Vol.2016-CSEC-74 No.31
Vol.2016-SPT-19 No.31
2016/7/15

k-匿名性による特定可能性分析に基づいたデータプライバシーのリスク分析

山岡 裕司[1,a)]　伊藤 孝一[1]

概要：匿名化されたパーソナルデータの二次活用では，そのデータが曝露するプライバシーのリスクアセスメントが必要である．リスクアセスメントの従来技術に JO モデルがあるが，綿密なリスク分析ができない問題があり，多くの場合で匿名化の効果を評価できない．本論文では，レコード特定可能性を k-匿名性で分析し，綿密なリスク分析を支援する方式を提案する．提案方式はまず，JO モデルの本人特定容易度を，匿名化の効果が反映されやすいように変更し，それによって算定した漏洩個人情報価値をリスクとする．そして，高リスクのレコードを高速に抽出し，具体的なプライバシー侵害シナリオを添えて提示する．k-匿名化のベンチマークであるデータ Adult やそれを匿名化したデータに適用した結果，提案方式は従来より綿密なリスク分析ができることを確認した．

キーワード：プライバシー，リスクアセスメント，k-匿名性

Risk Analysis for Data Privacy based on Identifiability Analysis with k-Anonymity

Yamaoka Yuji[1,a)]　Itoh Kouichi[1]

1. はじめに

パーソナルデータ（個人に関する情報）[14] の活用が注目されている．たとえば，カナダの CHEO (Children's Hospital of Eastern Ontario) は生誕に関する情報を匿名化して提供しており，製薬会社，保険会社，そして IT 企業などに活用されている [16]．また，日本では，適切な規律の下でのパーソナルデータの活用を促進するなどの目的で，2015 年 9 月に改正個人情報保護法[*1]が成立した．改正個人情報保護法では本人の同意がなくても事業者による第三者提供を可能とする枠組み「匿名加工情報」が導入されている．

事業者がパーソナルデータを適切に活用するには，データプライバシーのリスクアセスメント[*2]が必要である．なぜなら，パーソナルデータの漏洩や不適切な利用などの事故は，データ主体本人のプライバシーを侵害しかねず，慰謝料や損害賠償を請求されたり，その後のパーソナルデータ収集が困難になったりと，重大な損害をもたらしかねないためである．データプライバシーのリスクとは，それらの事故によるプライバシー侵害のリスクのことであり，事故時の事業者の想定損害に相当する．データプライバシーのリスクを，以降では単にリスクという．リスクは，プライバシー侵害の起こりやすさと，起こったときの侵害の度合いの組み合わせで表現される．セキュリティ機能の導入，データ取扱者の教育，そして適切な契約などで事故の可能性は低減できるが，それらの実施はコストとのトレードオフがあり，コストがリスクに見合っているか判断するためにはリスクアセスメントが必要である．また，リスクを低減する有効な手段に，データの匿名化がある．匿名化

[1] 株式会社富士通研究所
FUJITSU LABORATORIES LTD.
[a)] yamaoka.yuji@jp.fujitsu.com

[*1] 個人情報の保護に関する法律及び行政手続における特定の個人を識別するための番号の利用等に関する法律の一部を改正する法律

[*2] ISO 31000 / JIS Q 31000

とは，パーソナルデータを誰のデータかわからないようにし，不必要な情報を減らすデータ加工である．匿名化はリスク低減と活用性維持のトレードオフである．許容できるリスクの範囲で活用性を最大限維持したいので，匿名化した場合にリスクがどう低減したかを評価する必要がある．

匿名化のためのリスクアセスメントは，あらかじめ具体的な匿名化方法が決まっている場合は不要な可能性も考えられるが，そのような場合は現状では少ない．改正個人情報保護法では，個人情報を匿名加工情報の加工基準に沿って匿名化すれば，個人情報ではなく匿名加工情報として扱えるとしている．そのため，加工基準に沿って匿名化すれば，事業者のリスクを十分低減でき，アセスメントの必要がなくなるという考え方もできる．しかし，その基準は未公開である．法案を作成した「パーソナルデータに関する検討会」は汎用的な匿名化方法は存在しないと報告しており [15]，包括的かつ具体的な匿名化基準が作られるには時間がかかりそうである．よって，匿名化のためのリスクアセスメントも必要な状況である．

以降では，パーソナルデータは，過去の多くの匿名化の研究 [5], [6], [8], [9], [10], [12] に倣い，1 人 1 レコードの 2 次元表とする．表の属性は個人の各情報を示し，各セル値は当該レコードの個人の当該属性の情報である．また，必要に応じて，その表に含まれていることの意味を属性として表現する．たとえば，表が政治的な集会の出席者名簿の場合，表に記載されているデータだけを分析するのは不適切であり，表に記載されていることの意味を属性として表現しその機微性を考慮できるようにすべきである．

包括的な匿名化基準の候補に k-匿名化 [9] があるが，属性の一部にしか適用されない場合が多く，その場合のリスクはアセスメントが必要である．k-匿名化は，QI (Quasi-Identifier) と呼ばれる対象属性について，どのレコードも k レコード以上で QI の全属性値が同じになるよう，各セル値を必要に応じて一般化する方法である．一般化は，セル値の抑制（任意の値の可能性を示す値への一般化）や属性の削除も含む．「パーソナルデータに関する検討会」は，全属性を QI として k-匿名化すれば「非識別非特定情報」になると報告している [15]．そのため，k-匿名化は，匿名化前のデータを保持しても，個人情報でないパーソナルデータに加工する方法として有望である[*3]．しかし，QI とする属性が少ないほど活用性は維持されるため，従来の多くの研究 [8], [9], [10] では全属性を QI としていない．一方で，QI を適切に設定することの難しさが指摘されており [7]，QI の設定が不適切なことによるリスクは k-匿名化では対応できない．従って，全属性を QI としない限りはリスクアセスメントは必要である．

リスクアセスメントの従来技術に JO モデル [13] がある．

[*3] 第 189 回国会の答弁から，保持しているデータとの照合により特定個人が識別できるデータは個人情報の可能性が高い．

表 1 パーソナルデータの例
Table 1 An Example of Personal Data

メールアドレス	年齢	職業	本籍	B 社顧客
hanako@...	12	ピアニスト	大字	T
haruko@...	18	会社員	本町 1	T
natsuko@...	18	公務員	本町 1	T
taro@...	43	会社員	本町 1	T
jiro@...	43	会社員	大字	T
saburo@...	43	公務員	本町 1	T

JO モデルは，個人情報漏洩時の損害賠償金想定額を各レコードに対し算出する方法である．そのうち，データに関するリスクに相当する「漏洩個人情報価値」は，「基礎情報価値」と「機微情報度」と「本人特定容易度」の積で金額化される．JO モデルにより，俯瞰的で大雑把なリスク分析を実現できる．

しかし，JO モデルは，本人特定容易度の算定方法が大雑把過ぎるため，綿密なリスク分析，特に匿名化のためのリスク分析ができないという問題がある．具体的には，氏名と住所を含まないレコードは全て「特定困難」という一括りで算定される点が問題である．実際には，氏名や住所を含まなくても特定個人を識別できる可能性が高いレコードがある．たとえば，**表 1** はそのようなレコードを含むパーソナルデータの例で，事業者 B 社が保有していると想定したデータである．表 1 の各レコードは氏名も住所も含まないがメールアドレスを含んでおり，特定個人の友人などにとってその識別は難しくない．この場合，一定のリスク低減効果がある匿名化としてまずメールアドレスの削除が考えられる．しかし，その匿名化を実施しても JO モデルでは額が変わらない．また，年齢や職業などは一見すると特定個人の識別につながりにくいデータだが，特異値を持つ特定個人はそれらだけで識別できる可能性が高い．たとえば，表 1 の最初のレコードは 12 歳のピアニストであり，一般的に稀な人であるため，たとえメールアドレスを削除しても依然として特定個人の識別につながりやすい．しかし，JO モデルでは特異値かどうかは区別されないため，その他のレコードと同じ額と算定される．

本論文では，レコード特定可能性を詳細に分析することで，綿密なリスクアセスメントを支援する方式を提案する．提案方式はまず，JO モデルの「特定困難」な場合の本人特定容易度を，匿名化の効果が反映されやすいように変更し，それによって算定した漏洩個人情報価値をリスクとする．そして，プライバシー侵害シナリオを，リスクが高いものから優先的に，高速に抽出する．プライバシー侵害シナリオとは，パーソナルデータを見た攻撃者（プライバシー侵害を企てる者）が，一部の情報でレコードを特定でき，そのレコードの他の情報を知る事ができる，という具体的な因果関係である．たとえば，表 1 で，12 歳のピアニストのレコードは 1 つしかなく，その本籍は「大字」だとわか

る，といった関係である．リスクアセスメントにおいて，高リスクなレコードとそのプライバシー侵害シナリオを具体的に把握できれば，そのリスクを受け入れられるか評価しやすく，またリスク低減する場合の匿名化方針も立てやすくなる．たとえば，直前のプライバシー侵害シナリオ例が受け入れがたい場合，年齢を 10 歳階級に一般化したり，本籍を県レベルに一般化したりしてリスク低減を狙う，といった方針を立てられる．プライバシー侵害シナリオの抽出は本論文が初めて提案する．「特定困難」な場合の本人特定容易度は，JO モデルの機微情報度がより低い属性のより少ない組み合わせで，レコード特定可能なレコードほど高くなるように変更した．これは，機微情報度の低い属性の少数の組み合わせでレコードを特定できるほど，攻撃者は特定個人の識別が容易であるというモデルである．特異値は少ない属性組み合わせでもレコード特定性を高めるため，提案方式では特異値を含むレコードは本人特定容易度が高くなりやすい．提案方式のリスクは匿名化の効果が反映されやすいため，匿名化の比較のための指標として使いやすい．

ただし，各レコード対し，最も特定しやすい属性集合を抽出するには，計算量が大きい（指数関数的）という課題がある．そのため我々は，高リスクの可能性が高い属性集合から優先的に分析するアルゴリズムを開発した．これにより，高リスクのプライバシー侵害シナリオを高速に抽出できる．

我々は，k-匿名化のベンチマークとして使われているデータを使って実験し，提案方式の実用性を確認した．プライバシー侵害シナリオの抽出により，確かにリスクが高いと感じるレコードが抽出されたことを確認した．また，我々のリスクは JO モデルより匿名化の効果が反映されやすいこと，我々のアルゴリズムが実用的な性能であることを確認した．

2. 関連研究

関連研究について述べる．

レコード特定可能性の指標として，k-匿名性 [10] がある．k-匿名性は，どのレコードも k レコード以上で QI の全属性値が同じという，表の性質であり，k-匿名化 [9] は k-匿名性を達成するようなデータの一般化である．先述の CHEO では k-匿名性やその類似指標をリスクアセスメントで使用している．しかし，k-匿名性は特定可能性の指標であって，レコード特定による結果が考慮されておらず，リスク分析に使いづらい場合があるという問題がある．たとえば，表 1 に比べ，表 1 から本籍を削除したデータの方が低リスクなのは明らかだが，k-匿名性はどちらも同じ値（$k = 1$）と算定される．また，先述の通り，一部の属性を QI とする場合も多く，その妥当性のリスクは分析対象外である．k-匿名性の安全性向上の提案にあたる l-多様性 [8] なども，QI の設定に安全性が依存しているがそのリスクは分析対象外である．

特定個人のレコードがデータに含まれているかどうかの指標として，ϵ-差分プライバシー [2] がある．ϵ-差分プライバシーは，攻撃者にとって誰のレコードも，データに含まれている確率と含まれていない確率の比が $\exp(\epsilon)$ 以下であるような性質である．これも，レコードが含まれているか否かが推定されることによる結果が考慮されないという問題がある．また，ϵ-差分プライバシーは，偽のレコードを生成し得る加工をすることを前提としているが，Fung らが指摘している通りそのような加工は活用者が真正性を要求する場合には適用できない [4] ため，そのような場合には使用できない．

提案方式は，リスクを定量化するため，リスク分析に使いやすいという利点がある．リスクである漏洩個人情報価値は，本人特定容易度だけでなく機微情報度にも比例する．つまり，レコード特定による結果も考慮した指標となっている．全属性を対象に適用でき，k-匿名化における QI 設定のリスクも定量化される．一方，偽のレコードを含むデータには，適用しても適切なリスク分析がおこなえないという制限がある．

また，提案方式は，プライバシー侵害シナリオを抽出する初めての方式である．これにより，高リスクなレコードを具体的に確認でき，匿名化の方針を効率的に立てられるようになる．

3. JO モデルの漏洩個人情報価値

提案方式の元となる JO モデル [13] の漏洩個人情報価値について述べる．

JO モデルは，保険，ICT，そしてセキュリティなどの各専門家達により作られた，個人情報漏洩時の損害賠償金想定額の算定モデルである．実際の複数の漏洩事件の判決との比較による検証もされている．

JO モデルのうち，データから算定する金額が漏洩個人情報価値である．レコードと，各セル値が示す類型を用意すれば適用でき，そのレコードの金額を計算できる．次の式で詳細化される．

$$\begin{aligned}&\text{漏洩個人情報価値}\\ &= \text{基礎情報価値} * \text{機微情報度} * \text{本人特定容易度}\end{aligned}$$

基礎情報価値は 500 円である．

機微情報度 σ は各セル値の類型の「EP レベル」から計算できる．属性 a のセル値の類型の EP レベルは，経済的損失レベル $E_a \in \{1, 2, 3\}$ と精神的苦痛レベル $P_a \in \{1, 2, 3\}$ の組である．代表的な類型は JO モデルで EP レベルが定められている．たとえば，氏名，住所，生年月日，メールアドレス，そして職業などの EP レベルは $\{E : 1, P : 1\}$，年収・年収区分，所得などの EP レベルは $\{E : 2, P : 2\}$，

政治的見解，本籍などの EP レベルは $\{E:1, P:3\}$ である．σ は次式で計算される．

$$\sigma = s(A), \tag{1}$$

$$s(I) = 5^{\max_{a\in I}(E_a)-1} + 10^{\max_{a\in I}(P_a)-1} \tag{2}$$

ここで，式 1 の A は表の全属性である．

本人特定容易度 ι は，**表 2** により決められる．

たとえば，表 1 の各レコードの漏洩個人情報価値 ϕ は次のように算定される．まず，B 社顧客は高所得者が多いことが知られているとし，属性「B 社顧客」のセル値「T」の類型の EP レベルを，所得に準じ $\{E:2, P:2\}$ とする．その他のセル値の類型の EP レベルは，属性に対応させる（年齢は生年月日の EP レベルとする）と，$\max(E)=2, \max(P)=3$ となり，$\sigma = 105$ となる．また，各レコードは氏名も住所も含まないため，$\iota = 1$ となる．よって，

$$\phi = 500 * \sigma * \iota = 52{,}500 \text{ [円]}$$

となる．表全体の総額は，6 レコードあるので $52{,}500 * 6 = 315{,}000$ 円となる．

4. 提案方式

提案方式は JO モデルの漏洩個人情報価値のうち，「特定困難」な場合の本人特定容易度を 2 以下とするように変更したものである．変更後の本人特定容易度を ι' と，漏洩個人情報価値を ϕ' とする．変更の詳細は後述する．

提案方式は JO モデルと違い，各レコードの本人特定容易度の算定に表全体のデータを使用する．これは，表中でそのレコードが特定できるか分析するためである．

提案方式は，表と，各セル値が示す類型を用意すれば適用でき，各レコードのプライバシー侵害シナリオとそれに対応する ι' および ϕ' を出力できる．ただし，一般的には全てのプライバシー侵害シナリオを抽出するには計算量が多いため，閾値の設定次第でリスクの低いレコードの分析が省略される．より高いリスクを受け入れられる場合，より低いリスクも受け入れられるのが普通なため，リスクの低いレコードの分析は省略しても問題になりにくい．プライバシー侵害シナリオ抽出のアルゴリズムは後述する．

4.1 課題

本論文では，本人特定容易度の判定基準が「特定困難」な場合に，特定個人の識別しやすさに関して明らかと考えられる次の各性質について，JO モデルが全くあるいはほとんど対応できていないことを課題とする．

性質 1 攻撃者は特定個人を識別するために必要な知識が入手しやすいほど，特定個人を識別しやすい．たとえば，表 1 で 2 つ目のレコード（haruko@...）は年齢と職業で特定でき，5 つ目のレコード（jiro@...）は年齢と本籍で特定できる．攻撃者は普通は職業より本籍の方が知りづらいので，他の条件が一緒であれば，後者の方が特定個人を識別しづらいと考えられる．

性質 2 各レコードについて，特定に必要な属性の数が多いほど攻撃者は特定のために多くの知識が必要になり特定個人を識別しづらくなる．たとえば，表 1 で 4 つ目のレコード（taro@...）は，メールアドレスが削除された場合は，攻撃者がレコード特定するには特定個人について年齢と職業と本籍の 3 属性の情報を知っている必要があるため，たとえば年齢と本籍の 2 属性で特定できるレコードより特定が難しい．

性質 3 特異値を含むレコードは特定個人に識別されやすい．たとえば，表 1 で特異値と考えられる 12 歳ピアニストというデータを含む最初のレコードは特定個人の識別につながりやすい．

機微情報度は課題としなかった．ただし，いくつか課題になり得る項目が考えられる．まず，機微情報度は属性の数を考慮しないが，属性の数が多いほど高リスクとなる方が妥当とも考えられる．しかし，そうするとリスクの上限を決めるのが難しくなり，金額の規模が JO モデルから乖離するため，課題としなかった．また，機微情報度の算定対象に，特定個人を識別するために必要な属性まで含めない方が妥当とも考えられる．しかし，属性値より一般化された情報で特定個人を識別された場合に，その属性値のより詳細な情報がわかってしまうリスクもあるため，課題としなかった．たとえば，表 1 では，攻撃者が，43 歳会社員で本籍が「本町 1」以外の特定個人を知っている場合，表中ではそのレコードは 5 つ目（jiro@...）に特定でき，本籍が「大字」であるという未知の情報を得ることができるリスクがある．

4.2 提案方式の本人特定容易度

提案方式は，本人特定容易度 ι' を，性質 1〜3 に対応させるために次の式とした．

$$\iota' = 2 * \max_{I\in J} i(I), \tag{3}$$

$$i(I) = \frac{0.9^{|I|-1}}{\log_8(s(I)-1)+1} \tag{4}$$

ここで，式 3 の J は当該レコードを特定できる属性集合の集合，式 4 の s は式 2 である．なお，全属性において一意に特定できないレコードは，個人情報とされない可能性が高く，リスクを無視できるとみなし，$\iota' = 0$ とする．また，$\iota \geq 3$ の場合は $\iota' = \iota$ とする．

式 4 の取り得る値の範囲は $(0, 1]$ であるため，式 3 の ι' の取り得る値の範囲は $(0, 2]$ である．JO モデルの本人特定容易度は表 1 の通り，住所がなくても氏名があれば 3 で，氏名もなければ 1 だが，マイナンバーなどの ID や電話番号やメールアドレスがあれば「特定困難」とまではいえないと考え，最大で中間の 2 を取れるようにした．

表 2 本人特定容易度の判定表（[13]）
Table 2 The Decision Table of Identifiability ([13])

判定基準	本人特定容易度
個人を簡単に特定可能. 「氏名」「住所」が含まれること.	6
コストをかければ個人が特定できる. 「氏名」または「住所 + 電話番号」が含まれること.	3
特定困難. 上記以外.	1

式 4 の分母は，性質 1 のモデル化に対応する．普通，攻撃者は他人の機微な情報，つまり機微情報度が高い情報ほど入手が難しいと考えた．機微情報度の対数を取っているのは，人の感覚は刺激の対数に比例することが多いためである．各個人は自分の感覚でリスクを評価して情報を他者に知らせるため，機微情報度が高い情報でもそれほどリスクを感じずに知らせる傾向があると考える．また，機微情報度が EP レベルの指数で表現されているのも同様の考え方によると考えられ，対数を取ることで EP レベルと同様の尺度に戻せる．対数の底が 8 なのは，式が簡単になる整数のうちで，EP レベルの尺度に一番近づけられる値だからである．

式 4 の分子は，性質 2 のモデル化に対応する．各属性の情報の得やすさが独立だとすると，複数属性の情報を同時に得ることは属性数に応じて指数的に難しくなるため，指数関数とした．攻撃者が通常以上に多くの属性を知らないと特定できない場合，2 レコードから無作為に特定個人を識別するのと同程度に難しいというモデルを考え，指数の底を決めた．El Emam らによると攻撃者が知っている属性は高々 7 つだろうと報告されている [3] ため，$|I| > 7$ の場合に 0.5 程度にする指数が良い．指数の底が 0.9 なのは，$|I| > 7$ で初めて 0.5 未満になり，また式が簡単になるためである．

式を簡単にしているのは，方式を定量的に評価することは難しく，細かい値の違いは重要でないためである．

式 3 で $I \in J$ による最大値を算出していることが，性質 3 のモデル化に対応する．特異値を含むレコードはそれらの特異値の少ない組み合わせを I に含んだ場合に，式 3 で最大値を与える $I \in J$ となりやすい．そのため，特異値を含むレコードは本人特定容易度が高くなる傾向がある．

本人特定容易度以外は $I \in J$ に非依存なため，$\iota = 1$ の場合は式 3 で最大値を与える $I \in J$ が ϕ' を決める．

4.3 方式検証

提案方式が，性質 1〜3 に対応していて，それにより匿名化の効果が反映されやすいようになっていることを例で示す．

まず，表 1 での本人特定容易度を確認する．JO モデルでは，前述の通り，各レコードとも $\iota = 1$ となる．提案方式では，各レコードとも式 3 より ι' は $I = \{$ メールアドレス $\}$ などの場合に最大値 $\iota' = 2$ となる．このように，提案方式では「特定困難」とまではいえないとして，JO モデルより高く算定する．

表 3 表 1 からメールアドレスを削除したデータの本人特定容易度 ι'
Table 3 Identifiability ι' for Data That Email Addresses Were Removed from Table 1

レコード番号	$I \in J$（一例）の要素	ι'
1	年齢	2.0
2	年齢, 職業	1.8
3	年齢, 職業	1.8
4	年齢, 職業, 本籍	0.50
5	年齢, 本籍	0.56
6	年齢, 職業	1.8

次に，表 1 からメールアドレスを削除したデータの本人特定容易度を確認し，性質 1〜3 への対応を確認する．JO モデルでは，各レコードとも変化はなく，$\iota = 1$ となる．提案方式では，レコードによって ι' が異なり，**表 3** の通りとなる．提案方式では，JO モデルと違い，性質 1〜3 へ対応していることが見てとれる．本人特定容易度は，それぞれ，性質 1 よりレコード番号 2 より 5 の方が低く，性質 2 よりレコード番号 4 より 5 の方が高く，性質 3 よりレコード番号 1 より 2 などの方が低くあるべきだが，表 3 はその通りになっている．

最後に，表 1 のメールアドレス削除による漏洩個人情報価値の変化を確認する．JO モデルでは，削除前後で変わらず総額 315,000 円となる．提案方式では，削除前は総額 630,000 円となり，削除後は総額約 440,000 円と約 7 割に減った．表 1 からのメールアドレス削除は匿名化効果があると考えられるが，JO モデルではそれを示せていない一方，提案方式では示せている．

4.4 プライバシー侵害シナリオ

プライバシー侵害シナリオは，より高リスクなものを抽出することが重要である．

各レコードで最も高リスクなプライバシー侵害シナリオは，ι' および ϕ' で最大値を与える $I \in J$ に対応するものである．つまり，I によりそのレコードが特定され，そのレコードの情報が推定されるというプライバシー侵害シナリオである．

そのため，提案方式では ι' と共に，その算定の過程で抽出した最大値を与える $I \in J$ を高リスクなプライバシー侵

害シナリオとして出力する．

たとえば，表 1 で最も高リスクなプライバシー侵害シナリオは，12 歳という年齢（やピアニストという職業）でレコードが特定され，本籍が「大字」であることや B 社顧客であることなどがわかる，というものである．

4.5 高速計算アルゴリズム

式 3 を高速に計算するアルゴリズムを提案する．

式 3 は，全属性の冪集合の各要素を順に調べていけば求められる．しかし，それでは属性数に対し指数関数的な計算量が必要なので，高速化が望まれる．

式 4 は，I の要素に対する単調性がある．具体的には，$I' \supset I \Rightarrow i(I') < i(I)$ が成立する．よって，式 3 の算出にあたっては，より少ない属性でレコード特定できる I を見つければ，それらを含む I' の計算は省略できる．

提案アルゴリズムは，相関ルールの高速抽出方式である Apriori アルゴリズム [1] を応用したものである．Apriori アルゴリズムは条件を満たす属性集合を抽出するため，小さい属性集合から順に処理をし，そのときの集合に属性を一つ追加した集合に対する処理を単調性を利用してできるだけ省略する．提案アルゴリズムも同様の処理をおこなう．

単調性を利用した基本アルゴリズムを Algorithm 1 に示す．主な入力は，表 T，表の全属性 A，レコード番号 r，閾値 t である．正確には各セル値の類型の EP レベルの情報も必要だが，省略している．出力は，式 4 の最大値を与える $I \in J$ の集合である．$T(r, I)$ は，表 T のレコード r のうち属性集合 I に対応するセル値のリストを示す．関数 *update* は，J の候補である C から C' を取り除き，C に C' の要素を一つ増やした新たな候補を必要に応じて追加する処理で，H は履歴情報である．たとえば，$C = \{\{a_2\}, \{a_3\}\}, H = \{\{a_1\}\}, C' = \{\{a_2\}\}$ の場合，$update(C, H, C') = (\{\{a_3\}, \{a_1, a_2\}\}, \{\{a_1\}, \{a_2\}\})$ などとなる．

当該レコードが全属性で k-匿名性（$k \geq 2$）を（ほとんど）満たしている場合，基本アルゴリズムは $|A|$ に関し $\mathcal{O}(2^{|A|})$ の計算量となる．一方，基本アルゴリズムはレコード特定ができた場合に最もリスクが高くなる属性集合から優先的に調べていくため，途中で打ち切っても高リスクのプライバシー侵害シナリオを抽出し損ねることはない．

そのため，基本アルゴリズムは調べた属性集合の種類が閾値 t 以上になった場合に，それ以上の抽出を打ち切れるようにしている．

基本アルゴリズムのレコード数 $|T|$ に関する計算量は，$R' = \{\, r' \mid r' \neq r \wedge T(r', I) = T(r, I) \,\}$ の抽出に必要な計算量と等しくなる．特に工夫をしなければその計算量は $|T|$ に比例するため，全レコードのプライバシー侵害シナリオの抽出には $|T|^2$ に比例する計算量が必要になる．各属性集合 I についてセル値の度数分布を計算して記憶おけば R' の計算量を $|T|$ に非依存にできるが，$\mathcal{O}(2^{|A|})$ のオーダーの記憶容量が必要となる．

Algorithm 1 基本アルゴリズム

Input: A table T, a set A of attributes of T, a record index r of T, a threshold t
Output: $\arg\max_{I \in J} i(I)$

```
C ← ∅
for all a ∈ A do
  C ← C ∪ {a}
end for
H ← ∅
n ← 0
while C ≠ ∅ do
  J' ← ∅
  C' ← arg max_{I∈C} i(I)
  for all I ∈ C' do
    if { r' | r' ≠ r ∧ T(r', I) = T(r, I) } = ∅ then
      J' ← J' ∪ I
    end if
    n ← n + 1
  end for
  if J' ≠ ∅ then
    return J'
  end if
  if n ≥ t then
    return ∅
  end if
  (C, H) ← update(C, H, C')
end while
return ∅
```

そこで，我々は基本アルゴリズムを変更し，レコード番号を入力側で走査するのではなく，I が変わった時点でレコードを走査して度数分布を作成するようにした．これにより，I が変わった時点で古い度数分布が不必要になるため，少ない記憶容量で高速化を実現できる．これを提案アルゴリズムとし，実験をおこなった．

5. 実験

提案方式の実用性を示すため，実際のデータで次を確認する必要があると考え，実験をおこなった．

- リスクが高いと感じるプライバシー侵害シナリオが実際に抽出されること．
- JO モデルより匿名化の効果が反映されやすく，匿名化の方針を立てやすいこと．
- 性能面で問題がないこと．

使用したデータは，k-匿名化のベンチマークである UCI Machine Learning Repository[11] の Adult である．訓練データの後にテストデータを追加した 48,842 レコードを用いた．属性は，過去の匿名化の研究 [5], [6], [8] でよく用いられる 9 属性を用いた．各属性と，それらに紐付けた，類型および EP レベルと，そして k-匿名化や情報量算出で用いる一般化階層の一覧を**表 4** に示す．なお，一般化階層

表 4　Adult の属性の情報
Table 4　Attribute Information on Adult

属性名	類型	EP レベル	一般化階層の高さと出典
age	生年月日	E:1, P:1	4 [5]
workclass	職業	E:1, P:1	2 [12]
education	学歴	E:1, P:2	1
marital-status	家族構成	E:1, P:1	2 [12]
occupation	職業	E:1, P:1	1 [12]
race	人種	E:1, P:2	1 [12]
sex	性別	E:1, P:1	1 [12]
native-country	国籍	E:1, P:2	1 [12]
INCOME	年収・年収区分	E:2, P:2	1 [5]

の根の値（抑制したことを示す値）は欠損値「?」と同じとし，欠損値以外のセル値の類型はその属性の類型としたが，欠損値はその属性の情報がないものとして扱った．

提案方式を適用し，プライバシー侵害シナリオとして抽出されたもののうち，最も高リスクとされた 7 レコードを**表 5** に示す．なお，$\iota' = 1.8$ のレコードは他に 127 行抽出された．

表 5 を見ると，母集団が大きくなければ，確かに特定個人の識別ができそうなレコードが抽出されている．たとえば，1 行目のレコード番号 24028 は，年齢（86）だけで特定できる唯一のレコードであり，さらに性別（女）や婚姻状況（未婚）などを組み合わせると母集団でも稀な可能性が高い．また，3 行目のレコード番号 2697 は，年齢（18）と婚姻状況（離婚）だけで特定でき，さらに性別（男）などを組み合わせると母集団でも稀な可能性が高い．そして，5 行目のレコード番号 44169 は，職業（未就労）と婚姻状況（配偶者不在）だけで特定でき，さらに年齢（20）や性別（男）などを組み合わせると母集団でも稀な可能性が高い．こういったことから，年齢と婚姻状況の組み合わせには特定性の高い値が多いのではないか，といったことがわかる．なお，Adult は米国の国勢調査結果の標本であり，母集団が大きいため，実際には特定個人の識別はできないかもしれない．しかし，仮に母集団が小さいとすると，特定個人を識別できる攻撃者が存在する可能性は高そうである．

次に，数種類の匿名化の出力結果である各データに対し，JO モデルと提案方式を適用し，両者を比較した．匿名化は，次の 3 つの方式を用いた．

(1) age を 10 歳階級化．
(2) occupation と INCOME 以外を QI とし，k-匿名化．
(3) INCOME 以外を QI とし，k-匿名化．

k-匿名化は，Xu らのアルゴリズム（Top-Down 法）[12] を実装し，$k = 3$ で適用した．匿名化による活用性の低下は，情報量を計算して定量化した．情報量は，統計コミュニティで良く使用される情報利得（Kullback-Leibler ダイバージェンス）に基づいて計算した．具体的には，変換前後の表を T, T' とし，欠損値のみの表を $T_\emptyset$ とすると，情報量 U は次の式で算出した．

$$U = 1 - D_{\mathrm{KL}}(T||T')/D_{\mathrm{KL}}(T||T_\emptyset)$$

ここで，$D_{\mathrm{KL}}(P||Q)$ は Q に対する P の情報利得である．

適用した結果を**図 1** に示す．図 1 の凡例について，0 は無加工で，1〜3 は先述の方式 (1)〜(3) である．情報量は多い方が良く，漏洩個人情報価値は低い方が良いので，図の右下の方にプロットされるほど良い匿名化の傾向があるといえる．**図 1(a)** より，JO モデルではどの匿名化によっても漏洩個人情報価値が匿名化前と変わっていない．一方，**図 1(b)** より，提案方式ではリスク低減と活用性維持のトレードオフがみてとれる．これは，提案方式が匿名化の効果を適切に反映できているためである．そのため，提案方式を用いることで，トレードオフのバランスが良い匿名化を選びやすくなり，匿名化の方針が立てやすい．

最後に，提案方式の性能を確認した．適用環境は一般的な PC（Intel Core i7-2600 CPU @ 3.4GHz, RAM 16GB, 64bit Windows 7 Professional SP1）で，実装と適用は Java 8 でおこなった．k-匿名性があることはデータを一度走査するだけで検出できるので，提案方式が最も時間がかかるのは k-匿名性まであと少しのデータである．上記 (3) はそのようなデータである．9 属性が対象なので，全組み合わせを調べると $2^9 - 1 = 511$ 回の走査が必要だが，提案方式により (3) のデータに対して 378 回の走査で済み，20 秒で完了した．また，閾値によって走査回数を減らしても高リスクのプライバシー侵害シナリオは抽出されるため，十分に実用的な性能である．

6. まとめ

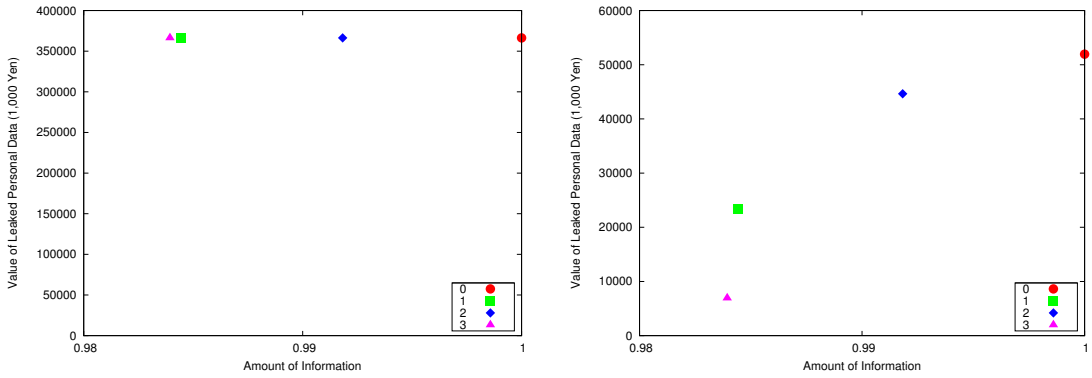
本論文では，JO モデルの本人特定容易度を匿名化の効果が反映されやすいように変更し，それによって算定されるデータプライバシーのリスクが高いレコードを優先的に抽出する新たな方式を提案した．本人特定容易度は，機微情報度の低い属性の少数の組み合わせでレコードを特定できるほど，攻撃者は特定個人の識別が容易であるというモデルに基づいて算定するようにした．この算定は計算量が大きいため，Apriori アルゴリズムと同様に単調性を利用して不必要な計算を省略する新たなアルゴリズムを提案した．k-匿名化のベンチマークであるデータ Adult やそれを匿名化したデータに提案方式を適用し，リスクが高いと感じるプライバシー侵害シナリオが実際に抽出されること，JO モデルより匿名化の効果が反映されやすいこと，性能面で問題がないこと，を確認した．

参考文献

[1] Agrawal, R. and Srikant, R.: Fast Algorithms for Mining Association Rules in Large Databases, *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, San Francisco, CA, USA, Morgan

表 5　Adult で最も高リスクとされたレコード
Table 5　The Highest Risk Records in Adult

レコード番号	ι'	特定に必要な属性集合の要素	age 値	workclass 値	marital-status 値	occupation 値
24028	2.0	age	86	Private	Never-married	Adm-clerical
15534	1.8	age, workclass	21	Without-pay	Never-married	Craft-repair
2697	1.8	age, marital-status	18	Private	Divorced	Other-service
1301	1.8	age, occupation	29	Federal-gov	Never-married	Armed-Forces
44169	1.8	workclass, marital-status	20	Never-worked	Married-spouse-absent	?
20074	1.8	workclass, occupation	65	Without-pay	Married-civ-spouse	Transport-moving
23502	1.8	marital-status, occupation	29	Private	Married-AF-spouse	Farming-fishing



(a) JO モデル　(b) 提案方式

図 1　各匿名化後データの，情報量と，両方式での漏洩個人情報価値との関係
Fig. 1　The Relation between Amount of Information and Value of Leaked Personal Data for Each Anonymized Data and Each Method

Kaufmann Publishers Inc., pp. 487–499 (1994).
[2] Dwork, C.: Differential privacy, *in ICALP*, Springer, pp. 1–12 (2006).
[3] El Emam, K. and Arbuckle, L.: *Anonymizing Health Data: Case Studies and Methods to Get You Started*, O'Reilly Media, Inc., 1st edition (2013).
[4] Fung, B. C. M., Wang, K., Chen, R. and Yu, P. S.: Privacy-preserving data publishing: A survey of recent developments, *ACM Comput. Surv.*, Vol. 42, pp. 14:1–14:53 (2010).
[5] 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).
[6] LeFevre, K., DeWitt, D. J. and Ramakrishnan, R.: Mondrian Multidimensional K-Anonymity, *Proceedings of the 22nd International Conference on Data Engineering*, ICDE '06, Washington, DC, USA, IEEE Computer Society, pp. 25– (2006).
[7] Li, N., Qardaji, W. and Su, D.: On Sampling, Anonymization, and Differential Privacy or, K-anonymization Meets Differential Privacy, *Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security*, ASIACCS '12, New York, NY, USA, ACM, pp. 32–33 (2012).
[8] Machanavajjhala, A., Kifer, D., Gehrke, J. and Venkitasubramaniam, M.: L-diversity: Privacy beyond k-anonymity, *ACM Trans. Knowl. Discov. Data*, Vol. 1 (2007).
[9] Sweeney, L.: Achieving k-anonymity privacy protection using generalization and suppression, *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, Vol. 10, pp. 571–588 (2002).
[10] Sweeney, L.: k-anonymity: a model for protecting privacy, *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, Vol. 10, pp. 557–570 (2002).
[11] UCI: Machine Learning Repository, University of California, Irvine (online), available from ⟨`http://archive.ics.uci.edu/ml/`⟩ (accessed 2016-05-27).
[12] Xu, J., Wang, W., Pei, J., Wang, X., Shi, B. and Fu, A. W.-C.: Utility-based Anonymization Using Local Recoding, *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, New York, NY, USA, ACM, pp. 785–790 (2006).
[13] 特定非営利活動法人日本ネットワークセキュリティ協会セキュリティ被害調査ワーキンググループ：２００３年度情報セキュリティインシデントに関する調査報告書，特定非営利活動法人日本ネットワークセキュリティ協会（オンライン），入手先 ⟨`http://www.jnsa.org/houkoku2003/incident_survey2.pdf`⟩（参照 2016-05-27）.
[14] パーソナルデータの利用・流通に関する研究会：報告書，総務省（オンライン），入手先 ⟨`http://www.soumu.go.jp/main_content/000231357.pdf`⟩（参照 2016-05-27）.
[15] パーソナルデータに関する検討会技術検討ワーキンググループ：報告書，首相官邸高度情報通信ネットワーク社会推進戦略本部（ＩＴ総合戦略本部）（オンライン），入手先 ⟨`http://www.kantei.go.jp/jp/singi/it2/pd/dai5/siryou2-1.pdf`⟩（参照 2016-05-27）.
[16] 独立行政法人情報処理推進機構：パーソナル情報保護と IT 技術の調査，独立行政法人情報処理推進機構（オンライン），入手先 ⟨`https://www.ipa.go.jp/files/000024428.pdf`⟩（参照 2016-05-27）.