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

千田 浩司† 菊池 亮† 濱田 浩気† 五十嵐 大† 廣田 啓一† 冨士 仁†
高橋 克巳†

†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
chida.koji@lab.ntt.co.jp

あらまし パーソナルデータの利活用が注目される中，プライバシ保護は一層重要な課題となっている．パーソナルデータの開示リスクを抑制するプライバシ保護手法の一つに，k-匿名性やその派生指標に基づく集合匿名化が知られる．本稿では，要求に応じて動的に生成される複数の集合匿名化データの有用性評価と開示リスクの考察を行う．先ず複数の集合匿名化データの開示リスク指標である多重 k-匿名性と多重 Pk-匿名性について，研究用公開データセットを用いて実験的に有用性の比較評価を行う．次に複数の集合匿名化データの開示リスクについて，公的統計における集計表の開示リスクの考え方に基づき考察する．

Utility and Disclosure Risk of Dynamic Group Based Anonymized Data

Koji Chida† Ryo Kikuchi† Koki Hamada† Dai Ikarashi†
Keiichi Hirota† Hitoshi Fuji† Katsumi Takahashi†

†NTT Secure Platform Laboratories
3-9-11 Midori-cho, Musashino, Tokyo 180-8585, JAPAN
chida.koji@lab.ntt.co.jp

Abstract Privacy protection has been much attention with a change of the value of personal data. A group-based anonymization method such as k-anonymization is known as one of privacy enhancing technologies that control disclosure risk of personal data. In this paper, we estimate a utility of data satisfying multiple k-anonymity/Pk-anonymity. Besides, we consider the disclosure risk of multiple k-anonymized/Pk-anonymized data according to the theory of statistical disclosure control for tabular data.

1 はじめに

ICT の発達に伴い多種多様な大量の情報が容易に収集できるようになり，情報の利活用による新たな価値創造への期待が高まっている．しかしパーソナルデータ（個人に関する情報）[1]を扱う際はプライバシ保護に十分配慮する必要がある．

パーソナルデータの利活用に資するプライバシ保護技術として，SDC (Statistical Disclosure Control)[2]や PPDM (Privacy Preserving Data Mining) が近年注目を集めている．SDC は元々公的統計の研究であり，個体 (個人や組織) に関

[1]総務省「パーソナルデータの利用・流通に関する研究会」[1] 等，個人情報保護法で定義される個人情報に限らずより広い意味で用いられる．

[2]SDL (Statistical Disclosure Limitation) ともいう．

する情報を開示するリスクを抑制するための方法と定義される [2]．個体に関する情報はセンシティブなデータを含む場合があり，SDC では情報の開示によって特定の個体のセンシティブなデータが知られてしまう「開示リスク」の問題を扱う．一方 PPDM は，パーソナルデータそのものを他者に明かさず特定のデータマイニング結果を得る方法として，暗号とデータベースの各研究分野で独立に提案された [3, 4]．以降 PPDM は，パーソナルデータを扱うデータマイニングや統計等の分析においてパーソナルデータを保護する技術の研究として，SDC の研究も取り入れつつ発展してきた [5]．

SDC や PPDM では，パーソナルデータから得られるミクロデータ，集計表，基本統計量，およびその他の分析結果について開示リスクを抑制するための方法が研究されている．特に SDC では古くから集計表や基本統計量の開示リスクの問題を扱っている．これはミクロデータや各種分析結果の開示リスクがより複雑な問題であることが背景に挙げられるだろう．

ミクロデータの開示リスクについて，PPDM の代表的な指標として k-匿名性 [6] が知られる．k-匿名性はミクロデータから特定個人のデータを k 個未満に絞り込めないことを保証するための指標であり，これまで ℓ-多様性 [7] や Pk-匿名性 [8] 等様々な派生指標が提案されている．k-匿名性やその派生指標を満たすための開示リスク抑制手法を本稿では「集合匿名化」と呼び，集合匿名化によって作成されるミクロデータを集合匿名化データと呼ぶ．

開示リスクの抑制は一般にデータの変造や損失を伴いデータの有用性が低下するため，開示リスクを抑制しつつデータの有用性を高めることが課題となる．Kifer らは，k-匿名性や ℓ-多様性を満たすミクロデータに「マージナル」と呼ばれる情報を付加する方法を提案し，ミクロデータとマージナルの組に適用できる k-匿名性の拡張指標を与えた [9]．また筆者らは，k-匿名性および Pk-匿名性を複数のミクロデータの組に適用するための自然な拡張である，多重 k-匿名性 [10] および多重 Pk-匿名性 [11] を提案した．依存関係がある複数の集合匿名化データを開示できれば，有用性の向上が期待される．

本稿では，多重 k-匿名性や多重 Pk-匿名性に基づき，利用者の要求に応じて複数の集合匿名化データを開示するリスクの考察と有用性評価を行う．先ず多重 k-匿名性と多重 Pk-匿名性をそれぞれ満たすデータを研究用公開データセット [12] から作成し，情報量の指標である L1 距離，L2 距離，および Kullback-Leibler 情報量を用いて実験的に有用性の比較評価を行う．次に複数の集合匿名化データを開示するリスクについて，公的統計における集計表の開示リスクの考え方 [13] に基づき考察する．

2 準備

2.1 ミクロデータと集計表

ミクロデータは各個体の情報を表すデータである．表 1 に示すミクロデータの例は，1 行目の「e メールアドレス」「年齢」「学歴」「年収」が属性名を表し，2 行目以降の各行 (レコード) が各個体の属性値を表す．

表 1: ミクロデータの例

e メールアドレス	年齢	学歴	年収
aaa@xx	24	Bachelor	$40K
bbb@xx	25	Bachelor	$50K
ccc@yy	30	Master	$50K
abc@xx	30	Master	$50K
abb@zz	32	Master	$60K
bcc@xx	32	Doctorate	$100K

本稿では，ミクロデータの各属性を次のように分類する．

- **正識別子**: 個人を一意に識別できる属性，または属性の組．氏名，住所の組み合わせは無視できない確率で正識別子となる [2]．
- **準識別子**: 間接的に個人を識別できる属性．性別や年齢は間接的に個人の識別に利用できる [14]．
- **センシティブ属性**: 正識別子，準識別子以外で，他人にむやみに知られたくないセンシティブな属性．
- **非センシティブ属性**: 上記以外の属性．

なお記述の簡略化のため，以降，非センシティブ属性はミクロデータに含まれないものとする．

集計表はミクロデータの特定の属性値を集計したデータである．表 2 に示す例は,「年収」「年齢」の組の集計表となる．集計表は度数表と数量表に分けて考えることができる．度数表は特定の属性値を含むレコードの数 (度数) を表し，数量表は特定の属性値の合計 (数量) を表す．表 2 は度数表となる．数値属性は区間表示されることが多い．

表 2: 集計表の例

年収 / 年齢	[20-24]	[25-29]	[30-34]
[$0-$50K]	1	1	2
[$51K-$80K]	0	0	1
[$81K-$100K]	0	0	1

2.2 開示抑制

開示リスクは，個体の情報が正識別子と対応付く「識別リスク」，および特定の個体のセンシティブ属性の値が知られる「属性開示リスク」とされる [15]．したがって識別リスクを避けるためには，正識別子を開示しないことが前提となる．また集計表はミクロデータから作成されるため，一般に集計表の方が開示リスクは低いと考えられる．なお開示リスクは一般に，狭い範囲で真の値が推定される場合も考慮する．

ミクロデータや集計表の開示リスクを抑制するための基本的な処理として，撹乱的/非撹乱的なデータ操作が知られる．撹乱的な処理は置換や丸め法 (例えば値の繰り上げや四捨五入) 等があり，非撹乱的な処理は大域的再符号化，局所再符号化，局所秘匿等がある．大域的再符号化は表 2 の年齢の操作のように属性値の粒度を一律に粗くし，局所再符号化は一部の属性値の粒度を粗くする．局所秘匿は一部の属性値を非開示とする．詳細は [13],[5, 3 章] 等を参照されたい．

2.3 集合匿名化

k-匿名性は識別リスクの指標であり，k-匿名性を満たすための基本処理として，ミクロデータに同じ準識別子の値の組を持つレコードが他に $k-1$ 個以上含まれるよう非撹乱的なデータ操作を行う．一方 Pk-匿名性は k-匿名性を確率的なモデルに拡張した指標であり，ミクロデータの各レコードが $1/k$ 以上の確率で識別されないよう撹乱的なデータ操作を行う．すなわち k-匿名性や Pk-匿名性は何れも，特定個人のレコードを k 個未満に識別されないための指標である．また ℓ-多様性は属性開示リスクを扱った k-匿名性の拡張指標である．

依存関係がある複数のミクロデータの開示リスクの指標として，筆者らは多重 k-匿名性 [10] および多重 Pk-匿名性 [11] を提案した．これらは k-匿名性や Pk-匿名性の自然な拡張であり，依存関係がある複数のミクロデータが開示されても，任意のミクロデータについて特定個人のレコードを k 個未満に識別されないための指標である．

[10] では，複数のミクロデータについて，全てのミクロデータが k-匿名性を満たし，同一のセンシティブ属性を持たなければ，多重 k-匿名性を満たすと定義としている．するとどのミクロデータも同一の準識別子の値の組を持つレコードを k 個以上含み，センシティブ属性は識別に影響しない前提であるため，形式的には識別できない．ただし同一のセンシティブ属性が複数のミクロデータに含まれると，センシティブ属性の値がキーとなり識別されるリスクがある．

[11] では，全ての属性値について撹乱的なデータ操作を行い，全てのミクロデータが Pk-匿名性を満たせば，任意のミクロデータの各レコードを $1/k$ 以上の確率で識別できないことを示している．

3 実験評価

多重 k-匿名性や多重 Pk-匿名性を満たすデータの有用性について，UCI Machine Learning Repository[12] の Adult Data Set を用いて実験的に比較評価を行った．実験シナリオは，先ずミクロデータ利用者 (以降，単に利用者) は欲しい属性と属性値の粒度をミクロデータ保持者 (以降，単に保持者) に要求する．これを受けて保持者は，利用者の要求に応じて動的に多重 k-

匿名性や多重 Pk-匿名性を満たすミクロデータを作成し，利用者に提供する．

本実験では，属性と属性値の粒度を変えた複数のミクロデータを Adult Data Set から抽出し，多重 k-匿名性や多重 Pk-匿名性を満たすデータを作成して有用性の評価を行う．データの有用性の指標として，L1 距離，L2 距離，および Kullback-Leibler （KL) 情報量を用いた．

Adult Data Set は，1994 年のアメリカのセンサスデータベースから抽出された 15 の属性からなる 32,561 件の公開データセットである．本実験では ‘age’ (17 から 90 までの整数値)，‘education’ (16 通り)，‘occupation’ (15 通り)，‘race’ (5 通り)，‘sex’ (2 通り)，‘income’ (‘<=$50K/year’ と ‘>$50K/year’ の 2 通り) を用いた．income をセンシティブ属性とし，残りは準識別子とした．そして 3 人の利用者がそれぞれ以下のミクロデータ (要求データ 1〜3) を保持者に要求すると仮定した．

要求データ 1 age を 20 歳刻みの 5 種類 (20 歳未満，20-39 歳，40-59 歳，60-79 歳，80 歳以上) とし，age, education, occupation, sex, income の 5 属性 (属性値の組合せは 4,800 通り)．

要求データ 2 age を 5 歳刻みの 15 種類 (20 歳未満，20-24 歳，25-29 歳，...，85-90 歳) とし，age, education, occupation, sex, income の 5 属性 (属性値の組合せは 14,400 通り)．

要求データ 3 age を 20 歳刻みの 5 種類とし，age, education, occupation, race, sex, income の 6 属性 (属性値の組合せは 24,000 通り)．

k-匿名性を満たすデータと多重 k-匿名性を満たすデータの違いは，2.3 節で述べたようにセンシティブ属性の扱いである．多重 k-匿名性を満たすためにはセンシティブ属性を繰り返し用いることができず，ミクロデータの利用が制限される．多重 Pk-匿名性についても同様である．すると上記の 3 要求は同一のセンシティブ属性 income を含むため，そのままでは多重 k-匿名性や多重 Pk-匿名性を満たさない．そこでセンシティブ属性も準識別子と同様に撹乱的/非撹乱的なデータ操作を行うこととした．

上記の 3 要求は何れも，利用者の要求によって属性 age が大域的再符号化されている．しかし要求条件が多重 k-匿名性を満たさない場合，2.2 節で述べたような非撹乱的な処理がさらに必要となる．本実験では，利用者が要求した大域的再符号化をミクロデータに適用した後，同一の準識別子の値の組が k 個未満のレコードは削除 (局所秘匿) し，k-匿名性を満たすミクロデータを作成した．なお L1 距離，L2 距離，および KL 情報量 (以降，まとめて指標値とよぶ．指標値が小さいほど有用性が高い) は集計表に基づき計算されるため，レコードを削除しても比較できるよう集計表の度数を割合に正規化して評価する．また KL 情報量は度数の差が等しいとき (0 となるとき) 定義できないため，本実験では度数の差が等しいときは 0.1 とした．

Pk-匿名性や多重 Pk-匿名性を満たすミクロデータの作成は [17] に記載の方法にしたがう．すなわち，属性値ごとに確率的に別の属性値に変化させる撹乱的なデータ操作を行い，ベイズ推定によって集計表の度数を推定する．各度数は実数となるため，小数点以下を r としたとき，確率 r で繰り上げ，そうでなければ繰り下げることで整数に置き換える．ベイズ推定は反復的な処理を行うが，1 回前の処理結果との差が 0.001 以下になったとき収束したとみなして処理を終了させる．そして最後にベイズ推定によって得られた集計表から，ミクロデータを再設計する「疑似ミクロデータ」を作成する．なおセンシティブ属性については属性値を変化させていない．

Pk-匿名性や多重 Pk-匿名性を満たすデータは確率的な操作を伴うため，一般に実行ごとに値が異なる．そこで本実験では，各データを 5 個ずつ作成し，各指標について平均の指標値を求めた．

実験結果を図 1〜9 に示す．図 1,2,3，図 4,5,6，および図 7,8,9 はそれぞれ要求データ 1,2,3 の L1 距離，L2 距離，KL 情報量を示している．

先ず多重 k-匿名性/多重 Pk-匿名性を満たすデータをみると，k-匿名性/Pk-匿名性を満たすデータよりも指標値が何れも大きい．これはセンシティブ属性も撹乱的/非撹乱的なデータ操

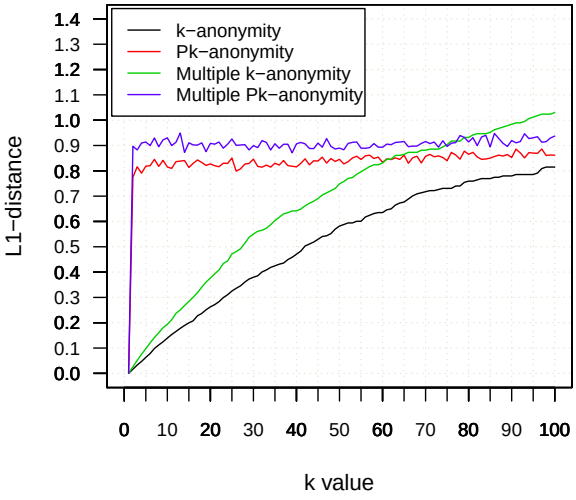

図 1: 要求データ 1 の L 1 距離

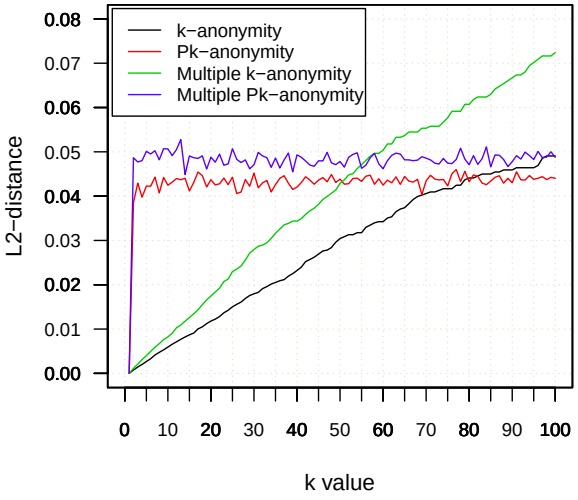

図 2: 要求データ 1 の L 2 距離

作を行っていることからも明らかであり，正しく結果に反映されていることが分かる．

次に k-匿名性と Pk-匿名性を比較すると，k の値が小さいときは概ね k-匿名性を満たすデータの方が指標値は小さいが，k の値が大きくなると逆転し，その差は広がる傾向がみられる．また Pk-匿名性を満たすデータは k の値が大きくなっても，指標値がほとんど変化していない．

多重 k-匿名性と多重 Pk-匿名性の比較についても，k-匿名性と Pk-匿名性の比較と同様の傾向がみられるが，何れも指標値が逆転する k の値がより小さい，興味深い結果が得られた．また多重 Pk-匿名性は，多重 k-匿名性と異なり指標の種類によらず同様のグラフとなっていることが分かる．これは集計表の各度数の誤差の偏りが少ないためと考えられる．

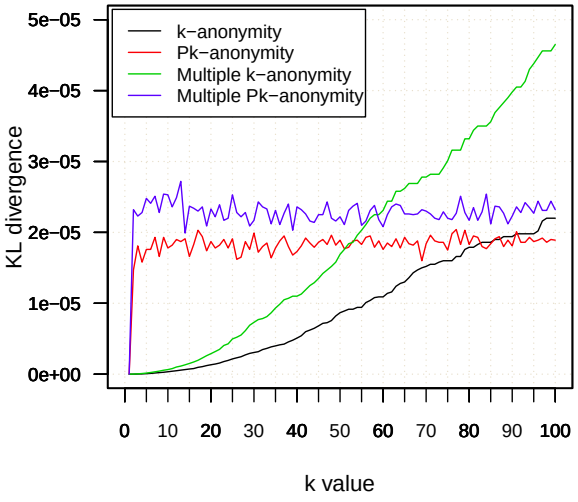

図 3: 要求データ 1 の KL 情報量

4 考察

複数の集合匿名化データの開示リスクについて，公的統計における集計表の開示リスクの考え方に基づき考察する．今回実験に用いたデータは，age を除き全てカテゴリ属性であり，age も区間に大域的再符号化している．したがって本実験で作成した集合匿名化データは多次元の度数表 (例えば 5 属性のミクロデータであれば 5 次元の度数表) とみなせる．

代表的な度数表の開示リスク指標として，閾値ルールが知られる [13]．閾値ルールは閾値未満の度数をセンシティブな値とする指標であり，撹乱的/非撹乱的なデータ操作により開示を抑制する．しかし度数表ではデータの有用性の観点から度数の合計値 (マージナル) も開示される場合があり，閾値未満の度数を秘匿することはそれほど容易ではない．例えば表 3 の度数表に閾値 3 の閾値ルールを適用することを考える．このとき，年収=‘[$81K-$100K]’，年齢=‘[30-34]’ の度数がセンシティブとなる．しかし表 4 のように単純にセンシティブな度数を局所秘匿する (1 次秘匿とよばれる) だけでは，マージナルの値との差分により復元される場合がある．そこで表 5 のように，センシティブでない度数の一部も秘匿し，センシティブな度数の復元を困難

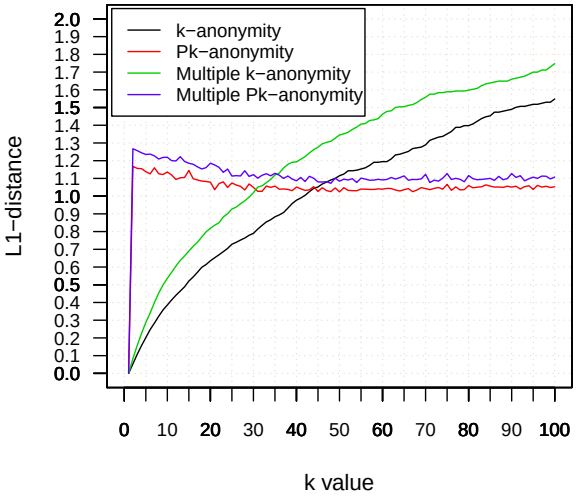

図 4: 要求データ 2 の L 1 距離

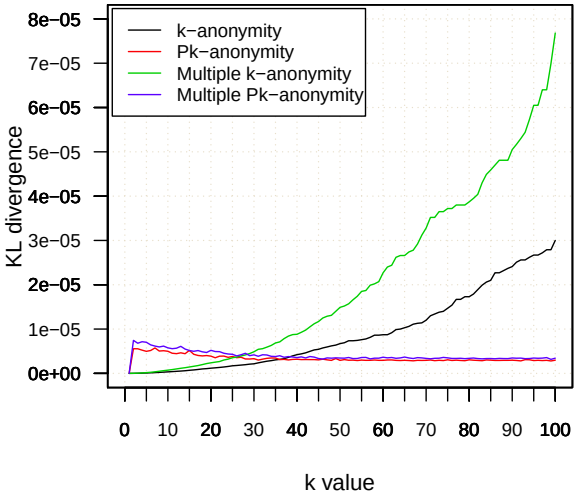

図 6: 要求データ 2 の KL 情報量

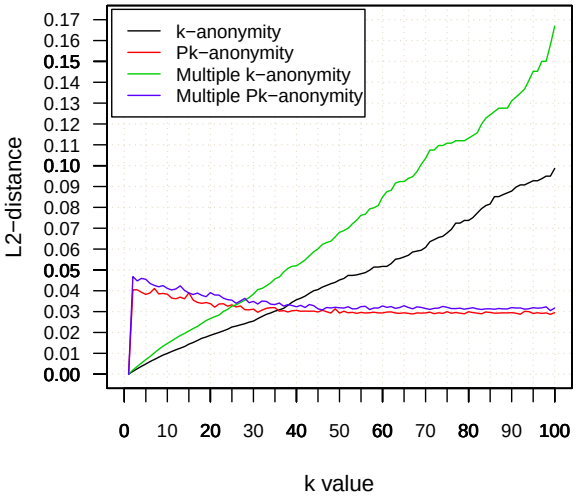

図 5: 要求データ 2 の L 2 距離

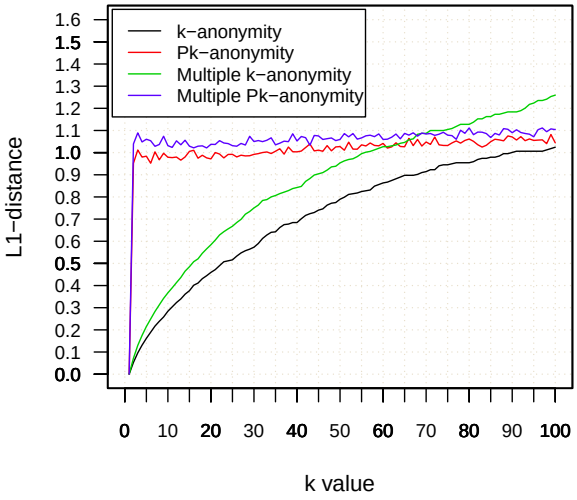

図 7: 要求データ 3 の L1 距離

とするアプローチ (2 次秘匿とよばれる) がしばしば用いられる．

1 章で述べた Kifer らのアプローチは，マージナルを用いた度数表の考えに基づいている．また本稿における実験シナリオでも，年収と年齢の組，年収，および年齢の 3 種類のミクロデータを要求することで，マージナルを含む度数表を作成できる．先に述べたように，本実験では利用者が要求した大域的再符号化をミクロデータに適用した後，同一の準識別子の値の組が k 個未満のレコードは削除 (局所秘匿) している．このデータ操作は度数表の 1 次秘匿であり，表 4 のようにマージナルの値との差分により削除したレコードが復元されるリスクがあることが分かる．

一方，利用者が要求した大域的再符号化では k-匿名性を満たさない場合，局所秘匿ではなく，更に大域的再符号化や局所再符号化を行い，確率的な操作で利用者が要求する属性値の粒度に合わせる方法も考えられる．すると度数に誤差が生じ，有用性は低下するものの，マージナルの値との差分を求めることが困難になる可能性がある．Pk-匿名性を満たすデータも同様のことがいえる．この種の撹乱的なデータ操作は，度数表の有望な開示抑制として期待される．例えば [13] では丸め法を用いた度数表の開示抑制の効果について考察されている．また Differential Privacy[18] も撹乱的なデータ操作を行う近年注

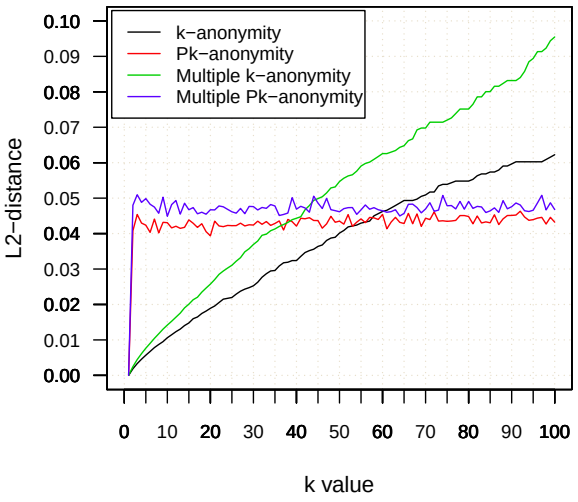

図 8: 要求データ 3 の L2 距離

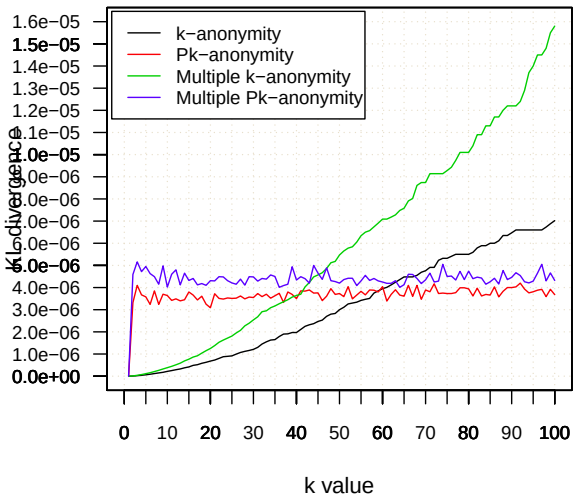

図 9: 要求データ 3 の KL 情報量

目されているアプローチである．

まとめると，多重 k-匿名性を満たすための局所秘匿は度数表における 1 次秘匿に相当し，マージナルの値との差分により度数が復元されるリスクがある．一方，多重 Pk-匿名性を満たすデータや，多重 k-匿名性を満たす確率的な操作を含んだデータは，度数表に誤差が生じるため，度数の復元可能性は自明でない．これらの理論的な解明は今後の課題だが，多重 Pk-匿名性を満たすデータは自明に度数を復元できず，かつ k の値が大きくなるにつれ多重 k-匿名性を満たすデータよりも有用性がより高まるという興味深い結果が得られた．

表 3: マージナルを含む度数表

年収 / 年齢	[20-24]	[25-29]	[30-34]	計
[$0-$50K]	15	10	6	31
[$51K-$80K]	8	17	4	29
[$81K-$100K]	10	5	2	17
計	33	32	12	77

表 4: マージナルを含む度数表の 1 次秘匿

年収 / 年齢	[20-24]	[25-29]	[30-34]	計
[$0-$50K]	15	10	6	31
[$51K-$80K]	8	17	4	29
[$81K-$100K]	10	5	x	17
計	33	32	12	77

5 今後の課題

本研究では，k-匿名性や Pk-匿名性を満たすミクロデータに着目し，依存関係がある複数のミクロデータを開示する有用性とリスクの評価および考察を行ったが，今回扱ったデータと開示リスクは限定的であり，今後さらなる検討が必要である．例えば，ミクロデータに数値属性が含まれる場合に，大域的再符号化を適用せず数値として開示することを想定していない．撹乱的なデータ操作には数値として開示する方法もいくつか知られており，数量表の開示リスクの考え方に基づき多重 Pk-匿名性の開示リスクを評価できる可能性がある．また ℓ-多様性等，属性開示リスクを扱った指標を動的集合匿名化に適用できるかどうかについても今後の大きな検討課題である．

謝辞

NTT ドコモの寺田雅之氏には，SDC の研究動向について色々とご教示頂いた．この場を借りて感謝の意を表したい．

参考文献

[1] 総務省「パーソナルデータの利用・流通に関する研究会」, http://www.soumu.go.jp/main_sosiki/kenkyu/parsonaldata/.

表 5: マージナルを含む度数表の 2 次秘匿

年収 / 年齢	[20-24]	[25-29]	[30-34]	計
[$0-$50K]	15	10	6	31
[$51K-$80K]	x	17	x	29
[$81K-$100K]	x	5	x	17
計	33	32	12	77

[2] 独立行政法人 統計センター: 統計データ開示抑制に関する用語集 改訂版 (対訳) 2005 年 8 月, http://www.nstac.go.jp/services/pdf/skk-yogosyu2.pdf.

[3] R. Agrawal and S. Srikant: Privacy-preserving data mining, *Proc. SIGMOD 2000*, pp.439–450, ACM, 2000.

[4] Y. Lindell and B. Pinkas: Privacy preserving data mining, *Proc. Crypto 2000*, LNCS 1880, pp.20–24, Springer-Verlag, 2000.

[5] C. Aggarwal and P. Yu: Privacy-preserving data mining: Models and algorithms, Springer-Verlag, 2008.

[6] L. Sweeney: k-anonymity: A model for protecting privacy, *Int'l Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, Vol.10, No.5, pp.557–570, World Scientific Publishing, 2002.

[7] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam: ℓ-diversity: Privacy beyond k-anonymity, *Trans. on Knowledge Discovery from Data*, Vol.1, No.1, ACM, 2007. and ℓ-diversity, *Proc. ICDE2007*, pp.106-115, IEEE, 2007.

[8] 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, CSS2009 論文集, pp.763–768, 情報処理学会, 2009.

[9] D. Kifer and J. Gehrke: Injecting utility into anonymized datasets, *Proc. SIGMOD 2006*, pp.217–228, ACM, 2006.

[10] 千田浩司, 五十嵐大, 高橋克巳, 濱田浩気, 菊池亮, 冨士仁: 集合匿名化クラウドの課題と対策, 電子情報通信学会論文誌, Vol. J96-A, No.4, 2013.

[11] 五十嵐大, 千田浩司, 濱田浩気, 菊池亮: 秘匿計算とランダム化によるハイブリッド匿名化システム, SCIS2012 論文集 (CD-ROM), 電子情報通信学会, 2012.

[12] UCI repository of machine learning databases, 1998, http://archive.ics.uci.edu/ml/datasets/Adult.

[13] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, Rainer Lenz, J. Naylor, E. S. Nordholt, G. Seri, and P.-P. De Wolf: Handbook on statistical disclosure control (version 1.2), Jan. 2010, http://neon.vb.cbs.nl/casc/handbook.htm.

[14] 竹村彰通: 個票開示問題の研究の現状と課題, Vol.51, No.2, pp.241–260, 統計数理, 2003.

[15] D. Lambert: Measures of disclosure risk and harm, *Journal of Official Statistics*, Vol.9, No.2, pp.313–331, Statistics Sweden, 1993.

[16] 千田浩司, 木村映善, 濱田浩気, 五十嵐大, 高木康彦, 冨士仁, 高橋克巳, 石原謙: 撹乱手法を用いたプライバシー保護医療情報分析の実験評価, 医療情報学 31(Suppl.). pp. 689–692, 2011.

[17] 永井彰, 五十嵐大, 濱田浩気, 松林達史: クロネッカー積を含む行列積演算の最適化による効率的なプライバシ保護データ公開技術, SCIS2010 予稿集 (CD-ROM), 電子情報通信学会, 2010.

[18] Dwork, C.: Differential Privacy, *Proc. ICALP 2006*, pp.1–12, Springer-Verlag (2006).