２０２１年度　先端数理科学研究科

博士学位請求論文（要旨）

個人情報の識別リスク評価に基づいた匿名化に関する研究

先端メディアサイエンス専攻
伊藤　聡志

１　問題意識と目的

機械学習や AI 技術の発展により，ビックデータの利活用が民間企業・医療機関・金融機関・行政など多様な場面で盛んになっている．移動履歴や購買履歴などのビッグデータを分析することによって我々は様々な知見を得ることができ，それを利活用することによって大きな利益を生むことができる．しかし，ビッグデータは有用であるのと同時に危険なものでもある．例えば，2006 年には，Sweeny らによって ZIP コード, 生年月日, 性別の 3 つの属性の組み合わせから米国居住者の 87%を一意に識別できることが報告された．また，2019 年には Rocher らによってデータからランダムにサンプリングされた 0.5%のデータからでも，個人の一意性を誤差平均 0.028 で求められることが示され，大きな反響をよんだ．そのため，企業や組織は収集したビッグデータを利活用する際，そのデータから個人が識別されてしまうリスクを評価し，そのリスクを低減するために**匿名化**を行う必要がある．

匿名化とは，PII（personal identifiable information）から個人が特定されないようにデータの一部の削除やランダムなノイズを追加（摂動化）することであり，匿名化されたデータから特定の個人を識別する攻撃を再識別と呼ぶ．健康情報の匿名化の標準化文書 ISO/TS 25237 では，匿名化（anonymization）が「データ管理者（data controller）が単独，または他者と協力して，データ主体（data subject）を直接的または間接的に識別できないように, 個人データを不可逆的に変更する手順」と定義されている. 国内においては, 匿名化技術は 2017 年の個人情報保護法改正で導入された「匿名加工情報」に関わっている．匿名加工情報は,「特定の個人を識別することができないように個人情報を加工し，当該個人情報を復元できないようにした情報」であり，病歴などの要配慮個人情報を第三者に提供する際には，データに含まれる本人の同意をあらかじめとる（オプトイン）か，個人情報の第三者提供とならないようにデータを匿名加工情報とすることが必要となった．この匿名加工情報を作成するために，匿名化技術は顧客のビッグデータを有する多くの企業や組織から注目を集めている．

一般に，データを匿名化すると個人が識別されるリスクが低くなるため安全性は上がるが，データ中の値を加工するため有用性は下がる．しかし，匿名化による安全性や有用性の変化度合いには，匿名化に用いる加工手法，データの安全性や有用性を評価する指標，加工の対象となるデータ，などの様々な要因が影響するため評価が困難であり，これまで明らかになっていなかった．本研究の目的は，このような**データに対する匿名化の影響を明らかにすること**である．

２　構成及び各章の要約

本論文では研究目的を達成するために，以下の 4 つの課題に取り組んでいる．

（**課題 1：既存の攻撃者モデルの問題点**）匿名化されたデータの安全性を評価するためには，そのデータから個人を識別しようとする攻撃者の想定をする必要がある．攻撃者はデータについての何かしらの背景知識

を持っており，それを手がかりにして個人の識別を試みることが考えられるが，この背景知識によって攻撃者の危険度は大きく変わる．これまで提案されてきた攻撃者モデルの例として，Domingo-Ferrer らによって提案された最大知識攻撃者モデルや, El Emam によって提案された Dumber 数モデル（[K. El Emam and L. Arbuckle, “Anonymizing Health Data”] p.28, T2: Inadvertent Attempt at Re-identification）がある．しかし，最大知識の仮定は強力すぎて現実的でない．一方の El Emam のモデルは人間が安定して維持できる知人の数として知られている Dumber 数に基づいており，楽観的すぎる．

（**課題 2：履歴データの識別リスクの問題点**）既存の匿名化手法のほとんどは，時刻情報の無い静的なデータを仮定しており，動的に変化するデータから個人が識別されるリスクについては，限られた研究しか行われていない．その理由には，商品の購買履歴や位置情報のように，仮名のもとに結びつけられた時系列データから，即時に個人が識別されることはないという誤解に加えて，動的に変化するイベントを正確に定式化することの技術的な困難性があると考える．

（**課題 3：*k*-anonymity の問題点**）データの安全性を評価する指標の代表的なものとして，Sweeny らによって提案された *k*-anonymity があり, これはデータ中のいかなる個人のデータも少なくとも *k* 人の個人が同じ値を持っており，区別がつかないことを保証するものである．データを *k*-anonymity を満たすように加工する手法を *k*-匿名化と呼ぶ．しかしこの指標には課題も多く，Tamir らは「*k*-anonymity のためにできてしまう *k*+1 人以上のグループ」が過度な加工であると指摘している．

（**課題 4：実験データへの依存性**）多くの匿名化の既存研究では, *k*-anonymity の *k* などのパラメータを様々な値に変化させることにより，実験的に加工コストを求めている．しかし，加工コストや最適な *k* を求めることは，対象となるデータや想定するユースケースシナリオ等に大きく依存するので，非常に困難である．

これらの課題に対して，本論文では次の方法で解決を試みている．

課題 1 を解決するために, 本研究では, 部分的な背景知識を有する新たな攻撃者モデルを提案する. また, その攻撃者が個人を識別できる確率の期待値（平均識別確率）を近似する数理モデルを用いてデータの安全性を理論的に評価する．

課題 2 を解決するために,「*x* レコードのデータが与えられた時, その中のユニークな値の数（種類）*y* は, 全 ℓ 種類のうちいくらか？」という問題（**商品種類数問題**）を考え，それを解決する履歴データの新たな数理モデルを提案する．本モデルを用いることにより，データの基本統計量や加工パラメータといった加工前に手に入る値から加工コストの見積りをできるようになる．

課題 3 を解決するために，Tamir らによって提案された *k*-concealment という指標に着目する．本研究では，履歴データが *k*-concealment を満たすための新たな匿名化手法や，データ中の全ての個人が等しく *k* 人と区別がつかない状態（完全 *k*-concealment）に加工する新たな手法を提案する．

課題 4 のデータ依存の問題は，データサイエンスにおいては不可避な命題であろう．ここでは，データに特有のふるまいによって想定外の評価をしてしまう危険性を最小化するため，購買履歴データや健康診断データなどの，出来るだけ多種多様な分野のオープンデータや匿名加工情報を評価して対処する．

本論文は 11 個の章で構成されており，各章の要約は以下の通りである．

1 章では，本研究の背景，研究目的，既存の研究における問題点と，それに対する本研究のアプローチと新規性を述べる．

2 章では，*k*-anonymity, *k*-concealment や最大知識攻撃者モデルなどの，本稿と大きく関わる主要な先行研究について解説し，課題を整理する．

3 章では，課題 1 を解決するために，新たな攻撃者モデルを提案している．データ中のある一つの値を確率的に得る攻撃者モデルを提案し，その攻撃者の平均識別確率によってデータ中のどの属性が危険であるかを評価する．また，平均識別確率を近似する 3 つの数理モデル（平均モデル，最小コストモデル，サンプリングモデル）を提案し，これらを用いて 4 つのデータ（購買履歴データ，糖尿病患者データ，世帯収入データ，ローン借り入れデータ）に対して安全性の理論的な評価結果を示している．購買履歴データの時刻属性

の値 1 つのみから平均 32%の確率で個人が識別されることや，データによっては提案モデルでも精度よくリスクを評価できることなどを明らかにしている．

4 章では，課題 2 を解決するために，履歴データのふるまいを数理モデル化する．提案データモデルでは，履歴データの値が一様に生起する仮定の下，履歴データ中に登場する項目の種類数の確率分布とその期待値が与えられる．また，提案モデルを応用することにより，履歴データを k-匿名化するために必要なダミーレコード数の期待値を，元の履歴データの統計量やパラメータ k などから求めることができることを示している．さらに，匿名化の加工コストを理論的に評価した結果に基づき，従来は加工前に算出することが困難であった k-匿名化における k の最適値を算出している．

5 章では，課題 4 を解決するために，購買履歴に対する匿名化の影響を実験的に評価している．購買履歴データの個人が購買商品特徴から識別されるリスクを想定し，そのリスクへの耐性を高めるためにかかる加工コストを実験的に評価している．購買履歴データを 50 個のクラスタに分割して k-匿名化をするためには，約 18 万のダミーレコードの追加が必要であることなどを明らかにしている．

6 章では，課題 4 を解決するために，健康診断データと傷病/医薬品レセプトデータを匿名化することによって，データの安全性と有用性がどのように変化するかを実験的に評価している．病歴/処方歴を k-匿名化することによって，識別される人数の割合が平均 2.9%まで減少すること，高血圧に対する相対リスクが相対誤差で 0.073 しか変化しないことなどを明らかにしている．

7 章では，課題 4 を解決するために，複数用途の含まれるデータから個人が識別されるリスクを実験的に評価している．乗降履歴や購買履歴などの複数用途の履歴が含まれている交通 IC カードデータから個人が識別されるリスクを，エントロピーを用いて定量化している．単一の乗降履歴によって個人を識別できる確率が 3.3%から 28.4%まで上がることや，購買履歴と乗降履歴を 1 つずつ知られた場合は，識別率が 88.1%まで上がることなどを明らかにしている．

8 章では，課題 4 を解決するために，世帯収入に対する匿名化の影響を実験的に評価している．世帯支出データの個人がレコード間のユークリッド距離から識別されるリスクを想定し，そのリスクへの耐性を高めるための加工手法を検討している．ノイズ付加や値のスワップのような単純な摂動化では再識別を全く防げないことや，k-匿名化によって再識別率を 17%まで下げられることなどを明らかにしている．

9 章では，課題 3 を解決するために，履歴データに対する k-concealment 化手法を提案する．Tamir らが提案した k-concealment 指標は，レコード数が個人数より多い履歴データは想定されていなかった．本章では，仮名の一般化とレコード間の k-concealment という新しい方式を提案し，個人によってレコード数の異なる履歴データを k-concealment を満たすように加工する手法を実現した．提案手法を用いることにより，従来手法では避けられなかった顧客やレコードの追加/削除をすることなく，履歴データを最低でも k 人の区別がつかない状態にすることができる．

10 章では，課題 3 を解決するために，新たな完全 k-concealment 化アルゴリズムを提案する．データを全ての個人が等しく k 人と区別がつかない状態に加工することを完全 k-concealment 化と定め，巡回セールスマン問題の近似解法やクラスタリングを応用して完全 k-concealment 化におけるコストを低減している．提案手法によって，k-匿名化によって生じる過度な加工や，k-匿名化されたデータ内の個人識別リスクの不公平さを解消することができる．

11 章で，本研究を結論づける．本論文では，データに対する匿名化の影響を明らかにするため，4 つの課題を設定してそれらを解決した．本研究の貢献は以下の 3 つである．（**貢献１：匿名化による安全性・有用性変化の理論的評価**）攻撃者や履歴データをモデル化することにより課題 1,2 を解決し，データの安全性や有用性を理論的に評価した．（**貢献 2：匿名化による安全性・有用性変化の実験的評価**）課題 4 を解決するために多種多様なデータセットに対する識別リスクを想定し，それらを匿名化した際の影響を実験的に評価した．（**貢献 3：新たな匿名化手法の提案**）課題 3 を解決するために k-匿名化の際に生じる過度な加工を解決する k-concealment 指標に注目し，これを満たすための新たな匿名化手法を提案した．このように本研究では，4 つの課題を解決することによって，データに対する匿名化の影響を明らかにした．