連続的なデータ公開のための m-不変性に基づく一般化手法の安全性向上

張 扶蘇† 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: †zhangfusu@lcs.info.hiroshima-cu.ac.jp, ‡{yoko, wakaba}@hiroshima-cu.ac.jp

あらまし 本研究では挿入と削除より動的に変化する連続的なデータ公開表における一般化手法について安全性を考察する．従来使われている動的データに対する m-不変性を利用する一般化手法を用いて動的データを一般化したとき，m-不変性の定義によれば 1/m 以下になる機密値の特定確率が実際には 1/m を超えてしまう可能性が指摘されている．本研究では動的データの連続した一般化公開表を使ってこのような場合が実際にあることを明らかにし，この問題を解消するための改良方法を提案する．

キーワード 動的データ，データの一般化， m-不変性，プライバシー

1. はじめに

データ公開に対する個人情報データのプライバシー保護における研究は社会から注目されている．いままで，静的データを対象とした研究では，いくつかの個人情報を保護できる安全な匿名化保護手法が開発された．近年，連続的なデータ公開に対する個人情報の匿名化によりプライバシー保護が研究され始めた．

連続的なデータを対象とする手法の開発では，扱われているデータは動的に変化するデータのため，主にタプルの挿入と削除に注目されていた．従来研究 [1]では m-不変性という動的データのプライバシーが保護できる性質が定義されており，m-不変性という性質を保つことを目的とした一般化手法が提案された．手法の安全性を評価する方法として全射関数を利用した方法が提案されている．文献[1]で m-不変性の一般化は動的に更新するデータの再公開においても，個人の機密情報が特定される確率（リスク）を 1/m 以下に抑えられている安全な方法だと主張されている．しかし，文献[2]によりその安全性が成立する前提として，攻撃側も強制的に m-不変性の基準を考えている上で攻撃を行うことだと仮定している．この前提を置かないと，個人の機密情報が特定される確率（リスク）が 1/m 以上となる可能性がある．

本稿では，Xiao と Tao が提案した m-不変性の一般化手法[1]において個人の機密情報が特定されるリスクを 1/m 以下に抑えられない場合が存在することを明示し，そのような安全性の問題を克服するための情報提供方法と一般化手法を提案する．

2. 準備/定義

本章では，本研究で用いる用語の定義[1]を紹介する．

公開者が管理している元データからなるデータ表を T とする．T の要素を次のように分類する．

(i) T の主なキーとなる識別子(identifier)属性A^{id},
(ii) d 個の準識別子(quasi-identifier)属性$A_1^{qi}, \ldots, A_d^{qi}$
(iii) 機密値(sensitive value)属性A^s.

今回使用するデータでは，A^{id}が名前，A_1^{qi}とA_2^{qi}が年齢と郵便番号，A^sが病気の形となっている．

表 T はデータの挿入や削除によって更新される．j 番目の公開のタイムスタンプを時刻 j で表す．時刻 j の時の元データ表 T を T(j)とする．公開者はタプルの挿入や削除を行うことができるが，一旦，削除されたタプルが再び追加されることはないものとする．時刻 j で公開される T(j)を一般化した表を T*(j)と表す．

[定義 1] (QI グループ) データ表 T(j)に対して, QI グループは T(j)内のタプルの部分集合である．表 T(j)のタプルを複数の QI グループタプルに分割して，すべての QI グループの和集合が表 T(j)となる．各 QI グループには 1 から ID を割り当てる．タプル$t \in T(j)$に対して，$t.QI(j)$は t を含む QI グループを示している．

[定義 2] (偽造入りの一般化) 表 T(j)が匿名化された公開表 T*(j)は T(j)の属性に基づいて作成されて，次の性質を持っている．

(1) 表 T*(j)はA^{id}を除いて T(j)のすべての属性を含み，さらにはグループ ID と名付けられたA^gを持っている．
(2) 各元タプル$t \in T(j)$ と一般化されたタプル t*∈ T^*(j)の機密値に対して，以下の式を満たしている．

$$t^*[A^s] = t[A^s]$$

(3) 同じグループ ID, つまり, 同じA^gを持っている表 T*(j)のすべてのタプルはすべて同じ QI 属性値範囲の$A_1^{qi}, \ldots, A_d^{qi}$を持つ. 元データ表 T(j)にあるタプルの QI 属性の値は, 表 T*(j)にある一般化されたタプルの QI 属性の値の範囲に含まれる.
(4) 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. タプルt_c^*の機密値$t_c^*[A^s]$と準識別子$t_c^*[A_d^{qi}]$は所属する QI グループ$t_c^*[A^g]$の範囲内となる.

本研究の内容を説明するために, ある病院が患者の記録データを 1 時刻ごとに公開する場合を考える. 各公開には公開時点から六月分の診断結果のみが含まれる.

表 1 時刻 1 の元データ T(1)

Name	Age	Zip.	Disease
Bob	21	12k	dyspepsia
Alice	22	14k	bronchitis
Andy	24	18k	flu
David	23	25k	gastritis
Gary	41	2k0	flu
Helen	36	27k	gastritis
Jane	37	33k	dyspepsia
Ken	40	35k	flu
Linda	43	26k	gastritis
Paul	52	33k	dyspepsia
Steve	56	34k	gastritis

表 2 時刻 2 の元データ T(2)

Name	Age	Zip.	Disease
Bob	21	12k	dyspepsia
David	23	25k	gastritis
Emily	25	21k	flu
Jane	37	33k	dyspepsia
Linda	43	26k	gastritis
Gary	41	2k0	flu
Mary	46	3k0	gastritis
Ray	54	31k	dyspepsia
Steve	56	34k	gastritis
Tom	60	44k	gastritis
Vince	65	36k	flu

この病院が表 T の診断記録から属性 Name が削除され, 個人情報保護するための匿名化した上で表 T*を医学研究者に公開する. 手法の具体的なアルゴリズムは後の第四章で詳しく説明する. ここで, Age 列と Zip.列は病気を分析するため, 準識別子として公開する患者のデータである. Disease 列は, 保護すべき患者の個人データを含むため, 機密性が最も高い機密値である.

表 3 時刻 1 の公開データ T*(1)

G.ID	Age	Zip.	Disease
1	[21,22]	[12k,14k]	dyspepsia
1	[21,22]	[12k,14k]	bronchitis
2	[23,24]	[18k,25k]	flu
2	[23,24]	[18k,25k]	gastritis
3	[36,41]	[20k,27k]	flu
3	[36,41]	[20k,27k]	gastritis
4	[37,43]	[26k,35k]	dyspepsia
4	[37,43]	[26k,35k]	flu
4	[37,43]	[26k,35k]	gastritis
5	[52,56]	[33k,34k]	dyspepsia
5	[52,56]	[33k,34k]	gastritis

表 4 時刻 2 の公開データ T*(2)

G.ID	Age	Zip.	Disease
1	[21,23]	[12k,25k]	dyspepsia
1	[21,23]	[12k,25k]	gastritis
2	[25,43]	[21k,33k]	flu
2	[25,43]	[21k,33k]	dyspepsia
2	[25,43]	[21k,33k]	gastritis
3	[41,46]	[20k,30k]	flu
3	[41,46]	[20k,30k]	gastritis
4	[54,56]	[31k,34k]	dyspepsia
4	[54,56]	[31k,34k]	gastritis
5	[60,65]	[36k,44k]	gastritis
5	[60,65]	[36k,44k]	flu

[定義 3] (ライフスパン) 時刻 n まで, 公開表 T*(x)から T*(y) $(1 \le x \le y \le n)$まで存在する任意のタプル t において, t のライフスパンが[x, y]と定義する. 公開側ではライフスパンを公開しない.

[定義 4] (データの統合表) 時刻 n≥1 において, 統合表 U(n)はタイムスタンプ 1, 2, ..., n における T のすべてのタプルを統合してできる表である. 正式には次の式で表す.

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

各タプル t ∈ U(n)は暗黙のうちにライフスパン[lsf, lsb]と関連付けられ, t[lsf]と t[lsb]は t が T(lsf)または T(lsb)に現れる最小 (最大) の整数 lsf と lsb とする. U(n)は T(n)と同じように, 時刻を持つ表で表すことができる.

時刻 2 まで, T(1)と T(2)による作成された表 U(2)は, 表 T(1), 表 T(2)と完全に同じ表と考えられる. ここで, U(n)を用いて本研究で想定する敵対者の背景知識を定義する.

表 5 時刻 2 まで元データの統合表 U(2)

Name	Age	Zip.	Disease	Lifespan
Bob	21	12k	dyspepsia	[1,2]
Alice	22	14k	bronchitis	[1,1]
Andy	24	18k	flu	[1,1]
David	23	25k	gastritis	[1,2]
Gary	41	2k0	flu	[1,2]
Helen	36	27k	gastritis	[1,1]
Jane	37	33k	dyspepsia	[1,2]
Ken	40	35k	flu	[1,1]
Linda	43	26k	gastritis	[1,2]
Paul	52	33k	dyspepsia	[1,1]
Steve	56	34k	gastritis	[1,2]
Emily	25	21k	flu	[2,2]
Mary	46	3k0	gastritis	[2,2]
Ray	54	31k	dyspepsia	[2,2]
Tom	60	44k	gastritis	[2,2]
Vince	65	36k	flu	[2,2]

[定義 5] (敵対者が把握している背景知識表) 本稿では，敵対者が n 回まで公開した表を攻撃したいとき，T(n)まですべての元データ表 T(j)に対して，敵対者は機密値以外すべての情報の把握という最悪な状況を想定する．敵対者が機密値以外すべての情報と公開表の系列から機密値を推定して，真の機密値に特定するリスクについて考察する．

表 6 時刻 2 まで敵対者が把握できる背景知識表

Name	Age	Zip.	Lifespan
Bob	21	12k	[1,2]
Alice	22	14k	[1,1]
Andy	24	18k	[1,1]
David	23	25k	[1,2]
Gary	41	2k0	[1,2]
Helen	36	27k	[1,1]
Jane	37	33k	[1,2]
Ken	40	35k	[1,1]
Linda	43	26k	[1,2]
Paul	52	33k	[1,1]
Steve	56	34k	[1,2]
Emily	25	21k	[2,2]
Mary	46	3k0	[2,2]
Ray	54	31k	[2,2]
Tom	60	44k	[2,2]
Vince	65	36k	[2,2]
c1	∅	∅	[2,2]
c2	∅	∅	[2,2]

表 T*(j)は任意の数の偽造されたタプルt_c^*を含む場合もあるため，第三者が把握している情報は U(n)から機密値だけ取り除いた表と偽造タプルの和集合となる．

3. 従来研究

従来研究[1]では，時刻 n まですべての公開表が m-不変性という性質を満たせば，連続的なデータ表をプライバシー保護した上で公開可能と提案された．

本節では，m-不変性を理解するための定義を示し，n 番目のデータ表 T(n)から公開表 T*(n)を作成する m-不変性を満たすための一般化手法を詳しく説明する．

3.1 m-不変性

[定義 6] (シグネチャ) 時刻 n まで，任意の時刻 j に対して，公開表 T*(j)における QI グループの機密値の集合を，その QI グループのシグネチャと定義する．

例として，表 3 では，QI グループ[G.ID=1]のシグネチャが{dyspepsia, bronchitis}となり，QI グループ[G.ID=2] のシグネチャが{flu, gastritis}となる．

[定義 7] (m-ユニーク) 時刻 j のときの公開表 T*(j)において，以下の性質を満たすときに，公開表が m-ユニークであると定義する．

(1) T*(j)の各 QI グループが最小でも m 個のタプルを含む．

(2) 各 QI グループに存在するタプルがすべて異なる機密値を持っている．

[定義 8] (m-不変性) 時刻 n まで，タイムスタンプ 1, 2, …, n におけるすべての公開表 T*(1), T*(2)…T*(n)において，以下の性質を満たすときに，公開表 T*(1), T*(2)…T*(n)が m-不変性を満たすと定義する．

(1) 公開した n 個の公開表 T*(1), T*(2)…T*(n)がすべて m-ユニークである

(2) ライフスパン[x, y] ($1 \le x < y \le n$)で任意のタプル t において，t のシグネチャが変わらない

文献[1]では公開表 T*(1), T*(2)…T*(n)が m-不変性を満たすとき，任意のタプル t の機密値が特定される確率(特定リスク)が 1/m 以下に抑えられることが主張されている．

3.2 具体的なアルゴリズム

元データ表 T(n)と T(n+1)と元データ表 T(n)に対する公開表 T*(n)が与えられたときに，m-不変性を満たす公開表 T*(n+1)を作成する具体的なアルゴリズム[1]を紹介する．まずは，アルゴリズムで用いる用語を定義する．

[定義 9] (m-エリジブル) 元データ表 T において，T にあるすべてのタプルの中，最も多い機密値A^sの要素数の割合が 1/m 以下の場合，このデータ表 T を m-エリジブルを満たすと定義する．m を式で表すと:

$$m = \text{最も多い機密値} A^S_{max} / \text{全タプル数}$$

元データ表 T が m-エリジブルであるときのみ，m-不変性を満たす公開表 T*が作成できる．

[定義 10] (積集合$S_{n\cap}$と差集合$S_{n_}$) 元データ表 T(n)と表 T(n-1)に対して，表 T(n)と表 T(n-1)の両方に存在するタプルの集合を$S_{n\cap}$とし，$T(n)$のみに存在する新タプルの集合を$S_{n_}$とする．正式には次の式で表す．

$$S_{n\cap} = T(n) \cap T(n-1)$$
$$S_{n\_} = T(n) - T(n-1)$$

実例では，T(1)と T(2)から作られた$S_{2\cap}$と$S_{2_}$を表 7 と表 8 に示す．T(1)に対して，T(0)が存在しないため，$S_{1\cap} = \emptyset$，$S_{1_} = T(1)$となる．

表 7　T(1)と T(2)から作られた$S_{2\cap}$

Name	Age	Zip.	Disease
Bob	21	12k	dyspepsia
David	23	25k	gastritis
Gary	41	2k0	flu
Jane	37	33k	dyspepsia
Linda	43	26k	gastritis
Steve	56	34k	gastritis

表 8　T(1)と T(2)から作られた$S_{2_}$

Name	Age	Zip.	Disease
Emily	25	21k	flu
Mary	46	3k0	gastritis
Ray	54	31k	dyspepsia
Tom	60	44k	gastritis
Vince	65	36k	flu

以上の定義を用いて，データ表 T(n) を 4 つのフェーズにより公開表 T*(n)を作成する．

[Phase 1] 最初は，$S_{n\cap}$の各タプルを T*(n-1)にあるシグネチャにより分類する．このフェーズ 1 では，単純に$S_{n\cap}$を複数のバケットにシグネチャごとに分配する．T(n-1)が存在しないとき，このフェーズを飛ばす．

実例では，$\mathbf{S_{2\cap}}$の中のタプルが表 3 の T*(1)において，Bob のシグネチャが{dys,bro}，David と Gary のシグネチャが{flu,gas}，Jane と Linda のシグネチャが{dys,flu,gas}，Steve のシグネチャが{gas,dys}で，$S_{2\cap}$のタプルを分配した結果が図 1 に示す．

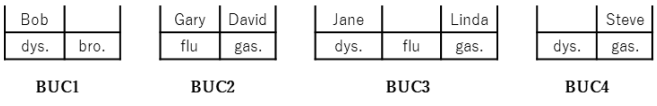

図 1 実例のバケット

[Phase 2] そしてフェーズ 2 では，$S_{n_}$からタプルをできるだけ取り出して，フェーズ 1 で作成したバケットに埋め込んでバランスを取ることである．

バケット BUC は，そのシグネチャのすべての機密値が BUC 内の同じ数のタプルによって所有されている場合，バランスが取れていると言う．各バケット BUC を順番に検査し，バランスが取れていない場合，その BUC にはいくつかの不足が存在する．この場合，タプルが取り出された$S_{n_}$を m-エリジブルを満たす前提で，$S_{n_}$から不足なところと同じ機密値のタプルを取り出して BUC に移動する．どうしても取り出せるタプルがない場合，準識別子を持たない偽造タプルを埋め込む．すべてのバケットがバランスを取るまで，フェーズ 2 を繰り返す．

実例では，$\mathbf{S_{2_}}$の中のタプルを取り出して，フェーズ 1 で作成した四つのバケットに埋め込む．$\mathbf{S_{2_}}$では，個数が最も多い機密値の割合が 2/5 で 2-エリジブルを満たす．2-エリジブルが壊れないうちに，機密値が flu のタプルを一つ取り出して BUC3 に埋め込む．最後に$\mathbf{S_{2_}}$に存在しない bro.のところに偽造タプルc_1を入れる．バランスを取ったバケットが図 2 に示す．

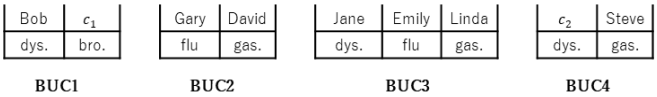

図 2 バランスを取ったバケット

[Phase 3] 次にフェーズ 3 では，フェーズ 2 でタプルがいくつ取り出された後の$S_{n_}$のタプルを引数として，繰り返し実行して，$S_{n_}$が空になるまでタプルをバケットに分配する．

まずは$S_{n_}$に残る機密値の種類数λと$S_{n_}$に残るタプルの総数$\gamma = |S_{n_}|$を計算し，すべての機密値を個数が多い順から$v_1, v_2, \ldots$のように並べ，それぞれの数$n_1, n_2, \ldots$を記録する．(v_1, v_2)そして m-エリジブルの$S_{n_}$に対し，バケットの幅βの値をmとし，$S_{n_}$が m-エリジブルを満たせる最大のバケットの高さαを探す．正式には以下の三つの式で表す．

$$\alpha \le n_\beta$$
$$n_1 - \alpha \le (\gamma - \alpha \cdot \beta)/m$$
$$n_{\beta+1} \le (\gamma - \alpha \cdot \beta)/m$$

そして$S_{n_}$からタプルを取り出し，シグネチャが$\{v_1, \cdots, v_\beta\}$のバケットに分配する．$\{v_1, \cdots, v_\beta\}$のバケッ

トが存在しない場合，新しいバケットを作成する．

実例では，S_2に残った{Mary, Ray, Tom, Vince}をバケット BUC2 と BUC3 に割り当てる．結果を図 3 に示す．

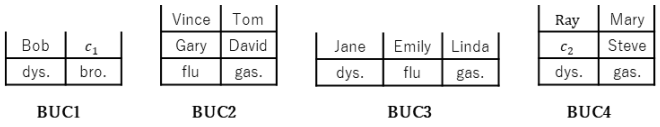

図 3 T*(1)と T(2)から作ったバケット

[Phase 4] 最後に，各バケット BUC を個別に処理する．一つのバケット BUC を|BUC|/α個の QI グループに分割する．一般化されたタプルより公開された T*(n)の QI グループ QI*を形成する．バケットの中のタプルから，各準識別子において最小の区間とするように QI*を作成する．

実例では，バケットからタプルを取り出して作成した公開表 T*(2)を表 9 に示す．

表 9 m-不変性を満たすように一般化した時刻 2 の公開表 T*(2)

G.ID	Age	Zip.	Disease
1	[20,21]	[11k,12k]	gastritis
1	[20,21]	[11k,12k]	bronchitis
2	[23,41]	[20k,25k]	flu
2	[23,41]	[20k,25k]	gastritis
3	[60,65]	[36k,44k]	flu
3	[60,65]	[36k,44k]	gastritis
4	[25,43]	[21k,33k]	dyspepsia
4	[25,43]	[21k,33k]	flu
4	[25,43]	[21k,33k]	gastritis
5	[45,46]	[29k,30k]	dyspepsia
5	[45,46]	[29k,30k]	gastritis
6	[54,65]	[31k,34k]	dyspepsia
6	[54,65]	[31k,34k]	gastritis

4. 提案手法

本節ではまず，文献[1]における安全性に関する問題点[2]を指摘する．そのため，文献[1]での特定リスクの定義をまず紹介する．

4.1 確率的リスク評価

[定義 11] (特定リスク)

タプル t∈T*に対して，公開表 T*から t の真の機密値を正しく特定する確率を特定リスクと言い，一般化手法の一つの評価基準である．

例えば，敵対者が表 6 の情報を把握していると考える．病院側から時刻 1 と 2 で表 3 と表 7 を公開した.タプル<Bob, 21, 12k, [1,2]>に対して，敵対者から表 3 と表 7 の QI グループ 1 に存在すると判断できるが，タプル<Bob, 21, 12k, [1,2]>の機密値が gastritis だと断言できない．そのため，タプル<Bob, 21, 12k, [1,2]>の特定リスクは 1/2 である．

4.2 問題点

文献[1]では，m-不変性という性質を定義し，m-不変性を満たす公開表の特定リスクが 1/m 以下になると示した．

ただし，この結論には敵対者が m-不変性を考えるという前提が不可欠である．すなわちすべてのタプルを m-不変性の「すべての公開表にはシグネチャが変わらない」という条件を前提とした上で特定しなければならない．

時刻 1 と 2 の元データを表 10 と表 11 にして，それらを m-不変性の一般化より一般化した公開表が表 12 と表 13 である．

表 10 時刻 1 の小さな元データ T(1)

Name	Age	Zip.	Disease
Bob	21	12k	dyspepsia
Alice	22	14k	bronchitis
Andy	24	18k	dyspepsia
David	23	25k	gastritis

表 11 時刻 2 の小さな元データ T(2)

Name	Age	Zip.	Disease
Bob	21	12k	dyspepsia
David	23	18k	gastritis
Emily	21	12k	dyspepsia

表 12 時刻 1 の 2-不変性を満たした小さな公開表 T*(1)

G.ID	Age	Zip.	Disease
1	[21,22]	[12k,14k]	dyspepsia
1	[21,22]	[12k,14k]	bronchitis
2	[23,24]	[18k,25k]	dyspepsia
2	[23,24]	[18k,25k]	gastritis

表 13 時刻 2 の 2-不変性を満たした小さな公開表 T*(2)

G.ID	Age	Zip.	Disease
1	[21,22]	[12k,14k]	dyspepsia
1	[21,22]	[12k,14k]	bronchitis
2	[21,23]	[12k,25k]	dyspepsia
2	[21,23]	[12k,25k]	gastritis

敵対者の視点から，タプル<Bob, 21, 12k, [1,2]>の機密値を特定したいとき，<21, 12k>より T*(1)の QI グループ 2 を排除でき，さらにタプル Bob のライフスパン[1,2]から，T*(1)の QI グループ 1 に存在しない gastritis を排除できる．残った三行から，dyspepsia の可能性が 2/3 で 1/2 を超えてしまい，bronchitis の可能性が 1/3 となってしまい，特定リスクが 1/m 以下に抑えられない問題が発生する．

この問題を起こす原因として，m-不変性の「すべての公開表で連続して存在するタプルのシグネチャが変わらない」という前提で従来手法では特定リスクの解析が行われ，シグネチャが異なるグループへの割当ての可能性の排除にある．この前提の上では，一般化した QI グループの準識別子属性の範囲が重なっても，どの公開表も同じシグネチャなので，特定可能な機密値の割合が変わらない．ただし敵対者が m-不変性を知らない前提で攻撃を行うと，準識別子属性の範囲が重なる上でシグネチャも重なることで特定リスクが 1/m 以上になってしまう可能性もある．つまり，２つのグループの間で，準識別子の範囲かシグネチャのどちらかが互いに素であれば，問題は解消できる．

4.3 問題に対する改良策

情報追加の公開方法

前で指摘したように，敵対者の視点からも，m-不変性，または「すべての公開表で連続して存在するタプルのシグネチャが変わらない」を考える前提で攻撃を行うと，特定リスクが文献[1]の予想通りに 1/m 以下に抑えることができる．

そのため，公開表とともに，m-不変性,または「すべての公開表で連続して存在するタプルのシグネチャが変わらない」の知識を追加情報として一緒に第三者に公開すれば，この問題を解消できる．ただしデメリットとして，情報不足の敵対者に便利な情報を送る可能性もある．本研究ではできるだけ最小限の情報提供となるように追加する情報の最小化を目的とする．そのため，提案手法では，QI グループの準識別子属性の範囲が重なるところだけに，追加情報を与える．

例として，小さな公開表 T*(1)と T*(2)を作成する時，QI グループの構成タプルが一つ時刻前の公開表に属する QI グループのグループ id を追加情報 Last として追加する．この方法で作成した公開表を表 14 と表 15 に示す．

表 14 追加情報が付いた時刻 1 の小さな公開表 T*(1)

G.ID	Last	Age	Zip.	Disease
1	-	[21,22]	[12k,14k]	dyspepsia
1	-	[21,22]	[12k,14k]	bronchitis
2	-	[23,24]	[18k,25k]	dyspepsia
2	-	[23,24]	[18k,25k]	gastritis

表 15 追加情報が付いた時刻 2 の小さな公開表 T*(2)

G.ID	Last	Age	Zip.	Disease
1	1	[21,22]	[12k,14k]	dyspepsia
1	1	[21,22]	[12k,14k]	bronchitis
2	-	[21,23]	[12k,25k]	dyspepsia
2	-	[21,23]	[12k,25k]	gastritis

アルゴリズム

問題を起こす根本的な原因は，準識別子属性の範囲が重なる上でシグネチャも重なることである．そのため，一般化を行うとき，シグネチャと準識別子属性の範囲を同時に重ならないようにするよう(元データ数が少ない時に完全に起こさないよう)に改良することで，問題の発生率を低下できると考える．

従来手法では，一般化に最小限の情報損失を求めるため，表 13 の QI グループ 2 のような偽造データを入れる QI グループを作成するとき，QI グループの準識別子範囲を実在するタプルの準識別子における最小限の曖昧化により決める．例えば，表 13 では，QI グループ 1 に含まれる実在タプルは<Bob, 21, 12000, dyspepsia>だけなので，最小限の情報損失に抑えるため，<21, 12000>を最小限に曖昧化した<[21, 22], [12000, 14000]>を QI グループ 1 の準識別子範囲に代入した．

ただし最小限の情報損失だけを求めるなら，4.2 節で紹介した問題を引き起こす可能性がある．提案手法では，問題を起こさないため，偽造タプルを持つ QI グループの準識別子範囲を決める 3.2 節で紹介した一般化手法のフェーズ 4 に対する改良を行う．

提案手法では，公開データの安全性を最も優先すべきだと考え，安全性を確保した上で情報損失をできるだけ小さくする．安全性を確保するには，準識別子の値を最小限でも歪めることがある．そのため提案手法では，一回目の公開前に各準識別子の最小範囲を決める．

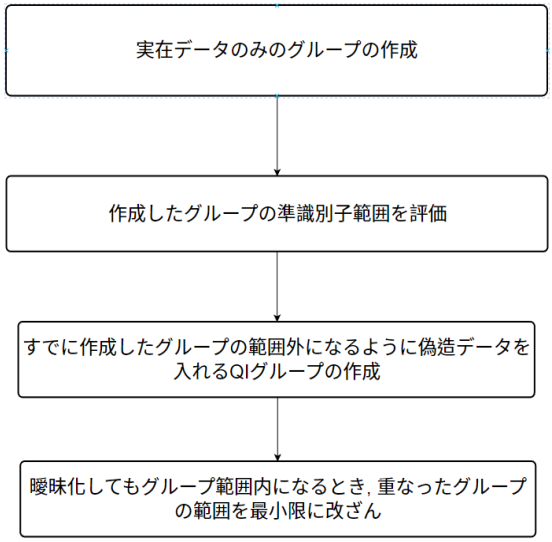

図 4 提案手法の枠組み

流れとしては，図 4 に示したように，提案手法はまず実在データのみのグループを作成する．偽造データを入れる QI グループを作成するとき，偽造データ以外の実在データの準識別子がすでに作成したほかの QI グループの準識別子の範囲内にあるかどうかを判断し，範囲内の場合に，再びシグネチャが重なったかどうかを判定する．両方が範囲内でない限り，従来手法でも安全に一般化できる．シグネチャを構成するのは一番機密度の高い機密値のため，両方が範囲内の場合では，準識別子だけを範囲外になるように以下の手順でグループを作成する．

実在データの準識別子がすでに作成したほかの QI グループの準識別子の範囲内にあるかどうかを正しく判断するために，準識別子範囲に対する評価基準が必要である．提案手法では，重なった範囲を<[x1, x2], [y1, y2]>に変形し，一般化された準識別子の最小区間を<d1, d2>にする．ここで，d1 は第 1 番目の座標のおける一般化における最小区間を，d2 は第 2 番目の座標における一般化での最小区間を示すとする．一番範囲に近い範囲外の点を<x1-d1, y1-d2>, <x1-d1, y2+d2>, <x2+d1 y1-d2>, <x2+d1, y2+d2>と定義する．

例として，図 5 が表 13 の 2 つのグループの準識別子範囲を二次元座標に変換した結果となる．

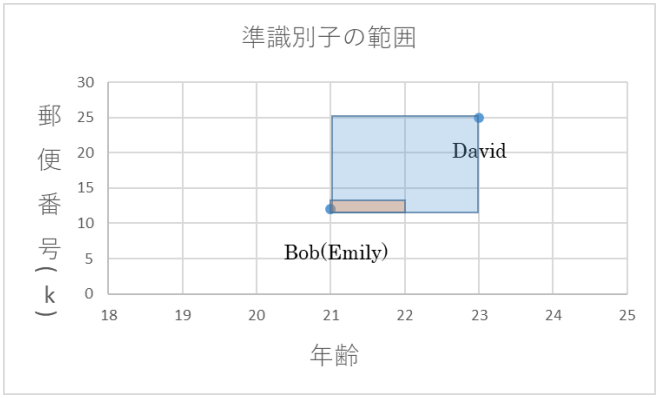

図 5 提案手法を適用前のグループの準識別子範囲

4 点(３つの準識別子の場合は 8 点)がほかの QI グループの範囲内にあるかどうかを判断し，ほかのシグネチャが重なった QI グループを検出し，それらの範囲内の点を排除し，全部排除した場合，このグループを安全に一般化することが困難と判断し，グループを削除する．

残った点とグループ内のすべての実在タプルの距離を計算し，一番近い点を候補点とする．候補点に最も近い範囲内の点と実在タプルの点を交換し，元のグループの準識別子範囲を最小限に改ざんし，新しいグループを作成する．

ただし，単に数値化と座標変換した準識別子でタプルの距離を計算するには，良い評価基準とならない．例えば，<a, 21, 12000>, <b, 20, 21000>, <c, 35, 14000>の三つのタプルから，a に最も近いデータを判定したい場合，準識別子を数値に変換し，二次元座標で考える場合，c の方にが a に最も近いが，準識別子 1 の年齢から見れば b の方が a に最も近い．

このような曖昧な評価を防ぎ，タプルの距離を正しく評価するために，提案手法では，一番目の準識別子(またはすべてのタプルで均等に分布している準識別子)を基準とし，ほかの準識別子を比例的に変形してから(例えば，順識別子年齢の範囲が[10, 90]，その差が 80 で，順識別子郵便番号の範囲が[10000, 90000]，その差が 80000 とすると，郵便番号の数値を 80/80000 にしてから距離の計算を行う)，多次元の数値としてタプル間の準識別子の近さを評価する．

例えば，表 5 を元に公開表を作成するとき，準識別子年齢の範囲が[21, 65]で差が 44 となり，郵便番号の範囲が[12000, 44000]で差が 32000 となる．後者が前者の約 727 倍のため，一般化のフェーズ 4 で情報損失を評価するとき，郵便番号の数値を 1/727 にしてから距離を計算する．

表 11 の T(2)に提案手法を適用すると，タプル Bob <21, 12000>が既に作成した QI グループ 1 の範囲内<[21, 22], [12000, 14000]>にある．準識別子の最小区間を<1, 2000>にすると，一番範囲に近い範囲外の点が<20, 10000>, <20, 27000>, <23, 10000>, <23, 27000>となる．実在タプル Bob <21, 12000>に一番近いほかのグループの範囲外の点が<20, 10000>となり，候補点と設定する．候補点<20, 10000>に最も近い範囲内の点<21, 12000>と実在タプル Bob <21, 12000>を交換し，さらに元のグループの一番目の準識別子を最小区間<1, 2000>により改ざんする．最後に公開データに改ざんが存在することを追加情報として公開する．提案手法を適用し一般化した公開表を表 16 に示す．

表 16 提案手法を適用した
時刻 2 の小さな公開表 T*(2)

G.ID	Age	Zip.	Disease
1	[21,23]	[14k,25k]	dyspepsia
1	[21,23]	[14k,25k]	gastritis
2	[20,21]	[10k,12k]	dyspepsia
2	[20,21]	[10k,12k]	bronchitis
追加情報			
データ分析に影響がない改ざんが グループ 1 に存在			

提案手法を適用した一般化を行うと，図 6 のように，準識別子属性の範囲を重ならないようにできる．ただしデメリットとして，データ数が多く，元データで重なりが多い場合には，情報損失が増加することが考えられる．

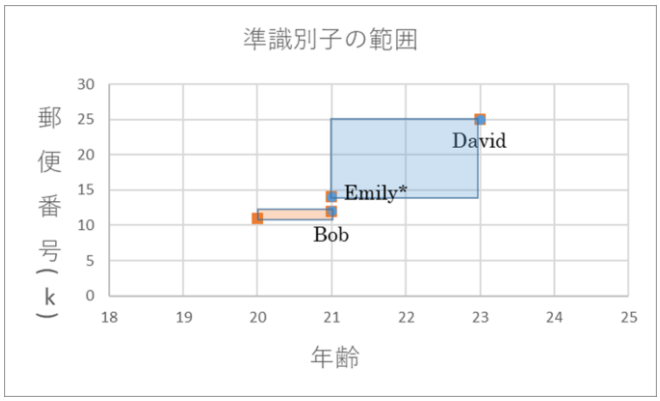

図 6 提案手法を適用したグループの準識別子範囲

5. まとめ

本稿では，m-不変性の一般化手法により作成された連続的な公開表の安全性を考察した．m-不変性の一般化手法に存在する問題に対する改良手法を提案した．

今後の課題として，特定リスクの計算を計算機上で実現し，一般化の途中で特定リスクを低下させるようにすることがある．実用性方面では，提案手法を実在タプルのみのグループにも適用できるように拡張することもある．データの有効性に対し，従来手法で挿入する偽造タプルの数を最少に抑えているが，提案手法ではタプルの削除とデータの改ざんも発生するので，タプルの削除数と挿入数による公開データの実用性の低下についての考察もある．さらに，安全性を確保するために必要な追加情報量の比較を行うことなどが挙げられる．

参 考 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] 坂田 奈々子，上土井 陽子，村上 頼太，若林 真一，“動的データセットにおけるプライバシ保護の厳密な安全性評価と妥当な安全性評価について”, DEIM Forum 2018, F7-5 (2018).

[3] 柚木 塁，上土井 陽子，若林 真一，“動的データテーブルの連続的匿名化の安全性について”, DEIM Forum 2014, E5-3 (2014).