ポリシーを使ったプライバシー保護要求の柔軟な指定

湯浅　佑介†　　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: †yuasa@lcs.hiroshima-cu.ac.jp, ††{yoko,wakaba}@hiroshima-cu.ac.jp

あらまし　データベースにおける個人のプライバシーとデータ解析の有用性のトレードオフを調整することはビッグデータアプリケーション分野で重要である．上記のトレードオフ調整に関する既存の研究では、データ所有者がデータ作成者に機密にしたい情報と解析者にあらかじめ知られている知識を詳細に指定する手段としてポリシーを用いる方法が提案された．ポリシーを用いる方法は他の一般化の基準をプライバシー保護要求として与える場合に比べ、保護したい個人情報を柔軟に指定することができると考えられている．しかし、従来ポリシーではあらかじめ知られている知識等を詳細に指定する、もしくは必要以上に機密にしたい情報を指定しなければならないプライバシー保護要求が存在する．本研究では従来のポリシーに要求の指定方法を追加することによって従来手法と比べ，より柔軟に強力なプライバシー保護要求を指定できることを示す．

キーワード　プライバシー保護要求, ポリシー

1. はじめに

個人レベルの情報を収集したデータ (データセット) は解析を行うことで高い価値の情報を生み出すことができるので様々な分野で応用されている．一方で，個人情報のプライバシーを保護しなければならない．データ解析におけるデータの有用性と個人のプライバシー保護はトレードオフの関係にある．文献 [1] の研究ではこの問題に対し，データの所有者がデータ作成者に機密にしたい情報と解析者にあらかじめ知られている知識の仮定を詳細に指定するポリシーと呼ばれる手法を用いて上記のトレードオフを調節し，元のデータを一般化する方法が提案された．個人のプライバシーを保護するためにデータを加工する手法として k-匿名化などがあるが，ポリシーを用いた方法では他の一般化の基準をプライバシー保護要求として与える場合に比べ, データのプライバシー保護の重要度に応じて，保護したい個人情報を柔軟に指定することができると考えられている．しかし，文献 [1] のポリシーを用いたプライバシーの保護要求では解析者にあらかじめある個人の情報が知られていた場合に，その情報を利用して他の個人情報を推測されることを防ぐことを要求として指定したいとき，必要以上にプライバシー保護を強めなければならないという問題があった．

本稿では文献 [1] のポリシーを拡張し，上記の要求に対しデータの有用性を考慮しつつ解析者にある個人の情報が知られていた場合でも，その情報によって他の個人情報が推測される可能性を最小限に抑えることができる指定方法を考察する．

2. 準　　備

本研究において考察の対象としているデータベースにおけるプライバシーの保護方法とデータの有用性を述べるにあたり，想定している状況，扱う用語について説明する [4].

2.1 想定する対象の状況

本研究では 1 のような状況を想定している．まず病院の入院記録やオンラインショップでの購入記録などの個人データを所有している者をデータ所有者と呼ぶ．データ所有者は個人データをデータ公開者に提供する際，プライバシーの保護を要求する．データ公開者はデータ所有者から収集したデータをこの保護要求に基づきデータを加工し解析者に公開する．本研究で考察するプライバシー保護要求とはデータ所有者からデータ公開者への個人データの保護方法の要求を表している．加工されたデータを活用する者をデータ解析者と呼ぶ．データ解析者はより元の個人データに近いデータを解析したいので，データ公開者はデータ所有者の要求に基づき公開表を作成する際，データの有用性を失わないために情報損失を少なくすることを考慮する必要がある．

2.2 個人データのプライバシー保護

データ公開者が収集した個人データは表 1 のようなものを想定し元のデータセット D と定義する．元のデータセット D は有限個のタプル (行に対応) と属性 (列に対応) からなり，個人を直接特定できる属性を識別属性，複数の属性を組み合わせることで個人を特定できる可能性のある属性を準識別属性，機密としたい属性を機密属性と呼ぶ．表 1 において識別属性は個人の名前，準識別属性は年齢，郵便番号，機密属性は病気である．また本研究の例において個人の名前を A_{id}, 病気を A_s とする．個人データのプライバシー保護とはデータ解析者に A_{id} と A_s が結び付けられることを防ぐことを意味する．このためにデータ公開者は元のデータセットを加工する必要がある．データセットの加工法としての一般化手法の代表的な例として k-匿名性を満たすように一般化する k-匿名化がある．

2.3 背景知識

解析者が個人を特定するため元のデータセットについてあらかじめ知っていると仮定する知識を背景知識と呼ぶ．背景知識

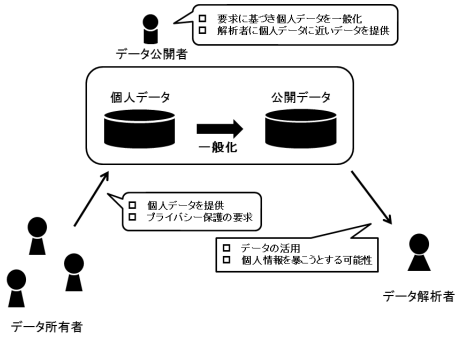

図 1　本研究で想定している状況

は元のデータセットから機密属性を削除したもので、元のデータセットの表 1 において解析者が知っている可能性のある背景知識は表 2 のように表現される．

また，2-匿名性を満たす表 3 では表 2 のような背景知識とタプルが一意に対応しないので，背景知識によるデータ推測は防止できているといえる．このように元のデータセットに一般化を行った表を公開表 T と呼ぶ．k-匿名化以外にもデータ所有者はプライバシー保護の方法を要求することができる．その表現方法としてポリシーを用いる方法を次節で説明する．

表 1　データセット D(例)

Name	Age	Zip.	Disease
Bob	21	12k	dyspepsia
Alice	22	14k	bronchitis
Andy	24	18k	flu
David	23	25k	gastritis
Gray	41	20k	flu
Helen	36	27k	gastritis

表 2　データ解析者の背景知識

Name	Age	Zip.
Bob	21	12k
Alice	22	14k
Andy	24	18k
David	23	25k
Gray	41	20k
Helen	36	27k

表 3　公 開 表 T

	Age	Zip.	Disease
t_1	[21,22]	[12k,14k]	dyspepsia
t_2	[21,22]	[12k,14k]	bronchitis
t_3	[23,24]	[18k,25k]	flu
t_4	[23,24]	[18k,25k]	gastritis
t_5	[36,41]	[20k,27k]	flu
t_6	[36,41]	[20k,27k]	gastritis

3. 従来ポリシー

本節では文献 [1]，[2]，[3] で提案されているポリシーを用いたプライバシー保護要求の指定方法について紹介する．

3.1 ポリシーを用いたプライバシー保護要求の指定

ポリシーは元のデータセット，機密グラフ，背景知識の 3 つから構成され，これらを指定することによりデータ公開者は公開表を作成することができる．従来ポリシーでは元のデータセット，背景知識は 2 節で述べた表 1，表 2 のようなものを想定している．また，機密グラフとはプライバシーを保護するために守るべき情報である機密情報について，何を機密情報とし，どのように保護するのかを指定するもので，その表現方法について次部分節で詳しく説明する．

3.2 機 密 情 報

ポリシーにおける機密情報の指定方法では (1) 個人データの情報で何を機密とするのか (2) 機密としたい情報をどのような方法で一般化するのかを表現する方法を定式化している．

3.2.1 機密情報の候補の指定

データ所有者が個人データに対し，何を機密情報としたいのかを指定する表現方法について述べる．機密情報を表現する方法として元のデータセット内の任意の命題記述を考え，そのうち機密としたい命題記述を要素とする集合を考えることで機密情報を指定する．まず，記述 s をデータセット内の任意の値の命題記述と定義し，特に記述 $s_x^i(t.id = i \wedge t = x)$ を個人 i に対応するタプルが x であるかどうかの命題記述とする．次に記述 s のうち，その真偽が機密情報の漏えいにつながる可能性のあるものを要素とする候補集合を S とする．

ここで，具体例を表 1 を用いて示す．ボブのタプルが t_1 であることを命題記述で表すと $s_{t_1}^{Bob}$ となる．$s_{t_1}^{Bob}$ は表 1 で真であり，解析者に真理値を知られないようにする必要があるため機密情報である．よって，$s_{t_1}^{Bob}$ を集合 S の要素とする．また，$s_{t_2}^{Bob}$ を考えたとき，表 1 で真理値は偽であるが解析者に Bob のタプルは t_2 でないと知られてしまうことはタプルの推測に利用されてしまう可能性があるので機密情報である．よって，$s_{t_1}^{Bob}$ も集合 S の要素とする．同様にして個人 6 人とタプル 6 個の組合せを S の要素とするので，少なくとも S は s_x^i で表現できる 36 個の要素を持つ．

3.2.2 機密情報の保護方法の指定

前部分節で説明した集合 S の要素である機密情報をどのように保護するのかを $S \times S$ の部分集合を用いて表現する [2]．これを区別不可能なペアの集合 S_{pairs} とする．$S \times S$ の要素 (s_i, s_j) のうち解析者が元のデータセット上では s_i が真であるのか s_j が真であるのかを区別できないようにすることを要求したいとき，その (s_i, s_j) を S_{pairs} の要素とする．例えば $(s_{t_1}^{Bob}, s_{t_2}^{Bob})$ を S_{pairs} の要素とするとき，これは Bob のタプルが t_1 であるのか t_2 であるのかを解析者がわからないように元のデータセットを一般化し公開表を作成してほしいという要求を示す．また，このとき (s_i, s_j) は $MutuallyExclusivepairs$ である必要がある．$MutuallyExclusivepairs$ とは機密情報の候補の二項組のうち同時に真とならない (同時に偽となる場合を含む) 二項組の集合のことである．例えば二項組 $(s_{t_1}^{Bob}, s_{t_2}^{Bob})$ は同時に真となることはないので $MutuallyExclusivepairs$ である．また二項組 $(s_{t_1}^{Bob}, s_{t_2}^{Alice})$ は同時に真となる可能性があるので $MutuallyExclusivepairs$ ではない．ここで「上記の解析者がわからないように」という表現を形式的に定義するため，区別不可能という用語を導入する．記述 s は一般化された公開表で解析者から見て、真理値がわかる場合, わからない場合がある．区別不可能なペアとは（s_i,s_j）の真理値が解析者から見てどちらもわからない，もしくはどちらも偽であるとわかるようなペアを表す．

区別不可能なペアの例を示す．公開表が表 3 が表 2 のように与えられたとする。

例 1 s の二項組 ($s^{Bob}_{t_1}$,$s^{Bob}_{t_2}$) は解析者から見るとタプル t_1，t_2 ともにボブの可能性があるがどちらかは見分けがつかないので区別不可能なペアである．
例 2 ($s^{Bob}_{t_1}$,$s^{Bob}_{t_3}$) はボブのタプルは t_1 であるかどうかはわからないがボブのタプルは t_3 でないことはわかる．よって二つの命題は区別可能なペアである．
例 3 ($s^{Bob}_{t_3}$,$s^{Bob}_{t_4}$) はボブのタプルはどちらでもないとわかる．しかし二つのタプルの見分けはつかない．よってこの場合も区別不可能なペアである．

文献 [1] では S_{pairs} の指定の例として以下の 3 つが紹介されている．それ以外にも S_{pairs} はデータ所有者の保護要求を自由に指定することができるが本研究では以下の 3 つの指定方法を元に議論する．

全領域

$$S^{full}_{pairs} = \{(s^i_x, s^i_y)|\forall i, \forall (x,y) \in T \times T\} \qquad (1)$$

(3.1) 式はすべての個人 i に対してその個人のタプルが x である命題と y である命題のペアが区別不可能であることの要求を表す集合である．
つまり，全領域の指定では解析者はすべての個人についてその個人がどのタプルなのか見分けがつかない．

属性

機密にしたいある属性を A とし，$x[A]$ をあるタプルの属性 A の値，$x[\bar{A}]$ をあるタプルの属性 A 以外のすべての値を示すとする．文献 [1] では以下の式を定義している．

$$S^{attr}_{pairs} = \{(s^i_x, s^i_y)|\forall i, \exists A, x[A], \neq y[A] \wedge x[\bar{A}] = y[\bar{A}]\} \quad (2)$$

S^{attr}_{pairs} の指定方法では秘密にしたい属性の値が異なりそれ以外の属性が同じのタプルが 1 つ以上存在するように元のデータセットを一般化する要求を示している．

分割

集合 $P = \{P_1, P_2, ..., P_p\}$ を考え元のデータセットを互いに素な p 個の領域に分割する．ここで，($\cup_i = T$ and $\forall i, j \leqq p, P_i \cap P_i = \emptyset$) が成り立っている．文献 [1] では以下の式を定義している．

$$S^{P}_{pairs} = \{(s^i_x, s^i_y)|\forall i, \exists j, (x,y) \in P_j \times P_j\} \qquad (3)$$

S^{P}_{pairs} の指定方法では，分割した p 個の領域内でタプルがそれぞれ区別不可能となるように元のデータセットを一般化する要求を示している．

3.2.3 要求のグラフ化

S_{pairs} をグラフを使って指定する方法が文献 [1] では提案されている．無向グラフ $G = (V, E)$ を考え，$V = T$, $E \subseteqq T \times T$ とする．グラフの頂点 V はタプルの集合，枝集合 E は解析者が区別不可能なタプルのペア集合を表す．このとき，S_{pairs} を $S^{G}_{pairs} = \{(s^i_x, s^i_y)|\forall i, \forall (x,y) \in E\}$ で表す．

ここで，それぞれのプライバシー保護要求のグラフ化の例を示す．

全領域

S^{full}_{pairs} の保護要求をグラフ化した例を図 2 に示す．

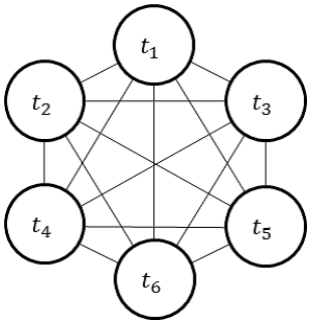

図 2 S^{full}_{pairs} のグラフ化

分割

S^{P}_{pairs} の保護要求を満たす例を以下に示す．
S^{P}_{pairs} の指定方法では，分割した p 個の領域内でタプルがそれぞれ区別不可能となるように元のデータセットを一般化する要求を示している．
要求をグラフ化した例を図 3 に示す．

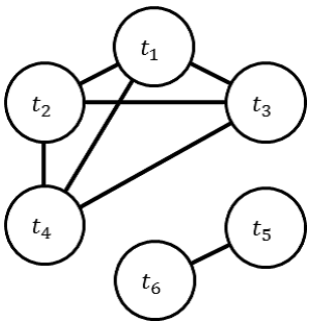

図 3 S^{P}_{pairs} のグラフ化

4. 従来ポリシーの問題点と新しい保護要求

本節ではデータ解析者の背景知識の仮定を強化した場合について考察する．文献 [1] の従来ポリシーではデータ解析者の背景知識は表 2 を仮定していた．しかし，実際は表 2 に加え「Bob の病気は胃炎ではない」などの軽微な背景知識は知られている可能性がある．このような軽微な背景知識の変化が起こったとき，従来ポリシー用いて一般化された公開表の解析に与える影響について考察し，より影響を少なくできるような新しい保護要求を提案する．

4.1 元データの再構築

追加の背景知識が公開表の解析に与える影響を調べるため解析者が一般化された公開表と背景知識から元データを再構築する方法について考察する．再構築とは解析者が背景知識と公開表のタプルを組み合わせ元のタプルを推測することである．

解析者が推測できるデータセットについて例を示す．
公開表に対応する表 3 が作成され，解析者に与えられたとす

る．また，背景知識に対応する表が表 2 だと仮定すると，解析者は公開表と背景知識の対応を考えることで元データを推測することができる．

このとき，解析者が可能な推測の数について考えると，8 通りの推測が可能である．

推測の数を行列を用いて表現する．個人を行，タプルを列とし i 行 j 列は個人 i がタプル j である推測の数を表す．また行を列の数字のそれぞれの和は可能な推測の数となっている．

公開表に対応する表 3 を背景知識表 2 を用いた推測の行列表現は表 4 のようになる．

また個人 i がタプル j である推測の数をすべての推測の数で割ったものを解析者が推測した場合の個人 i がタプル j である確率とする．表 4 を確率に変換した表は表 5 のようになる．

4.2 強化された背景知識と従来ポリシー

解析者に追加の背景知識があるとき，それが従来ポリシーで一般化された公開表の解析にどのような影響を及ぼすかについて例を用いて考察する．

従来ポリシーの 1 つである S^P_{pairs} を満たす要求が図 4 が示すグラフで与えられ，その要求を満たすように元のデータセット表 6 を一般化した公開表が表 8 だとし，背景知識は表 7 であるとする．このとき 4.1 節で示した行列表現を用いて元データ再構築したものを表 9 で示し，それを確率に変換した表を表 10 に示す．以降ではある個人とタプルを結びつける確率が 1/2 以下であるとき，機密情報は保護できているとする．表 4.3，表 4.4 からわかるとおり，解析者がある個人とタプルを正しく結び付けられる確率は高くとも 1/2 となっているので機密情報は保護できているといえる．

ここで，表 7 の背景知識に加え，解析者が David の病気は gastritis ではないということを知っていたとし，それを考慮した場合の元のデータセットの再構築について考える．背景知識の追加により推測可能な数が 40 通りから 36 通りに減少する．またこのときの推測の行列表現を確率に変換した表を表 11 に示す．

表 11 を見ると Helen のタプルが t_4 である確率が 5/9 となり，背景知識の追加により Helen の情報が知られてしまう確率が高まっていることがわかるので，背景知識が追加された場合，機密情報が保護されているとはいえない．

4.3 新しい保護要求と問題点

4.2 節で示したような危険性を解消また軽減できるような新しい保護要求の指定方法について考える．つまり，ある個人とタプルに関する軽微な背景知識があっても他のタプルが推測される確率がある値 (例えば 1/2) 以下であることが保障できるようにする要求を考える．

従来ポリシーの保護要求の指定は機密グラフによって表されるので，すべての個人に対して特定のタプルが区別不可能となるような要求しか表現することができない．

よって，ある個人とタプルに関する軽微な背景知識があった場合でも，他のタプルが推測される確率が 1/2 以下であることが保障できるような保護要求の指定は，すべてのタプルが区別不可能となる S^{full}_{pairs} であれば表現することができる．

しかし，S^{full}_{pairs} の指定では全てのタプルが同じ値の組み合わせをもつよう一般化され，データの有用性を失ってしまう問題点がある．そこで S^{full}_{pairs} の指定以外の方法で 4.2 節で示した危険性を解消できる保護要求について次節にて考察する．

表 4　元データ再構築における可能な推測の数の行列表現

	t_1	t_2	t_3	t_4	t_5	t_6
Bob	4	4	0	0	0	0
Alice	4	4	0	0	0	0
Andy	0	0	4	4	0	0
David	0	0	4	4	0	0
Gary	0	0	0	0	4	4
Helen	0	0	0	0	4	4

表 5　表 4 の確率表現

	t_1	t_2	t_3	t_4	t_5	t_6
Bob	1/2	1/2	0	0	0	0
Alice	1/2	1/2	0	0	0	0
Andy	0	0	1/2	1/2	0	0
David	0	0	1/2	1/2	0	0
Gary	0	0	0	0	1/2	1/2
Helen	0	0	0	0	1/2	1/2

表 6　元のデータセット

Name	Age	zip.	Disease
Bob	18	12k	dyspepsia
Alice	19	14k	bronchitis
Andy	21	18k	flu
David	23	20k	gastritis
Gary	25	25k	flu
Helen	30	27k	dyspepsia

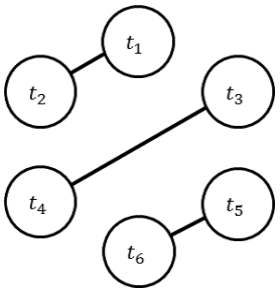

図 4　S^P_{pairs} の要求を満たすグラフ

5. 提案ポリシー

本節では第 3 節で紹介した従来ポリシーを拡張した提案ポリシーを提案する．従来ポリシーの機密情報の保護方法の指定では区別不可能なタプルを指定することで要求を表現していたが，第 4 節で述べたような背景知識の追加があった場合にもプライバシー保護できるように指定する要求を考察するため，提案ポリシーではタプルの指定に加え区別不可能な個人の組について

表 7　データ解析者の背景知識

Name	Age	zip.
Bob	18	12k
Alice	19	14k
Andy	21	18k
David	23	20k
Gary	25	25k
Helen	30	27k

表 8　図 4 を満たすように一般化した公開表

	Age	Zip.	Disease
t_1	[18,23]	[12k,20k]	dyspepsia
t_2	[18,23]	[12k,20k]	bronchitis
t_3	[19,30]	[14k,27k]	flu
t_4	[19,30]	[14k,27k]	dyspepsia
t_5	[21,25]	[21k,25k]	flu
t_6	[21,25]	[21k,25k]	gastritis

表 9　元データ再構築

	t_1	t_2	t_3	t_4	t_5	t_6
Bob	20	20	0	0	0	0
Alice	12	12	8	8	0	0
Andy	4	4	4	4	12	12
David	4	4	4	4	12	12
Gary	0	0	4	4	16	16
Helen	0	0	20	20	0	0

表 10　表 9 の確率表現

	t_1	t_2	t_3	t_4	t_5	t_6
Bob	1/2	1/2	0	0	0	0
Alice	3/10	3/10	1/5	1/5	0	0
Andy	1/10	1/10	1/10	1/10	3/10	3/10
David	1/10	1/10	1/10	1/10	3/10	3/10
Gary	0	0	1/10	1/10	2/5	2/5
Helen	0	0	1/2	1/2	0	0

表 11　表 8 において「David のタプルは t_4 ではない」という背景知識を追加した場合の確率表現

	t_1	t_2	t_3	t_4	t_5	t_6
Bob	1/2	1/2	0	0	0	0
Alice	5/18	5/18	2/9	2/9	0	0
Andy	1/9	1/9	1/9	1/9	5/18	5/18
David	0	0	1/9	0	1/3	1/3
Gary	0	0	1/9	1/9	7/18	7/18
Helen	0	0	5/9	4/9	0	0

も注目する．

5.1　機密情報の保護方法の指定

提案ポリシーも従来ポリシーと同様に機密情報を s_x^i を要素とする S で表し，S_{pairs} を用いて機密情報の保護方法を指定する．従来ポリシーでは $(s_{t_1}^{Bob}, s_{t_2}^{Bob})$ のような同じ個人に対する区別不可能なペアについてのみ考察していた．しかし S_{pairs} の要素には $(s_{t_1}^{Bob}, s_{t_1}^{Alice})$ のような同じタプルに対する区別不可能なペアも存在する．提案ポリシーではこのようなペアに注目し，従来ポリシーにこれを加えた保護要求の指定方法を考察する．

第 3 節で述べた S_{pairs} の指定の例を拡張した式を以下に示す．

全領域

$$S'^{full}_{pairs} = \{(s_x^i, s_y^i) | i \in ID \wedge (x,y) \in T \times T\} \cup \{(s_x^i, s_x^j) | (i,j) \in ID \times ID \wedge x \in T\} \quad (4)$$

提案ポリシーの全領域指定では，従来ポリシーのすべての個人に対してすべてのタプルの組が区別不可能である要求の指定に加え，すべてのタプルに対してすべての個人の組が区別不可能である要求の指定を追加している．

属性

$$[S'^{attr}_{pairs} = \{(s_x^i, s_y^i) | (i \in ID) \wedge (x,y \in T) \wedge x[A_s] \neq y[A_s] \wedge x[A_s] = y[A_s]\} \cup \{(s_x^i, s_y^i) | (i,j \in ID) \wedge (x \in T) \wedge x[A_s] \neq y[A_s] \wedge x[A_s] = y[A_s]]\} \quad (5)$$

この式では一般化後のデータセットで秘密にしたい属性の値が異なり，それ以外の属性が同じタプルが k 個以上存在する要求を指定している．また，k 人が同じタプルに入っているタプルが 1 つでもあれば，その k 人はすべてのタプルについて区別不可能であることを要求として追加している．

分割

集合 $P = \{P_1, P_2, ..., P_p\}$ を考え, 元のデータセットのタプルを互いに素な p 個の領域に分割する．また，集合 $Q = \{Q_1, Q_2, ..., Q_q\}$ を考え, 元のデータセットの個人について互いに素な q 個の領域に分割し以下の式を定義する．

$$S'^{P}_{pairs} = (\bigcup_{j \in \{1,...,p\}} \{(s_x^i, s_y^i) | i \in ID \wedge (x,y) \in P_j \times P_j\}) \cup (\bigcup_{y \in \{1,...,p\}} \{(s_x^i, s_x^j) | x \in T \wedge (i,j) \in Q_y \times Q_y\}) \quad (6)$$

提案ポリシーの分割指定では, 分割した p 個の領域内でタプルがそれぞれ区別不可能となることを指定する要求に加え，分割した q 個の領域内で個人がそれぞれ区別不可能となることを指定する要求を追加している．

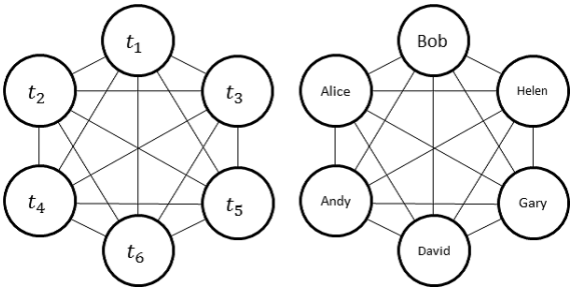

図 5　$S^{'full}_{pairs}$ のグラフ化

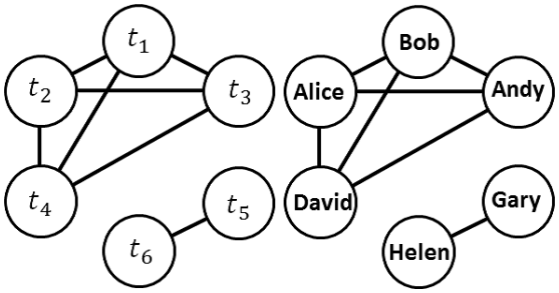

図 6　$S^{'P}_{pairs}$ のグラフ化

5.2　要求のグラフ化

提案ポリシーのグラフ化では要求を 2 つのグラフのペア (G_t, G_{id}) で表す．このとき $G_t = (V_t, E_t)$,$V_t = T$, $E_t \subseteqq T \times T$, $G_{id} = (V_{id}, E_{id})$,$V_{id} = ID$, $E_{id} \subseteqq ID \times ID$ である．また枝集合 E_s, E_v は以下の条件を満たす．

$$\forall i \in ID[(s^i_x, s^i_y) \in S_{pairs}] \Leftrightarrow (x, y) \in E_t \qquad (7)$$

$$\forall x \in T[(s^i_x, s^i_y) \in S_{pairs}] \Leftrightarrow (i, j) \in E_{id} \qquad (8)$$

ここで，提案ポリシーを用いたプライバシー保護要求の例を示す．

全領域

$S^{'full}_{pairs}$ の保護要求ををグラフ化すると図 5 のようになる．

この要求を満たすように表 1 を一般化した例を表 12 に示す．

分割

$S^{'P}_{pairs}$ の保護要求を満たす例を以下に示す．

提案ポリシーの分割指定では, 分割した p 個の領域内でタプルがそれぞれ区別不可能となるような要求に加え，分割した q 個の領域内で個人がそれぞれ区別不可能となるような要求なので,$P = \{P_1, P_2\}, P_1 = \{t_1, t_2, t_3, t_4\}, P_2 = \{t_5, t_6\}$,$Q = \{Q_1, Q_2\}, Q_1 = \{Bob, Alice, Andy, Gary\}, P_2 = \{David, Helen\}$ が与えられたとするとそれを満たす例は以下のようになる．

$$S^{'P}_{pairs} = \{(s^{Bob}_{t_1}, s^{Bob}_{t_2}), (s^{Alice}_{t_1}, s^{Alice}_{t_2}), ..., (s^{Helen}_{t_1}, s^{Helen}_{t_2}),$$
$$(s^{Bob}_{t_1}, s^{Bob}_{t_3}), (s^{Alice}_{t_1}, s^{Alice}_{t_3}), ..., (s^{Helen}_{t_1}, s^{Helen}_{t_3}), ...,$$
$$(s^{Bob}_{t_3}, s^{Bob}_{t_4}), (s^{Alice}_{t_3}, s^{Alice}_{t_4}), ..., (s^{Helen}_{t_3}, s^{Helen}_{t_4}),$$
$$(s^{Bob}_{t_5}, s^{Bob}_{t_6}), (s^{Alice}_{t_5}, s^{Alice}_{t_6}), ..., (s^{Helen}_{t_5}, s^{Helen}_{t_6}),$$
$$(s^{Bob}_{t_1}, s^{Alice}_{t_1}), (s^{Bob}_{t_2}, s^{Alice}_{t_2}), ..., (s^{Bob}_{t_6}, s^{Alice}_{t_6}),$$
$$(s^{Bob}_{t_1}, s^{Andy}_{t_1}), (s^{Bob}_{t_2}, s^{Andy}_{t_2}), ..., (s^{Bob}_{t_6}, s^{Andy}_{t_6}), ...,$$
$$(s^{Andy}_{t_1}, s^{David}_{t_1}), (s^{Andy}_{t_2}, s^{David}_{t_2}), ..., , (s^{Andy}_{t_6}, s^{David}_{t_6}),$$
$$(s^{Gary}_{t_1}, s^{Helen}_{t_1}), (s^{Gary}_{t_2}, s^{Helen}_{t_2}), ..., (s^{Gary}_{t_6}, s^{Helen}_{t_6})\}$$

また, この要求をグラフ化すると図 6 のようになる．

この要求を満たすように表 1 を一般化した例を表 13 に示す．

表 12　$S^{'full}_{pairs}$ を満たす公開表

	Age	Zip.	Disease
t_1	[21,41]	[12k,27k]	dyspepsia
t_2	[21,41]	[12k,27k]	bronchitis
t_3	[21,41]	[12k,27k]	flu
t_4	[21,41]	[12k,27k]	gastritis
t_5	[21,41]	[12k,27k]	flu
t_6	[21,41]	[12k,27k]	gastritis

表 13　$S^{'P}_{pairs}$ を満たす公開表

	Age	Zip.	Disease
t_1	[21,24]	[12k,25k]	dyspepsia
t_2	[21,24]	[12k,25k]	bronchitis
t_3	[21,24]	[12k,25k]	flu
t_4	[21,24]	[12k,25k]	gastritis
t_5	[36,41]	[20k,27k]	flu
t_6	[36,41]	[20k,27k]	gastritis

6.　提案ポリシーの有効性

本節では第 4 節で示した従来ポリシーの要求を満たすように一般化した公開表において，解析者に追加の背景知識があると，機密情報が保護できなくなる場合が存在した問題に対して，第 5 節で示した提案ポリシーを用いることでその問題がどのように解決できるかについて考察する．

6.1　提案ポリシーを満たす公開表の性質

4.2 節では従来ポリシーである分割指定の S^P_{pairs} を満たす要求を満たすよう表 6 を一般化したが，表 7 に加えて「David のタプルは t_4 ではない」という背景知識が解析者に知られていると考慮した場合，Helen のタプルが t_4 であると知られてしまう確率が高くなり，十分に機密情報が保護できているとはいえなかった．一方で提案ポリシーの分割指定の $S^{'P}_{pairs}$ の要求を満たすグラフを図 7 に示し，それを満たすよう表 6 を一般化した公開表を表 14 に示す．このとき，解析者が背景知識と公開表を組み合わせることで推測可能な個人とタプルの組み合わせを行列で表現したものを表 15 に示し，また表 6.2 を確率に変換した表を表 16 に示す．表 15，表 16 からわかるようにどの確率も 1/2 以下に抑えられており，提案ポリシー保護要求を満た

す公開表も機密情報が保護されているといえる．次に，背景知識が追加された場合について議論していく．

4.2 節で考察した表 7 の背景知識に加え，解析者が David の病気が gastritris ないと知っていた場合，推測の組合せの数は表 17 のようになり確率に変換した表を表 18 に示す．表 17，表 18 を見るとボブのタプルが t_4 である確率が 0 だとしても，他の個人の情報が知られてしまう確率は最大で 10/33 と低い値で抑えることができている．よって，従来ポリシーの要求を満たす公開表では Helen のタプルが t_4 である確率が 5/9 と高くなっていたが，提案ポリシーではそれが改善されたといえる．

次に，背景知識をさらに追加した場合を考える．解析者に背景知識について Bob の病気は dyspepsia，Alice の病気は bronchitis，Andy の病気は flu，David の病気は gastrirtis であると知られていたとする．このとき，推測の数を行列で表現したものを表 19，確率に変換したものを表 20 に示す．ここで，要求のグラフ図 7，表 19, 表 20 に注目すると，4 人の情報が知られていてもその 4 人と枝で結ばれていない Gary と Helen については情報が知られてしまう確率を 1/2 以下に維持することが出来ている．

上記の例による考察の結果から，提案ポリシーにおける要求のグラフとそれを満たす公開表について追加の背景知識を考慮した場合，以下の性質が成り立つといえる．

[性質]　グラフ G_{id} において個人 i を含む完全部分グラフ K_i が存在するとき，部分グラフ K_i に属する節点集合を V_i とすると，個人 i とあるタプルが結び付けられる確率は V_i に属する個人の情報が知られていない限り $1/|V_i|$ 以下に維持できる．

6.2 背景知識追加耐性を表現する場合の従来ポリシーとの比較

4 節で述べた，解析者にある個人について追加の背景知識が知られている場合に他の個人情報が推測されることを防ぎたいという要求に対して，提案ポリシーでは 6.1 節で述べたように個人 i とあるタプルが結び付けられる可能性は V_i に属す個人の情報が知られない限り $1/|V_i|$ に防ぐことができる．一方で，従来ポリシーのタプルに関するグラフだけで上記で示した提案ポリシーの要求を表現しようとすると，ある個人について情報が知られたときに，その情報によって影響を受ける個人が特定できないし，また，G_t が複数の連結成分をもっている場合には，少なくとも各成分に属するタプルに対応する 1 個人の情報は保護できない可能性がある．よって，全領域指定にせざるを得ない．また，全領域指定はデータの有用性を大きく失う指定方法なので，提案ポリシーの要求の方が有用性を維持できているのは明らかである．

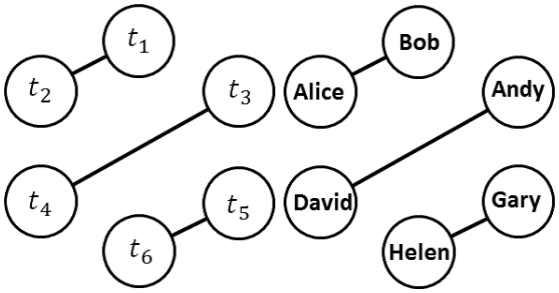

図 7　S'^{P}_{pairs} のグラフ化

表 14　S'^{P}_{pairs} を満たす公開表

	Age	Zip.	Disease
t_1	[18,19]	[12k,14k]	dyspepsia
t_2	[18,19]	[12k,14k]	bronchitis
t_3	[21,23]	[18k,20k]	flu
t_4	[21,23]	[18k,20k]	gastritis
t_5	[25,30]	[25k,27k]	flu
t_6	[25,30]	[25k,27k]	gastritis

表 15　表 14 の推測数の行列表現

	t_1	t_2	t_3	t_4	t_5	t_6
Bob	48	48	28	28	0	0
Alice	48	48	28	28	0	0
Andy	28	28	20	20	28	28
David	28	28	20	20	28	28
Gary	0	0	28	28	48	48
Helen	0	0	28	28	48	48

表 16　表 15 の確率表現

	t_1	t_2	t_3	t_4	t_5	t_6
Bob	6/19	6/19	7/38	7/38	0	0
Alice	6/19	6/19	7/38	7/38	0	0
Andy	7/38	7/38	5/38	5/38	7/38	7/38
David	7/38	7/38	5/38	5/38	7/38	7/38
Gary	0	0	7/38	7/38	6/19	6/19
Helen	0	0	7/38	7/38	6/19	6/19

表 17　表 14 において「David のタプルは t_4 ではない」という背景知識を追加した場合の推測数の行列表現

	t_1	t_2	t_3	t_4	t_5	t_6
Bob	40	40	24	28	0	0
Alice	40	40	24	28	0	0
Andy	24	24	16	20	24	24
David	28	28	20	0	28	28
Gary	0	0	24	28	40	40
Helen	0	0	24	28	40	40

7. まとめと今後の課題

本稿では，個人データを収集したデータセットを解析者に

表 18　表 17 の確率表現

	t_1	t_2	t_3	t_4	t_5	t_6
Bob	10/33	10/33	2/11	7/33	0	0
Alice	10/33	10/33	2/11	7/33	0	0
Andy	2/11	2/11	4/33	5/33	2/11	2/11
David	7/24	7/24	5/33	0	7/33	7/33
Gary	0	0	2/11	7/33	10/33	10/33
Helen	0	0	2/11	7/33	10/33	10/33

表 19　表 14 においてさらに背景知識を追加した場合の推測数の行列表現

	t_1	t_2	t_3	t_4	t_5	t_6
Bob	2	0	0	0	0	0
Alice	0	2	0	0	0	0
Andy	0	0	2	0	0	0
David	0	0	0	2	0	0
Gary	0	0	0	0	1	1
Helen	0	0	0	0	1	1

表 20　表 19 の確率表現

	t_1	t_2	t_3	t_4	t_5	t_6
Bob	1	0	0	0	0	0
Alice	0	1	0	0	0	0
Andy	0	0	1	0	0	0
David	0	0	0	1	0	0
Gary	0	0	0	0	1/2	1/2
Helen	0	0	0	0	1/2	1/2

公開する際の，データセットを一般化する手法について解析者にある個人の情報が知られていると仮定した場合でも，その情報が他の個人情報の推測に与える影響を最小限に抑える一般化基準の新しい指定方法を提案した．文献 [1] で提案された従来手法でも一般化基準を満たすことはできるが，提案手法の方がデータの有用性をより維持できることを示した．

今後の課題として，提案した一般化基準の要求を満たし，かつデータの情報損失が最小限となるようにデータセットを一般化し，公開表を作成するアルゴリズムを開発することがある．

文　献

[1] Xi He, Ashwin Machanavajjhala and Bolin Ding, “Blowfish privacy: tuning privacy-utility trade-offs using policies,” SIGMOD’14, pp.1447-1548, 2014.

[2] D. Kifer and A. Machanavajjhala, “A rigorous and customizable framework for privacy,” PODS’12, pp.77-88, 2012.

[3] D. Kifer and A. Machanavajjhala,“Pufferfish: A framework for mathematical privacy definitions,” ACM Transactions on Database Systems, Vol.39, Issue 1, 2014.

[4] 村本俊祐, 上土井陽子, 若林真一, “ データを極小歪曲し k-匿名性を保持したデータに変換するプライバシー保護アルゴリズム,” 日本データベース学会 Letters, Vol.6, No.1, pp.97-100, 2007.