動的データテーブルの連続的匿名化の安全性について

柚木　　塁† 上土井陽子†† 若林　真一††

† 広島市立大学情報科学部　〒 731–3194 広島県広島市安佐南区大塚東三丁目 4 番 1 号
†† 広島市立大学大学院情報科学研究科　〒 731–3194 広島県広島市安佐南区大塚東三丁目 4 番 1 号
E-mail: †rui@lcs.info.hiroshima-cu.ac.jp, ††{yoko,wakaba}@hiroshima-cu.ac.jp

あらまし　本研究では挿入や削除により時間の経過に従って変化する動的データテーブルの連続した匿名化における安全性について考察する．従来知られている静的データテーブルに対する匿名化手法を用いて動的データテーブルを匿名化した場合，複数の匿名化テーブルの情報を統合することで個人を特定できる可能性が指摘されている．本研究では動的データテーブルの連続した匿名化テーブルの列が匿名性および多様性を維持するために満たすべき性質を明らかにする．

キーワード　動的データテーブル，挿入，削除，m-不変性

1. はじめに

プライバシー保護データ公開では主に一度だけ公開する静的データテーブルを対象としてきた．近年，挿入や削除が行われる動的データテーブルの連続的匿名化が考察され始めた．しかし，動的データテーブルは挿入や削除を行うことが可能となったのだが，静的データテーブル向けのプライバシー保護手法で匿名化し再公開を行うとプライバシーが保護されないことがあった．そのため動的データテーブルの安全性は再公開では不確かなものとなっていた．従来研究 [1] では再公開を行ってもプライバシーが保護できる性質が導出されている．本研究では従来研究より匿名化に伴う情報損失を低減することを目標として，動的データテーブルの連続的匿名化の新しい安全性を提案する．

2. 準　　備

本節では本研究で必要となる基本的な定義を示す．

公開者によって保管されている元のデータテーブルを T とする．表 T の中の列を (i)T の主要なキーとなる識別子属性 A^{id}(今回の場合，名前)，(ii)d 個の準識別子属性 A_1^{qi},...,A_d^{qi}(今回の場合，年齢，ジップコード)，(iii) 機密属性 A^s(今回の場合，病名) で分類する．d 個の準識別子属性は数値属性か分類別属性のどちらかであるが，機密属性 A^s は分類属性とする．表 T の行である各タプル t に対して、t[A] はタプル t の属性 A の値を示す．

表 T は任意の順序で実行される挿入や削除によって更新される．公開者はプライバシーを侵害することなく，挿入や削除を行うことが出来き，いつでも T の匿名化された表を公開するかもしれない．そこで，j 番目の公開のタイムスタンプを示すために整数 j を使う．

時刻 j での T の表を $T(j)$ とする．表 $T(j)$ を匿名化した表を $T^*(j)$ とする．表 $R(j)$ を表 $T^*(j)$ についての統計のための補助の表とする．公開者は表 $T^*(j)$ と補助表 $R(j)$ をペアで公開する．特に，匿名化は後で述べる偽造された一般化を適用して達成される可能性がある．この概念 [1] を定義する前に，いくつかの基本的な概念を説明する．

[**定義** 1]（QI グループ/分割）　元のデータの表 $T(j)$ で、QI グループはタプルの部分集合である．表 $T(j)$ の分割は互いに素な交わりのない複数の QI グループタプルからなっていて，それらの集合和は表 $T(j)$ となる．各 QI グループは特定の ID をもつ．$t \in T(j)$ を満たす各タプル t で，$t.QI(j)$ は t を含む QI グループを示す．$t.QI(j)$ のことを表 $T(j)$ での t のホスティンググループと呼ぶ．

時刻 j で公開される表 $T^*(j)$ を以下に定義する．例として表 1 の元データ表と表 2 の一般化表を用いて，定義を詳細に説明する．

[**定義** 2]（偽造された一般化）　表 $T(j)$ の匿名化された表である表 $T^*(j)$ は表 $T(j)$ の分割に基づいて計算される．そして表 $T^*(j)$ は以下の性質を持っている．

（1）　表 $T^*(j)$ は”Group-ID”と名付けられた列 A^g を含み，A^{id} を除く表 $T(j)$ のすべての属性を含む (表 2 の場合，年齢とジップコードと病名が含まれる)．

（2）　同じグループ IDA^g をもつ表 $T(j)$ のすべてのタプルはすべて同じ QI 属性値を持つ．これらのタプルは表 $T^*(j)$ の QI グループを形成し，ID はグループ内の A^g の値に等しい．

（3）　$T(j)$ の各タプル t∈$T(j)$ は表 $T^*(j)$ に一般化されたタプル t^* を持つ．ここで t^* は以下を満たす．

- $t^*[A^s]$=t$[A^s]$
- グループ $t^*[A^g]$ は表 $T^*(j)$ の $t*$ のホスティンググループの ID である．
- 値 $t^*[A_i^{qi}]$ は，値 $t[A_i^{qi}]$ を含む範囲である．(t=(表 1 のボブのタプル),t^*=(表 2 のボブのタプル),$t^*[A^s]$=”消化不良”,$t[A^s]$=”消化不良”,$t[A^g]$=1, $t_c^*[A_1^{qi}]$=age[21,22],$t[A_1^{qi}]$=age21).

（4）　表 $T(j)$ の QI グループ QI に対して，表 $T^*(j)$ は任意の数の偽造されたタプル t_c^* を含む場合がある．ここで値 $t_c^*[A^s]$ は機密属性 A^s の領域の値になる．グループ ID$t_c^*[A^g]$ は QI の ID に等しい．$t_c^*[A_i^{qi}]$ は値の範囲である．

t∈T(j) に対して，t.$QI^*(j)$ を t の一般化されたホスティンググループとして示し，$t.QI^*(j)$=$t^*.QI(j)$ とする．

明らかに偽造を含む一般化は偽造がない従来の一般化の方法を特別な場合として含んでいる．次の定義は補助の表 R(j) が表 $T^*(j)$ と伴に公開されることを明らかにする．

表 1　1 回目の元データ

名前	年齢	コード	病気
ボブ	21	12k	消化不良
アリス	22	14k	気管支炎
アンディ	24	18k	風邪
デイビット	23	25k	胃炎
ガーリー	41	20k	風邪
ヘレン	36	27k	胃炎
ジェーン	37	33k	消化不良
ケン	40	35k	風邪
リンダ	43	26k	胃炎
ポール	52	33k	消化不良
スティーブ	56	34k	胃炎

表 2　表 1 の一般化テーブル

G.ID	年齢	コード	病気
1	[21,22]	[12k,14k]	消化不良
1	[21,22]	[12k,14k]	気管支炎
2	[23,24]	[18k,25k]	風邪
2	[23,24]	[18k,25k]	胃炎
3	[36,41]	[20k,27k]	風邪
3	[36,41]	[20k,27k]	胃炎
4	[37,43]	[26k,35k]	消化不良
4	[37,43]	[26k,35k]	風邪
4	[37,43]	[26k,35k]	胃炎
5	[52,56]	[33k,34k]	消化不良
5	[52,56]	[33k,34k]	胃炎

［**定義** 3］（補助の表） 補助の表 $R(j)$ は二つの列”Group-ID”と”Count”を持つ．表 $T^*(j)$ の中で偽造タプルを少なくとも一つ含む QI グループ QI^* に対して，補助の表 $R(j)$ には列 $[g, c]$ がある．g はグループ QI^* の ID であり，c はグループ QI^* の偽造されたタプルの個数である．偽造されたタプルが表 $T^*(j)$ に存在しないなら，補助の表 $R(j)$ は空である．

例 1. 表 $T(j)$(表 3)，表 $T^*(j)$(表 5)，表 $R(j)$(表 6) によって定義 2，3 を説明する．表 $T(j)$ のタプル ⟨ ボブ,21,12k, 消化不良 ⟩ を t として考える．ホスティンググループ $t.QI(j)$ は表 $T(j)$ の最初の二つの行となる．表 $T^*(j)$ でタプル t の一般化されたタプルは ⟨1,[21,22],[12k,14k], 消化不良 ⟩ である．t の一般化されたホスティンググループ $t.QI^*(j)$ は表 $T^*(j)$ で二つの行である Group-ID=1 であるタプルから成る．特に，$t.QI^*(j)$ でのタプル c_1 は偽造され，$T(j)$ には存在しない。ID=3 での QI グループにももう一つ偽造された c_2 がある．補助の表 $R(j)$ は各 QI グループで偽造された個数を示す．

表 3　2 回目の元データ

名前	年齢	コード	病気
ボブ	21	12k	消化不良
デイビット	23	25k	胃炎
エミリー	25	21k	風邪
ジェーン	37	33k	消化不良
リンダ	43	26k	胃炎
ガーリー	41	20k	風邪
メアリー	46	30k	胃炎
レイ	54	31k	消化不良
スティーブ	56	34k	胃炎
トム	60	44k	胃炎
ベニス	65	36k	風邪

表 4　表 3 の一般化テーブル

G.ID	年齢	コード	病気
1	[21,23]	[12k,25k]	消化不良
1	[21,23]	[12k,25k]	胃炎
2	[25,43]	[21k,33k]	風邪
2	[25,43]	[21k,33k]	消化不良
2	[25,43]	[21k,33k]	胃炎
3	[41,46]	[20k,30k]	風邪
3	[41,46]	[20k,30k]	胃炎
4	[54,56]	[31k,34k]	消化不良
4	[54,56]	[31k,34k]	胃炎
5	[60,65]	[36k,44k]	胃炎
5	[60,65]	[36k,44k]	風邪

表 5　動的データに対する一般化テーブル

名前	G.ID	年齢	コード	病気
ボブ	1	[21,23]	[12k,14k]	消化不良
c_1	1	[21,23]	[12k,14k]	気管支炎
デイビット	2	[23,25]	[21k,25k]	胃炎
エミリー	2	[23,25]	[21k,25k]	風邪
ジェーン	3	[37,43]	[26k,33k]	消化不良
c_2	3	[37,43]	[26k,33k]	風邪
リンダ	3	[37,43]	[26k,33k]	胃炎
ガーリー	4	[41,46]	[20k,30k]	風邪
メアリー	4	[41,46]	[20k,30k]	胃炎
レイ	5	[54,56]	[31k,34k]	消化不良
スティーブ	5	[54,56]	[31k,34k]	胃炎
トム	6	[60,65]	[36k,44k]	胃炎
ベニス	6	[60,65]	[36k,44k]	風邪

表 6　偽造テーブル

G.ID	個数
1	1
3	1

［**定義** 4］（履歴の統合） 時刻 n≧1 で、統合表 $U(n)$ はそれぞれの 1,2,...,n のタイムスタンプでの表 T のすべてのタプルを含む表であり，以下のように形式的に記述される．：

$$U(n) = \bigcup_{j=1}^{n} T(j) \tag{1}$$

$t \in U(n)$ を満たす各タプル t はライフスパン $[x, y]$ で関連づけられる．ここで x は表 T に現れた始めの時刻 j を表し，y は表 T に現れた最後の時刻 j を表す．

統合表 $U(n)$ は表 T と同じような表と見すことが出来る．もしタプルが異なる j でいくつかの表 $T(j)$ に現れるなら，統合表 $U(n)$ では，重複させないようにする．例として表 7 に示す．以降，我々は統合表を用いて攻撃者の背景知識を定義する．

表 7　統　合　表

名前	年齢	コード	病名
ボブ	21	12k	消化不良
アリス	22	14k	気管支炎
アンディ	24	18k	風邪
デイビット	23	25k	胃炎
ガーリー	41	20k	風邪
ヘレン	36	27k	胃炎
ジェーン	37	33k	消化不良
ケン	40	35k	風邪
リンダ	43	26k	胃炎
ポール	52	33k	消化不良
スティーブ	56	34k	胃炎
エミリー	25	21k	風邪
メアリー	46	30k	胃炎
トム	60	44k	胃炎
ベニス	65	36k	風邪

［**定義** 5］（背景知識）　時刻 n での相手の知識は以下とする．

- 利用される一般化の法則 (一般化の方法)
- 知識テーブル $B(n)$:列 A^g と名付けられた”グループ ID” と A^l と名付けられた”ライフスパン”と機密情報を除く表 $U(n)$ のすべての属性を持つ (今回の場合，名前と年齢とジップコード) 表であり，以下に内容を詳細に示す．

　– すべてのタプル $t \in$U(n) に対して，$1 \leqq$i$\leqq d$ の間のすべてにおいて $b[A^g]=*$，$b[A^{id}]$=t$[A^{id}]$，$b[A_i^{qi}]$=t$[A_i^{qi}]$ であり，$b[A^l]$ が t のライフスパンと等しいような行 b$\in B(n)$ をもつ．

　– $1 \leqq$j$\leqq n$ で各表 $T^*(j)$ の偽造された t_c に対して，$b_c[A^g]=t_c[A^g]$，$b_c[A_i^{qi}]$ が $B(n)$，$b_c[A_i^{qi}]=0$ すべての $1 \leqq$i$\leqq d$ の間で $b_c[A^l]=[j, j]$ において偽造の識別子であるような行 $b_c \in B(n)$ がある．

同じく，知識テーブル $B(n)$ は (1) 列 A^s を除く統合表 $U(n)$ のすべてのタプル，(2) 統合表 $U(n)$ で各タプルのライフスパン，(3) すべての偽造されたタプルの公開された細部を含んでいる．

例 2.n=2 で仮定することによって，表 $T(1)$(表 1)，表 $T^*(1)$(表 2)，表 $T(2)$(表 3)，表 $T^*2(2)$(表 5)，補助の表 $R(2)$(表 6) として定義 4 と 5 を説明する．

重複を除いた後で，統合表 $U(2)$(表 7) は $T(1)$ と $T(2)$ のすべてのタプルを含む．表 $T(1)$，表 $T(2)$ に存在するので，タプル ⟨ ボブ,21,12k, 消化不良 ⟩ のライフスパンは [1,2] である．一方で，⟨ アリス,22,14k, 気管支炎 ⟩ のライフスパンは [1,1] である．なぜなら，時刻 1 では挿入され，時刻 2 では削除されているからである．表 8 は表 $B(n)$ の一部を表している．例として，$U(2)$ のタプル ⟨ ボブ,21,12k, 消化不良 ⟩ に対して，対応する行 $b \in B(n)$ は表 2 の最初のタプルである．b[Group-ID]=*は相手が t の一般化されたホスティンググループについて確信がないことを意味する．

表 5 の偽造された c_1 を t_c とする．行 b$\in B(n)$ が表 2 の最後から 2 番目のタプルに対応する．相手は b_c[Group-ID]=1 を知っている．なぜなら，この値は $R(2)$ に明確に示されているからである．一般に偽造されたタプルは一般化以前に”特定の QI 値” を持たないので，QI 値は Ø である．すべての偽造タプルは一つの表にしか使えない．例えば，b_c[Lifespan] は t_c が $T^*(2)$ にあることを意味している．

もしグループ ID を取り除くなら，知識テーブル $B(n)$ が公開表に対するプライバシー攻撃 (機密属性などを推測すること) を実行する間，一般に想定された相手のもつ最大の知識を表している．現実に，相手の知識は知識テーブル $B(n)$ に示されるよりもっと少ないだろう．例えば，知識テーブル $B(n)$ は病院での受診者の情報の連続的公開の場合，(1) すべての患者の名前と歳とジップコード，(2) 病院に滞在する正確な日付などをすべて含むことになる．言い換えると，定義 5 の背景知識を利用した攻撃から守ることによって，我々は実際に出くわすより不利な状況でのプライバシー保護を目的としている．

補助の表 $R(1)$,...,$R(n)$ のすべての情報は知識テーブル $B(n)$ にあると考える．このとき，情報漏洩は以下のように定式化される．

表 8　相手の背景知識テーブル B(2)

G.ID	名前	年齢	コード	スパン
∗	ボブ	21	12k	[1,2]
∗	アリス	22	14k	[1,1]
∗	アンディ	24	18k	[1,1]
∗	デイビット	23	25k	[1,2]
∗	ガーリー	41	20k	[1,2]
∗	ヘレン	36	27k	[1,1]
∗	ジェーン	37	33k	[1,2]
∗	ケン	40	35k	[1,1]
∗	リンダ	43	26k	[1,2]
∗	ポール	52	33k	[1,1]
∗	スティーブ	56	34k	[1,2]
∗	エミリー	25	21k	[2,2]
∗	メアリー	46	30k	[2,2]
∗	トム	60	44k	[2,2]
∗	ベニス	65	36k	[2,2]
1	c_1	Ø	Ø	[2,2]
3	c_2	Ø	Ø	[2,2]

［**定義** 6］（プライバシー違反）　もし表 $T^*(1)$,...,$T^*(n)$ と知識テーブル $B(n)$ を利用して．相手が正確に任意のタプルの機密情報を見つけ出すことが出来るなら，プライバシー違反が起こる．

例えば，例 2 でもし相手がテーブル 1b,3a,4 からボブの病気が消化不良だと推測することができるならプライバシー違反である．

以下に再公開問題の再帰的な定義を示す.

[**定義** 7]（再公開） 公開者が元のデータ表 T について n-1 回匿名化した表を公開していると想定する. $\{T^*(1),R(1)\}$...,$\{T^*(n-1),R(n-1)\}$ や $\{T^*(j),R(j)\}$ は定義 2 と 3 で定義されている公開表とする. このとき, プライバシー保護再公開の目的はプライバシー違反を最小限にしながら, できるだけ多くの元データの情報を含む表のペア $\{T^*(n),R(n)\}$ を作成することである.

以降では, 特別な場合を除き, 公開表 $T^*(1)$ と補助の表 $R(1)$ を合わせて公開表 $T^*(1)$ と呼ぶ.

3. 従 来 手 法

従来研究 [1] では, 動的データテーブルの匿名化表の列が満たすことにより, 連続的なプライバシー保護公開可能とする性質である m-不変性という概念が提案されている. 以下で, m-不変性を考える上で必要な定義を示し, 続いて m-不変性の概念を示す.

3.1 m-不変性

[**定義** 8]（シグネチャ） 1 から n の間の任意の j に対する公開表 T^*(j) における QI グループを QI^* とする. グループ QI^* のシグネチャは QI^* での機密情報の集合とする.

次に, 従来研究 [1] で再公開でのプライバシー保護のため, 公開テーブルが満たすべき性質と提案された m-不変性について定義する.

[**定義** 9]（m-不変性） 一般化されたテーブル $T^*(j)$ が (1) $T^*(j)$ の各 QI グループが少なくとも m 個のタプルを含み, かつ (2) QI グループ内のすべてのタプルが異なる機密情報を持つなら, $T^*(j)$ は m-ユニークであると言う. このとき, 公開された表 $T^*(1)$,...,$T^*(n)$ の列が以下の条件を満たすなら m-不変性であると言う：

（1） すべての表 T^*(j) は m-ユニークである.

（2） ライフスパン $[x,y]$ で任意のタプル t に対して, $t.QI^*$(x),$t.QI^*(x+1)$,...,$t.QI^*(y)$ は同じシグネチャを持つ. ここで $t.QI^*(j)$ は時刻 j$\in[x,y]$ での公開表 $T^*(j)$ でのタプル t の一般化されたホスティンググループである.

m-ユニーク性は各機密情報がすべての QI グループの中で高々 1 回だけ現れるべきであると要求する. 明白に, m-ユニーク性は m-多様性を意味する. しかし各 QI グループに m 種類の機密属性が存在すればよい m-多様性の定義の場合, 反対はそうではない. m-不変性の原理では, もしタプル t が数回にわたって連続で公開されるなら, タプル t を含むすべてのホスティンググループは同じ機密情報を持たなければならない.

[補題 1] もし $\{T^*(1),...,T^*$(n)$\}$ が m-不変性であるなら, そのとき risk(t)$\leqq 1/m$ となる. ここで risk とは, 任意のタプル t の機密属性が特定される確率である.

それゆえ, 公開者はプライバシー保護を達成するために十分な値を単に m を定めることができる.

[補題 2] もし $\{T^*(1),...,T^*$(n-1)$\}$ が m-不変性を満たすなら, そのとき $\{T^*(1),...,T^*(n-1),T^*(n)\}$ は以下の 2 つの条件を満たすとき, かつその場合に限り, m-不変性を満たす：

（1） $T^*(n)$ は m-ユニークである.

（2） $T(n-1)\cap T(n)$ に含まれるタプル t に対して, 一般化されたホスティンググループである $t.QI^*(n-1)$ と $t.QI^*(n)$ は同じシグネチャを持つ.

補題 2 は再公開を行うため追加的なアプローチを取ることができることを示す. 特に $T^*(n)$ を準備するため, 公開者は表 $T(n-1)$ と $T(n)$ と最後に公開された表 $T^*(n-1)$ を考えるだけでよい. 古い元データである表 $T(1)$,...,$T(n-2)$ と同様に公開された表 $T^*(1)$,...,$T^*(n-2)$ は保持する必要がない.

3.2 問 題 点

従来研究 [1] での m-不変性はプライバシーを保護することに関しては今のところ問題はない. しかし m-不変性を保つために偽造タプルを挿入しなければならない等, 情報損失が大きくなるという問題がある. そこで情報損失を低減させるために本研究では新しい安全性の提案を行う.

3.3 公開データ作成方法

従来研究 [1] では, m-不変性を満たす 1 回目の公開データ作成時に考察する性質も定義されている.

[**定義** 10]（m-エリジブル） 元のデータの表の中で最も多い機密属性 A^s の要素数/元データのタプルの個数が $1/m$ 以下なら, このデータ集合は m-エリジブルと呼ぶ. また元データが m-エリジブルであるなら, m-不変性を満たす公開データを作成することが出来る.

m-エリジブル性質に基づく公開データ作成アルゴリズムについて述べる.

3.4 1 回目の公開データ作成アルゴリズム (提案 1)

元のデータを入力データとして, m-ユニークな公開データを作成する. 公開者によって保管されている元のデータの表を T とする. また元データのタプルの総数を $tupleN$ とする. 表 T の中の列を識別子属性 A^{id}(今回の場合, 名前), d 個の準識別子属性 A_1^{qi},...,A_d^{qi}(今回の場合, 年齢, ジップコード), 機密属性 A^s(今回の場合, 病名) で分類する. 機密属性 A^s の中で, 同じ機密属性 A^s は n 種類あると想定できるのでそれぞれの種類の個数を数える. そして個数の多いものから順に並べる. このとき一番多いものから順に機密属性 A_1^s,...,A_n^s(表 T の場合, A_1^s=胃炎, A_2^s=消化不良, A_3^s=風邪, a_4^s=気管支炎) とする.

以下に公開データ作成アルゴリズムを示す.

[**Step** 1] 最初に元のデータが m-エリジブルを満たすかどうか確認する. その後に最も多い要素数個の QI グループが存在すると考えて $|A_1^s|$ 個数分バケットを作成する (表 T(1) の場合, A_1^s=胃炎の 4 つが最大なので 4 つバケットを作成する). 作成したバケットに A_1^s のタプルを振り分ける.

次にまだ振り分けていないタプルの振り分けを行う.

[**Step** 2] 振り分ける前に m-エリジブルかどうか確認する. 確認後, 次に要素数の多いタプルの 1 つをバケットに振り分ける. 次に残りの振り分けられていないタプルの要素数の多いものを選択しておく.

すべての残りの要素数が 1 になるまで Step2 を繰り返す.

［**Step** 3］ 新しくバケットを作り残りのすべてのタプルを入れる.

このアルゴリズムによって初期解ではあるが公開データを作成できた．元データを表 9 として作成した公開データを表 10 とする．しかし，このアルゴリズムでは情報損失について考慮していないので情報損失を改良するためにタプルの交換を行う必要がある.

表 9　1 回目の元データ

名前	年齢	コード	病気
ボブ	22	15k	消化不良
アリス	22	14k	気管支炎
アンディ	24	18k	風邪
デイビット	23	25k	胃炎
ガーリー	40	20k	風邪
ヘレン	36	27k	胃炎
ジェーン	37	33k	消化不良
ケン	40	25k	風邪
リンダ	43	26k	胃炎
ポール	52	33k	消化不良
スティーブ	56	34k	胃炎

表 10　1 回目の公開データ (提案 1)

名前	G.ID	年齢	コード	病気
ボブ	1	[22,24]	[15k,25k]	消化不良
デイビット	1	[22,24]	[15k,25k]	胃炎
アンディ	2	[24,36]	[18k,27k]	風邪
ヘレン	2	[24,36]	[18k,27k]	胃炎
ジェーン	3	[37,43]	[26k,33k]	消化不良
リンダ	3	[37,43]	[26k,33k]	胃炎
ガーリー	4	[40,56]	[20k,34k]	風邪
スティーブ	4	[40,56]	[20k,34k]	胃炎
アリス	5	[22,52]	[14k,33k]	気管支炎
ケン	5	[22,52]	[14k,33k]	風邪
ポール	5	[22,52]	[14k,33k]	消化不良

4. 新しい安全性の考察

本節では始めに m-ユニークな公開データを m-エリジブルという概念を使わずに作成する公開データ作成アルゴリズムを提案する．次に提案した公開データ作成アルゴリズムを用いて作成した公開データの列は m-不変性の制約条件の緩和が可能か否かを考察する.

4.1 1 回目の公開アルゴリズム (提案 2)

元のデータを入力データとして，公開しても機密情報が特定できないような公開データを作成する．公開者によって保管されている元のデータの表を T とする．また元データのタプルの総数を $tupleN$ とする．表 T の中の列を識別子属性 A^{id}(今回の場合，名前)，d 個の準識別子属性 A_1^{qi},...,A_d^{qi}(今回の場合，年齢，ジップコード)，機密属性 A^s(今回の場合，病名) で分類する．機密属性 A^s の中で，同じ機密属性 A^s は n 種類あると想定できるのでそれぞれの種類の個数を数える．そして個数の多いものから順に並べる [2]．このとき一番多いものから順に機密属性 A_1^s,...,A_n^s(表 T の場合，A_1^s=胃炎，A_2^s=消化不良，A_3^s=風邪，a_4^s=気管支炎) とする.

このとき，以下の定理が成り立つことを新たに証明する.

[定理] $|A_1^s| \leqq \lfloor tupleN/m \rfloor$ の時，$\lfloor tupleN/m \rfloor$ 個の m-ユニークなグループへの表 T の分割が存在する.

[証明] $\lfloor tupleN/m \rfloor$ 個が最大バケット数であり，そのバケット数を持ち，m-ユニークを満たす解が存在することを示す．バケットとは，$\lfloor tupleN/m \rfloor$ 個のバケットを用意し，公開データの QI グループの 1 つにあてはまる各タプルを振り分けたものである.

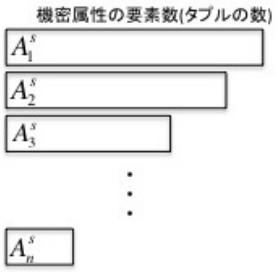

図 1　(a)

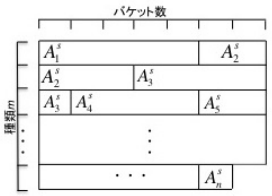

図 2　(b)

実際に，機密属性の要素数の多い順に 1 つずつバケットに振り分けると考える．このとき，命題の仮定よりすべての機密属性の要素数が $\lfloor tupleN/m \rfloor$ 以下であることがわかるから，同一バケットに同じ機密属性を持つタプルが振り分けられることはない (図 2(b) 参照)．したがって，バケット数 $\lfloor tupleN/m \rfloor$ で m-ユニークな解が存在するといえる.

上記の定理に基づいて，以下の m-ユニークなグループを作成する手法を提案する.

［**Step** 1］ それぞれの QI グループに分割するためにバケット BUC_1,...,BUC_l を作成する．定理を用いてバケットの個数 l を $\lfloor$tupleN$/m\rfloor$ とする．次に l 個の空のバケットを作成する．以後はバケットにタプルを振り分けていく.

［**Step** 2］ 最初に l 個のタプルを振り分けていく．このとき，順に並べておいた機密属性 A_1^s から振り分ける．例えば BUC_1 には機密属性 A_1^s である胃炎のタプルの一つを振り分ける．これをバケット数 l 回分行う.

［**Step** 3］ 次に振り分けていない残りのタプルを振り分けていく．振り分けるタプル t は，Step2 の振り分けが終わったタプルの次を選択する．振り分けを行うタプル t は各バケット BUC の中に入れたときに自信の値が情報損失の最も少ないバケットを選択する．この選択を行う際，以下の条件を満たす.

- タプルの個数が最小のバケットの中で選択する.
- バケット内はすべて異なる機密属性を持つので，同じ機密属性が存在する場合はそのバケットは選択しない.

Step3 をバケットにタプルがすべて振り分けられるまで行う．元データを表 9 として作成した公開データを表 11 とする． 表

表 11　1 回目の公開データ (提案 2)

名前	G.ID	年齢	コード	病気
アリス	1	[22,24]	[14k,25k]	気管支炎
アンディ	1	[22,24]	[14k,25k]	風邪
デイビット	1	[22,24]	[14k,25k]	胃炎
ジェーン	2	[36,37]	[27k,33k]	消化不良
ヘレン	2	[36,37]	[27k,33k]	胃炎
ガーリー	3	[40,43]	[20k,26k]	風邪
リンダ	3	[40,43]	[20k,26k]	胃炎
ポール	4	[52,56]	[33k,34k]	消化不良
スティーブ	4	[52,56]	[33k,34k]	胃炎
ケン	5	[22,40]	[15k,25k]	風邪
ボブ	5	[22,40]	[15k,25k]	消化不良

1 の元データを以上の Step で公開データにする．このとき，節 3.4 での手法との情報損失の比較を表 12 で示す．ここで，情報損失は一般化表での準識別子属性値の各区間の長さの総和とする．

4.2 比　　較

1 回目の公開データ (提案 1) と 1 回目の公開データ (提案 2) の情報損失の量を比較する．各準識別子 (今回の場合，列年齢と列コード) を正規化した値を用いる．結果を以下の表 9 に示す．

表 12　情報損失の比較

	年齢	ジップコード
提案 1	3.231	4.005
提案 2	1.730	3.278

これより 1 回目の公開アルゴリズム (提案 2) の方が情報損失が低減されているので以後使用する．

4.3 m-不変性を満たさない公開データの安全性

次に 2 回目の公開データ作成を考える．その際，2 つの方法を使用してデータの作成を試みた．1 つ目は 1 回目の公開アルゴリズム (提案 2) を使用する方法，2 つ目は m-不変性の定義に従って，m-不変性を満たす公開データを作成した．1 回目の元データを表 9，公開データを表 11，2 回目の元データを表 13 として 2 回目の公開データを作成した．1 つ目の 1 回目のアルゴリズム (提案 2) を使用し公開データを作成した表を表 14 に示す．2 つ目の従来研究 [1] で定義されている m-不変性を満たすように作成した表を表 15 に示す．まず，表 14 が再公開可能かどうか検討する．

表 13　2 回目の元データ

名前	年齢	コード	病気
ボブ	22	15k	消化不良
デイビット	23	25k	胃炎
エミリー	25	21k	風邪
ジェーン	37	33k	消化不良
リンダ	43	26k	胃炎
ガーリー	40	20k	風邪
メアリー	46	30k	胃炎
レイ	54	31k	消化不良
スティーブ	56	34k	胃炎
トム	60	44k	胃炎
ベニス	65	36k	風邪

表 14　2 回目の公開データ

番号	名前	G.ID	年齢	コード	病気
1	ベニス	1	[60,65]	[36k,44k]	風邪
2	トム	1	[60,65]	[36k,44k]	胃炎
3	レイ	2	[54,56]	[31k,34k]	消化不良
4	スティーブ	2	[54,56]	[31k,34k]	胃炎
5	ジェーン	3	[37,46]	[30k,33k]	消化不良
6	メアリー	3	[37,46]	[30k,33k]	胃炎
7	ガーリー	4	[40,43]	[20k,26k]	風邪
8	リンダ	4	[40,43]	[20k,26k]	胃炎
9	エミリー	5	[22,25]	[15k,25k]	風邪
10	ボブ	5	[22,25]	[15k,25k]	消化不良
11	デイビット	5	[22,25]	[15k,25k]	胃炎

表 15　2 回目の公開データ

番号	名前	G.ID	年齢	コード	病気
1	偽造タプル c_1	1	[22,25]	[14k,25k]	気管支炎
2	エミリー	1	[22,25]	[14k,25k]	風邪
3	デイビット	1	[22,25]	[14k,25k]	胃炎
4	ジェーン	2	[37,46]	[30k,33k]	消化不良
5	メアリー	2	[37,46]	[30k,33k]	胃炎
6	ガーリー	3	[40,43]	[20k,26k]	風邪
7	リンダ	3	[40,43]	[20k,26k]	胃炎
8	偽造タプル c_2	4	[52,56]	[33k,34k]	消化不良
9	スティーブ	4	[52,56]	[33k,34k]	胃炎
10	ベニス	5	[22,65]	[15k,36k]	風邪
11	ボブ	5	[22,65]	[15k,36k]	消化不良
12	レイ	6	[54,60]	[31k,44k]	消化不良
13	トム	6	[54,60]	[31k,44k]	胃炎

具体的には，始めに (場合 1)T(2)-T(1) の集合 { エミリー, メアリー, レイ, トム, ベニス } に属する 2 回目の表で追加された各要素の機密情報が特定される可能性を調べる．次に，(場合 2)T(1)-T(2) の集合 { アリス, アンディ, ヘレン, ケン, ポール } に属する 2 回目の表で削除された各要素の機密情報が特定される可能性を調べ，最後に，(場合 3)T(1)(表 9)∩T(2)(表 13) の集合 { ボブ, デイビット, ジェーン, リンダ, ガーリー, スティーブ } の各要素 t に関して推測される機密情報が特定される可能性 risk(t)≦1/m であることを確認する．例では m=2 であると

する．また表 11 と表 14 での任意のタプルが属する可能性のあるシグネチャをまとめた表を表 16,17 に示す．

表 16　表 11 の視点 1

名前	成り得る機密属性の候補
アリス	{ 気, 風, 胃 }
アンディ	{ 気, 風, 胃, 消 }
デイビット	{ 気, 風, 胃, 消 }
ジェーン	{ 消, 胃 }
ヘレン	{ 消, 胃 }
ガーリー	{ 風, 胃, 消 }
リンダ	{ 風, 胃 }
ポール	{ 消, 胃 }
スティーブ	{ 消, 胃 }
ケン	{ 風, 消, 胃 }
ボブ	{ 気, 風, 胃, 消 }

表 17　表 14 の視点 1

名前	成り得る機密属性の候補
ベニス	{ 風, 胃 }
トム	{ 風, 胃 }
レイ	{ 消, 胃 }
スティーブ	{ 消, 胃 }
ジェーン	{ 消, 胃 }
メアリー	{ 消, 胃 }
ガーリー	{ 風, 胃 }
リンダ	{ 消, 胃 }
エミリー	{ 風, 消, 胃 }
ボブ	{ 風, 消, 胃 }
デイビット	{ 風, 消, 胃 }

(場合 1) 始めに $T^*(1)$(表 11) では存在しなかったが $T^*(2)$(表 14) で挿入されたタプルについて考える．例えば，エミリーのタプルは属する QI グループのタプルが 3 つあるのでエミリーの機密属性値を推測されるリスク risk(エミリー) は $1/|\{$ エミリー, ボブ, デイビット $\}|$ より $1/3$ となり，$1/m$ 以下となる．メアリー, レイ, トム, ベニスの QI グループのタプルは各々2 つずつある．上記 4 名に関しても推測のリスクは $1/m$ となる．この方法を使用した場合，偽造タプルが必要なく再公開が可能となる．

(場合 2) 次に $T^*(1)$(表 11) には存在したが $T^*(2)$(表 14) では削除されたタプルについて考える．タプルを削除する際に，$T^*(1)$ での QI グループの中の任意のタプル t が $T^*(2)$ で m-不変性を満たされていないならば，同じ QI グループのタプルが m 個以上削除されるなら再公開可能である．今回の場合，アリスのタプルがあてはまる．$T^*(1)$ でアリスの QI グループで $T^*(2)$ でも公開されるデイビットは m-不変性ではない．しかし，同じ QI グループのタプルが m 個以上削除されるので $T^*(1) \cap T^*(2)$ によってアリスの機密属性値を推測されるリスク risk(アリス) は $1/|\{$ アリス, アンディ$\}|$ より $1/m$ となる．

(場合 3) 最後に $T^*(2)$(表 14) においてスティーブの場合，m-ユニークであり同じスティーブが属するグループは同じ機密属性値の集合を持つので m-不変性の性質を満たす．ジェーン，ガーリー，リンダも同様に m-不変性の性質を満たす．よって上記 4 名に関して推測のリスクは $1/m$ 以下となる．またボブの場合，$T^*(1)$ のボブが属するグループの機密属性値は { 消化不良, 風邪 } であり，$T^*(2)$ でのボブが属するグループの機密属性値の集合は { 風邪, 消化不良, 胃炎 } となるので，$T^*(1) \cap T^*(2)$ によって機密属性値を推測されるリスク risk(ボブ) は $1/|\{$ 消化不良, 風邪 $\}|$ となり，$1/m$ となる．またデイビットの場合は，表 $T^*(1)$ でデイビットが属するグループの機密属性値の集合は { 気管支炎, 風邪, 胃炎, 消化不良 } であり，表 $T^*(2)$ では { 風邪, 消化不良, 胃炎 } なので，G.ID=5 の QI グループが公開可能かどうかわかっていない．それを確かめるために表 18 を作成した．この表の作成方法としてはまずボブの属する機密属性は { 風, 消 } であるので属する risk は 1/2 ずつになることがわかる．そしてボブの risk の和は 1 となっている．また表 16 よりタプル番号 11 の胃炎に属するのは { エミリー, デイビット } なので risk は 1/2 ずつになる．そして胃炎の risk の和は 1 となる．残りの risk については機密属性とタプルの両方の risk の和が 1 となるように分ける．これによってすべての risk の和が 1 になることがわかる．そのなかでこの表のすべてが $1/m$ 以下になるので公開可能である．またすべてのタプルの成り得る機密属性の確定した表を表 19 とする．

表 18　属性分布確率

	風邪	消化不良	胃炎	
エミリー	1/3	1/3	1/3	1
ボブ	1/3	1/3	1/3	1
デイビット	1/3	1/3	1/3	1
	1	1	1	

表 19　統合表のタプルの成り得る機密属性

名前	成り得る機密属性
ベニス	{ 風, 胃 }
トム	{ 風, 胃 }
レイ	{ 消, 胃 }
スティーブ	{ 消, 胃 }
ジェーン	{ 消, 胃 }
メアリー	{ 消, 胃 }
ガーリー	{ 風, 胃 }
リンダ	{ 消, 胃 }
エミリー	{ 風, 消, 胃 }
ボブ	{ 風, 消 }
デイビット	{ 風, 消, 胃 }
アリス	{ 気, 風, 胃, 消 }
アンディ	{ 気, 風, 胃, 消 }
ヘレン	{ 消, 胃 }
ポール	{ 消, 胃 }
ケン	{ 風, 消, 胃 }

従来研究 [1] で定義されている m-不変性を保つように作成した公開データ (表 15 の場合) の場合すべてのデータについてのリスクが $1/m$ 以下になるので公開可能となる．しかし偽造タプルは 2 つ必要となる．

この 2 つの公開データを情報損失と偽造タプルの個数から比

較する．比較した表をまとめたものをそれぞれ表 20,21 をして公開する．これによると情報損失は低減していることがわかる．偽造タプルの重さを大きいものと考えると本研究の提案を使用して公開データを作成する方がよい．

表 20　情報損失の比較

	年齢	ジップコード
表 14	1.017	3.803
表 15	1.218	4.625

表 21　偽造タプルの個数の比較

	個数
表 14	0
表 15	2

4.4　安全性の提案

4.3 節の議論より従来の安全性の定義である m-不変性の制約が厳しすぎる場合があると考えられる．本研究では，偽造タプルの数を減らすことを目的として，背景知識テーブルを背景知識としたときに元データ表のタプルと一般化表のタプルとを連結する攻撃 (record-linkage[3]) と元データ表のタプルと機密属性値とを連結する攻撃 (attribute-linkage[3]) の双方が成功する確率を $1/m$ 以下に抑制するために，連続的匿名化結果である一般化表の列が満たすべき安全性の性質を新たに提案する．

［**定義** 11］ (m-連続安全性)

1 から l までの各時刻における元データ表の系列 T_1,T_2,...,T_l とそれぞれに対応する一般化表の系列 T_1^*,T_2^*,...,T_l^*，および，時刻 l までの元データ表の統合表 $U(l)$ が与えられたとき，以下の条件を全て満たすリスク変数行列の列 $Risk_1$,$Risk_2$,...,$Risk_l$(図 3 参照) の各要素への 0 以上 1 以下の実数値の割り当てが存在するなら，一般化表の系列 T_1^*,T_2^*,...,T_l^* は m-連続安全である．

ここで，ある時刻 $h(1 \leqq h \leqq l)$ でのリスク変数行列 $Risk_h$ は

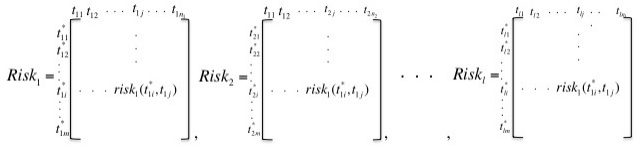

図 3　リスク変数行列

時刻 h での元データ表 T_h のサイズ n_h=$|T_h|$ と一般化表 T_h^* のサイズ n_h=$|T_h^*|$ において，m_h 行と n_h 行の行列であり，i 行 (i≦i≦m_h)j 行 (i≦j≦n_h) の要素を $risk_h(t_i^*,t_j)$ 変数とする．$risk_h(t_i^*,t_j)$ 変数での t_i^* は一般化表 T_h^* での i 番目のタプルを表し，t_j は元データ表での j 番目のタプルを表し，$risk(t_i^*,t_j)$ は一般化表 T_h^* のタプル t_i^* が元データ表のタプル t_j の一般化である可能性 (確率) を示す変数となる．

[条件 1] 各時刻 h において，各元データ表の各タプルは対応する一般化表のいずれかのタプルに該当する．

$\forall h \in [1,l], \forall t_{org} \in T_h [\sum_{t_g \in T_h^*} risk_h(t_g,t_{org})=1]$

[条件 2] 各時刻 h において，各一般化表 T_h^* の各タプルは対応する元データ表 T_h のいずれかのタプルに該当する．

$\forall h \in [1,l], \forall t_g \in T_h^* [\sum_{t_{org} \in T_h^*} risk_h(t_g,t_{org})=1]$

[条件 3] 各一般化表 T_h^* の各タプルにおいて，そのタプルが一般化タプルである可能性のあるタプルは一般化タプルのもつ準識別子の値の範囲に含まれる値しか持ち得ない．

$\forall h \in [1,l], \forall t_{org} \in T_h$,

$\forall t_h \in T_h^* [\exists A^{qi} \in \mathrm{QI}\ t_{org}[A^{qi}] \notin t_g[A^{qi}] \Rightarrow risk_h(t_g,t_{org})=0]$

[条件 4] 各時刻 h において，各元データ表 T_h のタプルの機密属性値が特定される確率は $1/m$ 以下である．

$\forall h \in [1,l], \forall t_{org} \in T_h$

$[\max\{\sum_{t_g \in T_h^* \wedge t_g^*[A_s]=S} risk_h(t_g,t_{org}) | \mathrm{S} \in A_s\} \leqq 1/m]$

[条件 5] 統合表 $U(l)$ に含まれる各タプルにおいて，各一般化表で該当する可能性がある機密属性の候補集合はすべて同じでなければならない．ここで $lifespan(t_{union})$ は，t_{union} が存在した期間を表す関数とする．

$\forall t_{union} \in U(l), \forall \mathrm{x} \in \mathrm{lifespan}(t_{union})$

$[\mathrm{x\text{-}1} \in \mathrm{lifespan}(t_{union}) \Rightarrow \{t_g^*[A_s] | t_g \in T_{x-1}^* \wedge risk_{x-1}(t_g,t_{union}) \neq 0\}]$

$=\{t_g^*[A_s] | t_g \in T_x^* \wedge risk_x(t_g,t_{union}) \neq 0\}$

[条件 6] 各元データ表 T_h のタプルに対応する一般化タプルに対しては，特定される確率を 0 にすることはできない．

$\forall \in [1,l], \forall t_{org} \in T_h, \forall t_g \in T_h^* [(t_{org}^*=t_g) \Rightarrow (risk_h(t_g,t_{org}) \neq 0)]$

m-不変性を満たせば，m-連続安全性も満たされる．しかし，m-連続安全性を満たしても，m-不変性を満たさない場合がある．例として，先に 4.3 で議論した表 11,14 の場合である．表 11,14 の場合，m-連続安全性を満たすリスク変数行列の列 $Risk_1$，$Risk_2$ の各要素への 0 以上 1 以下の実数値の割り当てが存在する．よって，m-連続安全性はすべての一般化テーブルに対する record-linkage 攻撃と attribute-linkage 攻撃の成功確率を $1/m$ 以下に抑えながら，m-不変性の制約を緩和した安全性の定義である．

5.　ま　と　め

本研究では情報損失の低減を目指して新しい安全性の考察を行った．実際に m-不変性を満たさなくても，m-連続安全性を用いて検証するとプライバシー保護データ作成することができ，偽造タプルを必要とせず情報損失を低減できる例を発見することができた．今後の課題としては，一般化表系列が m-連続安全性を満たすかどうかを判定する手法の提案，また，m-連続安全性を満たす一般化系列の作成手法の提案が挙げられる．

文　　献

[1] X.Xiao,Y.Tao,"m-invariance: toward privacy preserving republication of dynamic datasets," Proc. of SIGMOD'07, pp.689-700 (2007).

[2] 村本俊祐, 上土井陽子, 若林真一,"プライバシー保護データ公開に向けた l-多様化適正の評価", 情報処理学会論文誌 データベース, Vol.4, No.2, pp.126-141 (July 2011).

[3] B.G.M.Fung, K.Wang, R.Chen and P.S.Yu,"Privacy-preserving data publishing:a survey on recent developments," ACM Computing Surveys, Vol.42, No.1, pp.14:01-14:53 (2010).