Vol.2015-CSEC-69 No.28
Vol.2015-IOT-29 No.28
2015/5/22

Pk-匿名化データの有用性評価実験

正木 彰伍[1,a)]　五十嵐 大[1]　菊池 亮[1]　齋藤 恆和[1]　千田 浩司[1]　廣田 啓一[1]

概要：パーソナルデータの安全な利活用には，データの安全性と有用性の両立が必要である．安全性については代表的な k-匿名性や，k-匿名性を確率的な指標に拡張した Pk-匿名性が提案されるなど，匿名化技術が広く研究されている．一方で，匿名化データの有用性についての議論は未だ限定的である．特に，レコード数などのデータの特徴と有用性の関係性を明らかにすることは，実用上非常に有益である．しかし，これまで行われてきた，ウェブ上で公開されている実データなどを用いた実験では，用いるデータの特徴が限定的になり，議論が困難となっていた．そこで本稿では，多くのデータを包含する一般的な模擬データモデルを利用した評価法を提案し，この模擬データに，Pk-匿名化を適用した実験を行う．さらに実験結果から，有用性と模擬データモデルのパラメーターの関係について調べ，特定のパラメーターから有用性を予測できることがわかった．

1. はじめに

ICT の発達に伴い，サービス事業者には，ユーザーの購買履歴や位置情報といったデータが豊富に蓄積されるようになってきた．昨今，そういったパーソナルデータを二次利用し，サービス開発や商品販売戦略などに活用したいというニーズが高まっている．一方で，ユーザー個々人に紐づくパーソナルデータの利活用にはプライバシー侵害のリスクがあり，慎重な取り扱いが求められる．対応策の 1 つとして，データを安全な形に加工する匿名化技術が広く研究されている．

匿名化技術の中でも代表的な指標として，k-匿名性 [1], [2] が知られている．k-匿名性は，1 つのデータ中に “同じ属性の値の組み合わせを持つレコードが k 個以上存在する” という直観的でわかりやすいプライバシー保護指標である．一方，データへのノイズ付加といった確率的な操作を伴う匿名化には適用できなかった．これに対し，五十嵐ら [3] は，k-匿名性を確率的な指標に拡張し，“ある人のデータを $1/k$ 以上の確率で当てることができない” ことを保証する Pk-匿名性を提案した．Pk-匿名性を満たす匿名化方法（Pk-匿名化）としては，カテゴリ属性に対する維持置換撹乱 [4]，数値属性に対するラプラスノイズ加算 [5]，有界ノイズ加算 [6]，データ依存型維持置換撹乱 [7] および属性間の相関を考慮した撹乱再構築法 [8] がある．

Pk-匿名化は，確率的なノイズ付加を伴う加工であるため，匿名化されたパーソナルデータは元と異なるデータになってしまう．匿名化データを利活用するうえで，Pk-匿名化によってどれほど変化するか，有用性はどれほど保たれるかは明らかにする必要がある．特に，データの特徴から匿名化データの有用性の関係が明らかになれば，Pk-匿名化を施さずとも有用性が予測でき，実用上非常に有益である．これまで，Adult Data Set [9] に代表されるウェブで公開されているいくつかのデータを使った，実験的な有用性評価が行われてきた [10], [11]．しかし，ウェブ上の公開データや購入可能なデータを用いるだけでは，利用可能なデータの特徴が限定的になり，データの特徴と匿名化データの有用性の関係を議論することが困難となっていた．

1.1 本稿の成果

本稿では，多くのデータを包含する一般的な模擬データモデルを利用した評価法を提案し，この模擬データに，Pk-匿名化を適用した実験を行う．実験結果から，匿名化データの有用性はデータ分布にあまり依存せず，特定のパラメーターから有用性を予測できることがわかった．

2. 準備

2.1 テーブルとクロス集計

本稿で扱うテーブルとは以下のようなものである．テーブルは情報提供者から提供されたデータの集合である．各情報提供者からのデータはテーブル上では 1 行に表現され，これをレコードと呼ぶ．各レコードはあらかじめ定められた項目に対する値から成り立っており，この項目を属性といい，取りうる値の範囲を属性の値域という．表 1 にテー

[1] 日本電信電話株式会社 NTT セキュアプラットフォーム研究所
180-8585 東京都武蔵野市緑町 3-9-11
[a)] masaki.shogo@lab.ntt.co.jp

ブルの例を示す．“性別” 及び “年代” が属性であり，各行 “女性，10 代” 等がレコードである．

表 1 テーブルの例

ID	年代	性別
1	10 代	女性
2	40 代	男性
3	20 代	男性
4	30 代	女性
⋮	⋮	⋮

クロス集計とは，テーブルの複数の属性に着目し，その全ての属性の値が等しいレコードの数を表すものである．このクロス集計は，アンケート集計で使われるほかに数多くのデータマイニング手法の中で使用される基本的かつ重要な統計量である．一般的にクロス集計は行ベクトルとして扱われる．その大きさは各属性の組み合わせの総数，つまり各属性の値域の積となり，ベクトルの要素の和はレコードの総数と一致する．表 2 が表 1 におけるテーブルでのクロス集計である．ベクトルの第 2 要素である“{10 代，女性 }”はテーブルの中にある 10 代かつ女性である人数の総和であり，右辺のベクトルの第二要素はそれが 45 人であることを示している．

表 2 クロス集計の例

({10 代，男性 }, {10 代，女性 },
{20 代，男性 }, {20 代，女性 }, . . .)
= (61, 45, 73, 50, . . .)

2.2 撹乱再構築法による Pk-匿名化

撹乱再構築法とは，データを改変することによりプライバシーを保護する撹乱と，撹乱データから元データを推定する再構築という 2 つのアルゴリズムからなる匿名化手法である．

2.2.1 維持置換撹乱

具体的な撹乱方法として，属性値を一定の確率で維持し，それ以外では取りうる値の中でランダムに置換する維持置換撹乱がある [4]．ある属性 i の値 v_i が，v'_i に遷移する確率を，サイズ $M_i \times M_i$ の行列 $(A_i)_{v_i,v'_i}$ を用いて，

$$(A_i)_{v_i,v'_i} = \begin{cases} \rho_i + \frac{1-\rho_i}{M_i} & \text{for} \quad v_i = v'_i \\ \frac{1-\rho_i}{M_i} & \text{for} \quad v_i \neq v'_i \end{cases} \tag{1}$$

と表す．ここで，ρ_i は属性 i の維持確率と呼ばれ $0 \leq \rho_i \leq 1$ であり，M_i は属性 i の値域である．属性が n 個あった場合，行列 $A = A_1 \otimes A_2 \otimes \ldots \otimes A_n$ を遷移確率行列と呼ぶ．ここで，$\otimes$ はクロネッカー積である．

この維持置換撹乱は，維持確率を適切に設定することで，Pk-匿名性を満たすことができる．$\mathcal{R}$ をテーブルのレコード集合とした場合，パラメーター間で

$$k = 1 + (|\mathcal{R}| - 1) \prod_{i=1}^{n} \left[\frac{1-\rho_i}{1+(M_i-1)\rho_i} \right]^2 \tag{2}$$

の関係が成立する必要がある [3]．

2.2.2 再構築

具体的な再構築方法としては，反復ベイズ推定法が知られている [12]．これは，撹乱テーブルのクロス集計に対し，遷移確率行列 A を用いてベイズ推定を繰り返し行うことで，元テーブルのクロス集計を推定するものである．

反復ベイズ推定法のアルゴリズムを Algorithm 1 に示す．ここで，j は反復回数を表し，ベクトル $\vec{v}_1$, $\vec{v}_2$ に対して $\vec{v}_1 * \vec{v}_2$ と $\vec{v}_1/\vec{v}_2$ はベクトルの成分ごとの積と商を表す．$|\vec{x}_{j+1} - \vec{x}_j|_{L_1}$ は $\vec{x}_{j+1}, \vec{x}_j$ の L_1 距離と呼ばれ，各要素の差の絶対値の総和であり，本稿では総レコード数で割ったものを用いる．ϵ はあらかじめ定める収束半径である．遷移確率行列 A と撹乱テーブルのクロス集計 $\vec{y}$ が入力であり，出力として推定した元テーブルのクロス集計 $\vec{x}_j$ を得る．

Algorithm 1 ベイズ推定法による再構築アルゴリズム

```
Input: A, y⃗
Output: x⃗_j
x⃗_0 := y⃗
j := 0
while |x⃗_{j+1} − x⃗_j|_{L_1} ≥ ε do
  x⃗_{j+1} := x⃗_j ∗ (A(y⃗/x⃗_j A)^t)^t
  j := j + 1
end while
```

3. 提案する評価法

3.1 概要

ウェブ上の公開データや購入可能なデータを用いるだけでは，利用可能なデータの特徴が限定的になり，網羅的に有用性評価実験を行うことは困難になる．そこで，多様なデータを用いて実験を行うために，データを特徴づけるパラメーターを設定し，値を動かして多数の模擬データを作成し，実験に用いる．

手順としては以下の通りである．

(1) データを特徴づけるパラメーターの設定
(2) パラメーターの値を変えながら，模擬データを作成
(3) 作成した模擬データに対して Pk-匿名化を実行
(4) 匿名化前後のデータを比較し，有用性を評価

ポイントとなるのは，データの特徴の 1 つである属性値の分布のモデル化およびパラメーター化である（3.3.3 を参照）．

3.2 1 属性データと属性の連結化を用いた Pk-匿名化

本稿では，属性が 1 つの模擬データを作成し，実験に用いる．この属性が 1 つのみの模擬データに Pk-匿名化を適用することは，属性が複数あるデータに属性の連結化を用

いた Pk-匿名化を適用することと等しい．

属性の連結化とは，複数の属性の値を一括りにして，1 つの属性の値として扱うことである．表 3 に，表 1 の “年代” と “性別” を連結化した例を示す．当該表の属性は，“年代・性別” の 1 つとなり，ID が 1 のレコードの属性値は，“10 代・女性” となる．

表 3 表 1 の属性を連結化したテーブル

ID	年代・性別
1	10 代・女性
2	40 代・男性
3	20 代・男性
4	30 代・女性
⋮	⋮

属性の連結化によるメリットは，Pk-匿名化されたデータの有用性向上が期待できる点にある．Pk-匿名化は撹乱を伴うため，属性を連結せずに Pk-匿名化を行うと，元データには存在しない属性値の組合せが，匿名化データに現れることがある．属性を連結することでこれを防ぐことができ，Pk-匿名化されたデータの有用性向上，特に連結された属性間の相関がより良く保存されることが期待できる．

3.3 パラメーター値の設定

模擬の 1 属性データを作成するにあたって，1 属性データを特徴づけるパラメーターとして，次の 3 つを考える．

- レコード数
- 値域
- 値の分布

3.3.1 レコード数 N_{rec}

現実的な値として，$1{,}000\ (=10^3)$ から $10{,}000{,}000\ (=10^7)$ の範囲を考える．具体的には，$10^3,\ 2\times10^3,\ 5\times10^3,\ 10^4,\ 2\times10^4,\ 5\times10^4, \ldots,\ 10^7$ の 13 個の値を取る．

3.3.2 値域 V_{att}

値域とは，属性値の種類の数である．レコード数が増えれば取りうる値域が増えるため，レコード数に対する比をパラメーターとする．本稿では，具体的には，$V_{\mathrm{att}}/N_{\mathrm{rec}} = 10^{-6}, 2\times10^{-6},\ 10^{-5},\ 2\times10^{-5}, \ldots,\ 10^{-1},\ 2\times10^{-1}$ の値を取る．ただし，実験の冗長さを防ぐため，利活用の価値が比較的低いと考えられる V_{att} が 10 に満たない模擬データは実験から除外する．

3.3.3 値の分布に関するパラメーター a

本稿では，1 つのカテゴリ属性を持つ模擬データを考える．数値属性と違いカテゴリ属性の分布は，値に順序が無いため，正規分布等のモデルで単純に特徴づけることは難しい．そこで，度数が大きい順に，値を並び替え順位をつける．そして，度数を順位の関数として表し，その関数にパラメーターを用いることとする．度数を y，順位を x とした時に，

$$y(x) \propto x^{-a} \tag{3}$$

と近似するモデルを用いる．このようなべき乗分布は，自然界，人間界における出現頻度分布に多く見られることが知られている．ここで，$x,\ y$ は正数であり，特に $1 \le x \le V_{\mathrm{att}}$ である．a がパラメーターである．ただし，

$$N_{\mathrm{rec}} = \sum_{x=1}^{V_{\mathrm{att}}} y(x) \tag{4}$$

という関係式が成り立つ．パラメーター a の値として，$0.05,\ 0.1,\ 0.2,\ 0.5,\ 1,\ 2$ の 6 個の値を取る．a が小さいほど分布は平坦で，大きいほど高順位と低順位の度数の差は大きい．

3 つのパラメーター値の組合せにより，計 602 通りの模擬データを作成し，実験に用いる．

3.4 模擬データの作成

式 (3) を確率関数とし，それに従う乱数を逆関数法を用いて生成し，模擬データを作成する．値域が与えられているため，各値で 1 レコードを生成させ，すべての値が最低でも 1 度は現れるようにする（つまり，$y \ge 1$）．その後，乱数生成により残りのレコードを生成する．

図 1 に作成した模擬データの例を示す．レコード数が

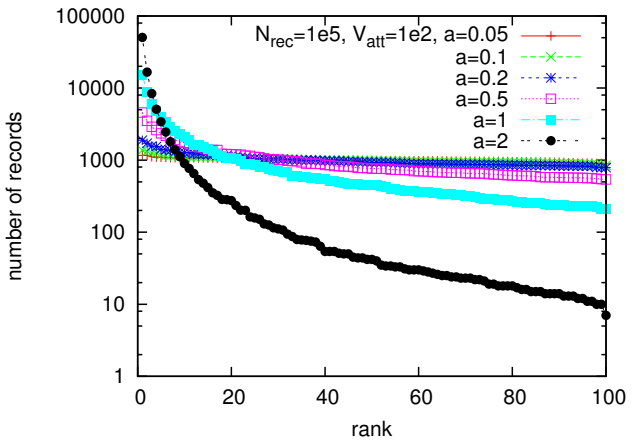

図 1 レコード数が 100,000，値域が 100 の模擬データ．横軸が度数の順位，縦軸が度数．分布のパラメーター a を 6 種類取っている．

100,000，値域が 100 の場合である．横軸が度数の順位，縦軸が度数となっている．分布のパラメーター a を 6 種類取っている．a が小さいほど，平坦な分布となっていることがわかる．

4. 実験結果

本稿では，匿名化前後の模擬データ間の L_1 距離を有用性の指標として用いる．1 つのパラメーター値のセットに対して，模擬データ作成と Pk-匿名化の実行を 10 回繰り返

し，それぞれの試行で L_1 距離を測る．以降の図では，10 個の L_1 距離の平均値をそのパラメーター値セットの L_1 距離とし，標準偏差を誤差として示す．

以下，L_1 距離が，模擬データを特徴づけるパラメーターを変化させた時，どのように影響を受けるか見ていく．匿名性指標 k の値は $k = 2$ で固定してある．

4.1 値域 V_{att} による影響

図 2 に，$N_{\mathrm{rec}} = 10^3,\ 10^4,\ 10^5,\ 10^6,\ 10^7$ の場合の，匿名化前後のデータの L_1 距離を示す．縦軸が L_1 距離である

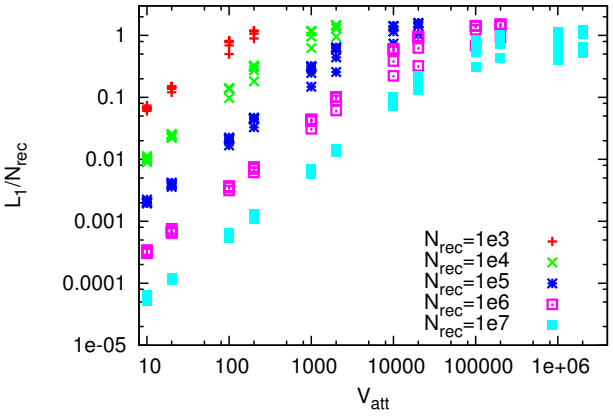

図 2 $N_{\mathrm{rec}} = 10^3,\ 10^4,\ 10^5,\ 10^6,\ 10^7$ の場合の，匿名化前後のデータの L_1 距離．縦軸は，レコード数 N_{rec} で正規化した L_1 距離．横軸は，値域 V_{att}．見やすさのため，誤差は示さない．

が，レコード数 N_{rec} で正規化している．横軸は値域 V_{att} である．レコード数を固定すると，値域が小さいほど L_1/N_{rec} は小さくなる．これはあるレコード数に対して，値域が小さいほど維持確率が大きくなるためである．N_{rec}，V_{att} が与えられた時に，わずかにバラつきがある．これは，分布パラメーター a の違いによるものである．

4.2 分布のパラメーター a による影響

図 3 には，分布のパラメーター a の影響を示す．レコード数が $N_{\mathrm{rec}} = 10^6$，レコード数と値域の比が $V_{\mathrm{att}}/N_{\mathrm{rec}} = 10^{-5},\ 10^{-4},\ 10^{-3},\ 10^{-2},\ 10^{-1}$ の結果を示している．分布のパラメーター a の影響は比較的小さいことが見て取れる．ただし，値域が大きい場合は，a が大きいと L_1 距離が下がる傾向にある．

5. 考察

5.1 L_1 距離とパラメーターの相関係数

データを特徴づけるパラメーターと L_1 距離との相関関係を評価するために，Spearman の相関係数を求めた．結果は，表 4 の通りである．L_1 距離は，値域 V_{att} と強い相関があることがわかる．レコード数 N_{rec} とは弱い相関が

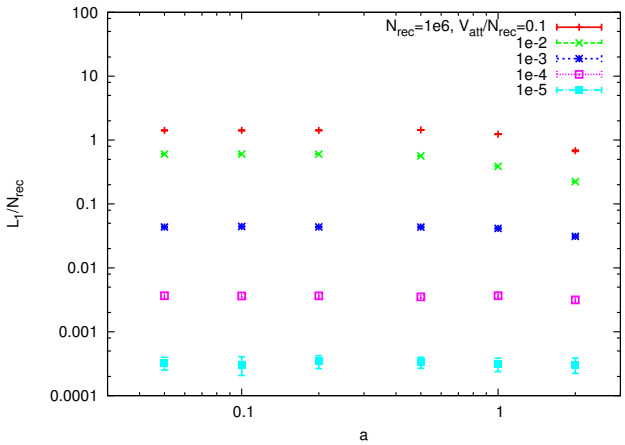

図 3 $N_{\mathrm{rec}} = 10^6$ の場合の，匿名化前後のデータの L_1 距離．縦軸は，レコード数 N_{rec} で正規化した L_1 距離．横軸は，分布のパラメーター a．レコード数と値域の比 $V_{\mathrm{att}}/N_{\mathrm{rec}}$ は，$10^{-6},\ 10^{-5},\ 10^{-4},\ 10^{-3},\ 10^{-2},\ 10^{-1}$ の場合を示す．

表 4 L_1 距離とデータを特徴づけるパラメーターの相関係数．値域 V_{att} と強い相関がある．

パラメーター	相関係数
レコード数 N_{rec}	-0.25
値域 V_{att}	0.75
分布のパラメーター a	-0.067

あり，分布のパラメーター a は，ほとんど相関が無い．

5.2 L_1 距離とパラメーターの関係

図 3 と表 4 で見たように a と L_1 距離との相関関係は非常に弱い．そこで，L_1 距離に影響を及ぼすのは，レコード数 N_{rec} と値域 V_{att} の 2 つだと考えられる．そのためこの 2 つの変数を使って L_1 距離をフィッティングして予測可能にすることを考える．ただし，分布パラメーター a による影響は全く無いわけではないので，これを誤差として評価する．具体的な評価法としては，a の値がいくつであっても，L_1/N_{rec} が $X \pm Y$ の範囲に収まるような X と Y を求める．例えば，図 2 中のある N_{rec}，V_{att} のペアに対して

$$L_1/N_{\mathrm{rec}} = 0.5 \pm 0.1 \ (a = a_0) \tag{5}$$

$$L_1/N_{\mathrm{rec}} = 0.8 \pm 0.2 \ (a = a_1) \tag{6}$$

$$L_1/N_{\mathrm{rec}} = 1.0 \pm 0.2 \ (a = a_2) \tag{7}$$

という a が異なる 3 点があったとする．a の影響を誤差として捉えるために，$1.0 + 0.2 = 1.2$ を上限，$0.5 - 0.1 = 0.4$ を下限，中央値 X を $(1.2 + 0.4)/2 = 0.8$，誤差 Y を $(1.2 - 0.4)/2 = 0.4$ とし，3 点の情報を

$$L_1/N_{\mathrm{rec}} = 0.8 \pm 0.4 \tag{8}$$

と 1 点にする．こうすることで，$a = a_0,\ a_1,\ a_2$ の誤差範囲をすべて包含する，誤差を最も大きく見積もった誤差範囲を決めることができる．この操作を行った時に，図 2 は，

図 4 となる．

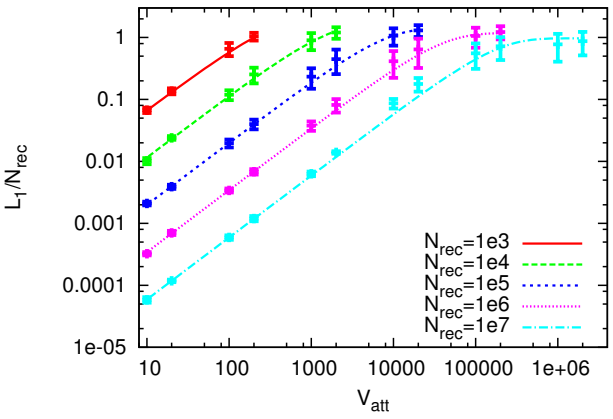

図 4 分布のパラメーター a による変化を誤差としてとらえた場合の図 2．線は，フィッティング関数である．

分布のパラメーター a による影響を誤差として扱った場合，L_1 距離に影響を及ぼすのはレコード数 N_{rec} と値域 V_{att} の 2 つとなる．ここで，実験で得られた L_1/N_{rec} の値を N_{rec} と V_{att} の 2 パラメーターの関数でフィットする．フィッティング関数として次のモデルを用いる．

$$L_1/N_{\mathrm{rec}} = A(N_{\mathrm{rec}}) \times [1 - \exp\{-B(N_{\mathrm{rec}}) \times V_{\mathrm{att}}\}] \quad (9)$$

$$A(N_{\mathrm{rec}}) = \alpha \times N_{\mathrm{rec}}^{\beta} \quad (10)$$

$$B(N_{\mathrm{rec}}) = \gamma \times N_{\mathrm{rec}}^{\delta} \quad (11)$$

あるレコード数 N_{rec}，匿名性指標 k に対して値域 V_{att} が増加すると，維持確率は減少し，一様ランダムなランダム化に近づいていくため，L_1 距離が一定値に漸近することが予想される．そこで，式 (9) のように一定値に漸近する関数を用いた．

最小二乗法によるフィッティングの結果，

$$\alpha = 3.69 \pm 0.30,\ \beta = -0.0830 \pm 0.0066 \quad (12)$$

$$\gamma = 0.358 \pm 0.036,\ \delta = -0.681 \pm 0.008 \quad (13)$$

の値を得た．図 4 中の線は，このフィッティング関数を示しており，実験結果と良い一致を見せることがわかる．

このフィッティング関数を用いることで，実際に処理をせずとも，Pk-匿名化を施したデータの有用性の予想ができ，実用上便利である．

6. まとめ

本稿では，1 属性のみ持つ模擬データを用いて，匿名化データの有用性検証実験を行った．模擬データ作成にあたって，レコード数，値域，データ分布といったデータを特徴付ける量をパラメーター化し，値を変えることで多数の模擬データを作成し，実験に用いた．特にカテゴリ属性のデータ分布については，度数の高い順に属性値の順位をつけ，度数を順位の関数として表し，その関数にパラメーターを用いてモデル化した．本稿で取り扱った範囲では，匿名化データの有用性指標（本稿では L_1 距離）に対するデータ分布のパラメーターの影響は，他の 2 つのパラメーターに比べて小さいという結果を得た．この結果に基づき，レコード数と値域によって，L_1 距離を近似的に表すフィッティング関数を求めた．得られたフィッティング関数を用いることで，実際に匿名化を行わなくとも，データを特徴付ける量によって，匿名化データの有用性を予測できる．

7. 今後の課題

7.1 連結化によるプライバシーリスク

本稿の有用性検証実験では，1 属性のみ持つデータを取り扱ったため，Pk-匿名化の中でも，属性の連結化を用いた Pk-匿名化を使用したことに対応している．属性の連結化によって，匿名化データの有用性向上や連結された属性間の相関関係がより良く維持されることが期待できる．

一方で，連結化によるプライバシーリスクが発生する場合も考えられる．Pk-匿名性は，“ある人のレコードがどれか，確率 $1/k$ 以下でしか当てられない” ことを保証するが，この匿名性だけで防ぐことができないプライバシー侵害からデータを守ることができないことも考えられる．1 つの例として属性の値域といったパーソナルデータのメタ情報にはプライバシー情報は含まれていないという前提に立っているため，メタ情報に対して何も保証していない．しかし，連結化した途端にメタ情報にプライバシー情報が含まれ，そこからプライバシー情報が流出する可能性がある[*1]．したがって，メタ情報にプライバシー情報が含まれないように連結化をすることが必要となる．本稿では，パーソナルデータ中の全属性を連結化することを想定したが，一部の属性のみを連結化するなどの方法が考えられる．

7.2 有用性検証実験の今後

模擬データ作成において，カテゴリ属性のデータ分布をべき乗型でモデル化した．本稿で扱った範囲では，データ分布と匿名化データの有用性の間の相関は非常に弱いという結果を得た．しかし，全ての実データの分布がべき乗型になる保証は無い．例えば，偏りのあるデータではステップ関数のような分布を持つことも考えられる．今後，べき乗型以外の現実的な分布の形を用いた実験を行い，匿名化データの有用性への影響をより網羅的に調べていく必要がある．

本稿では，1 属性データに Pk-匿名化，あるいは複数属性を持つデータに属性の連結化を用いた Pk-匿名化を適用した場合の有用性評価実験を行った．実際は，複数属性デー

[*1] ただし，そのような場合であっても Pk-匿名性が破られているわけではないことに注意．

タに対して，連結化を用いない他の Pk-匿名化[*2]を適用する場合も当然考えられる．したがって，模擬 1 属性データだけでなく，模擬複数属性データを用いた有用性評価実験を行うことも今後の課題の 1 つである．

参考文献

[1] Pierangela Samarati & Latanya Sweeney, “Generalizing Data to Provide Anonymity when Disclosing Information”, PODS, 1998.
[2] Latanya Sweeney, “k-Anonymity:A Model for Protecting Privacy”, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002.
[3] 五十嵐大， 千田浩司，高橋克巳，“k-匿名性の確率的指標への拡張とその適用例”，CSS，2009.
[4] Rakesh Agrawal, Ramakrishnan Srikant & Dilys Thomas, “Privacy Preserving OLAP”, SIGMOD, 2005.
[5] 五十嵐大，千田浩司，高橋克巳，“数値属性における，k-匿名性を満たすランダム化手法”，CSS，2011.
[6] 五十嵐大，長谷川聡，納竜也，菊池亮，千田浩司，“数値属性に適用可能な，ランダム化により k-匿名性を保証するプライバシー保護クロス集計”，CSS，2012.
[7] 菊池亮，五十嵐大，千田浩司，濱田浩気，“データ分布依存処理によって高い有用性を実現する確率的 k-匿名化”，SCIS，2013.
[8] 齋藤恆和，五十嵐大，菊池亮，廣田啓一，正木彰伍，“属性間の相関を考慮した撹乱再構築法の提案”，CSS，2014.
[9] UCI repository of machine learning databases, 1998, `http://archive.ics.uci.edu/ml/datasets/Adult`.
[10] 千田浩司，菊池亮，濱田浩気，五十嵐大，廣田啓一，冨士仁，高橋克己，“動的集合匿名化データの有用性評価と開示リスクに関する一考察”，CSS，2013.
[11] 正木彰伍，廣田啓一，齋藤恆和，菊池亮，濱田浩気，五十嵐大，千田浩司，“データ分布依存処理を用いた確率的 k-匿名化の有用性の検証”，SCIS， 2014.
[12] Rakesh Agrawal & Ramakrishnan Srikant, “Privacy-Preserving Data Mining”, SIGMOD, 2000.

[*2] 連結化を用いていないが，データ分布依存型撹乱 [7] も有用性維持が期待できる Pk-匿名化の 1 つである．