外部知識の影響を考慮した匿名化データベースの安全性の分析

鈴木諒子† 加藤遼† 西脇雄高† 越前功‡ 吉浦裕†

†電気通信大学 〒182-8585 東京都調布市調布ヶ丘 1-5-1

‡国立情報学研究所 〒101-8430 東京都千代田区一ツ橋 2-1-2

あらまし 近年，データベースからの個人情報の漏洩とともに，データベースの二次利用による情報漏洩が懸念されている．この問題を解決する代表的な手法は，データベースの匿名化である．しかし，匿名化された情報に，統計情報等の外部からの知識が加わると匿名性が破られ，個人情報が漏洩する可能性がある．本論文では，外部知識のもたらす危険性を明らかにするために，代表的な匿名化手法であるk-匿名性を取り上げ，保護対象者のSensitive情報についてk-匿名データベースと外部知識の双方から何が推定できるかを検討した．外部知識のモデルとして，保護対象者の属性とデータベースの属性の間の依存関係を取り上げた．このモデルに基づき，保護対象者の属性値とk-匿名データベースのレコード群が与えられたときに，外部知識によるSensitive属性の値のエントロピーの低下を，外部知識の影響とした．本手法は，保護対象者が複数の場合およびℓ-多様化，t-近傍化したデータベースにも適応できる． また，手法を実装し，外部知識の変化によるデータベースへの影響の変化を分析した．

On the Security of Anonymised databases

Ryoko Suzuki† Ryo Kato† Yutaka Nishiwaki† Isao Echizen‡ Hiroshi Yoshiura†

†The University of Electro-Communications
1-5-1 Chofugaoka, Chofu-city, Tokyo 182-8585, JAPAN

‡National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ward, Tokyo 101-8430, JAPAN

Abstract Personal information of many people is accumulated in databases. This information is not only used directly but also is put into secondary use after some anonymisation. This paper therefore analyzes the security of representative anonymisation methods of k-anonymity, l-diversity, and t-closeness.

1 はじめに

1990 年代の後半以降，企業や医療機関等が多数のユーザから個人情報を集めてデータベース化し，データマイニング等の分析を行うようになってきた．それに伴い，不正侵入による大規模な個人情報の漏洩や，内部不正による個人情報の持ち出しが問題になっている．

また，データベース化された個人情報は，集積した機関が利用するだけでなく，他の機関への貸与等を通じて二次利用される．2007 年に改正された統計法の第三十四条により，公共

の調査データを学術利用することも可能になっている. これらの二次利用において, 貸与先の機関からの情報漏洩が懸念される.

以上の問題を解決する代表的な方法は, データベース内の個人情報を匿名化することである. 統計法の第三十五条および第三十六条は, 匿名データの作成および提供についても定めており, この規定に基づいた調査データの提供が始まっている. ここでの匿名化とは, 情報の除去と抽象化である. 例えば, 氏名情報を除去し, 生年月日を年代(1970 年代 等)に抽象化することが挙げられる. 匿名化の手法としては, *k*-匿名性, ℓ-多様性, *t*-近傍性などが提案されている[1][2][3]. しかし, 匿名化データベースから得られる情報に統計情報等の外部の知識が加わると匿名性が破られ, 個人情報が漏洩してしまうことが問題となっている. この外部知識のもたらす危険性については, 著者らの知る限りでは検討が為されていない.

そこで本論文では, 外部知識による匿名化データベースへの影響を, エントロピーを用いて定量化する手法を提案する.

2 先行研究

2.1 データベースの匿名化

ある日本人 Hanako が郵便番号 13068 の地域に住んでおり, 36 歳であることを攻撃者が知っているとする. このとき, Hanako が含まれているデータベースから表 1 のようなデータテーブルが漏洩した場合, 攻撃者は Hanako の病気が癌であると特定できてしまう.

このような情報漏洩を防ぐために, *k*-匿名性, ℓ-多様性, *t*-近傍性といったデータベースの匿名化手法が提案されている.

表 2 のように匿名化されたデータテーブルにおいて, 郵便番号や年齢のように, 組み合わせることで個人の特定に繋がる属性の集合を QI(Quasi-Identifier), 病気のようにデータ解析者にとって重要な項目で匿名化を行いたくない属性を Sensitive データ, QI の等しいレコードの集合を QI グループと呼ぶことにする.

表 1 匿名化前のデータテーブル

郵便番号	年齢	国籍	病気
13053	28	ロシア	心臓病
13068	21	日本	ウイルス感染
13053	23	アメリカ	ウイルス感染
14853	50	インド	癌
14853	55	ロシア	心臓病
14850	49	アメリカ	ウイルス感染
13053	37	インド	癌
13068	36	日本	癌
13068	35	アメリカ	癌

2.1.1 *k*-匿名性

k-匿名性は, データテーブル内の全ての QI グループにおいて, レコードが k 個以上存在すること保証する手法である. 具体的には, 郵便番号の下 2 桁を隠したり, 年齢を抽象化(36 歳を 30 歳代)したりすることで匿名性を実現する.

表 2 では, 全ての QI グループ内にレコードが 3 つずつ存在しているおり, 3-匿名性を満たしている.

表 2 表 1 に *k*-匿名性を実装した例($k = 3$)

郵便番号	年齢	病気	
130**	<30	心臓病	QI グループ(1)
130**	<30	ウイルス感染	
130**	<30	ウイルス感染	
1485*	≥40	癌	QI グループ(2)
1485*:	≥40	心臓病	
1485*	≥40	ウイルス感染	
130**	3*	癌	QI グループ(3)
130**	3*	癌	
130**	3*	癌	
QI		Sensitive データ	

しかし, *k*-匿名性には以下のような攻撃が存在する[2].

・ 同種攻撃:

表 2 の QI グループ(3)のレコードは全て病気が癌のため, グループに属することが特定されると病気が癌であると特定されてしまう.

・ 背景知識攻撃:

ある日本人 Hanako が表 2 の QI グループ(1)に属することが分かったとする. このとき, 「日本人は心臓病の発病率が非常に低い」という知識が加わると, 病気の候補から心臓病が外れ,

Hanako の病気がウイルス感染であると推定されてしまう.

2.1.2 ℓ-多様性

k-匿名性に対する同種攻撃を防ぐために，ℓ-多様性という手法が考えられた．ℓ-多様性は，全ての QI グループ内に Sensitive データが少なくとも ℓ 種類存在することを保証する手法である.

ℓ-多様性の実装方式として，エントロピー ℓ-多様性がある．ある QI グループ q 対して，Sensitive データの値が s である確率を $P(q, s)$ としたとき，全ての QI グループにおいて以下の式が満たされていれば ℓ-多様性を満たすことになる.

$$-\sum_{s \in S} P(q,s) \log P(q,s) \geq \log(\ell)$$

（S は q における Sensitive データの母集合）

ℓ-多様性によって同種攻撃は解決したが，背景知識攻撃を防ぐことはできない．QI グループ内に Sensitive データが ℓ 種類以上存在したとしても，確率に関する知識が加われば Sensitive データの候補が絞られてしまうためである.

また，ℓ-多様性では Skewness 攻撃が問題となった[3].

・ Skewness 攻撃:

ある QI グループ内での Sensitive データの分布が，データテーブル全体での Sensitive データの分布と大きく異なっていたとする．例えば，陽性と陰性の比がデータテーブル全体では 99:1 であるのに対し，ある QI グループ内では 1:3 であった場合，QI グループを特定されることで，その特異性が明らかになってしまう.

2.1.3 t-近傍性

ℓ-多様性における Skewness 攻撃を防ぐために，t-近傍性が考えられた．t-近傍性とは，各 QI グループ内での Sensitive データの分布と，データテーブル全体での Sensitive データの分布の距離が，閾値 t より小さくなっていることを保証する手法である．分布間の距離を測る方法としては，EMD(Earth Mover's Distance)を用いる.

しかし，t-近傍性でも ℓ-多様性と同様に，背景知識攻撃は防ぐことができない.

2.2 プライバシーの定量化

プライバシーは，個人情報を識別できる程度として定量化することができる.

Agrawal らは，個人情報を含むデータにランダムな摂動を加えたデータマイニングにおいて，プライバシーを保護する方法を提案した[4]．この方法では，プライバシーは元のデータが区間[a,b]に存在する確信度によって定量化される.

Gruteser らは位置プライバシーに k-匿名性を導入し，プライバシー量として k を用いた[5]. Hoh らは位置プライバシーを攻撃者が対象の個人を追跡した期間として定量化した[6].

神山らは，SNS におけるプライバシー漏洩を，定量化した[7]. 個人のある属性情報について情報が開示される前と後のエントロピーを算出し，その差をプライバシー漏洩の量とした.

3 分析モデルの提案

3.1 分析モデル

外部知識による匿名化データベースへの影響を分析する簡単な例を図 1 に示す.

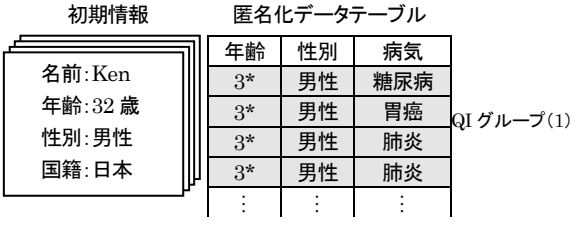

図 1 分析の例

攻撃者が保護対象者である Ken について，32 歳の男性の日本人であることを知っているとする．Ken の含まれている匿名化データテーブ

ルが漏洩し，QI グループ(1)に属することが推定されたとする．QI グループ(1)では Sensitive データである病気のデータが，糖尿病，胃癌，肺炎，肺炎，となっているため，Ken の病気の候補は糖尿病，胃癌，肺炎に絞られる．QI グループ(1)に含まれる他の 3 人についても同様である．

分析の都合上，QI グループ(1)の表記を表 3 のように変更する．count は QI グループ(1)内での病気の出現回数を示す．

表 3 表記を変更した匿名化データテーブル

年齢	性別	病気	count
3*	男性	糖尿病	1
3*	男性	胃癌	1
3*	男性	肺炎	2

ここで，初期情報によって明らかになった国籍と病気の関係を確率的に示す外部知識が加わった場合に，Ken の病気が推定される危険性を定量評価する．

この例に基づいて一般的なモデルを図 2 に示す．

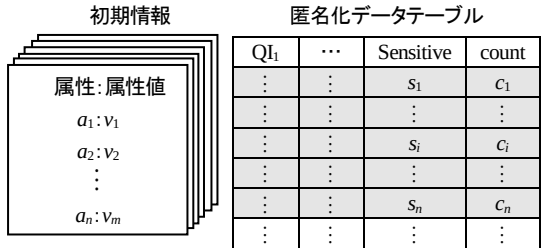

図 2 一般モデル

攻撃者が保護対象者(以下 *Target*)について m 個の属性情報を知っており，匿名化データテーブルから QI グループが特定されたとする．このとき，*Target* は複数であってもよい．

QI グループが特定されたとき，*Target* および同 QI グループに含まれる他の人物(以下 *Unknown*)の Sensitive データの候補は s_1～s_n となる．また，s_1～s_n のそれぞれの値が QI グループ内で出現する回数は count に記述するものとする．したがって，全ての行で s_i の値が異なる．

ここで，外部知識は *Target* のもつ属性 a_1～a_m の母集合 V_1～V_m，および匿名化データテーブルの QI 属性の母集合 W_1～W_q から与えられ，下記の式(1)となる関数である．

$$V_1\times\cdots\times V_m\times W_1\times\cdots\times W_q\times S\mapsto P=\{p\mid 0.0\le p\le 1.0\} \quad (1)$$

(SはSensitiveデータの母集合)

この関数は v_1, …, v_m, QI_1, …, QI_q の状況における s_i の生起確率を与える．外部知識が一部の属性について与えられる場合もあるが，これは上記のモデルの特別な例である．例えば，図 1 の外部知識は国籍情報のみで与えられているものである．また，初期情報で与えられた属性と QI 属性が独立であるとは限らない．図 1 では両方に性別があるが，このような場合には一方を無視する．図 1 のように初期情報の年齢が 32 歳，匿名化データテーブルの QI での年齢が 30 代となっている場合には，粒度の小さい方を取り上げることとする．

なお，攻撃者が入手するのは，ある匿名化データテーブルの QI グループの一部である場合もあるが，その場合もこのモデルに含まれる．例えば図 1 の場合，最初の 3 行だけが漏洩した場合にもこのモデルに含む．

3.2 分析方針

本論文では匿名化データベースの危険性を定量化する手法として，2.2 節で述べた神山らの手法と同様にエントロピーを用いる．

匿名化データベースから明らかになった *Target* の Sensitive データの候補集合を X とする．例えば，*Target* が図 1 の QI グループ(1)に属することが特定できた場合，

$$X=\{\text{糖尿病, 胃癌, 肺炎}\}$$

となる．このとき，外部知識による匿名化データベースへの影響を以下のように定量化する．

$$H(X)-H'(X)$$

($H'(X)$は外部知識取得後のエントロピー)

外部知識による影響が大きいほど，Sensitive データを推定される危険性が高いことになる．

4 定量化手法

3 章で述べた分析モデルおよび方針に基づいて，本章では外部知識の影響を定量化する手法を検討する.

4.1 例題の検討

攻撃者が初期情報として，Ken という男性について，32 歳の日本人であることを知っていたとする(図 3)．ここでは Ken を *Target* とする.

名前:Ken
年齢:32 歳
性別:男性
国籍:日本

図 3 *Target* の初期情報

Target の含まれている病院の 2-匿名データテーブルから QI グループ(表 4)を特定できたとする．このとき，同 QI グループに含まれるもう一人の人物を *Unknown* とする．*Target* と *Unknown* の病気として考えられるのは糖尿病と胃癌の 2 つである.

$Target$の病気$X_t = \{$糖尿病，胃癌$\}$

$Unknown$の病気$X_u = \{$糖尿病，胃癌$\}$

表 4 *Target* が含まれている QI グループ

年齢	性別	病気	count
3*	男性	糖尿病	1
3*	男性	胃癌	1

このとき， *Target* の病気が推定される危険性を考える.

外部知識がないとき，*Target* の病気についての情報はないので，病気の確率はそれぞれ count 数から計算し，エントロピーは以下のように計算される.

$$H(X_t) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1.0000$$

ここで，外部知識として各国における罹患率が分かったとする(表 5)．

表 5 各国における罹患率

	糖尿病	胃癌
日本	1%	4%
⋮	⋮	⋮
世界平均	4%	2%

さらに，表 4 より前提として X_t と X_u の組み合わせは以下の 2 通りである.

- X_t = 糖尿病，X_u = 胃癌
- X_t = 胃癌，X_u = 糖尿病

このデータテーブルから得られた前提と外部知識から得た罹患率を組み合わせることで，Ken の病気の確率 $P'(X_t)$ は以下のように計算できる．ただし，*Unknown* は国籍が不明のため，世界平均値を利用している．また，計算時には糖尿病と胃癌の併発の確率を除いている.

$$P'(X_t = \text{糖尿病}) = \frac{P(X_t = \text{糖尿病}, X_u = \text{胃癌})}{P(X_t = \text{糖尿病}, X_u = \text{胃癌}) + P(X_t = \text{胃癌}, X_u = \text{糖尿病})} = 0.1061$$

$$P'(X_t = \text{胃癌}) = 0.8939$$

以上より，*Target* の病気のエントロピーは

$$H'(X_t) \cong 0.4882$$

となり，例題における外部知識の影響は以下のようになった.

$$H(X_t) - H'(X_t) = 1.0000 - 0.4882 = 0.5118$$

4.2 一般式

攻撃者が初期情報として，1 人の *Target* について m 個の属性情報を知っていたとする(図 4)．

属性:属性値
$a_1 : v_1$
$a_2 : v_2$
⋮
$a_n : v_m$

図 4 *Target* の初期情報

また，病院の k-匿名データテーブルから初期情報により QI グループ(表 6)を特定できたとする．これより *Target* と *Unknown* の Sensitive データの候補は n 個に絞られる.

$$X_t = \{s_1, s_2, \ldots, s_n\}$$
$$X_{u_1} = \{s_1, s_2, \ldots, s_n\}$$
$$\vdots$$
$$X_{u_\ell} = \{s_1, s_2, \ldots, s_n\}$$
$$\vdots$$
$$X_{u_{n'-1}} = \{s_1, s_2, \ldots, s_n\}$$

（$n'=\sum_{j=1}^{n} c_j$: *Target*と*Unknown*の人数の和）

表 6 特定された QI グループ

QI_1	…	Sensitive データ	count
⋮	⋮	s_1	c_1
⋮	⋮	s_2	c_2
⋮	⋮	⋮	⋮
⋮	⋮	s_n	c_n

このとき，*Target* の Sensitive データが推定される危険性を考える．

外部知識がないとき，確率はそれぞれ count から計算し，エントロピーは以下のように計算される．

$$H(X_t)=-\sum_{i=1}^{n}\frac{c_i}{n'}\log\frac{c_i}{n'}$$

ここで，外部知識となる関数を得たとする．このとき，*Target* の Sensitive データ s_1～s_nについて得られた確率データを P_{t1}～P_{tn}，*Unknown* の Sensitive データ s_1～s_nについて得られた確率データを $P_{u_\ell 1}$～$P_{u_\ell n}$とする．

さらに，表 6 より前提としてデータベース内に含まれているn'人の Sensitive データの組み合わせを考える．組み合わせの数は

$$n'!/(c_1!\times c_2!\times\cdots\times c_n!)$$

となる．

このデータベースから得られた前提と外部知識を組み合わせることで，*Target* の Sensitive データの確率 $P'(X_t)$ は以下のように計算できる．

$$P'(X_t=s_\alpha)=\frac{P(X_t=s_\alpha)\sum_{d_2\in D_2}\prod_{\ell=1}^{n'-1}P(X_{u_\ell}=s_{d_2})}{\sum_{d_1\in D_1}\prod_{\ell=1}^{n'-1}P(X_t=s_{d_1})P(X_{u_\ell}=s_{d_1})}$$

$$=\frac{P_{t\alpha}\sum_{d_2\in D_2}\prod_{\ell=1}^{n'-1}P_{u_\ell d_2}}{\sum_{r=1}^{n}P_{tr}\sum_{d_1\in D_1}\prod_{\ell=1}^{n'-1}P_{u_\ell d_1}}$$

$$\left(\begin{array}{c} D_1=\{d_1\mid d_1は r を除いた1～n の順列\} \\ D_2=\{d_2\mid d_2は \alpha を除いた1～n の順列\} \end{array}\right)$$

以上より，Ken の病気のエントロピーは

$$H'(X_t)=-\sum_{r=1}^{n}P'(X_t=s_r)\log P'(X_t=s_r)$$

となり，外部知識の影響は以下のようになる．

$$H(X_t)-H'(X_t)$$
$$=-\sum_{i=1}^{n}\frac{c_i}{\sum_n c_j}\log\frac{c_i}{\sum_n c_j}+\sum_{r=1}^{n}P'(X_t=s_r)\log P'(X_t=s_r)$$

5 シミュレーション

4.2 で求めた一般式を用いて，外部知識のデータやデータベース内の Sensitive データの比が変化したときに外部知識の影響がどのように変化するかをシミュレートする．

まず，表 7 のような匿名化データテーブルを想定する．ここで，*Target* は日本人男性の Ken とアメリカ人男性の John の 2 人とする．

表 7 想定する匿名化データテーブル（1）

年齢	性別	病気	count
3*	男性	糖尿病	1
3*	男性	胃癌	1
3*	男性	肺炎	1

外部知識がない場合，Ken と John の病気のエントロピーは以下のように計算される．

$$H(X_{\text{Ken}})=-\sum_{3}\frac{1}{3}\log\frac{1}{3}\cong 1.5850$$

$$H(X_{\text{John}})=-\sum_{3}\frac{1}{3}\log\frac{1}{3}\cong 1.5850$$

外部知識（1.1）（表 8）のような各国における罹患率を得た場合，外部知識の影響は以下のようになった．

$$H(X_{\text{Ken}})-H'(X_{\text{Ken}})=1.5850-1.0960=0.4890$$
$$H(X_{\text{John}})-H'(X_{\text{John}})=1.5850-1.0960=0.4890$$

以降，外部知識の一部データを変更した場合の変化を分析する．

表 8 外部知識（1.1）

	糖尿病	胃癌	肺炎
日本	5%	3%	1%
アメリカ	1%	5%	3%
世界平均	3%	1%	5%

外部知識（1.2）（表 9）は，日本における糖尿病と胃癌の罹患率を等しくした場合である．

$$H(X_{\text{Ken}})-H'(X_{\text{Ken}})=1.5850-1.2061=0.3789$$
$$H(X_{\text{John}})-H'(X_{\text{John}})=1.5850-1.2801=0.3049$$

Ken の病気について糖尿病と胃癌の区別が曖昧になったことで，両者とも外部知識の影響

は小さくなった.

表 9 外部知識(1.2)

	糖尿病	胃癌	肺炎
日本	5%	5%	1%
アメリカ	1%	5%	3%
世界平均	3%	1%	5%

外部知識(1.3)(表 10)は世界平均において糖尿病と肺炎の罹患率が等しくなった場合である.

$$H(X_{\text{Ken}}) - H'(X_{\text{Ken}}) = 1.5850 - 1.3458 = 0.2392$$

$$H(X_{\text{John}}) - H'(X_{\text{John}}) = 1.5850 - 1.3814 = 0.2036$$

KenとJohnの国籍におけるデータに変化はないが, 両者とも外部知識の影響は小さくなった.

表 10 外部知識(1.3)

	糖尿病	胃癌	肺炎
日本	5%	3%	1%
アメリカ	1%	5%	3%
世界平均	5%	1%	5%

以上, 3 つの外部知識によるデータテーブルへの影響の変化を図 5 に示す.

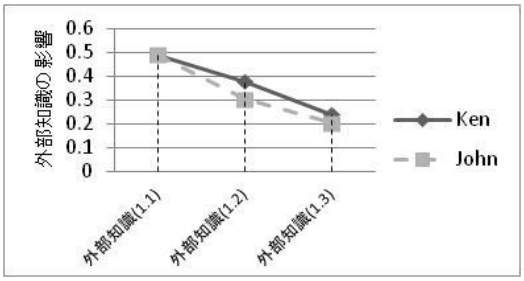

図 5 外部知識による影響の変化

次に, 表 11 のように Sensitive データの比が 2:1 であるような匿名化データテーブルを想定する. ここで, *Target* は日本人男性の Ken とする.

表 11 想定する匿名化データベース(2)

年齢	性別	病気	count
3*	男性	糖尿病	2
3*	男性	胃癌	1

外部知識がない場合, *Target* の病気のエントロピーは以下のように計算される.

$$H(X_{\text{Ken}}) = -\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3} \cong 0.9183$$

外部知識(2.1)(表 12)は, 外部知識の影響は以下のようになった.

$$H(X_{Ken}) - H'(X_1) = 0.9183 - 0.1305 = 0.7878$$

データテーブル内で糖尿病と胃癌の比が 2:1 であるのに対し, 外部知識から得られた *Target* の罹患率も糖尿病の確率が高くなっているため, 外部知識の影響は大きくなった.

表 12 外部知識(2.1)

	糖尿病	胃癌
日本	5%	1%
世界平均	1%	5%

外部知識(2.2)(表 13)を得た場合, 匿名化データベースの危険性は以下のようになった.

$$H(X_1) - H'(X_1) = 0.9183 - 0.3607 = 0.5576$$

外部知識(2.1)とは逆に, *Target* の胃癌の確率が高くなったため, 外部知識(2.1)よりも外部知識の影響は小さくなった.

表 13 外部知識(2.2)

	糖尿病	胃癌
日本	1%	5%
世界平均	5%	1%

6 2 つのデータテーブルへの拡張

匿名化データテーブル 2 つの場合について検討する(図 6).

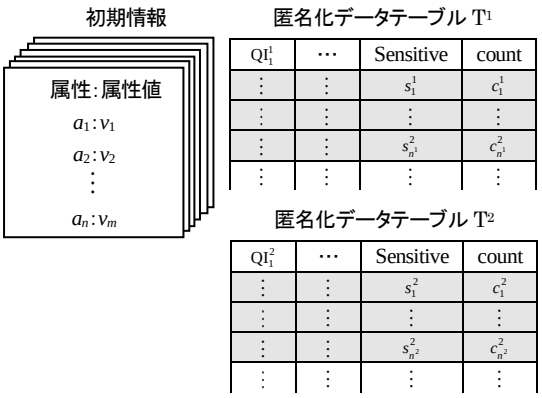

図 6 複数テーブルのモデル

Target が匿名化データテーブル T^1 と T^2 の両方に含まれていることが分かっており, それぞれ QI グループが特定できたとする. 両テーブルに含まれているのは *Target* のみで, *Unknown* は一方のテーブルにしか含まれないものとする.

ここで, 外部知識は 3 種類考えられる. *Target* に対する外部知識は, *Target* のもつ属性 a_1～a_m の母集合 V_1～V_m, および匿名化データテーブ

ル T^1 の QI 属性の母集合 W_1^1～$W_{q^1}^1$ と T^2 の QI 属性の母集合 W_1^1～$W_{q^2}^1$ から与えられ，下記の式(2)となる関数である．

$$V_1 \times \cdots \times V_m \times W_1^1 \times \cdots \times W_{q^1}^1 \times W_1^2 \times \cdots \times W_{q^2}^2 \times S^1 \times S^2 \mapsto P = \{p \mid 0.0 \le p \le 1.0\} \quad (2)$$

(S^1, S^2 は T^1, T^2 の Sensitive データの母集合)

この関数は $v_1, \cdots v_m, QI_1^1, \cdots, QI_{q^1}^1, QI_1^2, \cdots, QI_{q^2}^2$ の状況における $s_{i^1}^1, s_{i^2}^2$ の同時生起確率を与える．

Unknown に対する外部知識は，以下の 2 つの関数である．

$$V_1 \times \cdots \times V_m \times W_1^1 \times \cdots \times W_{q^1}^1 \times S^1 \mapsto P = \{p \mid 0.0 \le p \le 1.0\} \quad (3)$$

$$V_1 \times \cdots \times V_m \times W_1^2 \times \cdots \times W_{q^2}^2 \times S^2 \mapsto P = \{p \mid 0.0 \le p \le 1.0\} \quad (4)$$

この 2 つの関数は 3.1 節の式(1)と同様の意味をもつ．なお，式(3)は T^1 の *Unknown* に，式(4)は T^2 の *Unknown* に適用する．

以上の外部知識と T^1 と T^2 の直積から考えられる組み合わせを考慮して，外部知識の影響を定量化する．

7 結論

外部知識によって匿名化データベースの匿名性が破られるという問題について，著者らの知る限りでは定量的な検討が行われていなかった．本論文では，*k*-匿名データベースを取り上げ，保護対象者の Sensitive 情報について *k*-匿名データベースと外部知識の双方から何が推定できるかを検討した．

まず，匿名性が破られることを，保護対象者の Sensitive 属性のエントロピーが低下することと定義した．外部知識として，保護対象者の属性とデータベースの属性の間の依存関係を取り上げた．外部知識のモデルは，保護対象者の既知の属性値とデータベースの属性値を入力とし，Sensitive 属性の値の生成確率を出力とする関数とした．

このモデルに基づき，保護対象者の属性値と *k*-匿名データベースのレコード群が与えられたときに，外部知識による Sensitive 属性の値のエントロピーの低下を，外部知識の影響とした．本手法は，保護対象者が複数の場合および ℓ-多様化，*t*-近傍化したデータベースにも適応できる．また，手法を実装し，外部知識の変化によるデータベースへの影響の変化を分析した．

今後の課題としては，外部知識を前提とした新しいデータベースの匿名化手法を提案することが挙げられる．

参考文献

[1] Sweeney, L.: *k*-anonymity: a model for protecting privacy, Int. J., Fuzziness and Knowledge-based Systems, 10 (5), 2002, pp.557-570

[2] Machanavajjhala, et al.: ℓ-diversity: Privacy Beyond k-Anonymity, 22nd Intnl. Conf. Data Engg. (ICDE), 2006

[3] Li, N., Li, T. and Ventakasubramanian, S.: *t*-closeness: Privacy Beyond *k*-Anonymity and ℓ-Diversity, 23rd International Conference on Data Engineering(ICDE'07), 2007, pp.16-20

[4] Agrawal, D., Aggarwal, C.: On the design and quantification of privacy preserving data mining algorithms, 20th ACM SIGACT-SIGMOD-SIGART Symposium on Principle of Database System, 2001, pp. 247–255

[5] Gruteser, M. , Grunwald, D.: Anonymous Usage of Location-Based Services Through Spatial and Temporal Cloaking, 1st ACM/USENIX International Conference on Mobile Systems, Applications, and Services., 2003, p31-42

[6] Hoh, B., et al.: Preserving Privacy in GPS Traces via Uncertainty- Aware Path Cloaking, 14th ACM Conference on Computer and Communication Security, Alexandria , 2007

[7] Kamiyama, K., Ngoc, T., Echizen, I., Yoshiura,H.: Measuring Accumulated Revelations of Private Information by Multiple Media, 10th IFIP Conference on e-Business, e-Services, and e-Society, 2010