7 クラウドストレージにおける個人情報の利活用とプライバシー保護

佐久間淳◉筑波大学，高橋克巳◉ NTT 情報流通プラットフォーム研究所

さまざまなオンラインサービスの発展とともに，個人の生活や行動にまつわる情報が収集されつつある．このような個人情報はその取り扱いに注意を要することから，その保管や解析は in-house で行うことが当然とされてきた．しかし個人情報の保管や解析のコストはその規模が膨大である場合には無視できない．近年，個人情報の漏えいリスクをコントロールしながら，その保管や解析をクラウドストレージに委託するプライバシー保護技術が注目を集めつつある．本稿では，クラウドストレージを中心とした計算モデルのもとで実現可能なプライバシー保護データ解析技術について概観する．

クラウドストレージにおける個人情報の意味とその利活用技術の分類

クラウドと個人情報

クラウドコンピューティングでは業務とデータの委託が発生する．クラウドに委託される業務には個人情報が含まれ得るので，クラウドにおいても個人情報は正しく保護されなければならない．クラウド環境で個人情報の保護は重要な課題である．さらにクラウドへは，業務の委託だけでなく，そこをハブとしたビジネス上の新たな価値創出への期待もある．

その１つが，個人情報の利活用である．クラウドでコンシューマ向けのサービスを行えば，利用者の履歴という個人情報が蓄積される．この情報は直接サービス改善に用いることができるだけでなく，さまざまなサービス間で活用できるのではないかという期待がある．クラウド環境における個人情報の活用のためには，委託と活用の両方の観点から問題の有無を考える必要がある．

プライバシー保護活用技術の分類

クラウド環境におけるプライバシーが保護された状態は以下の通りである．

- クラウドは預けられたデータからプライバシー情報を得ない(入力プライバシーの保護)
- 情報利用者はクラウドからプライバシー情報を得ない(出力プライバシーの保護)

この状態を満たすため，クラウドに機密性の高いハードウェアを用意し，その中で個人情報の計算を行う方法もあり得るが，この記事ではソフトウェアによる保護アプローチを紹介する．なお，プライバシー保護は単純なデータの入出力への対策のみで達成できないため，狭義のストレージだけでなくアプリケーション層まで含んだ議論を行う．

クラウドでのプライバシー保護活用技術を表 -1 に示す．入力プライバシーを守るためには，大きく分けて，個人情報を平文のまま安全な形式に加工する方法と，暗号による方法がある．出力プライバシーは，クラウドストレージの出力を監査し続けるという基本的な方法があるが，この記事では注目されている数学的なモデルを紹介する．

入力プライバシー		出力プライバシー
平文による保護	匿名化	差分プライバシー
	ランダム化	
暗号による保護	秘匿関数計算	
	高機能暗号	

表 -1　クラウドストレージのプライバシー保護活用技術の分類

7. クラウドストレージにおける個人情報の利活用とプライバシー保護

＊用語

ここで本稿で用いる用語を整理する．本稿が対象とする**個人情報**（personal information）とは個人と結びつけることができる情報である．我が国の「個人情報の保護に関する法律」で定められる個人情報が「個人を識別できる」もの（personal identifiable information）とされているよりは広い概念である．本稿の個人情報はパーソナル情報とも呼ばれる．

本稿では，データの保管やデータ解析処理の委託を受ける主体を**クラウド**，個人情報を有し，その情報をクラウドに委託するものを**情報保有者**，クラウドが個人情報を用いて行う操作を，**計算**または**データ解析**，データ解析の結果を利用するものを**情報利用者**と呼ぶ．ここでクラウドは委託された個人情報を特権的に閲覧可能とする．**プライバシー保護**とはクラウドに預けられた情報から個人のプライバシーに関する情報が漏れないようにすることである．

個人情報は各レコードが各個人に対応する関係データベースに保管されることが多いため，ここでは各行が１人の個人の情報に対応する表形式データを想定する．情報利用者がこの表形式データを得たときに，ある行が特定の個人に対応する情報であると知ることを**識別**，各個人をそれ単体で一意に識別可能な情報を**直接識別情報**（identifier）と呼ぶ．直接識別情報は，運転免許番号などの ID や顔写真，指紋などの生体情報が該当する☆1．

それ単体では必ずしも個人は識別されないが，複数を組み合わせることによって個人の識別に至る情報を**間接識別情報**（quasi-identifier）と呼ぶ．これには年齢，性別，住所など，個人に関する基本的情報が該当する．間接識別情報の属性値の組合せが表中で一意であるならば，直接識別情報と同等の識別力を持ち得ると認識する必要がある．

必ずしも識別力を持たないが，直接識別情報あるいは識別力の高い間接識別情報の組合せと結び付いた状態での公表が望ましくないとされる情報を**センシティブ情報**（sensitive information）と呼ぶ．これには持病，支持政党，行動／購買履歴などがある．たとえば持病 = 糖尿病，という属性値そのものの公表はプライバシー侵害を構成しないが，直接識別情報と結び付いた形での公表はプライバシーの侵害のおそれがある．

☆1 氏名は必ずしも一意識別性はないが，識別性はきわめて高く，多くの場合，直接識別情報として扱われる．

匿名化技術を用いた保護

個人情報はサービスの提供者が顧客に個別の対応を行う手がかりとして収集されることがある．顧客情報やサービスの利用履歴を利用したサービスの個人化（personalization）はその代表例といえよう．このような個人情報はマーケティングなどの商業目的や研究目的などにも利用可能なため，その情報を収集した情報保有者以外にも利用価値は高い．しかしプライバシー保護の観点からは，情報保有者が個人情報をそのまま情報利用者に開示することは問題がある．情報保有者が情報を得たときに，それぞれの情報が各個人と識別できないよう修正を加えることで，情報利用者に情報提供を可能にする匿名化技術の研究が進められてる．

クラウドストレージは，匿名化の観点からは情報保有者と情報利用者の仲介の役割を担う．情報保有者が保持する個人情報を匿名化された形式でクラウドストレージに保存し，クラウドはこれを情報利用者に引き渡す（図 -1）．本章ではこのような状況を

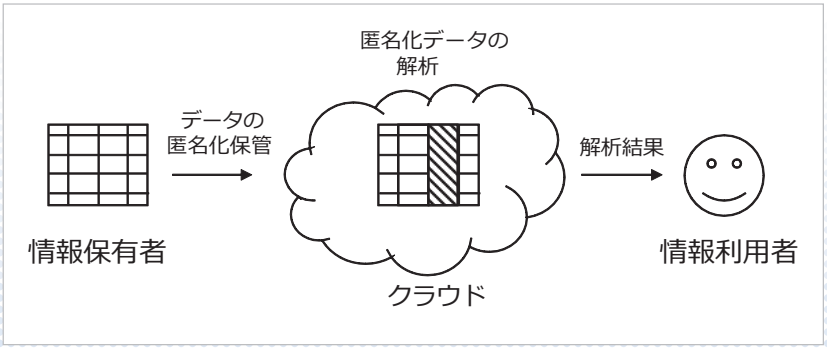

図 -1　クラウド環境における匿名化

想定し，匿名性についての定義を与えた上で，匿名化を達成するために必要な要件について考察する．

＊ クラウドストレージにおける個人情報の匿名性

情報保有者がクラウドストレージ上に個人情報を蓄積する動機は主に以下の 2 つである．

1 つは情報保有者の保持する個人情報が大規模であり，その保管および処理コストを下げるためにクラウドストレージを利用するケースである．このケースにおけるリスクは，クラウド自体による個人の識別である．もう 1 つは，情報保有者がクラウドストレージを介して個人情報を情報利用者に引き渡す場合である．このケースでは，個人情報を得た情報利用者による個人の識別リスクも同時に考慮する必要がある．いずれのケースも個人情報を手にした者による識別のリスクが問題であり，本質的には両者は同じ問題である．以降は，表形式データを得た情報利用者による識別のリスクについて検討する．表形式データにおける識別リスクに対応するために，クラウド上に保管するデータが満たすべき**匿名性**について考察する．匿名性とは，直感的には表形式データの各行が特定の個人と識別できないことを意味する．また表形式の個人情報が匿名性を満足するよう改変することを**匿名化**と呼ぶ．匿名性の達成には当然ながら直接識別情報を取り除く必要があるが（表 -2（左）），これだけでは不十分である．たとえば情報利用者が郵便番号 232-0011 の地域の 26 歳の住人について事前知識を持つ場合，情報利用者はこの住人がヘルニアを患っているという識別されたセンシティブ情報を取得する．このように直接識別情報が除去されていても識別や識別されたセンシティブ情報の漏えいは起こり得る．

これを防ぐためには，データの正確性を犠牲にして間接識別情報やセンシティブ情報を改変し推測を困難にする必要がある．**大域的符号化**はある属性のすべての値について，複数の変数のカテゴリを 1 つのカテゴリに統合する（例：表全体について喘息と結核を肺病に置き換える）．**局所的抑制**は特定の属性値について値を削除する（例：特定の属性値についてヘルニアを N/A に置き換える）．その他，属性値を行間で入れ替える**スワップ**，数値属性や階層構造のある離散属性の値を抽象化する**一般化**など，さまざまな操作が知られている [1]．

これらの操作を闇雲に適用しても匿名化が適切に達成されるわけではない．匿名化は一定の匿名性定義を達成するよう設計される必要がある．代表的な匿名性定義には，k-**匿名性** [2] や ℓ-**多様性** [3] が知られる．k-匿名性とは間接識別情報やセンシティブ情報からの**識別推定**に対する耐性を保証する．表形式データについて，間接識別情報の属性値の組合せが同じである行が，少なくとも k（>1）行存在していることを k-匿名性と呼ぶ．表 -2（左）第 1 行目の間接識別情報の組合せは（郵便番号 =232-0011，年齢 =26）であるが，この組合せはこの表においては唯一であり，情報利用者が間接識別情報について何らかの知識を持っていた場合，識別のリスクがあ

郵便番号	年齢	疾病
232-0011	26	ヘルニア
232-0015	34	腰痛
232-0017	27	腰痛
232-0012	45	鼻炎
232-0013	43	ぜんそく
232-0014	42	結核
232-0014	23	糖尿病
232-0014	24	糖尿病
232-0014	26	糖尿病

郵便番号	年齢	疾病
232-001x	[20-39]	ヘルニア
232-001x	[20-39]	腰痛
232-001x	[20-39]	腰痛
232-001x	[40-49]	鼻炎
232-001x	[40-49]	ぜんそく
232-001x	[40-49]	結核
232-0014	[20-29]	糖尿病
232-0014	[20-29]	糖尿病
232-0014	[20-29]	糖尿病

郵便番号	年齢	疾病
232-001x	[20-39]	ヘルニア
232-001x	[20-39]	腰痛
232-001x	[20-39]	腰痛
232-001x	[40-49]	鼻炎
232-001x	[40-49]	ぜんそく
232-001x	[40-49]	結核
232-0014	[20-29]	糖尿病
232-0014	[20-29]	成人病
232-0014	[20-29]	糖尿病

表 -2 （左）直接識別情報が削除された個人情報，（中）3- 匿名化された個人情報，（右）3- 匿名化／ 2- 多様化された個人情報

る（識別推定）．表 -2（中）は，郵便番号の下 1 桁の抑制および年齢の丸め（いずれも一般化操作）を行うことで 3- 匿名化を達成している．

ℓ - 多様性とはセンシティブ情報の**属性推定**に対する耐性を保証する．具体的には，k - 匿名性を持つ表形式データの，間接識別情報の属性値の組合せが同じである k 行について，そのセンシティブ情報の属性値のバリエーションが少なくとも ℓ （$1<\ell\leq k$）存在していることを ℓ - 多様性と呼ぶ．表 -2（中）の第 7-9 行目の間接識別情報の組合せは，いずれも（郵便番号 =232-0014，年齢 =[20-29]）であるが，この 3 行においてセンシティブ情報である疾病の属性値はすべて糖尿病である．もし，情報利用者が郵便番号 232-0014 に居住する 20 歳代の住人 3 人すべてを知っていたならば，情報利用者はどの行が誰かを識別することなく，全員がいずれにせよ糖尿病であることを知る（属性推定）．表 -2（右）は第 8 行目の疾病属性を一般化することにより，2- 多様性を達成している．

匿名性の達成は，データの効用とトレードオフの関係にある．過剰な一般化は強い匿名性を達成するが，そのようなデータの効用は低い．一方，データの効用を高く保つには，匿名性を犠牲にする必要がある．大規模データにおける最適な k - 匿名化の達成は自明ではなく難しい問題である．情報大航海プロジェクトの個人情報匿名化基盤[☆2]は，大規模データを処理可能な k - 匿名化フレームワークを含む．

ランダム化と再構築法

匿名化と隣接する概念にデータのランダム化がある．**ランダム化（撹乱）**とは，プライバシーにかかわる個人のデータに対してランダム性を与える変換を施すことで，変換前の個人のデータの推定を困難にする技術である．ランダム化には，ランダム値の加算（ノイズ付加），ランダムに選択した他の個人のデータとの交換（スワップ），ランダム値との置換などの方法がある．ランダム化は一般に非可逆操作であり，ランダム化したデータから元のデータは復元できないことから，平文のままデータを扱いつつ，プライバシーが守られることとなる．ランダム化技術の中には，個人のデータの復元は困難だが，統計量に関しては，特定の操作により，精度良い復元を可能とするものもある．この統計量の復元は統計学の分野で以前より議論されており，古くは 1960 年代から提案がある．

この方法をデータベース分野の研究から明らかにしたものが**再構築法（reconstruction method）**である．再構築法とは，ランダム化したデータに特定の操作を行うことにより，統計結果を比較的精度よく得る方法である．データベースの分野で Rakesh Agrawal らが 2000 年に提唱した，「プライバシー保護データマイニング（privacy preserving data mining, PPDM）」[4]を実現するためのモデルである．

再構築法ではデータベースのランダム化を行い，ランダム化したデータベースについてデータマイニングを実行した後，統計的推定によってデータマイニング結果の復元を行う（図 -2）．この操作を再構築と呼ぶ．再構築法では，データマイニング結果について，逆行列の計算やベイズ推定等のランダム化の影響を除くような操作を行い，真の値に近い結果を得ることができる．これはランダム化がなされたテーブルに属する個々のデータはノイズの影響を受けているが，テーブルの持つ統計量はランダム化アルゴリズムの性質に基づいて真の統計量に漸近するからである．

再構築法の安全性，すなわち個人識別の「されにくさ」の定量化は現在でも完全に解明されていないが，

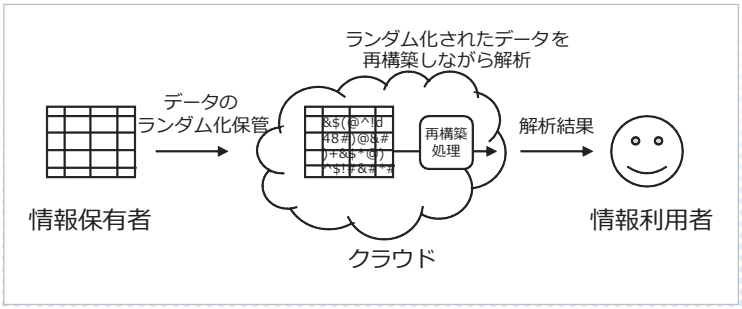

図 -2　ランダム化と再構築法

☆2 http://www.meti.go.jp/policy/it policy/daikoukai/igvp/cp2 jp/common/024/010/post-9.html

k-匿名性の尺度で評価できる再構築法が発見されている[5]．この研究ではあるランダム化を用いることで，k-匿名性と同等のプライバシーを持ちながら，再構築の計算を行うことができる．これは「ランダム化したデータベースにおいて，各レコードがある個人に対応する確信度（すなわち攻撃者から見た確率）が $1/k$ 以下である」ことを保証する．これらの研究により，再構築法の実用的な可能性が開拓されることが期待される．

暗号技術を用いた保護

クラウドのデータはストレージ，ファイル，データベースなどのレベルで暗号化することが可能で，ファイルやディスク装置の盗難などからデータを守ることができる．ただしクラウド上でデータ処理を行う場合は，暗号化データも一度は復号されるので，厳密な意味での入力プライバシーは保護できない．通常個人情報には厳重な運用管理を行って，セキュリティとプライバシーを守っている．しかし，データを暗号化するのであれば，その仕組みだけで入力プライバシーの保護の実現を望むのは自然な発想である．ある種の暗号プロトコルには，データを暗号化したまま一定のデータ操作を許すものが存在する．この性質を活かして，入力プライバシーを保護しながらデータを活用するクラウド上のサービスを設計することができる．

＊秘匿関数計算

秘匿関数計算（秘密計算）とはデータを暗号化したまま計算を行う技術で，計算のプロセスにおいてもデータを誰かに明かすことがない．このため，たとえばクラウドの管理者にも見られたくないプライバシー情報をクラウドに預けることができる．この原型は 1980 年代から知られているもので[☆3]，暗号化または秘密分散等で秘匿されたデータを秘匿したまま計算する．この方法は複数の計算主体による結託をしない協調計算を前提とし（マルチパーティプロトコルと呼ぶ），それぞれの計算主体が暗号の性質（準同型性等）や秘密分散の分散データの特性を使って部分的な計算を行いながら，最終的な計算結果のみを復元する．秘匿関数計算は論理回路演算の実行までを可能とするものがあり，これを秘匿回路計算と呼ぶ．

従来，秘匿関数計算は，情報保有者も協調計算の一端を担う前提で考えられてきたが，クラウド上で実現する場合は，クラウド上に結託をしない複数の計算主体を設置して，それらの複数主体全体を仮想計算システムとしてとらえることで実現することが可能である（図 -3）．情報保有者は仮想計算システムに対してデータを秘匿して入力し，システムは（あらかじめポリシーで合意された）計算結果のみを復元して解析者に出力する．このモデルを委託型秘匿関数計算と呼ぶ．

＊秘密分散に基づく秘匿関数計算

秘密分散とマルチパーティプロトコルに基づいて秘匿関数計算を構築することができる．この手法では情報保有者のデータが秘密分散されクラウドのマルチパーティに保存され，解析時にはクラウドにデータを明かすことなく処理が行われ，入力プライバシーが保護される．図 -3 および次ページのアルゴリズムは，クラウド上に 3 つの計算主体が存在するケースを想定している．マルチパーティに預けられたデータは 3 主体が結託しない限り秘匿される．掲出したアルゴリズム例は，秘密分散に基づく秘匿回路計算のプロトコルで，分散データに対する演算で加算と乗算が定義され，これを元に秘匿論理回路が

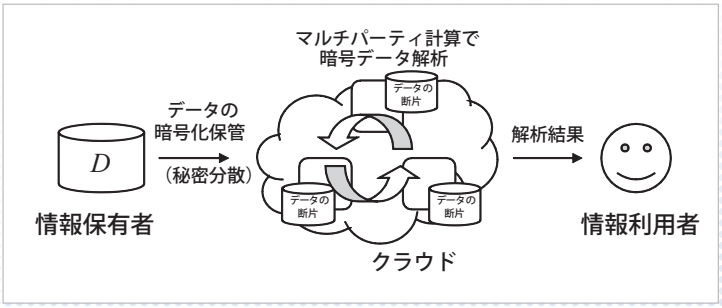

図 -3　秘密分散に基づく秘匿関数計算（委託型）

☆3 Yao “Protocols for secure computations”（1982）．

実現される[6]．秘匿関数計算は，データの分散保管ができる可用性の利点もあり，実用化に向けた段階の研究が，我が国や欧州で行われている．

アルゴリズムの直感的な説明

情報保有者を D，計算主体を P_i，$i=0, 1, 2$，情報利用者を U とする．

分散　D は入力 x を x_0, x_1, x_2 に分割し分散 (x_i, x_{i+1}) を作成，P_0, P_1, P_2 に送信．ただし x_0, x_1 は乱数で，$x_2=x-x_0-x_1$

復元　U が P_i のうち 2 者から分散を共有し，$x=x_0+x_1+x_2$ を用い復元．

加算　それぞれの P_i が加算結果の分散 $(a_i+b_i, a_{i+1}+b_{i+1})$ を計算．

乗算　$ab=(a_0+a_1+a_2)(b_0+b_1+b_2)$ であるので，P_i がそれぞれ a_ib_{i+1} を計算して，分散時と同様に乱数でマスクして共有，乗算結果の分散を得る．

0/1 の加算と乗算から AND/OR/NOT が容易に構成可能で，これを元に任意の論理演算を実現する．上記に計算の正当性検証も加える．

＊準同型性公開鍵暗号に基づく秘匿関数計算

公開鍵暗号の準同型性，すなわち暗号文のままの加算あるいは乗算ができる性質を利用しても，秘匿関数計算を実現することができる．

準同型性公開鍵暗号によって暗号化された数値は，その複号のためには秘密鍵が必要であるが，秘密鍵を知らなくても，その暗号化された数値に任意の値を加算することが可能である☆4．

この性質を活かし，複数のデータ保有者が持つデータを暗号化したまま，安全に決定木学習や k-means を計算するプロトコルが提案されている．このような準同型性公開鍵暗号を用いたデータ解析のためのプロトコルについては文献 7）に詳しく紹介されている．なお，近年に加算と乗算に関する準同型性を同時に有する暗号（完全準同型性）が発表され☆5，マルチパーティによらず単独の計算主体が秘匿回路計算を行えることへの可能性が示された．ただしこの技術はまだ理論上のもので，実用へはさまざまなブレークスルーが必要である．

＊高機能暗号を用いたプライベートストレージ

高機能な暗号を用いてプライベートなストレージを実現する研究が進められている．典型的なシナリオとしては，情報保有者がクラウドに高機能暗号で暗号化したデータを保管委託したときに，情報利用者はデータを復号することなくデータを検索をすることができ，またクラウドはデータの中身を見ることがない，といったことが可能になる．

図 -4 の例☆6 では，情報利用者の検索クエリ自体が鍵として働き，自分の暗号化データをクエリ鍵を用いて復号化せずに検索できる．ただしこの技術には利用方法や性能に制限があり，アーキテクチャ構成の検討を含めて興味深いチャレンジである．

出力のプライバシー保護

これまでに解説してきた匿名化，ランダム化，暗号化などは，個人情報自体の内容を直接開示せず計算を実行するための技術であった．これらは入力である個人情報漏えいの保護（入力プライバシーの保護）は問題視するが，計算の結果として得られる出力が引き起こす個人情報漏えいの保護（出力プライバシーの保護）は問題視しなかった．次の節で例示するが，出力による漏えいリスクも考慮する必要がある．

☆4 たとえば Paillier の加法準同型（1999）．
☆5 Gentry "Fully homomorphic encryption"（2009）．
☆6 Boneh "Public Key Encryption with Keyword Search"（2004）．

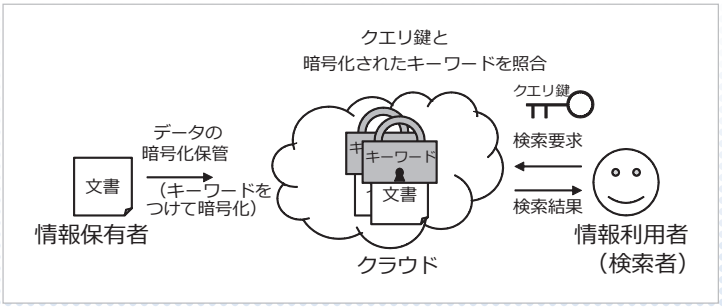

図 -4　キーワード検索暗号

出力による情報漏えいのリスクを理論的に扱う枠組みとして，差分プライバシー（differential privacy）が注目を集めている[8]．この章は差分プライバシーを中心に出力プライバシーの問題について議論する．

＊出力が引き起こす情報漏えい

ある会社Xの社員年収データベースを例に考察しよう．「この会社の入社3年目の社員の平均年収はいくらか？」という問合せに対し，「386.3万円」という応答があったとする．出力プライバシーの問題では，この「386.3万円」という応答が，各社員の給与情報をどれだけ漏えいするかに注意を払う．

会社Xの社員について何ら知識を持たない情報利用者にとっては，この応答からはクエリが示す内容以上の情報を取得することはできないため，この応答が特にプライバシー侵害を引き起こしたとは言えない．一方，もし会社Xに入社3年目の社員が3人のみ在籍しており（A，B，Cとする），かつクエリの発行者（＝情報利用者）がAさんだった場合はどうだろうか？　Aさんは当然自分の年収の正確な値（x_Aとする）を把握しており，よってBさんとCさんの年収の和は$x_B+x_C=386.3$万円$\times 3-x_A$であることを知るため，この開示はある種のプライバシー侵害を引き起こしている．では入社3年目の社員が100人いる場合はどうであろうか？　あるいは発行したクエリが「入社3年目の社員の最大年収はいくらか？」である場合はどうであろうか？　直感的にはデータベースサイズが大きいほうがプライバシー侵害の度合は弱く，平均値よりも最大値クエリのほうがプライバシーの侵害の度合いが強いように思える．

＊差分プライバシー

前節で得たプライバシー侵害の度合いに関する直感はどのように正当化されるだろうか？　出力からのプライバシー侵害は，データベースの規模，クエリの種類，情報利用者が持つ背景知識に強く依存しており，これらを考慮する必要がある．差分プライバシー[8]はこれらの疑問に一定の回答を与える．本稿では，クラウドが保持する統計データベースについて，情報利用者がその応答された統計値から知り得る情報を制限する方法について考察する．

データベースの出力が秘密の漏えいを引き起こすことを防ぐには，クラウドがその応答値にランダムなノイズを加えればよさそうである．ただし，ヒューリスティックにノイズを加えた場合には，だれのどのような情報がどのように守られたのかは不明確なままである．差分プライバシーはある安全性定義の下で，応答値に加えるランダムノイズの種類と分散について一定の理論的な基礎を与える．

差分プライバシーは直感的には以下のように説明される．「Aさんのデータがデータベースに含まれていようがいなかろうが，出力される統計値が大して変化しないのであれば，統計値を開示すること自体はAさんのプライバシーを侵害しない」．逆に言えば，Aさんがデータベースに含まれているかいないかを判別できないくらいの強さのランダムノイズを応答に加えることによって，応答がAさんのプライバシーを大して侵害しないことを保証しよう，というアイディアである．この大してという概念は，形式的には「Aさんのデータがデータベースに含まれている場合といない場合について，任意の統計値が返される確率の比が，ある数ϵについて，たかだか$\exp(\epsilon)$である」と定量化される．

＊ラプラスメカニズム

この差分プライバシーを達成するためには，応答値にラプラス分布で生成したノイズを加え，またそのラプラス分布の分散を計算対象である統計値の大域的敏感度（global senitivity）☆7に比例させればよいことが知られている（ラプラスメカニズム）．

D_1をAさんが含まれているデータベース，D_2をAさんが含まれていないデータベースとする．図-5（左）は情報利用者が平均年収をD_1，D_2に問い合わせたときに，応答する平均値にノイズを加えるプロセス（ラプラスメカニズム）を示している．

☆7 データベースサイズが大きい場合は小さい場合に比べ平均に対する敏感度が低いことから，前者は後者よりもプライバシー侵害の度合いが低いことが説明できる．また一般に平均関数はmax関数よりも敏感度が低いことから，やはり前者は後者よりもプライバシー侵害の度合いが低い．

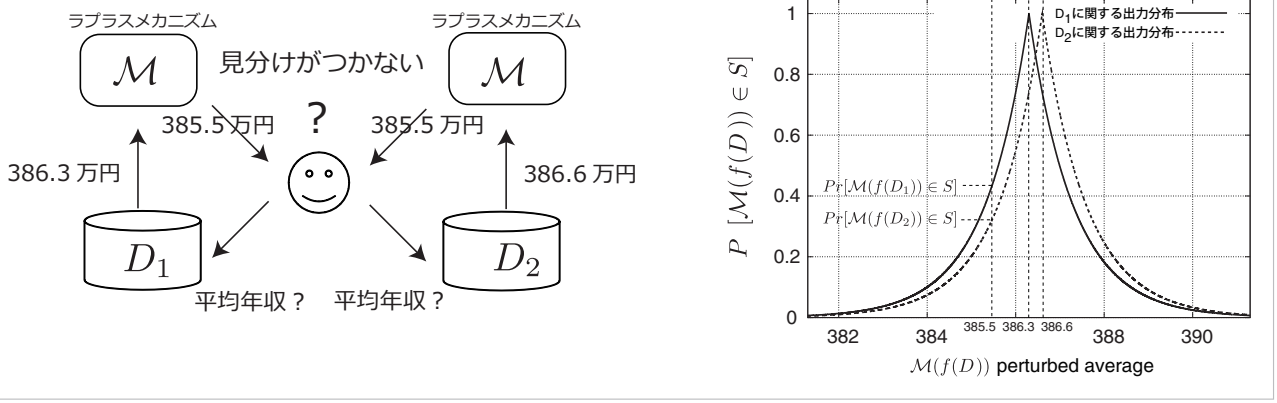

図 -5　(左) データベース D_1 および D_2 に対するラプラスメカニズムを通した問合せ，(右) データベース D_1 (問合せ"平均年収"に対する応答が 386.3 万円) とデータベース D_2 (問合せ"平均年収"に対する応答が 386.6 万円) について，ラプラスメカニズムを通した後の応答の分布.

図 -5（右）は情報利用者が D_1，D_2 からラプラスメカニズムを通じて得た応答(平均値)の確率分布を示している．大きい分散のノイズが加えられた場合，この 2 つの分布が与える確率密度の比は小さくなり，両者の区別はつきづらくなることからプライバシー保護の度合いはより強くなるが，応答値の正確性は低下する．一方小さい分散のノイズが加えられた場合，確率密度の比が大きく 2 つの応答値の区別がつきやすいことから，プライバシー保護の度合いは弱いものの応答の正確性は向上する．このように差分プライバシーは出力プライバシー保護の理論的枠組みを与えるが，カテゴリカルな値の応答への対応や対話的にクエリを発行する adversary への対応など，多くの open question が残されており，今後の発展が期待される．

クラウドにおけるプライバシー保護計算の委託モデル

本稿では個人情報のクラウドへの保管委託において，そのリスクを (1) クラウドあるいは情報利用者によるデータ解析計算が引き起こすプライバシーの侵害に注目する入力プライバシーと，(2) 情報利用者が受け取るデータ解析結果が引き起こすプライバシーの侵害に注目する出力プライバシーの 2 つに分類し，これを低減する技術について議論した．

個人情報を用いた多者間計算において，そのデータ解析計算をクラウドに委託することは，個人情報の漏えいリスクを増大させることから，伝統的なセキュリティ／プライバシー研究ではその必要性があまり認識されてこなかった．しかし現実問題として大規模個人情報を取り扱うデータ解析では，その大規模さゆえに in-house での保管／処理コストが高く，また高度なデータ解析計算の in-house での実現が困難であるなどの理由から，あえて個人情報の処理をクラウドへ委託する試みが模索されつつある．この章では，クラウド上に保管される個人情報を用いた計算全体を，その計算に関与するエンティティの観点から分類し，必要となる技術を整理する．

情報保有者が 1 人の委託モデル

最も単純なモデルは，情報保有者が 1 人であり，情報利用者と情報保有者が同一のエンティティであるケースである（図 -6（左））．この場合，情報保有者がデータ解析を in-house で行う限りプライバシー侵害のリスクは一切ない．しかし前述のように，情報の保管コストとデータ解析処理の維持コストがプライバシー侵害リスクを上回る場合は，これをクラウドに委託することに合理性があるといえる．

情報利用者と情報保有者が異なるエンティティであるケース（図 -6（中））もこれとほぼ同様のモデルで扱うことができるが，この場合クラウドは委託先としての役割のほかに，情報利用者と情報保有者の仲介役としての役割を持つ．より形式的には，これらモデルではあらかじめ定められたデータ解析計算

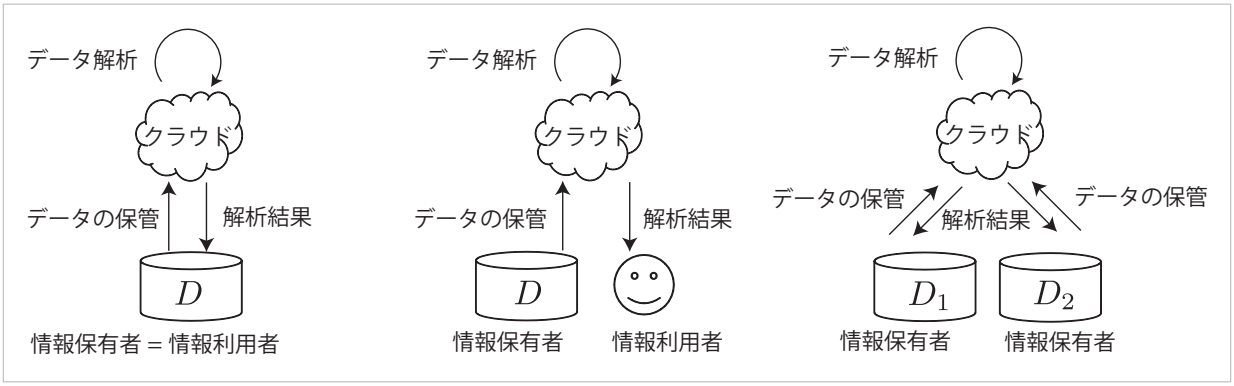

図 -6 （左）情報保有者が 1 人で，情報保有者＝情報利用者である委託モデル，（中）情報保有者が 1 人で，情報保有者，情報利用者である委託モデル，（右）情報保有者が 2 人以上の委託モデル．

について，以下の 2 条件の達成を目指す．

1. 情報利用者は，情報保有者の保有する個人情報について，（理想的には）データ解析の結果以外の情報を得ない
2. クラウドは，情報保有者の保有する個人情報について，（理想的には）何ら情報を得ない

ただし，図 -6（左）のケースでは，情報利用者＝情報保有者であり 1 番目の条件は考える必要がない．これを実現するには，匿名化，ランダム化，暗号化，いずれの技術も適用可能である．匿名化はクラウドが比較的多くの情報を取得することを許すが，クラウド上ではデータが平文で保管されるため任意のデータ解析計算を委託可能である．一方，ランダム化および暗号化では委託可能なデータ解析計算のクラスが制限される．ランダム化では，データ解析に要する処理時間はさほど大きくないことが多いが，処理結果の正確さは統計的にしか保証されない．暗号化は処理時間の増大が問題となるが，データ解析結果の正確さを保障できることが多い．

図 -6（中）のケースでは，情報利用者≠情報保有者であるため，上記に加え，情報利用者が得た出力結果が情報保有者のデータのプライバシーを侵害する可能性があり，差分プライバシーなどの利用が必要になる場合がある．

✻ 情報保有者が 2 人(以上)の委託モデル

このモデルでは，互いに開示できない情報を保持する 2 人以上の情報保有者がおり，両者が互いに情報を開示することなく，両者のデータのユニオンに対してデータ解析を行う（図 -6（右））．この問題設定は伝統的な二者間（あるいは多者間）プロトコルとして定義されるプライバシー保護データマイニングの問題として盛んに研究が行われてきたが，クラウドを仲介役として利用することはあまり意識されてこなかった．しかしこれまで議論してきたように，データが大規模である場合や情報保有者が解析技術を持たない場合には，クラウドを利用することに合理性がある．これに加え，クラウドはデータ解析計算について何ら情報を得ないエンティティとして振る舞い得るため，二者間プロトコルとして定式化するよりも自由度の高い設計が可能になる．形式的には，このモデルでは，以下の 2 条件の達成を目指す．

1. 情報保有者は，自分以外の情報保有者の保有する個人情報について，（理想的には）データ解析の結果以外の情報を得ない
2. クラウドは，すべての情報保有者の保有する個人情報について，（理想的には）何ら情報を得ない

ここでは，すべての情報保有者が情報利用者としても振る舞うことを想定したが，両者が別に独立している場合でも，議論はほとんど変わらない．

クラウドを仲介者としない多者間計算としてのプ

ライバシー保護データマイニングには多くの研究例がある[7]．これらの多くは情報保有者が2人(以上)の委託モデル上で直ちに利用できるが，情報保有者同士が常に通信可能であることを想定している点に難がある．情報保有者が2人(以上)の委託モデルにおいて想定するシナリオにおいて，クラウドへの委託を行う動機の大部分が，データ解析処理の委託にあることを考えれば，情報保有者同士が常にオンラインであるという想定は望ましいものではない．

繰り返しになるが，匿名化はデータを平文でクラウドに保管するため，情報利用者がオンラインであることを必要としないという意味で望ましい．匿名化は漏えいする情報量が比較的多いが，それが許容範囲内であれば，このモデルにおいては有力なプライバシー保護手法である．

より強い安全性を求める場合には，データを暗号文として保管する暗号化アプローチが有効であろう．準同型性暗号は比較的単純な計算には対応可能であるが，複雑なデータ解析への対応は難しい．また情報保有者と情報利用者はオンラインであることを要求する．クラウドが「結託しない」複数の計算主体から構成されている，という仮定を置くことができる場合，これらの計算主体間にデータを分散して委託し，計算を秘匿関数計算で行うことによって，情報保有者のオンライン性が不要となる．この仮定が許容可能であれば，秘匿関数計算も有望なシナリオとなる．

個人情報の高度活用に向けて

本稿では，クラウドストレージに保管される個人情報の利活用を巡って，さまざまな計算モデルとそのリスクについて議論してきた．プライバシーという概念は見る人によってとらえ方が異なるカメレオンのような存在であるといわれる．個人情報の扱いには確かに慎重さが求められるが，個人化サービスの発展には必要不可欠な資源でもある．漏えいリスクに敏感になりすぎるあまり，それを死蔵するのではなく，個人情報の提供者の信頼を失わない範囲で有効に活用することによって，個人情報の提供者と利用者が互いに利益を得ることは十分に可能であろう．本稿では入力プライバシーと出力プライバシーという2つの観点から，クラウドストレージに保管される個人情報の漏えいリスクをコントロールし，安全に利活用するためのさまざまな技術を解説した．実際のサービスの現場では，より多様な形態での活用が想定され，それに現実的に対応できる計算モデルと技術の展開が期待される．

参考文献

1) Aggarwal, C. C. and Yu, P. S. : *Privacypreservingdata Mining : Models and Algorithms*, Springer-Verlag New York Inc.(2008).
2) Sweeney, L. : *k*-Anonymity : A Model for Protecting Privacy, *World*, Vol.10, No.5, pp.557-570(2002).
3) Machanavajjhala, A., Kifer, D., Gehrke, J. and Venkitasubramaniam, M. : ℓ -diversity : Privacy Beyond *k*-anonymity, *ACM Transactions on Knowledge Discovery from Data(TKDD)*, Vol.1, No.1, pp.3(2007).
4) Agrawal, R. and Srikant, R. : Privacy-preserving Data Mining, *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, pp.439-450(2000).
5) 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，コンピュータセキュリティシンポジウム(2009).
6) 千田浩司，濱田浩気，五十嵐大，高橋克巳：軽量検証可能3パーティ秘匿関数計算の再考，コンピュータセキュリティシンポジウム(2010).
7) 佐久間淳，小林重信：プライバシー保護データマイニング，人工知能学会誌，Vol.24, No.2, pp.283-294(2009).
8) Dwork, C., McSherry, F., Nissim, K. and Smith, A. : Calibrating Noise to Sensitivity in Private Data Analysis, *Theory of Cryptography*, pp.265-284(2006).

(平成23年3月8日受付)

佐久間淳 ▌jun@cs.tsukuba.ac.jp

筑波大学コンピュータサイエンス専攻准教授．JSTさきがけ研究員（兼任）．機械学習・データマイニング研究と，セキュリティ・プライバシ研究の接点において，便利でフェアなサービスのあり方を探っている．博士(工学)．

高橋克巳(正会員) ▌takahashi.katsumi@lab.ntt.co.jp

日本電信電話（株）ＮＴＴ情報流通プラットフォーム研究所情報セキュリティプロジェクト“セプＧＬ”主幹研究員．情報検索とログデータマイニングの研究をし，社会科学と暗号を体験しプライバシー保護データ処理に熱中．博士(情報理工学)．