解　説

応般

# プライバシー保護データパブリッシング

南　和宏（統計数理研究所）

## 安全なデータ公開に向けて

近年 IT がさまざまな分野の公共サービス，ビジネスの情報基盤となり，個人レベルでもインターネットでのオンラインショッピングやモバイル端末を介したソーシャルメディアへの膨大な情報発信が行われている．我々の行動の詳細が膨大なディジタルデータとして記録され，その大量のデータにさまざまな分析を加えることで初めて実現できる高度な意思決定への期待が高まっている．しかし現状では組織間の壁で流通が阻まれ，莫大なビッグデータはそのポテンシャルを活かしきれていない．主な原因の 1 つは情報の機密性の問題であり，医療データ，商品の購買履歴など個人に関する情報を含んだデータセットを不用意に第三者に公開するとプライバシーの侵害につながる危険性がある．実際に 2006 年に起きた 2 つの事例，Netflix の映画のレーティングと AOL のサーチログ [3] の公開されたデータセットからそれぞれ個人情報の特定が報告されており，データ公開によるプライバシーの侵害が実際に起き得る危険性があることを広く一般に認識させた．

この問題を解決するため，近年プライバシー保護データパブリッシング（以下，PPDP）の研究が盛んになってきている．PPDP は個人情報を漏洩することなく有益な情報を公開することを実現する匿名化等のデータ加工技術である．もちろん従来より国政調査等，広く社会で統計データを共有することは政府，公共の組織を中心に長年行われてきたが，プライバシー保護の施策としては公開可能なデータに関するポリシーや利用目的を定めたガイドラインの策定およびそのコンプライアンス遵守が中心であった．そのため公開するデータが必要以上に制限され，またデータの供給先に対しデータ利用に関する信頼関係の構築が必要な場合が多い．それに対して PPDP では不特定多数に対して公開することを目的とし，厳密に定式化されたプライバシー指標を満足する技術的解決策を提供する．

本稿では，$k$- 匿名性，$l$- 多様性，$t$- 近似性，差分プライバシー等の代表的な PPDP のプライバシー指標およびその実現手法を解説し，後半で通常の $k$- 匿名化の手法の適用が困難であるトランザクションデータ，位置情報等の多次元データに対する PPDP 実現への課題，提案手法をまとめる．

## システムモデル

本稿で対象とする PPDP のシステムモデルを**図 -1** に示す．データ収集時にデータ加工者（パブリッシャー）は各個人からそれぞれの属性情報を収集する．属性の例としては，「年齢」，「性別」といった一般的属性に加え，「商品の購入」，「位置情報」といった個人の行動も属性となり得る．ここでは関係データベースのテーブルを想定し，各個人の属性値は割り当てられたレコードに格納される．次のデータ公開時にはテーブルを管理するデータ加工者がプライバシー保護のための匿名化等の処理を行い，不特定多数のデータ利用者に加工したデータセットを渡す．この PPDP のモデルにおいて，データマイニング等の高度なデータ分析を行うのはデータの受け手であるデータ分析者である．このモデルでは

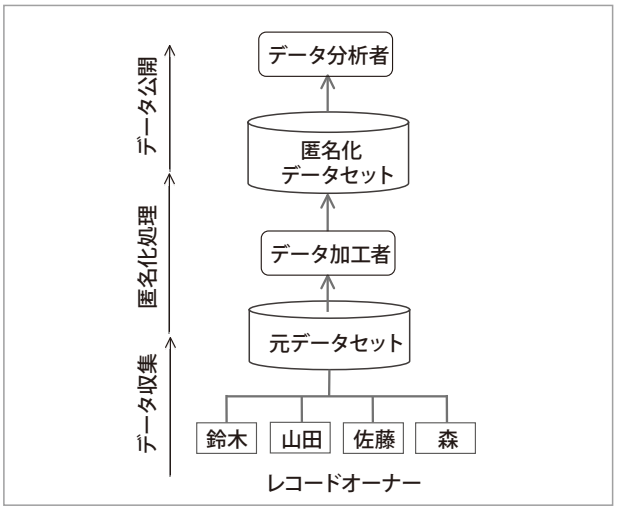


図 -1　プライバシー保護データパブリッシング（PPDP）のシステムモデル

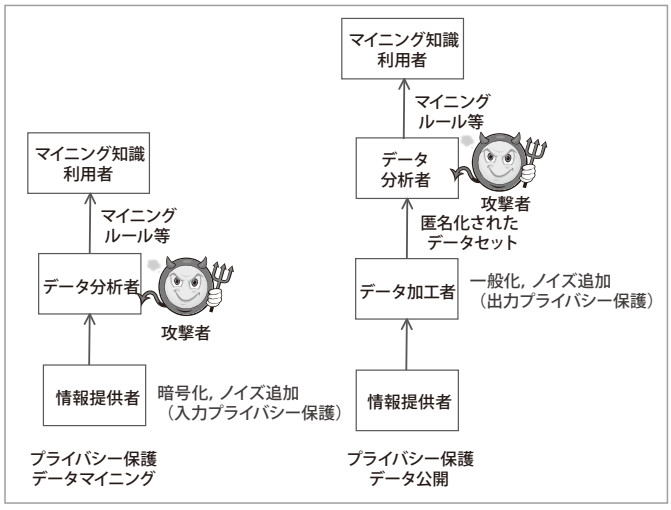


図 -2　プライバシー保護データマイニング（PPDM）とプライバシー保護データパブリッシング（PPDP）の攻撃モデルの比較

データ加工者は高度なデータ分析を自前で行う必要がなく，さまざまな外部のデータ分析者の分析スキルを享受できるという利点がある．その代わり，データ加工者はさまざまなデータ分析の対象となるように可能な限り元データに近いデータセットを公開する必要がある．

セキュリティの観点から見ると，データ加工者は信頼できる主体であり，情報を提供する個人はデータ加工者がプライベートな情報が漏洩しないように適切な処理をすることに関して信頼している．それに対してデータ利用者は悪意の攻撃者である可能性があり，入手したデータセットから個人の機密の情報を取得しようとする．ただしどのデータ利用者が攻撃者であるかデータ公開時に判別することは不可能であり，したがって信頼できる受け手だけに情報を渡す暗号化やアクセスコントロールの従来手法は適用できない．つまり PPDP の主要な研究課題は，与えられたプライバシーの要件を満足するという拘束条件のもと，いかに公開するデータセットの情報の有用性を最大化するかという最適化問題として定式化される．

PPDP にはプライバシー保護データマイニング（PPDM）[2] という類似の関連技術が存在し，そちらも活発な研究領域である．したがって両者の混乱を防ぐため**図 -2** に両者の攻撃モデルの比較を示す．PPDM の主目的は個人情報を秘匿したままデータマイニングの計算を行うことにあり，この場合に想定される攻撃者は中間のデータ分析者である．PPDM では暗号化またはノイズ追加の手法を用いてデータ分析者からの入力データのプライバシー保護を行う．データ分析者からの出力データであるアソシエーションルール等のマイニング結果は抽象度が高く，通常そのプライバシー保護は問題とはならない．それに対して前述のように，PPDP における攻撃者は匿名化されたデータセットを受け取るデータ分析者であり，データ加工者からの出力データのプライバシー保護が焦点となる．

## 匿名化手法

本稿ではプライバシー保護のためのデータ加工を行う技術を広く匿名化手法と呼ぶことにする．匿名化とは特定の個人とその機密の属性の関連（リンク）を断ち切る技術と考えてよい．この章では代表的なプライバシー指標を概説するが，PPDP では暗号化されていない平文のデータを公開するという性質上，統計的推論による機密情報の漏洩を完全に防ぐことは困難である．したがってこの漏洩のリスクのしきい値を指定することで安全性を定義するのが一般的である．この点は暗号技術のように計算論的安全性を保証する分野とは大きく異なる．また攻撃者の能力を定義する場合は，外部知識（たとえば利用可能

| 名前 | 職業 | 性別 | 年齢 | 病名 |
|---|---|---|---|---|
| 鈴木太郎 | 技術者 | 男 | 35 | 肝炎 |
| 木村次郎 | 技術者 | 男 | 35 | ねんざ |
| 高橋三郎 | 弁護士 | 男 | 38 | エイズ |
| 田中優子 | 作家 | 女 | 30 | インフルエンザ |
| 上田聡子 | 作家 | 女 | 30 | エイズ |
| 岡本英子 | ダンサー | 女 | 30 | エイズ |
| 中村和子 | ダンサー | 女 | 30 | エイズ |

**表 -1　医療データの例**
**（この表に現れる名前はすべて架空の名前である）**

な他の公開されたデータセット）を明確にすることが必須となる．

## □ 識別子削除による匿名化

PPDP を実現する上での考慮点を理解するために，一番単純な識別子削除による匿名化手法を最初に紹介する．説明には**表 -1** の医療データの例を使う．説明を単純化するために，ここでは「名前」という属性は個人を一意に特定する識別子と仮定しよう．

もし表 -1 の名前の属性の列を削除して匿名化すれば，**表 -2** のように各個人と病名の関連は失われ，プライバシー保護が実現できると考えるかもしれない．もしこのテーブルを入手した攻撃者が他の情報（他のデータセット）を持っていなければ正しいと言える．しかし通常，攻撃者は他の公開されたデータセットを自由に参照することができる．たとえば米国では**表 -3** のような投票者リストが公開されている．もし攻撃者が表 -3 の 1 番目のレコードの投票者の「高橋三郎」が表 -2 の匿名化された医療データにも現れるということを知っていれば，｛職業，性別，年齢｝の 3 つの属性を照合することで表 -2 の 3 番目のレコードと各属性値が一致することを見つけ，「高橋三郎」の病名が「エイズ」であることが判明してしまう．このように複数のデータセットに共通して現れる属性をマッチングすることでデータセット間のレコードを関連付ける攻撃を「レコードリンク攻撃」と呼ぶ．

実際にこれに似たレコードリンク攻撃による個人の情報漏洩は 1997 年米国マサチューセッツ州が医療データを公開する際に起きている．名前，ソーシャルセキュリティ番号，住所，電話番号等個人を特定する属性は取り除かれたにもかかわらず，誕生日，郵便番号，性別といった情報を州の投票者のリストと付き合わせることで当時の州知事の病名が特定されてしまったのである．当時のマサチューセッツ州の場合，生年月日と郵便番号の組合せで 97％の住人を特定することが可能であった．このように個人の特定のために間接的に使われる属性は「準識別子」と呼ばれる．

| 職業 | 性別 | 年齢 | 病名 |
|---|---|---|---|
| 技術者 | 男 | 35 | 肝炎 |
| 技術者 | 男 | 35 | ねんざ |
| 弁護士 | 男 | 38 | エイズ |
| 作家 | 女 | 30 | インフルエンザ |
| 作家 | 女 | 30 | エイズ |
| ダンサー | 女 | 30 | エイズ |
| ダンサー | 女 | 30 | エイズ |

**表 -2**
**識別子が削除された医療データ**

| 名前 | 職業 | 性別 | 年齢 |
|---|---|---|---|
| 高橋三郎 | 弁護士 | 男 | 38 |
| 岡本英子 | ダンサー | 女 | 30 |
| 木村次郎 | 技術者 | 男 | 35 |
| 上田聡子 | 作家 | 女 | 30 |
| 鈴木太郎 | 技術者 | 男 | 35 |
| 田中優子 | 作家 | 女 | 30 |
| 中村和子 | ダンサー | 女 | 30 |

**表 -3**
**投票者リスト**

最後にレコードリンク攻撃が成立しない場合でも個人の機密情報が漏洩する危険性があることを指摘しておく．表 -3 の 2 番目のレコードの「岡本英子」については表 -2 の 6 番目，7 番目の 2 つのレコードと準識別子の値が一致するため，2 つのレコードのどちらと実際に関連づけられるかは分からない．しかしどちらと対応するかにかかわらず 2 つのレコードの「病名」が「エイズ」であるため，機密情報である病名が漏洩してしまう．このように特定の個人とその機密属性を対応付ける攻撃は「属性リンク攻撃」と呼ばれる．

## □ $k$- 匿名化

$k$- 匿名化は前節で紹介したレコードリンク攻撃への対抗策として考案された．この場合の攻撃者はすでに流通しているデータセットから標的とする個

| 識別子 | 準識別子 | 機密情報 | 非機密情報 |
|---|---|---|---|

表 -4　レコードの属性の分類

人の属性情報を入手し，公開されたデータセットから候補となるレコードを絞り込む．$k$- 匿名化では攻撃者が候補のレコードを $k$ 個以下に絞り込めないことを保証する．

$k$- 匿名化では攻撃者が利用し得る外部知識を準識別子の概念に基づき明確に規定している．公開するデータセットのテーブルの属性は**表 -4** のように 4 つに分類される．識別子は米国のソーシャルセキュリティ番号のように直接個人を特定する情報，準識別子は住所等，間接的に個人を特定するような属性の集合，機密情報は，収入や病名といった個人のプライバシーに関する情報であり，これら 3 つのグループに属しない属性が非機密情報となる．$k$- 匿名化では攻撃者が各個人の準識別子のデータをすべて外部知識として知っていると仮定する．これはかなり保守的な仮定であるが，どのようなデータセットがすでに流通しているか想定することが困難であり，最悪の場合を想定する必要があるためである．

$k$- 匿名化の最初のステップとして，データセットの各レコードから識別子に相当する属性を取り除く．次にレコードリンク攻撃を防ぐため，準識別子の属性に対して一般化の処理を行い，各準識別子のユニークな組合せに対して，$k$ 個以上のレコードが存在するようにする．この準識別子に対する一般化の処理により，レコードリンク攻撃で対象となるレコードの数を $k$ 個以下に絞り込むのを防ぐことができる．つまり攻撃者がすべての個人の準識別子属性を知っていたとしても，$1/k$ 以下の確率でしか標的とするユーザのレコードを特定することができないことになる．ただし機密情報および非機密情報の属性値については一切変更する必要がないことに留意されたい．

通常，準識別子の属性値を一般化するため，準識別子の各属性についてドメイン一般化階層を定義する．**図 -3** に「職業」と「年齢」の属性の例を示す．「職業」のような分類属性は階層の上位にいくに従って

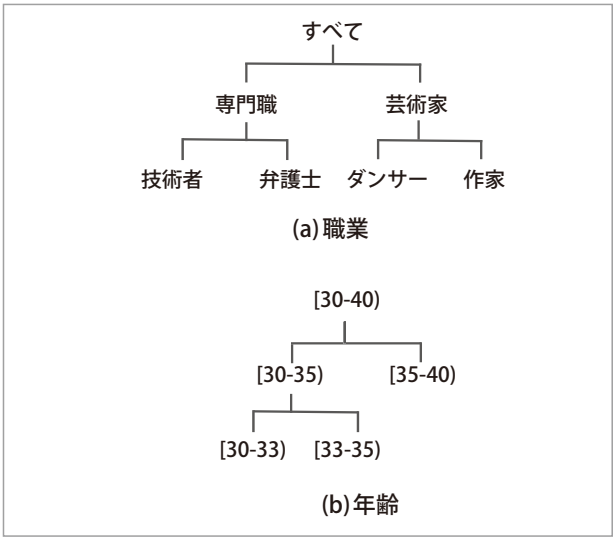


図 -3　ドメイン一般化階層の例

| 職業 | 性別 | 年齢 | 病名 |
|---|---|---|---|
| 専門職 | 男 | [35-40) | 肝炎 |
| 専門職 | 男 | [35-40) | ねんざ |
| 専門職 | 男 | [35-40) | エイズ |
| 芸術家 | 女 | [30-35) | インフルエンザ |
| 芸術家 | 女 | [30-35) | エイズ |
| 芸術家 | 女 | [30-35) | エイズ |
| 芸術家 | 女 | [30-35) | エイズ |

表 -5　3- 匿名化された医療データ

より一般的な分類名となる．また「年齢」のような数値データは値の取り得る範囲に従い，階層を構成することになる．

**表 -5** は表 -1 の医療データを 3- 匿名化した表である．この場合の準識別子は｛職業，性別，年齢｝の 3 つの属性の組合せであり，3 つのレコードと 4 つのレコードのグループに分割されている．準識別子の属性値に対して十分な一般化処理を行えば，$k$- 匿名性を満たすテーブルを作成することは容易であるが，データの有用性が失われてしまう．したがって $k$- 匿名性を満足するデータセットの中で定められたデータ有用性の指標を最大化するデータセットを選択する最適化アルゴリズムの研究が長年にわたり活発に行われている．データの有用性に関しては，もちろん想定するデータ分析方法ごとに指標が変わってくる．しかし PPDP では特定の分析手法を仮定せず，「最小のゆがみ (Minimal Distortion)」と呼ばれるデータの劣化の一般化処理の数で定量化する比較的単純な指標が用いられることが多い．

| 職業 | 性別 | 年齢 | 病名 |
|---|---|---|---|
| 芸術家 | 女 | [30-35) | エイズ |
| 芸術家 | 女 | [30-35) | エイズ |
| 芸術家 | 女 | [30-35) | エイズ |

**表 -6**
**病名が同一の 3- 匿名化された医療データの例**

| 職業 | 性別 | 年齢 | 病名 |
|---|---|---|---|
| 芸術家 | 女 | [30-35) | エイズ |
| 芸術家 | 女 | [30-35) | はしか |
| 芸術家 | 女 | [30-35) | 胃潰瘍 |

**表 -7**
**3- 多様性を満足する医療データの例**

$k$- 匿名化の最適化アルゴリズムは大まかに以下の 2 つに分類される．1 つはすべての一般化処理の組合せの中からデータ有用性の最大のものを見つける最適匿名化 (Optimal Anonymization) アルゴリズム [7] のグループであるが，一般に指数時間の計算が必要である．そこで局所的最適性のみを保証する局所最小匿名化 (Locally Minimal Anonymization) アルゴリズムも活発に研究されている．手法としては，効率性を重視した欲張り (Greedy) アルゴリズム，最も一般化された状態から特殊化するトップダウン型アルゴリズム，元データの状態から一般化していくボトムアップ的手法が検討されている．

最後に準識別子の選択に関して注意を喚起したい．ここまでの議論ではデータ公開者が準識別子を適切に選択できるという前提で話を進めてきた．しかし現実に正しい選択をすることはそれほど容易ではない．一般には攻撃者があるユーザについて知り得る属性はすべて準識別子に分類する必要がある．もし準識別子とすべき属性 $A$ を機密の属性に分類した場合，準識別子の値で決まる $k$ 個のレコードは属性 $A$ の値を照合することでさらに絞り込まれてしまう．逆に機密の属性であるべき $A$ を準識別子に分類すると $A$ 以外の準識別子の属性を用いて属性 $A$ への属性リンクが可能になってしまう．

## □ $l$- 多様性

$k$- 匿名性は本章の冒頭で説明したレコードリンク攻撃の防護策にはなるが，もう 1 つの問題であった属性リンク攻撃には有効ではない．たとえば，**表 -6** は表 -5 の 1 つの準識別子グループを取り出したもので，3- 匿名性を満足する．しかしグループのすべてのレコードが機密属性として同一の病名「エイズ」を持つため，この準識別子グループに関連付けられる個人の病名が漏洩してしまうことになる．

上記の問題を克服するために，Machanavajjhala ら [9] は $l$- 多様性の概念を提案した．$l$- 多様性では，テーブルの中の同じ準識別子を持つグループが少なくとも $l$ 個の異なる機密の属性の値を持つことを要求する．この $l$- 多様性の定義は各グループに $l$ 個以上のレコードを持つことを要求するので，自動的に $l$- 匿名性の要件を満たすことになる．**表 -7** の医療データは，表 -6 の場合とは異なり，同一の準識別子グループに 3 つの異なる病名が含まれているので，3- 多様性を満足する．

$l$- 多様性で想定する攻撃者は外部知識として，標的とする個人の準識別子の値に関する知識に加え，どのような機密属性の値を取り得ないかという否定文の情報を持つと考えることができる．たとえば，もし攻撃者が表 -7 の準識別子グループに属するユーザ $A$ の病名に関して，「はしかでも胃潰瘍でもない」いう外部知識を持っていれば，$A$ の病名がエイズであると特定できる．

$l$- 多様性の実現には $k$- 匿名性の要件に加え，機密属性の多様性に関する追加の要件が加わる．したがって多くの研究者が $k$- 匿名性のアルゴリズムを拡張した形で $l$- 多様性のアルゴリズムを開発している．たとえば，Machanavajjhala らは Incognite [7] を拡張したボトムアップ型のアルゴリズム [9] を考案している．

## □ $t$- 近似性

$l$- 多様性では暗黙にすべての機密属性がそのドメインから値を均一に取り，各値の発生頻度は同じ程度と仮定している．しかし機密属性の発生頻度に大きな差がある場合，公開するデータの有用性は大きく損なわれる．たとえば，機密属性として，エイズ患者かどうかを示す Yes か No の 2 値のフィールドがあるとし，データセットに含まれる 1,000 人中，

5 人だけがエイズ患者だとする．その場合，2- 多様性を満たすためには，各準識別子のグループごとにエイズ患者のレコードを含める必要があるため，最大 5 つのグループしか作れない．各グループの準識別子はそのグループのレコードすべてに共通であるために著しく情報を曖昧にする必要がある．

また機密情報に関するあるグループの分散がデータセット全体の分散と著しく違う場合はそこで確率的に情報を得る歪度 (skewness) 攻撃が存在する．たとえば，患者の 95% がインフルエンザで 5% がエイズという医療データのテーブルがあったとする．そしてある準識別子のグループでは，その割合が 50% と 50% であったとする．この場合，テーブルは 2- 多様性を満足しているが，それでも全体で見た場合は 5% であったエイズの患者である確率よりもはるかに高い 50% の確率でこのグループに関連付けられる個人がエイズであると推定されてしまう．この問題を解決するために $t$- 近似性 [8] では，機密属性の全体での分散と各グループでの分散の距離を Earth Mover's Distance で定義し，その距離が与えられたしきい値 $t$ 以下であることを要求する．しかしながら $t$- 近似性にはデータの有用性を著しく低下させる欠点がある．

## □ 差分プライバシー

レコードリンクや属性リンク攻撃での攻撃者はターゲットのユーザのレコードが公開されるデータセットに含まれていることを知っていると仮定した．しかしデータセットにユーザのレコードが含まれていると分かること自体，そのユーザのプライバシーの侵害になる可能性がある．たとえば，病名を含む資料データのテーブルにユーザが含まれると少なくとも健康上何らかの問題があるということが分かってしまう．このように攻撃者がターゲットとなるユーザのレコードが公開されるテーブルに含まれるかどうかを決定しようとする攻撃を「テーブルリンク攻撃」と呼ぶ．Dwork [4] は公開するデータセットに各個人が提供するレコードの数が少数であることに着目して差分プライバシーを提唱した．

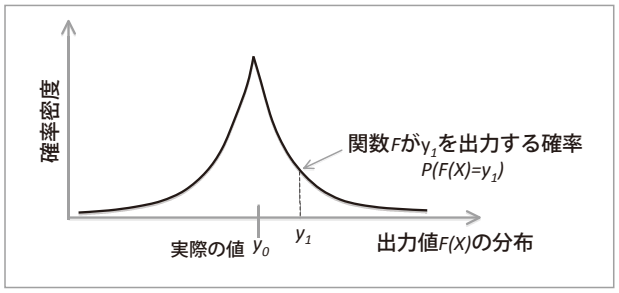


図 -4　データ加工のランダム関数 $F$

差分プライバシーでは主にデータセットに関する統計量（たとえば，ある条件を満足するレコードの数）に関する質問に答えるクエリー応答システムを想定し，システムの応答（PPDP におけるデータ加工処理）を確率的なランダム関数 $F$ で定式化する．**図 -4** にそのような関数 $F$ の例を示すが，入力として与えられるデータセット $X$ に対して，ある確率分布に従い出力 $F(X)$ を生成する．表 -1 の医療データのテーブル $X_0$ に対して，エイズ患者の数を正確に数えると 4 人という答えが得られる．それに対し，ランダム関数 $F$ の出力値 $F(X_0)$ は与えられた確率分布に従い，同じ質問に対してさまざまな異なる答え（たとえば 6 人）を生成する．

差分プライバシーではこのような出力関数 $F$ にランダム性を導入することにより，与えられたデータセット $X$ のレコードを 1 つ修正してもその違いが関数 $F$ の出力に際立った違いとして表れないことを目的とし，正式には下記のように定義される．

**定義 1（差分プライバシー）** 元のデータセットを変換して公開するデータセットを作成するランダム関数 $F$ は下記の条件を満たすとき差分プライバシーを保証するという．もしすべての 2 つのデータセットの組合せ $X_1$ と $X_2$ が最大でも 1 つのレコードしか違いがないとき，

$$\left| ln \frac{P(F(X_1)=S)}{P(F(X_2)=S)} \right| \leq \epsilon$$

の条件をすべての $S \in Range(F)$ に対して満たす．ここで $Range(F)$ は関数 $F$ の値域である．これは，元のデータセットに 1 つのレコードを追加，もしくはそのデータセットから 1 つのレコードを削除しても結果として関数 $F$ から出力される公開用のデータ

セットに際立った違いを生じさせないということを意味し，上記のテーブルリンク攻撃を不可能にする．

差分プライバシーではターゲットとなるレコード以外のすべてのレコードの情報を外部知識として持つ強力な攻撃者からの安全性を保証する．ただし差分プライバシーは前述のように主にクエリー応答型のシステムを考慮しており，データセットへの検索クエリーの数が全体のレコード数 $n$ に対して $o(n)$ のオーダに抑えないと安全性を保証することはできない．また公開するデータに二重指数分布であるラプラス分布からのノイズを加えることで元データの値を大きく変える可能性があり，現時点では可能な限り元データセットの生データを公開しようとするプライバシー保護パブリッシングの技術として適しているとはいえない．

## 多次元データへの匿名性の適用

前章で説明した $k$- 匿名化は安全性の面でいくつかの欠点はあるが，通常の定型的データの匿名化のための確立した手法と言える．しかし通常の $k$- 匿名化の手法は商品の購入などのトランザクションデータや人々の位置情報の軌跡のような多次元，疎，連続的といった性質を持つ移動軌跡データには適用できない．多次元データを公開する危険性は AOL が 2006 年に匿名化した検索ログを公開した際に No. 4417749 で表された個人が特定されたことで強く認識された [3]．

基本的に購買履歴や位置情報といった人々の行動に関する情報は攻撃者が観察することで容易に外部知識になるため準識別子として取り扱わなければならない．つまり各個人のレコードには莫大な数の準識別子の属性が含まれることになり，それらの属性は個人ごとにユニークな値の組合せをとることになる．したがってそのような多次元のテーブルに対して通常の $k$- 匿名化処理を行うと著しくデータの有用性が劣化してしまう．この問題を Aggarwal は「次元の呪い (the Curse of Dimensionality)」[1] という言葉で表現している．

| TID | 行動履歴 | 病歴 |
|---|---|---|
| $T_1$ | $a, c, d, f, g$ | 糖尿病 |
| $T_2$ | $a, b, c, f$ | 肝炎 |
| $T_3$ | $b, d, f, x$ | 肝炎 |
| $T_4$ | $b, c, g, y, z$ | エイズ |
| $T_5$ | $a, c, f, g$ | エイズ |

**表 -8　トランザクションデータの例．文献 12）から転記**

したがってこの分野の匿名性の研究では，攻撃者が知り得る準識別子の属性の数にある上限を設け，そのような限定した外部知識を持つ攻撃者に対する解決策を提案している．多次元の準識別子を持つ粗なデータセットにおいて，$k$- 匿名化における通常のデータ一般化手法を適用するのは困難であり，この分野での主な匿名化技法としては，ある属性値を省略するセル秘匿 (suppression) の手法が一般的である．以下多次元データの匿名性に関する代表的な研究を概観する．

### □ トランザクションデータ

トランザクションのデータの代表的な例は商品の購買履歴であり，準識別子の属性の集合は商品カタログに含まれる商品全体の集合となり超多次元のデータセットになり得る．ここでは Xu ら [12] によるトランザクションデータベースの匿名化手法を紹介する．Xu らの考慮するデータセットでは各ユーザのレコードは，準識別子となるユーザの行動の時系列と機密属性から構成される．

Xu らの提案する $(h, k, p)$ プライバシーは $p$ 個以下の長さの行動履歴の部分列を外部知識として持つ攻撃者に対し，行動履歴の $k$- 匿名性と機密属性の多様性に関して，どの値も $h$% 以下であることを要求するプライバシー定義である．つまり行動履歴に関する外部知識が長さ $p$ というパラメータで限定されていることを除けば $l$- 多様性に非常に似たプライバシー要件と言える．もし $k$=2, $p$=2, $h$=80% であれば，$\langle a, b \rangle$ という行動履歴の部分列は**表 -8** のレコード $T_2$ を特定するのでプライバシー要件が侵害されたことになる．

Xu らが考案した匿名化のアルゴリズムは長さ $p$ 以下のすべての行動履歴の部分列を考慮し，$k$ 個以

| 場所 | 時刻ごとの人口の推移（単位：千人） |
|---|---|
| A 町 | 11, 13, 15, 18, 20, … |
| B 町 | 8, 9, 12, 15, 16, … |
| C 町 | 21, 25, 28, 35, 33, … |

**表 -9**
**空間単位での位置情報データ匿名化の例**

上のレコードとその部分列が合致しない場合はその部分列に含まれる行動をすべてのレコードから削除する．たとえば，表 -8 において $<x>$ および $<y>$ という単一の長さの行動履歴はそれぞれ $T_3$ と $T_4$ をユニークに識別する．したがって行動 $x$ と $y$ は表 -8 のすべてのレコードから削除される．長さ $p$ 以下の部分列を列挙する手法は指数時間の計算量になるが，Xu らは効率的な欲張りアルゴリズムを考案している．

## □ 移動軌跡データ

移動軌跡の PPDP に関しても前節の Xu ら [12] に似た手法が提案されている．Fung ら [5] は位置情報のシーケンスを考慮し，攻撃者が外部知識として知り得る位置情報の軌跡の長さの上限を $L$ とし，$L$ 以下の長さの軌跡を共有するレコートが $k$ 個以上存在することを要求する「*LKC*- プライバシー」を提案している．また Terrovitis ら [11] は位置情報の測定が複数の管理者で分散的に行われる環境を想定し，各管理者が攻撃者として自身の管轄外の位置情報を推論する問題を考慮している．この場合には，攻撃者の外部知識となる移動軌跡の共通上限は存在しないが，それぞれの管理者が自分の管轄で取得した位置情報を外部知識として用いることを想定している．実現するプライバシー要件は Xu ら [12] に似た位置軌跡の $k$- 匿名性である．

## □ 位置情報データ

前節で紹介した手法では多くの位置情報データが秘匿されることになる．したがって**表 -9** のように位置名をレコードの主体としてその場所に存在する人数を公開するという手法に関する研究も非常に活発である．つまりテーブルの各レコードは位置空間の場所，そして属性は時間ごとにその場所に存在する人々の数となる．

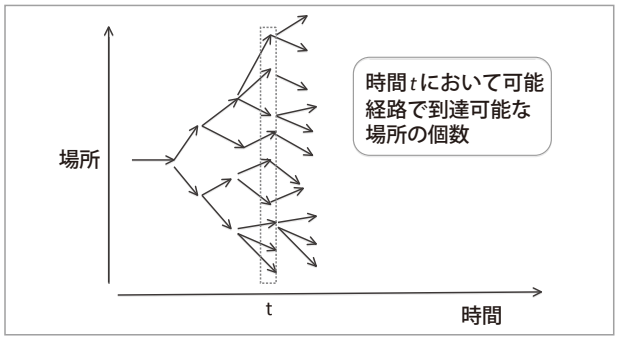


**図 -5　仮名交換による可能経路冗長性に基づく $(K, t)$ -プライバシー指標**

実際これまでの大多数の位置情報の $k$- 匿名化の研究は各個人に対するレコードの概念を捨て，空間を分割する領域単位での統計情報の公開手法を目指すものである．その代表例は Gruteser ら [6] による動的に位置情報の粒度を変えることで常に公開する位置情報には同時に $k$ 人以上が存在することを常に保証する手法である．ただしその数が $k$ 人以下の場合にはその情報は秘匿化される．この手法では次元の呪いの問題が回避できるので統計データに貢献できる個人の位置情報は飛躍的に増えるが，欠点としてはさまざまなデータ分析の対象となり得る軌跡情報が完全に喪失してしまう．

最近では次元の呪いの問題を回避しつつ軌跡情報を公開する手法として，真野ら [10] は位置情報の仮名化の手法を提案している．基本のアイディアはモバイルユーザの識別子を仮名に置き換え，その仮名を複数ユーザが同一場所で出会ったときにランダムに交換するという手法である．この手法により攻撃者から識別不能な各ユーザの代替移動経路を増やすことが可能であり，**図 -5** に示す「$(K,t)$ - プライバシー」では時間 $t$ における到達可能場所が $K$ 個以上あることが保証される．この手法では各ユーザの全軌跡を公開することはできないが，仮名交換を行うミックスゾーン間のセグメント単位での公開は可能であり，前節で述べた多くの位置情報が秘匿される問題がない．ただし現状のモデルでは攻撃者の外部知識が特定の場所におけるユーザの位置情報に限定されており，攻撃者モデルの一般化への対応が課題と言える．

## まとめと今後の課題

本稿ではプライバシー保護データパブリッシング (PPDP) の代表的プライバシー指標を概観し，その優位性と欠点を解説してきた．PPDP の想定するシステムモデルではデータの利用者と攻撃者が一致しており，従来の暗号技術，アクセス制御の手法では解決できない新しい研究課題が数多く存在する．PPDP における主な解決手法は匿名化と呼ばれるデータ加工技術であり，個人のプライバシー保護とデータの利得性の確保の両立が重要な研究目的となる．PPDP の研究を定式化するにあたり，2 つの重要な要素が存在する．1 つは攻撃者が利用可能な外部知識の厳密な定義であり，外部知識の性質により PPDP の実現手法は大きく異なる．もう 1 つはデータの利便性の指標である．匿名化アルゴリズムのゴールはデータ利便性を最大化することであり，想定するデータ分析に対して適切な指標を選択することが重要である．

最近重要性を増している多次元データに関しては「次元の呪い」という本質的な課題があり，既存の研究の多くは攻撃者の外部知識を限定することで対応しており，依然として根本的な解決策が見当たらないのが現状である．この分野での $k$- 匿名性の適用は本質的に困難に見え，本稿で述べた仮名化等，まったく別のデータ加工技術の考案によるブレークスルーが望まれる状況と言える．

**参考文献**


1) Aggarwal, C. C. : On K-anonymity and the Curse of Dimensionality, *Proceedings of the 31st international Conference on Very Large Data Bases*, VLDB'05, VLDB Endowment, pp.901-909 (2005).
2) Agrawal, R. and Srikant, R. : Privacy-preserving Datamining, *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, SIGMOD'00, New York, NY, USA, ACM, pp.439-450 (2000).
3) Barbaro, M. and Zeller, T. : A Face Is Exposed for AOL Searcher No. 4417749, *New York Times* (2006).
4) Dwork, C. : Differential Privacy, *Automata, Languages and Programming* (Bugliesi, M., Preneel, B., Sassone,V. and Wegener, I., eds.), Lecture Notes in Computer Science, Vol.4052, Springer Berlin Heidelberg, pp.1-12 (2006).
5) Fung, B. C. M., Cao, M., Desai, B. C. and Xu, H. : Privacy Protection for RFID Data, *Proceedings of the 2009 ACM Symposium on Applied Computing*, SAC'09, NewYork, NY, USA, ACM, pp.1528-1535 (2009).
6) Gruteser, M. and Grunwald, D. : Anonymous Usage of Location-Based Services Through Spatial and Temporal Cloaking, *Proceedings of Mobisys 2003 : The First International Conference on Mobile Systems, Applications, and Services*, San Francisco, CA, USENIX Associations, (online), available from <http://www.usenix.org/events/mobisys03/tech/gruteser.html> (2003).
7) LeFevre, K., DeWitt, D. J. and Ramakrishnan, R. : Incognito : Efficient Full-domain K-anonymity, *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, SIGMOD'05, New York, NY, USA, ACM, pp.49-60 (2005).
8) Li, N., Li, T. and Venkatasubramanian, S. : T-Closeness : Privacy Beyond K-Anonymity and L-Diversity, *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*, pp.106-115 (online), DOI:10.1109/ICDE.2007.367856 (2007).
9) Machanavajjhala, A., Kifer, D., Gehrke, J. and Venkitasubramaniam, M. : L-diversity: Privacy beyond K-anonymity, *ACM Trans. Knowl. Discov. Data*, Vol.1, No. 1 (2007).
10) Mano, K., Minami, K. and Maruyama, H.: Protecting Location Privacy with K-Confusing Paths Based on Dynamic Pseudonyms, *Proceedings of the 5th IEEE International Workshop on SEcurity and SOCial Networking* (*SESOC*) (2013).
11) Terrovitis, M. and Mamoulis, N. : Privacy Preservation in the Publication of Trajectories, *Proceedings of the The Ninth International Conference on Mobile Data Management*, MDM '08, Washington, DC, USA, IEEE Computer Society, pp.65-72 (2008).
12) Xu, Y., Wang, K., Fu, A. W.-C. and Yu, P. S. :Anonymizing Transaction Databases for Publication, *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD'08, New York, NY, USA, ACM, pp.767-775 (2008).


(2013 年 6 月 13 日受付)

**南　和宏**（正会員）minami.at.uiuc@gmail.com

統計数理研究所新領域融合研究センター特任准教授．2006 年 US ダートマス大学コンピュータサイエンス学科博士課程卒業．電子情報通信学会，IEEE，ACM 各会員．