Vol.2011-CSEC-53 No.21
Vol.2011-IOT-13 No.21
2011/5/13

集合匿名化クラウドの課題と対策

千田浩司[†1] 五十嵐大[†1] 高橋克巳[†1]
濱田浩気[†1] 冨士仁[†1]

k-匿名性及びその派生指標に基づく集合匿名化は，個人データの安全な利活用を支えるプライバシ保護技術として理論と実用の両面で近年注目を集めている．ただし現時点では，プライバシと有用性のトレードオフや計算量に関する課題が必ずしも解決されておらず，利用は限定的である．本稿では，クラウドコンピューティングに代表される外部の豊富な計算資源を用いて集合匿名化を行う際の利点と課題，そして既存技術による各課題の解決の可能性について考察する．特に撹乱・再構築と呼ばれるプライバシ保護技術の適用が，集合匿名化の包括的対策として有望であることを示す．

Challenges in Group-Based Anonymization on Cloud

Koji Chida,[†1] Dai Ikarashi,[†1] Katsumi Takahashi,[†1]
Koki Hamada[†1] and Hitoshi Fuji[†1]

The group-based anonymization, which is an anonymization method based on the k-anonymity or its variants, has been attracting attention as a privacy protection technology ensuring the safe utilization of individual data on both sides of the theory and practical use. However, the problem concerning the trade-off of privacy and utility and the computational complexity is not necessarily solved so far. In this paper, we consider the advantage and the specific problem if we use an abundant outside computer resource such as cloud computing. We also consider the potential solution by existing technologies and show the perturbation & reconstruction method is especially promising as a comprehensive solution.

1. はじめに

ICT の発達に伴い，組織や個人が所有するデータの管理や利用の形態が年々変化してきている．特に最近では，ネットワークを介してデータの管理やオペレーションを行い，豊富な計算資源を利用することで利便性の高いシームレスなサービスを提供する，クラウドコンピューティングが注目を集めている．各種データが容易に収集され，活用できるような環境が整えば，データの利活用における新たな価値創造が期待できる．例えば医療分野では，EHR(Electronic Health Record: 電子健康記録) や PHR(Personal Health Record: 個人健康記録) をクラウド上で実現しようとする動きが各所で見られるようになった．

現在では，センサデバイスや通信環境，データストレージ等の高度化によって，データ収集においては利活用の基盤が整いつつある．しかし収集したデータを価値あるデータに加工する技術 (データクレンジング，データマイニング等) は発展途上であり，更なる技術の進展が望まれる．またデータの利活用におけるより複雑な課題としてプライバシ保護が挙げられる．1) では，個人の行動データや医療健康データといったパーソナルデータの利活用を促進し，社会や産業の発展に資することを目的とした様々な取り組みが行われた．その中でパーソナルデータの利活用におけるプライバシ保護として**集合匿名化**技術が推進され，プロトタイプ開発を通じた実証実験が行われた．ここで集合匿名化技術 (または単に集合匿名化) とは，複数のパーソナルデータが与えられたとき，あるパーソナルデータに対応する特定個人を k 人未満に絞り込むことができるかどうかを匿名性の指標とする **k-匿名性** (k-Anonymity)[2]，またはその派生指標に基づく技術の総称とする．k-匿名性は直観的に理解しやすく，現在最も代表的な匿名性の指標といえよう．ただし k-匿名性を満たす最適なデータに加工する問題は NP-困難である等，技術的課題も存在する．

本稿では，多種多様な大量のデータの集約拠点となるクラウド上でパーソナルデータを即時に集合匿名化する基盤，**集合匿名化クラウド**の実現に向け，その利点及び課題，そして各課題の対策として有望な既存技術を俯瞰する．特に，撹乱・再構築と呼ばれるプライバシ保護技術の適用が，本稿で挙げた課題の包括的対策として有望であることを示す．

2. 準備

2.1 用語

1 節で述べた個々のパーソナルデータを**レコード**と呼び，以下の属性 (変数) の値からなるものとする．

- **正識別子** (Formal Identifier): 個人を一意に識別できる属性．例えば氏名，住所のような属性の組み合わせは無視できない確率で正識別子となる[3]．
- **準識別子** (Quasi-Identifier): 間接的に個人を識別できる属性．性別や年齢のような属性は間接的に個人の識別に用いることができる[4]．
- **センシティブ属性** (Sensitive Attribute): 正識別子または準識別子以外で，個人のプライバシに関するもの等，他人にむやみに知られたくない属性．以降，センシティブ属性の値をセンシティブデータと呼ぶ場合がある．
- **非センシティブ属性** (Non-Sensitive Attribute): 上記以外の属性．以降，非センシティブ属性の値を非センシティブデータと呼ぶ場合がある．

表 1 に示すように，先頭行の各列に属性の名称が記載され，各行には先頭行に記載の属

†1 日本電信電話 (株) 情報流通プラットフォーム研究所
NTT Information Sharing Platform Laboratories

Vol.2011-CSEC-53 No.21
Vol.2011-IOT-13 No.21
2011/5/13

表 1 テーブル
Table 1 A table

E-mail (正識別子)	Age (準識別子)	Education (準識別子)	Zip Code (準識別子)	Annual Income (センシティブ属性)	Smoking (非センシティブ属性)
aaa@xx.com	24	Bachelor	53711	40k	Yes
bbb@xx.com	25	Bachelor	53712	50k	No
ccc@yy.com	30	Master	53713	50k	No
abc@xx.com	30	Master	53714	80k	No
abb@zz.com	32	Master	53715	50k	No
bcc@xx.com	32	Doctorate	53716	100k	Yes

表 2 分割表
Table 2 A contingency table

Age / Annual Income	[0 - 50k]	[51k - 80k]	[81k - 100k]
[20 - 24]	1	0	0
[25 - 29]	1	0	0
[30 - 34]	2	1	1

性にしたがった同一レコードの値が記載されるような表形式のデータを**テーブル**と呼ぶ．表 1 の例では，(aaa@xx.com, 24, Bachelor, 53711, 40k, Yes) が一つのレコードであり，計 6 個のレコードによってテーブルが形成される．本稿で定義したレコードやテーブルを**ミクロデータ**と呼ぶ場合がある．また表 2 のように，テーブル (ミクロデータ) から得られる複数の属性の間の関係を表したものを**分割表** (Contingency Talbes) と呼ぶ．テーブルや分割表における最小のデータ領域，すなわち各々の枡目を**セル**と呼ぶ．

2.2 統計的開示抑制

統計学の分野では古くから，データ公開において個人識別やプライバシ侵害のリスク低減を図る SDC(Statistical Disclosure Control: 統計的開示抑制) と呼ばれる研究が進められている．SDC では，準識別子とセンシティブ属性の値からなるミクロデータ，またはそれから得られる分割表や各種統計量 (以降，まとめて「ミクロデータ等」と呼ぶ) の公開によって生じる個人識別やプライバシ侵害のリスクを扱う．5) によれば，ミクロデータ等を公開するリスクは，身元開示 (Identity Disclosure) 及び属性開示 (Attribute Disclosure) であるとし，これらを合わせて開示リスク (Disclosure Risk) と呼ぶ．身元開示リスクは，レコードが正識別子と対応付くリスクである．属性開示リスクは，ミクロデータ等から特定個人のセンシティブデータが知られるリスクであり，プライバシ侵害のリスクと捉えることができる．SDC におけるチャレンジは，情報損失を最小限に抑えつつ開示リスクが十分小さくなるようミクロデータ等を加工することである．特にミクロデータの加工を本稿では**匿名化**と呼び，匿名化されたミクロデータを**匿名化データ**と呼ぶ．6) [Section 3.1] によれば，SDC は以下の 3 つの領域においてそれぞれ研究が進められている[⋆1]．

- **分割表保護** (Contingency Table Protection): 分割表の公開における開示リスクの回避を目的とする．下の二つよりも古く確立した領域である．
- **統計データベース** (Statistical Databases): 合計 (sum) や平均 (average) といった統計クエリをデータベースに要求するモデルにおいて，開示リスクを回避したクエリ回答を返すことを目的とする．複数のクエリ結果から生じる開示リスクを考慮する必要がある．
- **ミクロデータ保護** (Microdata Protection): ミクロデータの公開における開示リスクの回避を目的とする．上の二つよりも研究の歴史が浅い領域である．

分割表はミクロデータより得られるため，一般に分割表保護よりもミクロデータ保護の方が難しい．しかし最近ではデータマイニングや高度な統計分析等，予め与えられた分割表だけではデータ処理の入力として不十分な状況が起こり得る．また統計クエリを繰り返すことで一般に開示リスクも増すことから，論理的には安全な統計データベースと安全な匿名化データはほぼ同等のものと理解して良いとの考え方もある[4)]．

データを確率的に偽の値に変換することで匿名化を図る手法を撹乱 (Perturbation) と呼ぶ．またデータを偽の値にすることなく匿名化を図る手法を非撹乱 (Non-Perturbation) と呼ぶ．統計データベースに対する SDC として，クエリ監査 (Query Auditing) とクエリ推論制御 (Query Inference Control) が古くから知られているが，クエリ監査はクエリを制限 (Restriction) することによって開示リスクを回避する手法であり，非撹乱に属する．過去のクエリに基づいてクエリ回答の可否を評価する．クエリ推論制御はクエリ回答の撹乱や区間の回答 (Interval Answers) といった手法により開示リスクを回避する．クエリ推論制御に関する最近の結果として，**差分プライバシ** (Differential Privacy) が注目を集めている．差分プライバシは,「ある個人のレコードが含まれていてもそうでなくても出力が変化しない」ことを一定の基準の下で保証するための指標であり，統計量に Laplase ノイズを付加して基準を満たす手法が提案されている[7)]．

匿名化データを利活用する際は，匿名化によって元のデータのプライバシがどの程度保護されているか (リスク指標)，また匿名化データが元のデータと比べてどの程度有用性を保っているか (有用性指標) が重要な指標となる．以下では，既に述べた代表的なリスク指標である k-匿名性，及びその派生指標の一つである ℓ-**多様性**[8)] について簡単に紹介する．

- k-**匿名性** (k-Anonymity): あるテーブルのレコードについて，準識別子の値が等しいレコードが他に $k-1$ 個以上存在するとき，そのレコードは k-匿名であると呼び，身元開示リスクに対して安全と考えるリスク指標である．準識別子の値が等しいレコードの集合を準識別クラス (QI-Class) と呼び，全ての準識別クラスが k-匿名であるとき，そのテーブルは k-匿名性を満たすという．また k-匿名性を満たす処理を k-匿名化と呼ぶ．表 1 のテーブルについて，2-匿名性を満たすように匿名化したテーブルの例を表 3 に示す (非センシティブデータは略記している)．

⋆1 6) [Section 3.1] では 3 つの領域をそれぞれ Tabular Data Protection, Dynamic Databases, Microdata Protection としているが，本報告書では他の用語との混同や，他の文献の用語を考慮し，それぞれ分割表保護，統計データベース，ミクロデータ保護とした．

Vol.2011-CSEC-53 No.21
Vol.2011-IOT-13 No.21
2011/5/13

表 3 2-匿名テーブル
Table 3 A 2-anonymous table

QI-Class	Age (準識別子)	Education (準識別子)	Zip Code (準識別子)	Annual Income (センシティブ属性)
g1	[24 - 25]	Bachelor	[53711 - 53712]	40k
g1	[24 - 25]	Bachelor	[53711 - 53712]	50k
g2	30	Master	[53713 - 53714]	50k
g2	30	Master	[53713 - 53714]	80k
g3	32	GradSchool	[53715 - 53716]	50k
g3	32	GradSchool	[53715 - 53716]	100k

- **ℓ-多様性** (ℓ-Diversity): k-匿名性の拡張であり，全ての準識別クラスについて ℓ 種類以上の異なる「良い」レコードが存在するとき，属性開示リスクに対して安全と考える指標である．ℓ-多様性を満たす処理を ℓ-多様化と呼ぶ．同種攻撃 (Homogeneity Attack) と呼ばれる準識別クラスの偏りに基づく攻撃と，背景知識攻撃 (Background Knowledge Attack) と呼ばれる背景知識によって属性値の候補を絞り込むような攻撃を考慮している．これらの攻撃を回避する，ℓ 種類以上の異なる良いレコードの指標は複数存在する．表 3 の例では，全ての準識別クラス (g1〜g3) について 2 種類の異なるレコードが存在していることが分かる．

3. モ デ ル

3.1 基本モデル

パーソナルデータを利用して統計分析やデータマイニングを行う場合，図 1 に示すようなモデルが考えられる．先ずパーソナルデータのレコードが複数存在し，複数のレコードからテーブルが形成される．パーソナルデータのレコードを提供する主体をデータ提供主体と呼ぶ．テーブルを形成した主体がそれを自ら利用する場合は通常一次利用である．当該主体をデータ一次利用主体と呼ぶ．データ一次利用主体が形成したテーブルを流通させる場合は，A) ミクロデータとして流通，B) 分割表や統計クエリ回答として加工して流通，C) 統計分析やデータマイニング結果を流通，とに分けて考えることができる．流通したデータの利用は二次利用である．当該利用主体をデータ二次利用主体と呼ぶ．A) または B) において流通データを受け取った二次利用主体は，当該データを利用して自ら統計分析やデータマイニングを行うことが想定される．

本稿ではミクロデータの匿名化を主な対象とし，A) に着目する．2.2 節で述べたように，匿名化データに基づき分割表やクエリ回答の生成，または統計分析やデータマイニングを行うものと仮定すれば，A) を考えれば十分である．A) におけるミクロデータの集合匿名化は基本的にデータ一次利用主体が行う必要がある．

3.2 クラウド利用モデル

次に 3.1 節で与えた A) において，クラウド上で集合匿名化を行う場合，図 2 に示すようなモデルが考えられる．先ずデータ一次利用主体である事業体は，データ提供主体である顧客等のパーソナルデータ (レコード) からなるテーブルを作成し，クラウドに提供する．

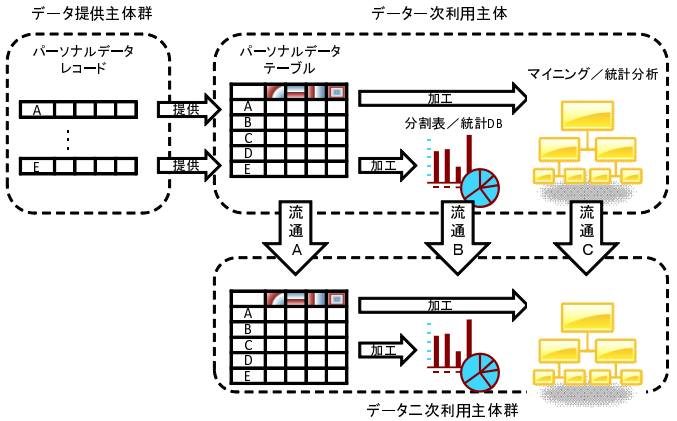

図 1 基本モデル
Fig. 1 A basic model

このときクラウドは受け取ったテーブルを集合匿名化し，生成された匿名化データを (事業体の指示に基づき) 流通させる． 流通した匿名化データを受け取った分析主体 (データ二次利用主体) は，当該匿名化データを利用して統計処理やデータマイニングを行う．

4. 利点と課題

図 1 の A) と図 2 では集合匿名化を行う主体が異なる．すなわち前者ではデータ一次利用主体であり，後者ではクラウドとなる．1 節で述べたように，ミクロデータを k-匿名性を満たす最適な匿名化データに加工する問題は NP-困難であり，一般に有用性を保つように集合匿名化を行う場合は計算コストが問題となり得る．したがって，集合匿名化を外部の豊富な計算資源であるクラウドに委託 (Outsource) することは，処理時間の短縮や計算資源の確保の面で利点になると考えられる．特に分析主体が対話的に集合匿名化データを要求する場合において，即時処理は必須要件と言える．

集合匿名化をクラウドに委託する利点は他にも考えられる．一つは 1 節で述べたように，クラウドは多種多様な大量のパーソナルデータの集約拠点となり，分析データの充実や分析精度の向上が期待できる．特に EHR や PHR で期待されているように，複数のデータ一次利用主体から受け取ったテーブルをプライバシを保護しつつ統合利用する環境として適していると言える．またレコードの追加や削除等，時間とともに動的変化するデータを扱うことも想定される．このような状況において，データ一次利用主体が作成するテーブルをクラウドが管理することで，分析主体は最新の状態のテーブルを分析できることも大きな利点になると考えられる．

一方，上記で挙げた利点はそれぞれ，集合匿名化特有の課題も潜在する．集合匿名化は一

Vol.2011-CSEC-53 No.21
Vol.2011-IOT-13 No.21
2011/5/13

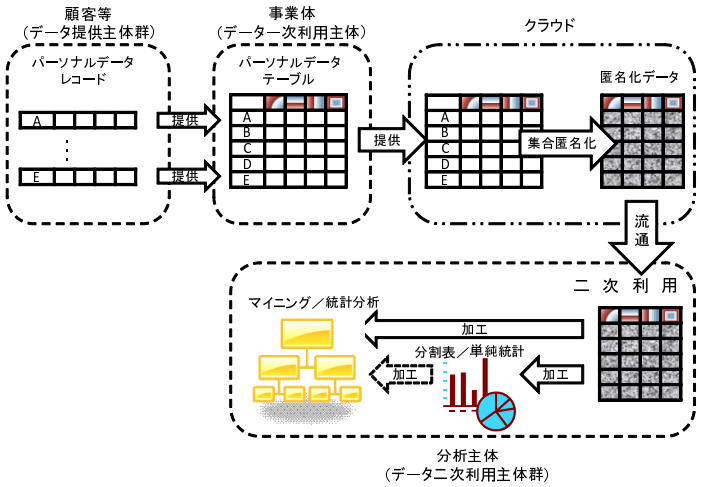

図 2 クラウド上の集合匿名化モデル
Fig. 2 A group-based anonymization model on cloud computing

般に，テーブルの属性 (列数) が多くなるほど準識別子の値の一般化の度合いが大きくなり，匿名化データの有用性が著しく低下する．具体的には，準識別子が多くなるほど，k-匿名性を満たす匿名化データの有用性が低下し，準識別子が少ない場合でも，センシティブ属性が多くなるほど，ℓ-多様性を満たす匿名化データの有用性が低下する．このような問題は**次元の呪い** (Dimentionality Curse)[9] と呼ばれる．すなわち，多種多様なパーソナルデータを集合匿名化する場合は，次元の呪いに直面しやすい．また動的変化するデータを扱う場合は，分析データ充実の反面，**逐次公開** (Sequential Release)[10] の問題が生じる．逐次公開は，動的変化するミクロデータを逐次匿名化して公開することで生じる開示リスクを扱った問題である．さらに集合匿名化をクラウドに委託する場合は，クラウドに対するデータ秘匿の問題も生じ得る．すなわち，データ処理を行う主体に対する守秘の問題である．これを本稿では**機密保持アウトソーシング** (Confidentiality in Oursorucing) の問題と呼ぶ．機密保持アウトソーシングは，クラウド上で集合匿名化を行う状況において，根本的に解決困難な問題と考えられる．しかし 5.3 節で後述するように，この一見相反する要件を解決し得る指標が最近になって提案された．

5. 対　　策

本節では，4 節で挙げた集合匿名化クラウド特有の課題について，既存技術による解決の可能性について考察する．

E-mail（正識別子）	Age（準識別子）	Education（準識別子）	Zip Code（準識別子）	AnnualIncome（センシティブ属性）	Blood Type（準識別子）	Disease（センシティブ属性）
aaa@xx.com	24	Bachelor	53711	40k	A	Flu
bbb@xx.com	25	Bachelor	53712	50k	O	Pneumonia
ccc@yy.com	30	Master	53713	50k	O	Dyspepsia
abc@xx.com	30	Master	53714	80k	B	Flu
abb@zz.com	32	Master	53715	50k	A	Flu
bcc@xx.com	32	Doctorate	53716	100k	AB	Dyspepsia

図 3 テーブル分解
Fig. 3 A table decomposition

5.1 次元の呪い

次元の呪いの有効な対策として，**効用ベースプライバシ保護** (Utility Based Privacy Preserving) と呼ばれる研究が 6) [Chapter 9] において紹介されている．これは特にサイズの大きなテーブルをなるべく有用性を損ねないように匿名化する，あるいは別のデータを補うことで有用性を向上させるアプローチである．効用ベースプライバシ保護の手法として，k-匿名性または ℓ-多様性の指標を満たしつつ，ミクロデータと匿名化データの分布の類似度向上を目的とした**マージナル付加** (Marginal Addition) が知られている．効用ベースプライバシ保護の多くは，与えられたテーブルと同じサイズの匿名化データを求めることを基本としているため，与えられたテーブルの属性数が多い場合は，有用性の維持があまり期待できない．これに対しマージナル付加は，匿名化データに別の有用データを追加することで有用性を高めることを特徴とし，特にサイズの大きいテーブルに対して効果的である．ただし属性数が特に多いテーブルは，マージナル付加を行う前に，テーブル分解を行うことが望ましい．すなわち図 3 に示すように，用途に応じてテーブルをいくつかの属性のグループに分解しておく．そしてそれぞれのグループについて集合匿名化を行う．

次に分解した各テーブルについて，マージナルの生成を含めた集合匿名化を実行する．マージナルは図 4 に示すように匿名化データを補完するデータ群であり，テーブルの一部や分割表を匿名化したデータである．図 4 の例では，匿名化テーブルとマージナルがそれぞれ 2-匿名性を満たしていることが分かる．マージナルは，何れも各レコードについてカウント数だけレコードのコピーを追加すれば，2-匿名性を満たす匿名化データとなる．しかしセンシティブデータが付加されている場合は注意が必要である．すなわち，図 4 の例では，匿名化テーブルとマージナル 2 で共通のセンシティブ属性である Annual Income から一意に結合できるレコードが存在し，全体として 2-匿名性を満たさないことが分かる．

上述のレコード結合の問題に対して，最近提案された興味深い関連研究として *ANGEL*[14] がある．ANGEL は，単調性 (Monotonicity) を持つ任意のリスク指標に適用可能かつ，有用性の高いテーブルの公開が可能な，マージナル公開 (Marginal Publication) アルゴリズムである．単調性は k-匿名性や ℓ-多様性が共通して持つ性質であり，テーブル T1 とテーブル T2 が当該リスク指標を満たせば，T1 と T2 を縦方向に統合したテーブル (水平統合テーブル) も (同一のパラメータで) 当該リスク指標を満たす，という性質である．一般化

Vol.2011-CSEC-53 No.21
Vol.2011-IOT-13 No.21
2011/5/13

匿名化テーブル(2-匿名)

Age (準識別子)	Education (準識別子)	Zip Code (準識別子)	Annual Income (センシティブ属性)
[27 – 30]	Bachelor	[53711 – 53713]	40k
[27 – 30]	Bachelor	[53711 – 53713]	50k
[27 – 30]	GradSchool	[53715 – 53716]	40k
[27 – 30]	GradSchool	[53715 – 53716]	80k
[27 – 30]	GradSchool	[537152– 53714]	50k
[27 – 30]	GradSchool	[537152– 53714]	100k

マージナル1(2-匿名)

Age (準識別子)	Count
27	2
28	2
30	2

マージナル2(2-匿名)

Education (準識別子)	Annual Income (センシティブ属性)	Count
Bachelor	40k	1
Bachelor	50k	1
Master	50k	2
Doctorate	80k	1
Doctorate	100k	1

図 4 匿名化テーブルとマージナル
Fig. 4 An anonymized table and marginals

を用いた匿名化アルゴリズムの多くはこの単調性を利用しており，ANGEL は k-匿名性や ℓ-多様性に加え，さらに多くのリスク指標，また未知のリスク指標にもアルゴリズムの修正なしに適用できる可能性がある．

5.2 逐次公開

逐次公開によって生じる属性開示リスクに関する指標の一つが *m-Invariance*[15] である．具体的には以下のようなものである．

- レコードが持つ属性を準識別子とセンシティブ属性に分ける．
- シグネチャとは準識別クラスのレコードが持つ特徴あるセンシティブ属性の集合とする．
- m-ユニークとは全ての準識別クラスが少なくとも m 個のレコードを持ち，全てのレコードは違うセンシティブデータを持つことをいう．
- m-invariance とは，逐次的な匿名化データの全ての公開が m-ユニークで，全てのレコードは毎回同じシグネチャを持つようにする (同じ準識別クラスに入れる)．
- m-invariance を満たせば，準識別子を知っている攻撃者もレコードの (ある個人の) センシティブデータを $1/m$ の確率でしか推定することができない．

m-invariance を満たす公開データは，属性開示リスクが生じる可能性があるデータの組み合わせに対して「偽のレコード」を挿入することにより，特定の個人が継続的に公開データに含まれていたときに起こる属性開示リスクを防ぐ．公開データは偽のレコードを含んだ匿名化データ (及び偽データを挿入した数を示す**補助関係表** (Auxiliary Relation)) となる．

15) では m-invariance の実現アルゴリズムが与えられている．これにより，レコードの削除や挿入があっても，逐次的な匿名化データの公開が可能になる．ただし，各公開のタイミング全てにおいてレコードの準識別子，及びセンシティブ属性は変更されないという制約がある．本アルゴリズムは k-匿名性や ℓ-多様性の指標に基づいて構築されている．

m-invariance の制約を補うために提案された手法が *HD-Composition*[16] である．これは不変センシティブデータ (犯罪歴や HIV が例に挙げられている) と変化するセンシティブデータが混在するときに，不変センシティブデータを手がかりに属性開示リスクが生じないように，不変センシティブデータを含む準識別クラスに対して L 種のおとり (Decoy) を常に配置しておく方法である．

5.3 機密保持アウトソーシング

機密保持アウトソーシングは，クラウド上で集合匿名化を行う状況において，根本的に解決困難な問題であることを 4 節で述べた．各主体が保持するミクロデータを互いに明かさず k-匿名化された統合テーブルを作成する分散 k-匿名化アルゴリズムもいくつか提案されているが (例えば 17))，クラウドの計算資源を利用する利点が損なわれてしまう．

本稿では機密保持アウトソーシングの対策として，**撹乱・再構築** (Perturbation & Reconstruction) に着目する．撹乱・再構築は，ノイズやデータ置換といった撹乱処理と，攪乱されたデータから特定の統計量のみを精度よく復元する再構築処理からなる．撹乱の度合いが大きいほど，撹乱前のデータを復元または推定することは困難となり，強いプライバシ保護を実現する．ただしトレードオフとして統計量の精度が一般に低下する．撹乱は，ランダムノイズの付加やランダムデータとの確率的な置換といった非常に軽量な処理である．その引き換えに，再構築処理が比較的計算コストのかかる処理となっている．この再構築処理をクラウド上で行うこととすれば，クラウドの計算資源を有効に活用することが可能となる．

ところで撹乱・再構築は k-匿名性に基づく手法ではないため，集合匿名化とは言えない．しかし近年，ある種の撹乱データの匿名性について，k-匿名性と等価である **Pk-匿名性**[18]，また ℓ-多様性と等価である **$P\ell$-多様性**[19] と呼ばれるリスク指標が提案されており，これに該当する撹乱・再構築も集合匿名化の一種と見なすことができるであろう．さらに撹乱・再構築はそのリスク及びデータ有用性における性能の高さにおいても注目されている．22) では，Pk-匿名性を満たす撹乱・再構築により，匿名性を保持しながら高い精度でのデータマイニングを行うことができることが報告されている他，20) では同じ撹乱・再構築が差分プライバシをも満たすことが示されている．また 21) においても，撹乱・再構築に相当する PRAM (Post Randomization) が k-匿名化と比較して高いプライバシ保護を示すことが報告されている．22) では，23) でも述べられている撹乱・再構築の欠点である，データ分析毎に特殊な再構築アルゴリズムを要するという欠点を補う，プライバシ保護疑似テーブル生成のアプローチも述べられている．プライバシ保護疑似テーブルは再利用可能であり，視覚的にもミクロデータと類似していることから，汎用的な分析が期待される．

5.4 包括的対策

最後に 5.1〜5.3 節で述べた個別の対策を包括的に実行することを考える．先ず次元の呪いの有望な対策であるマージナル付加と，逐次公開の有望な対策である m-invariance 等の指標を満たすアルゴリズムは，マージナルを匿名化データと見なすことで，m-invariance 等の指標を満たすよう各々の匿名化データを生成すれば良く，両立可能であると考えられる．一方，機密保持アウトソーシングの有望な対策である撹乱・再構築は，集合匿名化アルゴリズムではないため，基本的にはマージナル付加や m-invariance 等の指標を満たす従来のアルゴリズムと両立し得ない．そこで，次元の呪いや逐次公開の有望な対策となる集合匿名化

Vol.2011-CSEC-53 No.21
Vol.2011-IOT-13 No.21
2011/5/13

アルゴリズムを撹乱・再構築に置き換え，撹乱したデータを集合匿名化の指標で評価することを考える．図 2 のモデルを例に挙げると，先ず事業体によるテーブル分解は撹乱・再構築とは独立に実行可能である．匿名化テーブルとマージナルの作成は，事業体はミクロデータを撹乱してクラウドに送るだけで良く，クラウドは 22) で提案されている手法を用いれば，撹乱データから匿名化テーブル (プライバシ保護疑似テーブル) やマージナル (分割表) を再構築できる．ただし撹乱度合いについては，匿名化テーブル及びマージナルの構成に応じて適切に設定する必要がある．また m-invariance はセンシティブデータを $1/m$ の確率でしか推定できないという性質であり，$P\ell$-多様性も同様の性質を持つため親和性が高いと考えられる．ただし撹乱・再構築においてはレコード数が少ないほど撹乱度合いを大きくする必要があるため，予め最小レコード数を決めておき，それに合わせた撹乱を行う必要がある．

以上より，基本的には事業体が撹乱処理を行い，クラウドは再構築処理を行う構成を維持したまま，次元の呪いや逐次公開の対策を包括的に行えることが期待できる．ただし上述の考察は撹乱・再構築による包括的な対策の可能性についての示唆に留まっており，リスク指標を保証するものではない．また本稿ではページ数の制限で触れられなかったが，データ統合におけるリスク指標はより複雑なものとなる．これらの問題の解明は今後の課題である．

謝　　辞

本研究は，経済産業省 平成 22 年度産業技術研究開発委託費による「次世代高信頼・省エネ型 IT 基盤技術開発事業 (行動情報活用型クラウドサービス振興のためのデータ匿名化プラットフォーム技術開発事業)」の一環として行われた．この場を借りて，関係各位に感謝の意を表する．

参 考 文 献

1) 経済産業省：情報大航海プロジェクト，⟨http://www.meti.go.jp/policy/it_policy/daikoukai/igvp/index/index.html⟩ (参照 2011-04-05).
2) Sweeney, L.: 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).
3) 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⟩ (参照 2011-04-05).
4) 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 (2006).
5) Lambert, D.: Measures of Disclosure Risk and Harm, *Journal of Official Statistics*, Vol.9, No.2, pp.313–331, Statistics Sweden (1993).
6) Aggarwal, C. and Yu, P.: Privacy-Preserving Data Mining: Models and Algorithms, Springer-Verlag (2008).
7) Dwork, C.: Differential Privacy, *Proc. ICALP 2006*, pp.1–12, Springer-Verlag (2006).
8) Machanavajjhala, A., Kifer, D., Gehrke, J. and Venkitasubramaniam, M.: ℓ-diversity: Privacy beyond k-anonymity, *Trans. on Knowledge Discovery from Data* (TKDD), Vol.1, No.1, ACM (2007).
9) Aggarwal, C.: On k-anonymity and the Curse of Dimentionality, *Proc. VLDB 2005*, pp.901–909, ACM (2005).
10) Wang, K. and Fung, B.: Anonymizing Sequential Releases, *Proc. SIGKDD 2006*, pp.414–423, ACM (2006).
11) Xu, J., Wang, W., Pei, J., Wang, X., Shi, B. and Fu, A. W.: Utility Based Anonymization Using Local Recoding, *Proc. SIGKDD 2006*, pp.785–790, ACM (2006).
12) Fung, B. C. M., Wang, K. and Yu, P.: Top-down Specialization for Information and Privacy Preservation, *Proc. ICDE 2005*, pp.205–216, IEEE (2005).
13) Wang, K, Fung, B. C. M. and Yu, P.: Template-based Privacy Preservation in Classification Problems, *Proc. ICDM 2005*, pp.466–473, IEEE (2005).
14) Tao, Y., Chen, H., Xiao, X., Zhou, S. and Zhang, D.: ANGEL: Enhancing the Utility of Generalization for Privacy Preserving Publication, *Trans. on Knowledge and Data Engineering* (TKDE), Vol.21, No.7, IEEE (2009).
15) Xiao, X. and Tao, Y.: m-invariance: Towards Privacy-Preserving Re-publication of Dynamic Data Sets, *Proc. SIGMOD 2007*, pp.689–700, ACM (2007).
16) Bu, Y., Fu, A. W. C., Wong, R. C. W., Chen, L. and Li, J.: Privacy Preserving Serial Data Publishing by Role Composition, *Proc. VLDB Endowment 2008*, pp.845–856, ACM (2008).
17) Wang, K, Fung, B. and Dong, G.: Integrating Private Databases for Data Analysis, *Proc. ISI 2005*, pp.171–182, IEEE (2005)
18) 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).
19) 五十嵐大, 千田浩司, 高橋克巳: $P\ell$-多様性：属性推定に対する再構築法のプライバシーの定量化, CSS2010 論文集，pp.813–818，情報処理学会 (2010).
20) 五十嵐大, 千田浩司, 高橋克巳: Differential Privacy の数理的解析, CSS2010 論文集, pp.807–812, 情報処理学会 (2010).
21) Rebollo-Monedero, D., Forne, J. and Domingo-Ferrer, J.: From t-closeness-like Privacy to Postrandomization via Information Theory, *Trans. on Knowledge and Data Engineering* (TKDE), Vol.22, No.11, pp.1623–1636, IEEE (2010).
22) 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).
23) Aggarwal, C. and Yu, P.: A Condensation Approach to Privacy Preserving Data Mining, *Proc. EDBT 2004*, pp. 183–199, EDBT Association (2004).