を $UT^*(n)$ と表すこととする．

定義 4　全射関数による再構築

以下の条件を満たす場合，写像 $f : UT^*(n) \to UB(n)$ は再構築を行う全射関数である．ここで，各タプル $t \in UT^*(n)$ をあるタプル $b \in UB(n)$ に写像することを $f(t) = b$ と表現する．

- あるタプル $b \in UB(n)$ かつ $b \in B(i)$ なる時刻 i に対し，$f(t) = b$ となるような唯一のタプル $t \in T^*(i)$ が必ず存在する (f を公開表 $T^*(i)$ から背景知識表 $B(i)$ への関数と限定した場合には f は全単射関数)．
- $f(t) = b$ とする場合
  1　$t \in T^*(i)$ のとき，$b \in B(i)$ である
  2　各 QI 属性 A_i^{qi} $(i \in \{1, \ldots, d\})$ において，$t[A_i^{qi}]$ は $b[A_i^{qi}]$ を含むような範囲である (ただし，$b[A_i^{qi}] = 0$ の場合，任意の包含関係が常に成り立つとする)　□

このような全射関数において，機密属性値に着目した場合にも矛盾なく元のデータを再構築する全射関数を *reasonable* な (妥当な) 全射関数と言う．次に妥当な全射関数について定義を行う．

定義 5　妥当な全射関数

以下のような条件を満たす場合，定義 4 における関数 f は妥当である．

条件 (1)　あるタプルは $b \in UB(n)$ に対し，$f(t) = b$ を満たすようなタプル $t \in UT^*(n)$ の集合 $f^{-1}(b)$ が与えられた場合
(1) $f^{-1}(b)$ の全てのタプルは同じ機密属性の値をもつ
(2) 個人 b のライフスパン $Lifespan[b]$ の範囲内の各時刻 j に対して $t \in T^*(j)$ となる唯一のタプル $t \in f^{-1}(b)$ が必ず存在する

条件 (2)　f^{-1} は利用される一般化基準と一致するような一般化の候補のみを提示する　□

これらの概念から文献 [8] ではプライバシ漏洩のリスクが定義されている．

定義 6　Xiao らのリスク評価基準

b を背景知識表の和集合 $UB(n)$ に属する個人とすると，b のプライバシ漏洩のリスク $risk(b)$ は

$$risk(b) = \frac{n_{breach}(b)}{n_{total}}$$

で示される．ここで，n_{total} は妥当な全射関数の数であり，n_{breach} は妥当な全射関数のうち個人 b の本当の機密属性値に結びつける全射関数の数である．

定義 6 では個人 b と本物の機密属性値が結びつく確率を個人 b のリスクと定義している．しかし，個人 b と本物でない機密属性値 s が結びつく確率が高い公開表がプライバシ保護された公開表として公開されることで，個人 b の機密属性値は s ではないとわかってしまう可能性がある．そこで，以降では各個人 b と各機密属性値 s を結びつける確率 $risk(b, s)$ の中の最大値を公開表列 $UT^*(n)$ の確率に基づくリスク評価値 $risk(UT^*(n))$ とする．ここで, $risk(UT^*(n))$ は $risk(UT^*(n)) = \max\{risk(b, s) | b \in UB(n) \wedge s \in A^S\}$ を満たす．

Xiao らの定義 6 のリスク評価基準では定義 5 にある確率を計算するための母集合となる妥当な全射関数として条件 (2) にあるような公開表が作成された一般化基準を満たすものだけを考えるという特殊な条件がある．我々の関連研究 [5] では，条件 (2) がなければ，m-不変性を満たした公開表列でもプライバシ漏洩の確率的リスクを $1/m$ 以下に抑えられない場合が存在し，Xiao らのリスク評価基準は厳密な確率的リスクではないことがわかった．そこで，定義 5 の妥当な全射関数の定義から条件 (2) を除いた上で定義 6 と同様に定義されるリスク評価基準を厳密な確率によるリスク評価基準と呼ぶこととする．

定義 6, または，厳密な確率によるリスク評価基準に従って $risk(UT^*(n))$ を計算するためには，妥当な全射関数の数を導出する必要がある．妥当な全射関数の数を数えた上で，リスク評価を行うには組合せを列挙して確率を求めることが必要となり，$risk(UT^*)$ の正確な計算は入力サイズの指数時間の計算量を要し，規模の大きな公開表列では困難であると予測される．

3. c-可能性：提案安全性評価基準

厳密な確率としてリスクを正確に評価することは第 2 節で述べたように規模の大きな公開表では困難であると予測される．本研究では，どのような基準で各公開表が作成されたのかの仮定を用いずに，安全性を評価する基準を新しく提案する．さらに，評価値 c を計算する問題を線形計画問題として定式化する．

3.1 安全性評価基準の提案

まず，公開表の列に対する新しい安全性評価基準を以下に提案する．

定義 7　c-可能性 (c-*possibility*)

1 以上の整数 c, n において，n 個の公開表の列 $T^*(1)$, $T^*(2), \ldots, T^*(n)$ と各公開表に対応する n 個の背景知識表の列 $U(1), U(2), \ldots, U(n)$ が与えられたときに, 攻撃者が背景知識と公開表の列と個人の機密属性値の一貫性の仮定を用いて行える推測のすべてにおいて, $t \in T^*(i)$ $(1 \leq i \leq n)$ に属する各タプルと結びつけることが可能な個人 (候補者) の数が c 人以上であり，かつ，各個人 $b \in UB(n)$ と結びつけられる機密属性値の候補の数が c 種類以上であるなら，公開表の列は c-可能性を満たすとする．また，公開表の列が c-可能性であるが，$(c+1)$-可能性でないという条件を満たす整数 c を評価値と呼ぶ．　□

c-可能性の評価値 c を求めることは，公開表が与えられたとき，以下のような非零な値の確率を割り当て可能な変数を並べた行列を作成することで簡単にできると考えられるかもしれない．

定義 8　非零可能要素の行列表現 $X(1), X(2), \ldots, X(n)$
1 以上の整数 n と公開表の列 $T^*(1), T^*(2), \ldots, T^*(n)$ と各公

開表に対応する n 個の背景知識表の列 $B(1), B(2), \ldots, B(n)$ が与えられたときに, n 個の行列表現 $X(1), X(2), \ldots, X(n)$ を以下の方法で作成する．各タイムスタンプ i $(1 \leq i \leq n)$ において,

(1) 行列 $X(i)$ は背景知識 $B(i)$ に属する各個人 b に対応する行と公開表 $T^*(i)$ に属する各タプル t に対応する列をもつ.

(2) 行列 $X(i)$ の個人 b に対応する行とタプル t に対応する列の要素は個人 b の全ての準識別子属性の値がタプル t の準識別子属性の範囲内にあるなら，つまり，$\forall i \in \{1, \ldots, d\} b[A_i^{q_i}] \in t[A_i^{q_i}]$ なら変数 $x_{b,t}$ であり，そうでなければ定数 0 である．変数 $x_{b,t}$ のことを 0 または非零の値を取ることが可能な変数という意味で非零可能変数と呼ぶ．また，その行列表現を非零可能変数行列と呼ぶ．
□

上記の非零可能変数行列を作成し，各列にある非零可能変数の数や各行に変数をもつタプルの集合がもつ機密属性の種類を数えれば c-可能性の評価値 c は簡単に求められるように見える．しかし，個人 b の準識別子属性の値がタプル t の準識別子の範囲に含まれるからと言って，タプル t に結び付けられる個人の候補に b が該当すると言えない場合がある．以降では公開表の系列と背景知識の系列が与えられたときに，提案安全性評価基準で評価値 c を精確に計算する方法について考察する．

3.2 提案安全性評価基準による評価値の算出

非零要素となり得る変数の中から本当にタプル t に個人 b を割り当てることができる組合せ (b, t) を表現する変数 $x_{b,t}$ を抜き出すため，我々は実際に非零可能要素行列の系列 $X(1), \ldots, X(n)$ の非零可能変数へ確率を模倣した実数値を割り当てる問題を考える．以降では，変数に割り当てる確率を模倣した実数値のことを推測確率と呼び，変数に推測確率を割り当てる問題を推測確率割当て問題と呼ぶ.

定義 9 推測確率行列の系列 $R(1), R(2), \ldots, R(n)$

n 回目の公表が行われた場合, 1 から n 番目の公開表の系列 $T^*(1), T^*(2), \ldots, T^*(n)$ と背景知識表の系列 $B(1), B(2), \ldots, B(n)$ に基づき作成された非零可能要素行列の系列 $X(1), X(2), \ldots, X(n)$ の各変数 $x_{b,t}$ に以下の 2 つの条件を満たすように推測確率 (0 以上 1 以下の実数値) を割り当てた結果の行列の系列 $R(1), R(2), \ldots, R(n)$ を推測確率行列の系列とする．このとき，n 個の推測確率行列の系列 $R(1), R(2), \ldots, R(n)$ は以下のような条件を満たす必要がある.

(1) 2 つ以上の公開表にタプルが存在している個人は，その個人が存在する全ての時刻の公開表に関する推測確率行列において，各機密属性値に対し，その個人に対応するタプルに割り当てられた推測確率の総和は全ての時刻で等しい.

(2) 推測確率行列の各列に関して推測確率の総和は 1 であり，また，各行に関しても推測確率の総和は 1 である.
□

定義 9 での条件は一般的な確率の関係を模倣し，各列に対応する公開表のタプルに結び付けられる個人の推測確率の総和は 1 であり，各行に対応する個人に結び付けられるタプルの推測確率の総和も 1 であるという関係を成立させる．また，複数の時刻に存在するタプルの推測確率に不整合が生じないように，各機密属性値に割り当てられる推測確率の総和はどの時刻においても等しいという制約を追加している.

定義 9 を満たすように推測確率を非零可能要素行列の系列 $X(1), X(2), \ldots, X(n)$ に割り当てることで，実際には非零要素を割り当てることができない変数を各非零可能要素行列 $X(i)$ $(i \in \{1, \ldots k\})$ から削除できる.

一方で，非零可能要素行列の系列 $X(1), X(2), \ldots, X(n)$ に対して定義 9 の条件を満たす様々な推測確率の割当てが存在する.

背景知識だけを利用してできる推測で非零可能要素行列の系列 $X(1), \ldots, X(n)$ からタプル t に結びつけることができない個人 b の組合せ (b, t) に対する変数 $x_{b,t}$ を 0 に設定したいということが c-可能性の評価値を求める上での本来の目標である.

よって，可能な推測確率割当ての中でも各タプル t (各列) に非零な推測確率が割り当てられている要素数，各個人 b (各行) に非零な推測確率が割り当てられている機密属性値の種類数の最小値が候補数の下限となることから，候補数の下限を大きくする推測確率割当てを求めたい．これは推測確率行列の要素の最大値，もしくは，各個人の各機密属性値に割り当てられた値の総和の最大値を M を最小化することで達成できる．したがって，定義 9 の条件と M を最大値とする条件を制約条件とし, 目的関数 M を最小化する問題を提案安全性基準での評価値算出問題として定式化する.

上記の問題は，制約条件，目的関数とも線形式で表現でき，変数に割り当てられる値が実数であることより線形計画問題として定式化できる．この線形計画問題の最適目的関数値 M_{min} の逆数の切り上げは定義 7 での評価値 c となる．つまり，$c = \lceil 1/M_{min} \rceil$ により c-可能性の評価値が計算できる.

3.3 線形計画法を利用した安全性評価の実用性の検証

本部分節では公開表列 $T^*(1), \ldots, T^*(n)$ と背景知識表 $B(1), B(2), \ldots, B(n)$ が与えられたときに，前部分節で定義した c-可能性基準での評価値 c_{max} を精確に算出する安全性評価を以下の枠組みにて実現する．提案枠組みは公開表列，背景知識表を入力として，c-可能性基準での評価値を計算する問題である線形計画問題を自動で作成する線形計画問題自動作成手続きと商用混合整数計画問題ソルバである IBM ILOG CPLEX[3] より構成される.

計算機 (CPU:Intel Core i5 2.4GHz, メモリ:8GB) 上で線

形計画問題自動作成手続きを C 言語用いて実現し，線形計画ソルバとして混合整数計画ソルバ IBM ILOG CPLEX Optimization Studio 12.7.1 を用いることで提案安全性評価枠組みを実現した．タプル数が同じ 2 つの公開表の系列からなる 10 タプル，100 タプル，200 タプル，400 タプルをもつ 4 つの合成データを用いて提案安全性評価枠組みのシミュレーション実験を行った．合成データの機密属性の種類は 5 種類とした．また，全ての合成データにおいて，2 つの公開表に対応する元データ表で共通する個人のタプル数は 20 とした．例えば 100 タプルでの合成データは 50 タプルの 1 回目の公開表と 50 タプルの 2 回目の公開表をもち，50 人の個人データをそれぞれ含む元データ表も含む．ここで，両方の元データ表に存在する個人は 20 人である．よって，非零可能要素行列の系列は 50×50 の行列 2 つからなる．また，正確に線形計画問題を定式化できているか確認するため，実験に用いた合成データは予め全て 2-可能性をもつように作成した．各データに対し，線形計画問題の作成にかかった計算時間，および，作成された線形計画問題を入力とした CPLEX の計算時間，および，作成された線形計画問題の制約数をまとめた結果を表 7 に示す．表 7 より，制約数が数万程度であれば線形計画問題作成時間，CPLEX の計算時間とも，実用的であることが分った．

表 7 実験結果

データ タプル数	線形計画問題 作成時間 [秒]	CPLEX の 計算時間 [秒]	制約数	出力評価値 c_{max}
10	0.001	0.00	80	2
100	0.008	0.04	5300	2
200	0.018	0.03	20600	2
400	0.064	0.19	81000	2

4. c-可能性評価基準と確率的リスク評価基準の関係

本節では任意の公開表列において Xiao らの定義を基にした厳密な確率的リスク評価基準と提案安全性評価基準での評価で作成される推測確率行列の間に成り立つ関係を理論的に解析する．以下の定理により入力が特定の状況の場合には 2 つの評価基準の評価値が必ず等しいことを示す．

定理 1 任意の公開表の列 $T^*(1), \ldots, T^*(n)$ と対応する背景知識表の列 $B(1), \ldots, B(n)$ が与えられたとき，確率的リスク評価値 $risk(UT^*(n)) = 1$ なら提案安全性評価においても評価値算出問題の最適目的関数値は $M_{min} = 1$ であり，安全性評価値 c は 1 である．

[証明] 確率的リスク評価値 $risk(UT^*(n)) = 1$ となるとき，ある個人 b において結び付く全てのタプル t' の機密属性値が同一である（ここではその機密属性値を s とする）という条件が成り立つ．従って，個人 b と s とは異なる機密属性値をもつタプル t とを結び付ける全射関数は存在しない．このとき，以下の補題が成り立つことを示す．

補題 1 確率的リスク評価において，個人 b と s とはタプル t (タプル t は時刻 l の公開表 $T^*(l)$ のタプルとする）とを結び付ける全射関数が存在しないとき，提案安全性評価における非零要素可能変数行列 $X(l)$ への推測確率の任意の割当てにおいても，$x_{b,t} = 0$ である．

[補題 1 の証明] 確率的リスク評価値算出において，個人 b とタプル t を結び付ける全射関数が存在しない場合，以下のうちのどちらかが成立つ．

場合 (1) 背景知識表 $B(l)$ による非零可能変数行列 $X(l)$ において，変数 $x_{b,t}$ が存在しない

場合 (2) 非零可能変数行列 $X(l)$ において変数 $x_{b,t}$ は存在するが，どの全射関数においても b 以外の他の個人 b' がタプル t と結び付く

(1) のときの証明 提案安全性評価での非零可能変数行列 $X(l)$ への推測確率割当てにおいて，変数 $x_{b,t}$ が存在しないため，命題を満たす．

(2) のときの証明 命題が成り立つことを公開表の系列 $T^*(1), \ldots, T^*(n)$ に属するタプル数 k に関する数学的帰納法により証明する．

(基本ステップ）公開表の系列 $T^*(1), \ldots, T^*(n)$ と対応する背景知識表の系列 $B(1), \ldots, B(n)$ に関する非零可能変数行列の系列 $X(1), \ldots, X(n)$ において，ある時刻 j の公開表 $T^*(j)$ に属するあるタプル t' に対する非零可能変数行列 $X(j)$ の列に非零可能変数が存在しない場合，各列における推測確率の総和が 1 であるという制約条件を満たすことができない．よって，非零可能変数行列の系列 $X(1), \ldots, X(n)$ への制約条件を満たす推測確率の割当ては存在しない．したがって，任意の時刻 $j \in \{1, \ldots, n\}$ の非零可能変数行列 $X(j)$ の任意の変数 $x_{b',t'}$ の値は 0 である．よって，$x_{b,t} = 0$ となり命題を満たす．

上記以外の場合で，$k = 2$ のとき，背景知識表 $B(1) = \{b_1, b_2\}$, タプル集合 $T^*(1) = \{t_1, t_2\}$ に関する 2×2 の非零可能変数行列 $X(1)$ を考える．今，一般性を失うことなく，非零要素変数行列 $X(1)$ に変数 x_{b_1,t_1} は存在するが個人 b_1 とタプル t_1 を結び付ける全射関数が一つも存在せず，かつ，b_2 は l 番目の背景知識表 $B(l)$ 以外の背景知識表に存在しないと仮定する．このとき，確率的リスク評価基準で個人 b_2 とタプル t_1 を結び付ける全射関数しか存在しない．なぜなら，そうでないとすると，他には個人 b_1 とタプル t_1 を結び付けることができなくなる条件が存在しないため，個人 b_1 とタプル t_1 を結び付ける全射関数も存在するからである．確率的リスク評価基準で個人 b_2 とタプル t_1 を結び付ける全射関数が存在しない理由は，変数 x_{b_1,t_1} が存在することから，場合 (1) の理由である．よって，非零可能変数行列に変数 x_{b_2,t_2} は存在せず，任意の推測確率割当てにおいて $x_{b_2,t_1} = 1$ となることから，$x_{b_1,t_1} = 0$ となる．よって，命題が成り立つ．

(帰納ステップ）$k \geq 2$ を満たす k において, 公開表の系列の属するタプル数が k 以下であるときに，命題が成り立つ

と仮定する．このときに，公開表の系列に属するタプル数が $k+1$ である場合にも命題が成立することを示す．

公開表の系列 $T^*(1),\ldots,T^*(n)$ のタプルの数が $k+1$ であり，非零可能変数行列 $X(l)$ に変数 $x_{b,t}$ が存在し，確率的リスク評価においてどの全射関数においても，タプル t は他の個人 b' と結び付いていると仮定する．

今，公開表の系列 $T^*(1),\ldots,T^*(n)$ に対応する非零可能変数行列の系列 $X(1),\ldots,X(n)$ においてある時刻 j の公開表 $T^*(j)$ に属するあるタプル t' に対する非零可能変数行列 $X(j)$ の列に非零可能変数が存在しないなら，基本ステップと同様に非零可能変数行列の全ての変数には 0 しか設定できない．よって，命題が成り立つ．

そうでない場合，個人 b とタプル t を結び付ける任意の対応を考えた場合，少なくとも 1 つのタプル t' に結び付けられる個人が存在しない状況となり，かつ，タプル t' が属する非零可能変数行列 $X(j)$ $(1 \leq j \leq n)$ において設定されている変数が 1 つ以上存在する．なぜなら，このようなタプル t' が存在しなければ，b と t を結び付ける全射関数が存在することとなるため，矛盾を生じるからである．また，タプル t' が存在する時刻 j はタプル t が存在する時刻 l と同一でない可能性がある．

非零可能変数行列 $X(j)$ において，タプル t' と結び付けられる変数が $x_{b_1,t'},\ldots,x_{b_m,t'}$ という m 個であると仮定する．このとき，$i \in \{1,\ldots,m\}$ を満たす各 i において，$x_{b_i,t'}=1$ とし，その他の変数に 0 を設定する．つまり，非零可能変数行列 $X(j)$ においては $x_{b_i,t'}=1$,$x_{b_r,t'}=0$ $(r \in \{1,\ldots,m\}-\{i\})$ と制約し，機密属性値に関する整合性を満たすよう他の非零可能変数行列の制約を再調整した上で，タプル t' 以外の全体の推測確率を割り当てる推測確率割当て問題を考える．このとき，考える推測確率割当て問題での公開表の系列に属するタプル数は元の公開表の系列に属するタプルの集合から t' を除いた集合の要素数であるので k である．また，対応する確率的リスク評価では個人 b_i とタプル t' を結び付ける全射関数だけを抜き出し，確率を計算するが，元の全射関数の集合において個人 b とタプル t を結び付ける全射関数が存在しなかったので，その部分集合においても個人 b とタプル t を結び付ける全射関数は存在しない．よって，上記の問題での推測確率割当て結果で個人 b とタプル t を結び付ける推測確率を表す変数 $x^i_{b,t}$ の値は帰納法の仮定より 0 である．今，各限定された状況にて推測確率行列を求めたときに，任意のタプル b', 個人 t'' に割り当てられる最大値を $x^{i,max}_{b',t''}$ とする．このとき，元の推測確率行列の列の変数 $x_{b't''}$ に割り当てられる値は線形関係式と各変数に乗算される係数が 0 以上 1 以下であるという制約条件より，$x_{b',t''} \leq max\{x^{i,max}_{b',t''}|i \in \{1,\ldots,m\}\}$ が言える．一方，帰納法の仮定より，$x^i_{b,t}=0$ であるので $max\{x^{i,max}_{b,t}|i \in \{1,\ldots,m\}\}=0$ が言える．よって，任意の推測確率割当てにおいて，$x_{b,t}=0$ であり，命題が成り立つ． □

補題 1 より，確率的リスク評価において，個人 b とタプル t が結びつくリスク $r_{b,t}$ が 0 であるなら，提案安全性評価基準による推測確率行列 $R(1),\ldots,R(n)$ を求めたときにタプル t と個人 b を結び付ける変数 $x_{b,t}$ の値は 0 であることがわかった．よって，確率的リスク評価において，個人 b と結び付くタプルの集合の部分集合としか，提案安全性評価による推測確率行列においても個人 b は結び付かない．したがって，提案安全性評価基準での推測確率行列 $R(1),R(2),\ldots,R(n)$ でも個人 b に対しては非零の値が割り当てられる変数 $x_{b,t}$ に対応するタプル t は機密属性値 s しか持たないと言える． □

5. まとめ

本稿では，公開表が作成された一般化基準が不明な公開表列に関してもプライバシ漏洩のリスクを精確に，かつ，効率よく評価することを目的として，新しい安全性評価基準を定義した．そして，提案安全性評価基準での公開表列の安全性評価値を精確に算出する問題が線形計画問題に定式化できることを示し，従来の確率的リスク評価基準との定義上の関係性を明確にした．さらに線形計画ソルバを用いたリスク評価枠組みを提案し，計算機上で実現し，実用性を検証した．

参考文献

[1] M. M. Baig, J. Li, J. Liu, and H. Wang, “Cloning for privacy protection in multiple independent data publication,” Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKMM’11), pp. 885-894, 2011.

[2] J. Byun, T. Li, E. Bertio, N. Li, and Y. Sohn, “Privacy-preserving incremental data dissmination,” Journal of Computer Security, Vol. 17, No. 1, pp. 43-68, 2009.

[3] IBM ILOG CPLEX. [Online]. Available: http://www.ibm.com/software/integration/optimization/cplex-optimizer/ ver.12.7.1.

[4] B. C. M. Fung, K. Wang, A. W. Fu, and J. Pei, “Anonymity for continuous data publishing,” Proc. of the 11th International Conference on Extending Database Technology (EDBT), pp.264-275, 2008.

[5] 坂田奈々子, 上土井陽子，若林真一,“動的データセットのプライバシー保護再公開における厳密な安全性評価について”, 2017 年暗号と情報セキュリティシンポジウム, 3A4-1, 2017.

[6] E. Shmueli, and T. Tassa, “Privacy by diversity in sequential releases of databases,” Information Sciences: an International Journal, vol. 298, No. C, pp. 344-372, 2015.

[7] L. Sweeney, “Achieving k-anonymity privacy protection using generalization and suppression,” International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, Vol.10, No.5, pp.571–588, 2002.

[8] Xiaokui Xiao, and Yufei Tao, “m-Invariance: towards privacy preserving re-publication of dynamic datasets,” Proceedings of the ACM SIGMOD International Conference on Management of data (SIGMOD), pp.689–694, 2007.