情報処理学会論文誌 Vol.55 No.9 1956–1963 (Sep. 2014)

招待論文

有用な匿名化データ—経験からの学習

星野 伸明[1,a)]

受付日 2014年6月17日, 採録日 2014年6月28日

概要：情報保護研究では，完全な補助情報を持つ理想化された攻撃者が目につく．結果として最悪の場合でも安全なデータが生成されるが，そのデータに分析価値が残りにくい．実際，本論文で紹介されるように，過去に完全な攻撃者を想定して作成した統計は，科学研究で役に立たなかった．結局現在の官庁統計実務では，攻撃禁止の契約が前提で情報保護の程度を決める．つまり契約に縛られる不完全な攻撃者が暗黙に想定されている．このような官庁統計の経験を教訓とすれば，情報保護研究でも不完全な現実的攻撃者を想定すべきである．問題は現実の妥当なモデルだが，本論文では法的な制約を統計モデルで表す．主要な情報保護法制では，個体識別の有無で安全性を判断する．しかし既存研究ではこの判定を主観に委ねており，その点が情報保護の社会的利用において障害になっている．この判定を客観的に行うため，本論文では観測を利用する．観測は現実の契約や制度が攻撃者に与える影響の情報を持つので，そこから統計的に現実を推定できる．以上のような経験や観測による現実のモデル化は，機械学習などデータ駆動の知識形成と哲学を共有している．

キーワード：個体識別，プライバシー保護，統計的開示制限

Learning from the Experience of Publishing Useful Anonymized Data

Nobuaki Hoshino[1,a)]

Received: June 17, 2014, Accepted: June 28, 2014

Abstract: A research on information protection often assumes an idealized adversary who has perfect information. It follows absolutely safe data, which tend to be unusable for statistical analysis. In fact, under the assumption of a perfect adversary, anonymized statistics was useless historically for a scientific purpose, as described in the present article. Hence a statistical agency nowadays anonymizes data, depending on the ban of de-anonymization. In other words tacitly assumed is an imperfect adversary whose power is limited by a law. These experiences teach that a research on information protection should assume an imperfect adversary. A convincing model of such an adversary can be constructed by a statistical description of a law system; major laws regard identifiable data as unsafe. However, many researches leave the decision of an identifiable state subjective, which prevents the wide use of anonymization. The present article exploits observation to objectively discern an identifiable state. An observation carries information on the real effect of institutions or laws on an adversary, and thus it enables the statistical estimation of a reality. Our discovery of a reality from those experiences and observations shares the philosophy of data driven knowledge such as machine learning.

Keywords: identification, privacy preserving, statistical diclosure control

1. はじめに

情報保護の下でデータを分析する枠組みは，計算機科学では “Privacy Preserving Data Minig（PPDM）” として知られる．特に情報保護下のデータ公開は “Privacy Preserving Data Publishing（PPDP）” と呼ばれ，統計学における情報保護研究（“Statistical Disclosure Control（SDC）” もしくは “Statistical Disclosure Limitation（SDL）”）と同等と考えられている [9] ようだ．

本論文の著者は SDC の研究者である．日本でも 1990 年代の後半から，個体レベルのデータ（「個票データ（micro-

[1] 金沢大学経済学類
School of Economics, Kanazawa University, Kanazawa, Ishikawa 920–1192, Japan
[a)] hoshino@kenroku.kanazawa-u.ac.jp

data)」）利用促進のための官庁統計の改革が始まった．著者はこの時期にSDCの研究を始め，日本で初めて個票データを匿名化して公開する過程を詳しく観察することができた．近年はPPDM（P）に興味を持ち，計算機科学者と対話する機会にも恵まれた．その中でおぼろげに見えてきたのは，SDCとPPDM（P）の違いである．

情報保護を破ろうと試みる者（adversary）を「攻撃者」と呼ぶ．著者の理解では，攻撃者のモデルの立て方に，SDCとPPDM（P）の違いがある．端的にいえば，SDCは不完全な情報しか持たない攻撃者を想定し，PPDM（P）は完全な情報を持つ攻撃者を想定する．

たとえばSDCは母集団のk-匿名[17]と標本のそれを区別する．攻撃対象の標本が母集団の一部としか分からない攻撃者を，暗黙に仮定しているからだ．もし攻撃者が標本に含まれる個体の名簿を持っていれば，母集団から標本を抽出するうえでの不確実性は消える．つまりその名簿が母集団とみなせるので，標本のk-匿名のみ考えればよい．PPDM（P）ではそのような名簿を持つ攻撃者を想定するので，母集団のk-匿名という概念がない．

このようにPPDM（P）では，情報を持つ強い攻撃者を想定するように思われる．もう1つ例をあげよう．PPDM（P）で有名な差分プライバシー[5]は，攻撃者がいかなる補助情報を持っていても成立する情報保護の概念である．攻撃者を想定していないように見えるかもしれないが，すべての補助情報を持つ攻撃者の存在可能性が念頭に置かれている．

他方SDCは想定する攻撃者をモデルとして明示する意識が薄いのだが，現実的であろうとする態度が特徴のように思われる．現実の攻撃者は標本に含まれる個体の名簿を持たないし，統計当局が隠蔽した情報保護手法の詳細，母数は知らないというわけだ．

攻撃者が持つ情報を場当たり的に限るのは，美しい理論ではない．また攻撃者が知らないはずの情報が漏れれば，情報保護は破られるかもしれない．しかしそれでもなお，SDCが現実的な攻撃者にこだわるのは，データ分析者の要求に応えるためである．

隠した情報が攻撃者に漏れる最悪の場合を想定すると，情報保護を比較的強く施さなければならない．そして強く保護されたデータは，分析上の「有用性」が低くなりがちである．分析手法を限れば有用性を保つのは容易だが，SDCやPPDPでは多様な分析に耐えるデータを公開したい．そのためには強すぎない攻撃者の仮定が望ましい．このような仮定を正当化するのが，現実性である．

実は歴史的に，最悪の場合を想定して保護した統計データは，社会科学で要求される有用性を持たなかった．後述するように，最悪の場合を想定しても有用なデータが公開できると，SDCの黎明期では楽観されていた．しかし実務的にも理論的にも経験が蓄積されるにつれ，有用なデータが何らかの情報漏洩をともなうと理解されることになる．そして現状の官庁統計実務では，データの有用性が高いほど，より強く攻撃を禁じる契約を利用者に求める．契約によって攻撃者の能力を限定し，それを前提に情報を保護しているのである．

このような歴史的経験は，社会的に有用な情報保護を実現するうえで無視すべきでないと思われる．したがってSDCの理想が挫折して現状に落ち着いた経緯を紹介することが，本論文の第一の目的である．

本論文の第二の目的は，現実的な攻撃者について，ご都合主義でないモデルを提案することである．いかに現実を定式化するかは，情報保護の許容範囲を直接左右するにもかかわらず，主観的問題であるとして立ち入らない既存研究が目立つ．しかしこの問題は客観化する余地が残っていると著者は考える．

1つの方向性は，法的制約の技術的定式化であろう．法律は現実の攻撃者を縛ると同時に，情報が保護されている状態を定義している．実務は法律に縛られるので，その制約を技術的に表せば妥当な現実的モデルを得られるのではないか．

現実に接近するもう1つの方向性は，観測の利用である．情報保護の許容範囲の決定は，適当なリスク測度の閾値の決定に帰着させられる．この閾値を主観で決めるのではなく，統計的に推定してはどうか．観測には現実の情報が含まれており，統計モデルによってこれを利用できる．

このような視点から，本論文は以下のように構成した．まず2章では，官庁統計における情報保護実務の経験を概観する．2.1節では情報保護に関係する諸概念を紹介しながら，歴史的な試行錯誤を総括する．2.2節では現行法を参照し，現実の匿名化の目的が個体識別が不可能な状態ということを確認する．3章では，個体識別が可能か否かをデータから判別する方法を考察する．3.1節では個体識別行為を確率的にモデル化する．3.2節では個体識別のリスク測度を具体的に提示する．4章では結語を述べる．

2. 官庁統計の経験

2.1 情報保護概念の形成

統計当局が情報保護の重要性を認識したのは，Andersonら[1]によれば19世紀末のことである．徴税などの行政に必要な情報と統計を区別しなければ，脱税摘発などの不利益を恐れて統計調査への正直な回答は望めないと予想できる．実際Singerらの研究[15], [16]では，不利益の可能性があると調査への協力は減少した．

官庁統計は「統計目的」という文言で行政から区別される．米国ではすでに1910年の国勢調査において，回答は統計目的にのみ使われるとタフト大統領が宣言している．ただし実際には徴兵逃れの捜索に使われた[1]という．このように官庁統計は行政目的での使用について，たびたび

圧力を受けている．したがって現代的な法制では，官庁統計の統計目的外での利活用を原則的に禁止する．

データが本来作られた目的以外での利活用を「二次利用」と呼ぶ．たとえばある患者の容態管理目的でデータが生成されたとしよう．このデータを薬効の研究で用いるのは，望ましい二次利用といえる．一方，銀行がこの患者の経営する企業への融資判断にこのデータを用いるのも二次利用である．この場合は銀行に利益を，患者にはおそらく不利益をもたらす．このようにデータの二次利用は必ずしも望ましいとは限らない．ゆえに善い二次利用と悪い二次利用を定め，後者が起きないようにデータを管理すること，これが情報保護の目的である．

たとえば統計法によると統計調査の回答（調査票情報）の二次利用[*1]は，学術など限られた目的についてのみ許されている．統計当局は利用目的を審査し，法的に認められなければ調査票情報を渡さない．

しかし統計調査の回答が統計目的にのみ使われるとしても，公表した統計が統計目的外で利用されるとしたら，やはり統計調査への協力を阻害する．そして統計は万人に公表され，目的によって利用者を選別できない．

したがって統計は，統計目的でしか使えない形態で公表するのが望ましい．このように考えて初めて，データ変換による情報保護（sanitization）の必要性が生ずる．

遅くとも 1940 年代には「$n-k$ 占有ルール」に基づき，米国センサス局は集計表（table of magnitude）[*2]にデータ変換を施して情報を保護していた [3]．この手法を例で説明しよう．ある地域の小売店売り上げ総額について，上位 n 事業所の占有割合が k%以上とする．もし n が小さく k が大きければ，売上総額を公開するとその地域の大型小売店の売上高はおおよそ分かってしまう．したがってこの売上総額は情報保護の対象となり，公開されない．

$n-k$ 占有ルールで抑止されているのは，特定個体の秘密の暴露と考えられる．官庁統計の文脈では，このような個体の秘密保護の概念を「秘密性（confidentiality）」[*3]と呼ぶ．プライバシーが個人の権利であるのに対し，秘密性は法人にも保証される．

たとえば統計法（第 3 条第 4 項）には「公的統計の作成に用いられた個人又は法人その他の団体に関する秘密は，保護されなければならない」と明記されている．なお「公的統計」とは日本の官庁統計をさす．

秘密性と統計目的の関係は明瞭ではないが，おおよそ以下のように考えられている．まず統計目的でのデータ利用とは，分布の参照にほかならない（統計量は分布から導かれることに注意）．そして分布は，一回限りではなく繰り返す事象の性質である．一方，特定個体固有の性質は一回限り定まる．ゆえに特定個体固有の性質を利用することは，統計目的の利用ではない．したがって特定個体固有の性質のうち，その個体が隠したい性質が「秘密」であり，情報保護の対象となる．

特定個体固有の性質が分かるということは，性質のデータの主が特定個体と分かるということである．これを個体の「識別（identification）」と呼ぶ．そして識別ができないように施すデータ変換を「非識別化（de-identification）」もしくは「匿名化」と呼ぶ．

先ほどの小売店売り上げの例では，売上総額という性質の主が特定の大型小売店と（おおよそ）分かる．つまり個体識別が起きており，売上総額は統計目的外利用が可能な情報になる．したがって非公開という匿名化が施される．

個体識別が直接に問題となるのが，個票データの場合である．社会科学では集計量の解析を「マクロ分析」と呼び，集計表をデータとして用いる．他方，個体の挙動を解明するのが「ミクロ分析」であり，個票データを用いる．個票データのレコードが識別されると，その個体固有の属性が分かってしまう．

松田 [14] によれば，社会科学においてミクロ分析は遅くとも 1940 年代には始まっている．そして労働経済学など一部の分野では 1960 年代までに定着し，簡単な手続きで官庁統計の調査票情報が個票データとして提供されたといわれている．

しかし 1960 年代後半から欧米諸国においてプライバシー問題が顕在化する．宇賀 [18] によれば，プライバシー概念は「自己情報をコントロールする権利として理解する立場が有力」である．また「この考えの影響を受けた個人情報保護立法は 1970 年代からみられるようになった」とある．この時期に官庁統計の個票についても，情報保護が強く要請されるようになった．集計表と比べて個票の情報保護は論点が多く，データベース研究者の興味も引いた．以降，情報保護研究は本格化する．

この時期の情報保護研究は，理想主義的であった．たとえば Dalenius [4] は，統計データベースにアクセスしても，アクセスしない場合よりも多く個体について知ることができるべきではない，と述べている．また 1978 年に刊行された SDC に関する包括的な報告書 [7] の題目は “Report on Statistical Disclosure and Disclosure Avoidance Techniques” である（下線部は著者による）．ここで “Avoidance” とあるように，不要な情報の暴露は避けられるとの見通しが一般的であったように思われる．

理想主義的な見識は実務も左右する．一例をあげると 1980 年のドイツ統計法では，社会科学者が利用する非識別化された個票データについて「再識別」の可能性が疑う余

*1 統計法の文脈では「二次利用」と「二次的利用」を区別する．「二次的利用」という言葉は，統計法第 32 条から第 38 条までの規定に基づく統計の二次利用全体のことをさす．一方「二次利用」は，統計法第 32 条に基づく統計の二次利用のみをさす．

*2 “magnitude” と “frequency” の集計表は区別され，後者が「分割表」である．

*3 情報セキュリティの文脈では “confidentiality” は「機密性」と訳され意味が異なる．

地なく除かれていることを要求している．この状態は「絶対的な匿名」と呼ばれる．詳しくは濱砂 [10] を参照されたい．

ところが絶対的に匿名な個票データは，社会科学者にとってもはや分析価値を持たなかった．より弱い匿名性概念を前提としなければ，学術研究目的の調査票情報二次利用は存立し得なかったのである．

1983 年の EU 評議会勧告 [2] は，プライバシー保護と学術研究の両立を図る．そこでは，ある個体を識別するために非合理的な時間，コスト，人力が必要な場合，その個体を識別可能とみなすべきではない，と述べられている．この理屈は，暗号学において計算時間が限られる攻撃者を想定することと似ている．

勧告を受けて 1987 年に改正されたドイツ統計法では，再識別の費用が便益を上回る状態を「事実上の匿名」とみなす．この概念は打算的な攻撃者についての識別不可能性であり，絶対的な匿名概念が攻撃者の挙動を限定しなかったのと異なる．このように情報保護の実務は，理想主義を放棄せざるを得なかった．

情報保護研究における現実との妥協を象徴的に示すのが，先ほど例示した報告書 [7] の 1994 年における題目変更 [8] である．原題の “Avoidance” が消え “Report on Statistical Disclosure Limitation Methodology” となった．情報の暴露は避けられず，限定するものという理解が背景にある．

Dwork [5] によれば Dalenius [4] の理想実現は不可能であり，差分プライバシーは秘密の暴露を一定範囲に限定するツールである．

歴史的試行錯誤の結果，現在では官庁統計の個票データを利用する場合，データの匿名化処理が軽いほど，利用目的が審査されたり契約で再識別行為に罰則を科すなど，行為制限が重くなる．

たとえば基本的に誰でも利用可能な個票データを「一般目的汎用ファイル（Public Use File，PUF）」と呼ぶが，PUF にはかなりの匿名化処理が施される．なお日本法は PUF を制度化していない．それから「科学目的汎用ファイル（Scientific Use File，SUF）」は，利用目的の審査を通った者にのみ渡される．その多くは守秘義務契約をともなう．日本法で SUF に相当するのが「匿名データ」である．また「リサーチデータセンター（Research Data Center，RDC）」において，監視されながらデータ分析をする場合もある．RDC の利用にも通常は審査および契約がともなうが，匿名化処理が最低限のデータが使える．

このように匿名化と行為制限は合わせて必要な保護がなされている．官庁統計実務において匿名化水準を決める際，データ以外の制度的要因を無視することはできない．特に，有用な匿名化個票データの作成には，能力が限定された攻撃者の想定と，それを正当化する制度が必要ということが，歴史的教訓であるように思われる．

2.2 匿名化の現実

前節で確認したように，行為制限が少なく潜在的に危険な攻撃者には，より高度な匿名化を施したデータが提供されている．このような状況は，制度の効果を無視して一律に最強かつ完全な攻撃者モデルを想定すると説明しにくい．現実の匿名化では，制度に反応する攻撃者を暗黙に想定していると考えられる．そのような攻撃者を現実的とみなし，モデル化するべきではないだろうか．

能力が限られる攻撃者を想定したいのは，匿名化を減らすためである．できるだけ有用なデータを作成するには，匿名化を最小にしたい．この最適化には，データが公表可能な状態の定義が必要となる．そのような状態を定めている法律の参照から始めよう．

官庁統計の法制では多くの場合，データが公表可能か否かは個体識別の有無で判定される．たとえばドイツの絶対的な匿名性および事実上の匿名性概念では，個体の再識別が可能か否かを（異なる攻撃者の想定で）判断していた．また統計法（第 2 条第 12 項）から匿名データの定義を引用すると，「一般の利用に供することを目的として調査票情報を特定の個人又は法人その他の団体の識別（他の情報との照合による識別を含む．）ができないように加工したもの」である．

このように個体識別が可能か否かで公表可能か判断する方法は，統計外の情報保護法制でも採用される場合が多い．たとえばいわゆる個人情報保護法（第 2 条第 1 項）における「個人情報」とは「生存する個人に関する情報であって，当該情報に含まれる氏名，生年月日その他の記述等により特定の個人を識別することができるもの（他の情報と容易に照合することができ，それにより特定の個人を識別することができることとなるものを含む．）をいう．」

個人情報保護法は，個人情報について適正な管理を要求する．しかしデータが変換されて個人情報でなくなれば，個人情報保護法は適用されない．ゆえに個体識別を不可能にすること，すなわち非識別化が匿名化実務上の目的になる．

したがって統計であるかないかにかかわらず，個体識別が不可能な範囲で最も有用なデータを作成するのが法的に正しい最適化である．有用性が高いデータなら個体識別の可能性が高くてもよいという議論を見かけるが，現行法的には（制度的制約が一定なら）無理がある．

個体識別が可能な状態は，個体識別の方法に依存する．まず個体を識別する最も有力な方法といわれているのが，外部情報との「照合（matching）」（順攻撃ともいう）である．匿名データや個人情報の定義で，照合による個体識別を明記するのは当然であろう．照合では，特定個体の既知の性質と同じ性質のレコードを公表情報から探す．なお攻撃者にとって既知の性質を「キー変数」と呼ぶ．この概念は，名前のような（直接）識別子や性別のような疑似識別子を含む．他方，公表情報で珍しいレコードを選び，世の

中でそのような個体を探す方法を「逆攻撃」と呼ぶ．

既知でない秘密の変数は照合に使えない．また逆攻撃においても，秘密な変数では個体を限定できない．ゆえにそのような秘密変数を操作しても，個体識別の可能性に影響を及ぼさない．

したがって，キー変数でない秘密な変数にデータ変換を施す類いの情報保護は，法的には不要である．個体識別されても安心という理屈は，法の中にはない．そして秘密変数はデータ分析において興味の対象となることが多い．そのような変数のデータ変換は，有用性を確実に損なう．

たとえば差分プラバシーのように，秘密の推測精度の限定を目標とすることは研究レベルではあり得る．もう 1 つ例をあげると，ℓ-多様性 [12] は，キー変数の条件付きで秘密変数の（標本）分布が退化しないことを要求する基準である．この基準も研究レベルではあり得るが，実務ではいたずらに有用性を低下させるだけであろう．

さて，照合による個体識別の可能性は，キー変数の質と量に依存する．ここでの質の良さは，攻撃用情報と公開情報でコーディングの基準や時点の差，誤記などのノイズがないことを意味する．興味深いことに米国法の Health Insurance Portability and Accountability Act（HIPAA）では “safe harbor” 基準において，18 種類のキー変数を具体的に指定している．しかし一般に攻撃者像は明らかではないため，キー変数の設定が個体識別の研究では問題になる．

多くの研究では，匿名化される前のデータと外部情報が同じという完全な攻撃者が想定される．結果としてすべてでなくとも多くのキー変数が選ばれ，その質も最良となる．そして個体識別の可能性は過剰に評価される．

やはり有用なデータを作成するには，完全な外部情報ではなく，現実的な外部情報の想定が望ましい．たとえば Elliot ら [6] は，SNS などでの公開自己情報まで含めて網羅的に外部情報を調査し，系統的に蓄積することを主張している．調査を継続すれば，過去の任意の時点で現実的な外部情報が分かる．

Elliot らの主張はコストの点で無理があるとしても，現実を知るには観測するしかないのではないか．そして制度による攻撃者に対する行為制限の効果も，観測しなければ分からないはずだ．

ここまでの考察を総合すると，個体識別が可能か否かを観測データから判別するのが妥当と思われる．つまり我々は，データから個体識別の構造に関する知識を発見したい．データ駆動で攻撃者の現実的なモデルを構築することは，既存研究で欠けている態度のように思われる．

3. 個体識別可能性の判別

前節で我々の問題をデータ駆動で解くというアイデアを述べたが，話はそれほど単純にならない．最大の困難は，個体識別がほとんど観測されないことである．情報がなければサポートベクターマシンのような複雑なモデルは同定されないので，単純な判別モデルが望ましい．また観測情報を補うため，理論による情報の外挿は有効である．以下では実例として，星野 [11] の議論を紹介する．

3.1 個体識別の統計モデル

我々は個体識別が可能か否かを，データ要因と制度要因で説明したい．まずデータについて，匿名化による表現の変化は滑らかなのに対し，個体識別が不可能と可能の差は不連続である．このような状況をモデル化する場合，データの適当な実数特性値が閾値を超えれば個体識別が可能とみなすのが定石である．我々は，そのような特性値を個体識別の難易度と呼ぼう．そして個体識別が可能と不可能を分ける難易度の閾値を制度で説明できれば，合目的なモデルを得る．

この難易度 δ を引数とする関数 f は，個体識別が可能なら 1，不可能なら 0 を返すとする．つまり閾値が α として

$$f(\delta) = \begin{cases} 1 & \delta < \alpha \text{ の場合} \\ 0 & \delta \geq \alpha \text{ の場合} \end{cases}$$

ということになる．制度に依存する α は未知であろう．

個体識別の難易度に相当する様々な測度が提案されているが，その閾値 α の決定は主観の問題と主張されることが多い．そのため匿名化の実務では，「有識者会議」や「第三者委員会」のような権威に判断を投げるのが普通である．しかしそのような権威の判断が，透明に説明できなければ問題である．また権威の判断がつねに必要なら，コストなどの問題で権威にアクセスできない場合，匿名化は利用できないことになる．情報保護研究が主観の問題には立ち入らないと逃げれば，社会的責任を果たしているとはいいがたい．閾値 α を客観的に定める方法はないだろうか．

著者はデータ駆動の立場から，閾値 α の統計的推定を主張する．ただし問題は，個体識別が可能か否かは直接観測されないことだ．観測可能なのは，個体識別が起きたか否かという事実である．したがって，個体識別が可能という状態と観測結果を結ぶ論理が必要となる．1 つの方法として，以下のような確率モデルが考えられる．

確率変数 X が 1 なら個体識別が観測され，0 なら観測されないこととする．個体識別が不可能（$\delta \geq \alpha$）なら必ず $X = 0$ である．個体識別が可能（$\delta < \alpha$）の場合，難易度 δ に依存する確率 $p(\delta)$ で識別が観測されると考えよう．すなわち $\Pr(X = 1; \delta < \alpha) = p(\delta), \Pr(X = 0; \delta < \alpha) = 1 - p(\delta)$ とする．

もし $p(\delta) = 0$ の場合，攻撃者は個体識別が可能でも決してしないか，個体識別に成功しても必ず隠れることを意味する．前者なら個体識別は実質的に問題にならないし，後者はつねに隠れ通せるか疑問である．それから $p(\delta) = 1$ なら，α の推定誤差が問題にならない．ゆえに以下では

$0 < p(\delta) < 1$ と仮定する．

このような状況で，閾値が共通する n 件のデータ公開事例が存在するとしよう．その $i, i = 1, 2, \ldots, n,$ 番目について観測されるのは，少なくとも難易度 δ_i と識別の有無 x_i である．単純化のため $\delta_1 < \delta_2 < \cdots < \delta_n$ としよう．

もし個体識別が j 番目の事例で起きていれば，$\delta_j < \alpha$ と断定できる．ゆえに個体識別がこれまで起きていない（すべての i について $x_i = 0$）として考察を続ける．

この場合モデルの尤度 ℓ は，$\delta_i < \alpha \leq \delta_{i+1}$ のとき $\ell(\alpha) = \prod_{j=1}^{i}(1 - p(\delta_j))^j$ となる．そしてすべての δ について $0 < p(\delta) < 1$ なら，α の最尤推定値 $\hat{\alpha}$ は δ_1 以下である．つまり過去の事例で個体識別が起きていなければ，その最も低い難易度 δ_1 以下と閾値 α は推定される．もちろん，推定誤差が $p(\cdot)$ に応じて存在する．

強い仮定を置かなくても，このように難易度の閾値が δ_1 以下と推定できる．当然だが，δ_1 未満の難易度について個体識別が不可能な証拠はない．したがって新たに匿名化をする場合，その程度は δ_1 と同じにすればよい．このような考え方は，匿名性の程度について前例踏襲をするということであり，社会的にも受容されやすいのではないか．

閾値が共通する事例の範囲は，同制度の下での匿名化データ利用である．たとえば公的統計の匿名データは，同一プロトコルで利用が繰り返されており，閾値が共通するとみなしてよいだろう．個人情報保護法の下での匿名化データ利用については，現状で統一プロトコルが存在しない．しかし標準的な利用条件（契約のひな形）が示されて使われるなら，閾値の類似性が上がることを指摘しておく．

過去に事例がない，もしくは少ない場合はどうしたらよいか．そのような場合は，個体識別を起こさないように難易度 δ の高いデータを公開し，事例の蓄積を図るべきだ．この問題は，臨床試験において新薬候補の投与量を決める問題と似ている．人体に決定的な悪影響を与えないように，微量から始めて兆候を注視しながら増量するのである．

3.2 個体識別の難易度測度

前節では識別が観測されていないとして，個体識別の難易度が最低の事例と同等の難易度に匿名化するという指針が得られた．この指針の下で難易度測度 δ に要求されるのは順序の決定であり，値の決定ではない．ここで重要なのは，δ の具体型を（分からないので）間違えた場合に，順序は値よりも影響が少ないことであろう．以下では δ の順序付けの方法について，一案を示す．

難易度測度 δ の具体型を考察するため，まず個体識別行為を要因分解しよう．Marsh ら [13] によると

$$\begin{aligned}&\Pr(\text{識別が実際に起きる})\\&\quad = \Pr(\text{識別が起きる} \mid \text{識別を試みる})\\&\qquad \times \Pr(\text{識別を試みる}).\end{aligned} \tag{1}$$

式 (1) の各項を解釈すると，まず絶対的な匿名性概念が $\Pr(\text{識別が起きる} \mid \text{識別を試みる}) = 0$ という状態に対応する．また事実上の匿名性概念は $\Pr(\text{識別が実際に起きる})$ が低い状態に対応する．それから攻撃者が個体識別に成功しても観測されないかもしれないので，$p(\delta) \neq \Pr(\text{識別が実際に起きる})$ である．いずれにせよ，以下では絶対的な匿名性の評価に議論を絞る．

Marsh らは照合による攻撃を想定しており，それが成功する事態は 4 つの条件が成立する場合だという．すなわち

(a) 攻撃用ファイルの質が良い．
(b) 公開ファイルに個体が含まれている．
(c) 個体が「母集団一意」[*4]である．
(d) 個体が母集団一意と確証できる．

そしてこれらの条件が満たされる事象をそれぞれ a から d と書けば

$$\begin{aligned}&\Pr(\text{識別が起きる} \mid \text{識別を試みる})\\&\quad = \Pr(a)\Pr(b|a)\Pr(c|a,b)\\&\qquad \times \Pr(d|a,b,c).\end{aligned} \tag{2}$$

この議論では，あるレコードが特定個体の情報と断言できる場合のみを個体識別と考えている．一般に母集団 m 意は $1/m$ の確率で特定の個体と当たるといえるが，たまたま当たるケースは問題としていない．それから母集団 m 意について，自分のレコードが分かる $m-1$ 人が結託すれば，残りの個体を断言できる．しかしこのケースも（現実的でないから）問題としない．もしこれらの場合を問題と考えるなら，以下の議論で「母集団一意」を「問題の個体」と読み替えればよい．しかしそのような場合，有用なデータを作りにくい．

さて，実は式 (2) の定量評価に Marsh らは失敗している．しかし母集団一意概念は非専門家でも分かりやすく，匿名化処理に関するある種の単調性など良い性質を持つ．我々の問題でこの議論を活かす方法を考えよう．

式 (2) が正，つまり識別を試したときに識別が起きる確率が正ということは，個体識別が可能（絶対的な匿名性が成立しない）ということであった．

ゆえに式 (2) の右辺の要素のどれかが 0 なら，個体識別が不可能といえる．ただ多くの場合，2-匿名性を満たすような操作をしなければ $\Pr(a,b,c)$ は正である．その場合に式 (2) の右辺を書き換えると

$$\begin{aligned}&\Pr(\text{識別が起きる} \mid \text{識別を試みる})\\&\quad = \Pr(a,b,c)\Pr(d|a,b,c)\end{aligned}$$

なので，$\Pr(d|a,b,c)$ が 0 なら個体識別が不可能と考えられる．つまり Marsh らの枠組みにおいて多くの場合，個体識

[*4] 「母集団一意」とは，キー変数の条件を満たす個体が母集団で一意に定まることをいう．公開ファイルが 2-匿名なら，そのファイル中に母集団一意は存在しない．

別が可能か否かは $\Pr(d|a,b,c)$ が 0 か否かという問題に縮退する．しかし $\Pr(d|a,b,c)$ の直接的計量は，きわめて困難である．

我々は $\delta < \alpha$ なら $\Pr(d|a,b,c) > 0$ としよう．ここで条件付き確率 $\Pr(d|a,b,c)$ は，事象 (a,b,c) を所与としている．ゆえに事象 (a,b,c) を基準化した $\Pr(a,b,c)$ を，δ の引数とする．つまり適当な関数 h について，$\delta = h(\Pr(a,b,c)) < \alpha$ のとき $\Pr(d|a,b,c) > 0$ とすれば，これまでの議論と整合する．

確率 $\Pr(a,b,c)$ の増加は，照合が成功する母集団一意の個体がより多く公開されることを意味する．その場合，母集団一意の確証はより容易になるはずだ．ゆえに関数 h は以下の単調性を満たす．

$$q_1 \geq q_2 \Rightarrow h(q_2) \geq h(q_1).$$

この h の単調性が成立すれば，最も個体識別の難易度 δ の低い事例は，最も $\Pr(a,b,c)$ が高い事例と同じである．つまり h の具体型が分からなくても，$\Pr(a,b,c)$ が計量可能なら，最も個体識別の難易度が低い事例を選べる．そして本論文ではこれ以上議論しないが，$\Pr(a,b,c)$ の評価については研究が蓄積されている．

ゆえに所与のデータの $\Pr(a,b,c)$ を評価し，個体識別が観測されていない過去の事例で最大の $\Pr(a,b,c)$ 以下なら，そのデータが個体識別不可能ということの統計的証拠があるといえる．

結論を述べよう．標準的な $\Pr(a,b,c)$ の計算手法を定め，多くの事例で $\Pr(a,b,c)$ の評価値を蓄積すれば，客観的な匿名化はまったくの絵空事ではない．

4. 終わりに

本論文では官庁統計における過去から現在の匿名化実務を観測し，現実的な匿名化を定式化した．また観測データに基づいて，匿名化の程度を決める方法を提案した．我々の行ったことは，必ずしもデータ化されていない経験からの（非機械）学習にほかならない．

データが曖昧かつ限られていてもデータ駆動の考え方は有効で，観測事実を理論化して情報を補うことができる．泥臭く難しいが，現実的な攻撃者をモデル化するうえで妥当な方針ではないか．

現実性という制約をおけば，情報保護の技術革新において不利かもしれない．また最悪の攻撃者を想定すれば理想の議論ができる．しかしデータ分析で必要とされているのは，最悪の場合でも識別不可能なデータではなく，現実的に識別不可能なデータである．

現実的に識別不可能か判定する方法について，本論文はあまり強い仮定を置かない一例を示した．現実が曖昧で観測が限られているからこそ，現実的な匿名化の研究は広大な未開領域である．有用なデータを必要とする統計家の一員として，本領域における計算機科学の貢献を心待ちにしている．

参考文献

[1] Anderson, M. and Seltzer, W.: Federal Statistical Confidentiality and Business Data: Twentieth Century Challenges and Continuing Issues, *Journal of Privacy and Confidentiality*, Vol.1, pp.7–52 (2009).

[2] Council of Europe: Protection of personal data used for scientific research and statistics (Recommendation No.R(83)10) (1983).

[3] Cox, L.: Disclosure risk for tabular economic data, *Confidentiality, disclosure, and data access*, Doyle, P. et al. (Eds.), Elsevier, Amsterdam, pp.167–184 (2001).

[4] Dalenius, T.: Towards a methodology for statistical disclosure control, *Statistik Tidskrift*, Vol.15, pp.429–444 (1977).

[5] Dwork, C.: Differential privacy, *Proc. 33rd International Colloquim on Automata, Languages and Programming*, pp.1–12 (2006).

[6] Elliot, M. et al.: Data Environment Analysis and the Key Variable Mapping System, *Privacy in Statistical Databases*, Domingo-Ferrer, J. and Magkos, E. (Eds.), *LNCS*, Vol.6344, Springer, Berlin, pp.138–147 (2010).

[7] Federal Committee on Statistical Methodology: Report on Statistical Disclosure and Disclosure Avoidance Techniques, Statistical Policy Working Paper 2, U.S. Department of Commerce, Office of Federal Statistical Policy and Standards, Washington, D.C. (1978).

[8] Federal Committee on Statistical Methodology: Report on Statistical Disclosure Limitation Methodology, Statistical Policy Working Paper 22, U.S. Office of Management and Budget, Statistical Policy Office, Washington, D.C. (1994).

[9] Fung, B. et al.: *Introduction to Privacy-Preserving Data Publishing*, CRC Press, New York (2011).

[10] 濱砂敬郎：統計調査制度とミクロ統計の開示（松田芳郎他編），chapter 1.7, pp.109–128, 日本評論社 (2000).

[11] 星野伸明：エビデンスに基づいた匿名化，Discussion Paper 21, 金沢大学経済学経営学系 (2013).

[12] Machanavajjhala, A. et al.: ℓ-diversity: Privacy beyond kappa-anonymity, *Proceedings of the 22nd International Conference on Data Engineering*, p.24 (2006).

[13] Marsh, C. et al.: The Case for a Sample of Anonymized Records from the 1991 Census, *Journal of the Royal Statistical Society, Series A*, Vol.154, pp.305–340 (1991).

[14] 松田芳郎：統計調査制度とミクロ統計の開示（松田芳郎他編），chapter 1.1.5, pp.19–23, 日本評論社 (2000).

[15] Singer, E. et al.: The impact of privacy and confidentiality concerns on survey participation: The case of the 1990 U.S. Ceusus., *Public Opin. Q.*, Vol.57, pp.465–482 (1993).

[16] Singer, E. et al.: Attitudes and behavior: The impact of privacy and confidentiality concerns on participation in the 2000 Census, *Public Opin. Q.*, Vol.67, pp.368–384 (2003).

[17] Sweeney, L.: k-Anonymity: A model for protecting privacy, *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems*, Vol.10, pp.557–570 (2002).

[18] 宇賀克也：情報法（宇賀克也・長谷部恭男編），chapter 6, 有斐閣 (2012).

星野 伸明

1971 年生．1994 年東京大学経済学部経済学科卒業．1996 年同大大学院修士課程修了．1999 年同博士課程単位取得退学．1999 年金沢大学講師．2001 年カーネギーメロン大学客員教授．2003 年金沢大学助教授．2004 年東京大学経済学博士．2005 年同大客員准教授．2011 年金沢大学教授．統計的開示制限や離散分布論の研究に従事．日本統計学会，応用統計学会各会員．