Computer Security Symposium 2017
23 - 25 October 2017

集合値データに対する個人適応型匿名化手法

中川 拓麻[1,a)]　荒井 ひろみ[2,†1]　中川 裕志[3,4]

概要：集合値データは，購買履歴や Web 閲覧履歴など，各レコードがアイテムの集合として表される形式のデータである．集合値データにおける属性推定を防ぐことを目指した従来の研究では，守ることのできるアイテムは事前に指定した一部のアイテムに限られるという点が課題となっていた．これに対して本研究では，レコードごとに指定された異なるアイテムを機微属性と見なして守ることで，より個人に合ったプライバシー保護基準を与えることのできる新たなモデルを提案する．また，計算量の課題に対処するための，プライバシー保証の確率的緩和の考え方を導入し，現実的なデータに適用可能なアルゴリズムを提案する．

キーワード：PWS，プライバシー保護，匿名化，属性推定

1. はじめに

1.1 背景

ビッグデータの利活用が重視される中，特に幅広い活用が見込まれるデータ形式として，集合値データ（set-valued data）がある．これはデータに含まれる一つ一つのレコードが集合として表されるものであり，例えば時系列的な順序構造を除いた Web 閲覧履歴や，e-commerce サイトにおける各顧客の購買履歴などを記録する際に用いられるデータ形式の一つである．このようなデータを広くビジネスや研究の現場で活用するため，第三者にデータを渡したり，一般に公開したいという需要がある．しかし，これはデータ主体のプライベートな情報の漏洩に繋がる恐れがあり，安易に実行することはできない．この課題を解決するため，公開したいデータの有用性を維持しつつプライバシーを守るようなデータ加工を行うための手法が，プライバシー保護データ公開（Privacy-Preserving Data Publishing: PPDP）の文脈で広く研究されている [1]．本稿では集合値データのプライバシー保護に関連する既存の匿名化モデルに触れつつその課題と限界について考察し，これに対処するための新たなモデル，個人適応型 ρ-不確実性を提案する．

1.2 既存のモデル

ここでは，表の形式で表すことのできるデータベースを考える．各行はレコードと呼ばれ，1 つのデータ主体（個人）に対応する．また各列は属性と呼ばれ，そこに含まれる情報の種類を表す．一般に，関係データベースにおける属性は，

- **ID**：氏名などのように個人と一対一に対応し，レコードを特定する情報
- **擬似 ID**：性別，年齢，住所など，複数の属性を組み合わせることで個人と一対一に対応しうるため，擬似的に ID と見なせる情報
- **機微情報**：社会的身分，宗教，病歴など，個人と結びついて広く知られることが好ましくない情報

に分類される．これらの属性を含むデータの匿名化を考える際には，ID を削除するだけではなく，擬似 ID から個人のレコードが識別されて機微情報が明らかになってしまうことも防ぐ必要がある．代表的なモデルである k-匿名性 [2] は，まったく同じ擬似 ID の組み合わせを持つレコードが少なくとも k 個以上存在し，ある個人の擬似 ID についての背景知識を持つ第三者（攻撃者）がデータを見ても k 個以上のレコードの中でどれが対応するレコードなのかを識別できないことを保証する．これはつまり，データベースにおける個人識別（identity disclosure）のリスクを防ぐためのモデルである．しかしその場合も，擬似 ID によって区別できないレコード同士で機微情報の値が同じであるようなことがあると，レコードが識別できずとも個人の機微情報が高い確率で推定されてしまうという属性推定（attribute inference）のリスクが存在し得る．これに対応するモデルとしては例えば ℓ-多様性 [3] 等がある．

一方で集合値データにおいては，各レコードは ID に付

[1] 東京大学大学院 情報理工学系研究科
[2] 情報通信研究機構 サイバーセキュリティ研究所
[3] 東京大学 情報基盤センター
[4] 理化学研究所 革新知能統合研究センター
[†1] 現在，理化学研究所 革新知能統合研究センター
[a)] takuma_nakagawa@mist.i.u-tokyo.ac.jp

User	Contents
Alice	milk, bread, *medicine*
Bob	apple
Carol	*milk*, *coffee*, *bread*
Dave	milk, medicine
Ellen	coffee, bread, apple
Frank	orange, *medicine*

(a) 元データ

User	Contents
u_1	~~milk~~, bread, *medicine*
u_2	apple
u_3	*milk*, *coffee*, ~~*bread*~~
u_4	milk, medicine
u_5	coffee, bread, apple
u_6	orange, ~~*medicine*~~

(b) 加工データ

User	Contents
u_1	bread, medicine
u_2	apple
u_3	milk, coffee
u_4	milk, medicine
u_5	coffee, bread, apple
u_6	orange

(c) 公開データ

図 1: 集合値データの匿名化加工の例

随するアイテムの集合（例えば各個人が購入した商品集合など）からなり，一般に擬似 ID と機微情報との区別がないため，単純に k-匿名性を適用することはできない．このようなデータにおいて個人識別を防ぐためのモデルとして Terrovisit ら [4] は，k^m-匿名性を提案した．これは，攻撃者がある個人の持つアイテムのうちの高々 m 個までを背景知識として知っていたとしても，それらを含むレコードが k 個以上存在するようにデータを加工することで，レコードの特定ができないようにし，m 個以外のアイテムについて新たな情報が得られないようにすることを目指すものである．ここではすべてのアイテムを擬似 ID と見なして個人識別を防いでいると言える．

しかし，従来，このような機微情報が明確に区別されていない集合値データにおいて属性推定をも防ぐことは困難であると考えられてきた [5]．そのためこれまでには，どのアイテムが全員にとって守りたい機微情報であるかがあらかじめ特定されているという前提の上で，属性推定を防ぐためのモデルが数多く提案されている [6], [7], [8], [9]．例えば代表的なものとして，Cao ら [8] による ρ-不確実性のモデルでは，攻撃者がある人物の持つアイテムの一部を背景知識として知っていても，別の機微情報と見なされるアイテムを持っているということを ρ (≤ 1) を超える確信度で推定できないという保証を与える．

しかし，文献 [5] でも指摘されている通り，各アイテムが機微であるかどうかという情報は本来各個人の考え方に強く依存し得るものであり，これが個人によらず定まっていることを前提としている点は，現実的であるとは言い難い．プライバシーに対する意識に関しても，自らの持つアイテムの大部分をオープンにしてよいと考える個人もいれば，そのすべてを機微情報として守りたいと感じる個人もいるかもしれない．この問題に対して Cao ら [8] は，少なくとも 1 人以上が機微と考えるアイテムはすべて機微アイテムと見なせば対応できると述べているが，これは必要以上に大きい情報損失に繋がり得る．これまでの先行研究においてこの課題については意識されつつも具体的な対応策は見出されず，一部のアイテムを形式的に機微アイテムと見なし，それ以外と区別して扱えることを前提とするに留まっていた．

1.3 貢献

この課題に対し本研究では，ρ-不確実性をベースとしながら，個人ごとにまったく異なるアイテムを機微情報と見なすことを許し，これを「機微制約」として明示的に与えることで，現実的な状況において適用可能な匿名化モデルを提案する．また，データの加工方法としてアイテムの局所的抑圧を採用し，Jia ら [10] による手法を拡張して提案モデルを達成するためのアルゴリズムを導入する．

図 1 に，個人の購買履歴を表す集合値データとその匿名化の例を示す．**図 1a** は加工前のオリジナルデータ，**図 1b** は提案手法において $\rho = 0.5$ として匿名化加工を施したデータ，**図 1c** は実際に公開されるデータである．イタリック体で表示された部分は，各レコードにおいて機微であると見なされるアイテムであることを示す．いま，ある攻撃者が Alice が milk を買ったということを知っていてかつ元データを閲覧できるという状況を考える．このような場合，仮に ID がデータから削除されていたとしても，milk を買っている 3 レコードのうちの 2 レコードで medicine が買われていることから，この攻撃者は「Alice は 2/3 ($> \rho$) の確率で medicine も買っている」という推定ができてしまう．Alice は medicine を機微なアイテムと見なしているので，このような事態は避けなければならない．一方で，攻撃者が公開データしか閲覧できなければ，相関ルール {milk $\rightarrow$ medicine} の確信度は 1/2 ($\leq \rho$) にまで小さくなっているため，ρ を超える確信度でこのような推定をすることはできないと言える．同様にして，いくつかのアイテムを抑圧することで，公開データにおいては所望のプライバシー保護要件が満たされていることがわかる（詳細については 2.2 節で述べる）．

従来の研究におけるもう一つの課題として，集合値データの匿名化においては必要な計算量がデータのサイズに対して組み合わせ的に増大してしまうという問題があった．対策として，k^m-匿名化については，条件を確率的に緩和してサンプリング法によってこれを実現するという手法が提案されている [11]．本研究では同様のアイデアを用いて，

提案モデルに対する確率的緩和モデルを考案し，サンプリングを用いたアルゴリズムによってこれを保証することができることを示す．

これらの提案手法について，実データを用いた数値実験によってその性能を確かめる．実験の結果，個人が異なるプライバシー要件を持つ状況において，既存モデルによって対応する場合よりも匿名化による有用性損失を小さくできることを確認した．

2. 提案モデル

2.1 ρ-不確実性 [8]

まず，提案モデルのベースとなる ρ-不確実性 [8] について形式的に定義を行う．考える集合値データセットを D，データセットに現れる全アイテムの集合（ドメイン）を $\mathcal{I}$ とする．ここでは，全員にとって共通に，全アイテム中の一部が機微情報と見なされるアイテム（**機微アイテム**）であると仮定する．すなわち，機微アイテムのドメインを $\mathcal{I}_S$，機微でないアイテムのドメインを $\mathcal{I}_N$ としたとき，$\mathcal{I} = \mathcal{I}_S \cup \mathcal{I}_N$，$\mathcal{I}_S \cap \mathcal{I}_N = \emptyset$ であるとする．ユーザー集合を $\mathcal{U}$ とし，ユーザー $u \in \mathcal{U}$ のレコードを $D_u \in D$ と書く（$D_u \subseteq \mathcal{I}$）．$D$ においてアイテム集合 $Q \subseteq \mathcal{I}$ を含むレコード数を Q の **support** と呼び，$supp_D(Q)$ と書く．攻撃者があるユーザーの持つアイテムの部分集合 Q を知っているときに，ここに含まれない機微アイテムを含むアイテム集合 R について，高い**確信度**（confidence）を持つような相関ルール $Q \to R$ が存在することを防ぎたい．ただし，確信度は

$$conf(Q \to R) = \frac{supp_D(Q \cup R)}{supp_D(Q)} \qquad (1)$$

と定める．ρ-不確実性は次のように定義される．

定義 1. データセット D は，任意のアイテム集合 $Q \subseteq \mathcal{I}$ に対して次のいずれかが成り立つとき，ρ-不確実性を満たすという:

(1) $\forall e \in \mathcal{I}_S \backslash Q$，$conf(Q \to \{e\}) \leq \rho$.

(2) $supp_D(Q) = 0$.

この定義の 2 つ目の条件では，攻撃者の背景知識 Q に含まれるアイテムが匿名化加工によって抑圧され，Q を含むレコードが D 中に存在しなくなるケースを想定している．

2.2 個人適応型 ρ-不確実性

前節での定義では，守りたい機微アイテムが全員にとって等しいと仮定していた．ここでは，個人ごとに異なるアイテムを機微アイテムとして指定することで，より個人のニーズに合った匿名化処理を無駄なく行うことのできるモデルを定義する．

定義 1 においてと同様，データベースを D，アイテムのドメインを $\mathcal{I}$，データベースに含まれるユーザー集合を $\mathcal{U}$，ユーザー u のレコードを $D_u \subseteq \mathcal{I}$ とする．各ユーザー u に対し，ユーザー u が機微情報と見なすアイテムの集合を $E_u \subseteq \mathcal{I}$ とし，それらをまとめた $E = \{(u, E_u) : u \in \mathcal{U}\}$ を**機微制約**と呼ぶ．攻撃者 $adv = (u, Q)$ はユーザー u（ターゲット）がアイテム集合 $Q \subseteq D_u$ を持っているということを知っており，この情報を用いて，$e \in E_u \backslash Q$ を u が持つ確率を $conf(Q \to \{e\})$ によって推定する．事前に定めた，プライバシー強度の基準を表すパラメータ ρ について，満たしたい条件を次のように定める．

定義 2. データセット D は，攻撃者 $adv = (u, Q)$ について $\forall e \in E_u \backslash Q$，$conf(Q \to \{e\}) \leq \rho$ または $\forall u' \in \mathcal{U}$，$Q \nsubseteq D_{u'}$ が満たされているとき，adv に対して**安全**であるという．

この条件は，ターゲットの持つアイテムの一部を知る攻撃者に対し，そのターゲットにとっての機微情報が ρ を超える高い確信度をもって推定されない，または攻撃者はターゲットに対応し得るレコードをデータベース内に発見できないことを保証する．この条件を用いて，データベース全体のプライバシーを次のように定義する．

定義 3. データセット D，機微制約 E は，任意の攻撃者 $adv = (u, Q)$ に対して安全であるとき，個人適応型 ρ-不確実性を満たすという．

このモデルは，仮に従来のモデルのように全ユーザーについて機微アイテムが等しくなっている場合，定義 1 の ρ-不確実性と等価となるため，より汎用性の高いモデルとなっていると言える．ここで，各ユーザーは自分が持っていないアイテムも機微アイテムとして指定することができ，これによって実際は持っていないアイテムを持っていると誤って推定されてしまう濡れ衣への対策も可能となる．

2.3 アルゴリズム

個人適応型 ρ-不確実性を満たすよう，データセットを加工するためのアルゴリズムについて述べる．

集合値データの加工のための手法は，一般化と抑圧に大別される．しかし，機微アイテムとそうでないアイテムが混在する場面で一般化を用いると，一般化を行うこと自体が攻撃者に手がかりを与えることになるという危険性が生じる [8]．そこで今回は手法として抑圧を採用し，特にデータの情報損失に与える影響を鑑みて，局所的な抑圧を用いることとする．局所的な抑圧を用いたアルゴリズムとしては Jia ら [10] による手法が存在し，本稿ではこれを提案モデルに対して応用できるよう拡張したアルゴリズムを紹介する．

プライバシー保護のための加工をデータに施す際には，プライバシー保護の強度と，加工によるデータの有用性の損失との間にトレードオフの関係が生じる．今回扱うような，プライバシー保護のために満たすべき条件を与える場合には，この条件を満たしつつ，できる限りデータの有用性損失を最小化できるようなアルゴリズムを探ることとなる．データの有用性をどのように測るべきかは，加工後の

データをどのようなタスクを通して利用するかに依存するため，どんな場合にも適用できる汎用性のある基準は存在しない．そのため実際には，大抵のタスクに対して維持されるべきであると考えられるいくつかのある程度汎用的な基準を用意してアルゴリズムの性能を測ることとなる．そこで今回は，そのような基準として，抑圧されるアイテムの割合，元データと加工後データにおける各アイテムの出現頻度分布の差，の 2 つをなるべく小さくするようなアルゴリズムを考える．元のデータセットを D_0，加工後のデータセットを D とすると，抑圧されるアイテムの割合は

$$util_1(D_0, D) = \frac{\sum_{i \in \mathcal{I}} (supp_{D_0}(i) - supp_D(i))}{\sum_{i \in \mathcal{I}} supp_{D_0}(i)} \qquad (2)$$

となる．アイテムの出現頻度分布の差は KL-ダイバージェンスによって測ることとし，D におけるアイテム i の（正規化された）出現頻度を $D(i)$ と書いて，

$$util_2(D_0, D) = \sum_{i \in \mathcal{I}} D(i) \log \frac{D(i)}{D_0(i)} \qquad (3)$$

とする．

提案モデルも含め，多くの匿名化モデルにおいて，最適な加工結果を得る問題は NP 困難である．また特に今回扱う類のモデルは単調性（加工を加えるごとにプライバシー強度は単調に増加するという性質）を満たさないという点も問題を難しくしている．これについては，例えばアイテムの一般化によって個人識別を防ごうとする場合においては，加工を進めるほど識別リスクが減少していくという性質が成り立つ（Property 1 [4]）．しかし今回のように属性推定のリスクを考慮すると，加工によって新たなリスクが生まれるということが起こりうるため，このような性質は満たされない．こういったことをすべて防ぎながら加工を行うためには極めて大きな計算コストが必要となってしまい，効率的なアルゴリズムの構築が難しい．そこでここでは，安全でないようなある攻撃者に注目してそのリスクを除くということを，すべてのリスクが除かれるまで単純に繰り返すというシンプルな方針を採用することにする．

準備として，確信度が ρ 以下となっていない（安全でない）相関ルール $Q \to \{e\}$ に対して，アイテム $i \in Q \cup \{e\}$ の**抑圧数** N_s を

$$N_s(i,\ Q \to \{e\}) = \begin{cases} \lceil supp_D(Q \cup \{e\}) - supp_D(Q) \cdot \rho \rceil & \text{if } i = e \\ \left\lceil \frac{supp_D(Q \cup \{e\}) - supp_D(Q) \cdot \rho}{1 - \rho} \right\rceil & \text{if } i \in Q \end{cases} \qquad (4)$$

と定める．これは，$conf(Q \to \{e\}) \le \rho$ を満たすために，$i \in Q \cup \{e\}$ を 1 つ選んで $Q \cup \{e\}$ を含むレコードの中からこれを抑圧する際，抑圧する必要のある個数を表している．**図 2** に，この操作の例を示す．いま，ユーザー u_1 は $x \notin E_{u_1}$，$y \in E_{u_1}$ の 2 つのアイテムを持っており，$supp(\{x\}) = 4$，$supp(\{y\}) = 3$ であるとする（**図 2a**）．また，$\rho = 0.5$ とする．y を抑圧するアイテムとして選んだ場合は，$\{x, y\}$ を含むレコードの中から $3 - 4 \times 0.5 = 1$ 個を選んでそこから y を削除することで，$conf(\{x\} \to \{y\}) = 2/4 \le \rho$ とできる（**図 2b**）. また，x を選んだ場合は，$\frac{3-4\times0.5}{1-0.5} = 2$ 個のレコードからこれを削除すると，$conf(\{x\} \to \{y\}) = 1/2 \le \rho$ となる（**図 2c**）．これらの操作の結果，いずれの場合も，$conf(\{x\} \to \{y\}) \le \rho$ の条件を満たすことができる．

User	Contents
u_1	$x, \boldsymbol{y}$
u_2	x, y
u_3	x, y
u_4	x

(a) $conf = 0.75$

User	Contents
u_1	$x, \boldsymbol{y}$
u_2	$x, \not{y}$
u_3	x, y
u_4	x

(b) $conf = 0.5$

User	Contents
u_1	$\not{x}, \boldsymbol{y}$
u_2	$\not{x}, y$
u_3	x, y
u_4	x

(c) $conf = 0.5$

図 2: 抑圧の操作の例

Algorithm 1 Suppressor

Input: データセット D，機微制約 E，パラメータ ρ

Output: 個人適応型 ρ-不確実性を満たす加工済データセット D

```
D_0 ← D
while D, E が条件を満たさない do
  for ℓ = 1 to max_u(|D_u|) do
    for u in 𝒰 do
      while 安全でない adv = (u, Q), |Q| = ℓ が存在 do
        C ← {e | e ∈ E_u\Q, conf(Q → {e}) > ρ}
        (d, e) ← argmax_{d∈Q∪{e}, e∈C} F(d, e, Q, D_0, D)
        N ← N_s(d, Q → {e})
        Q ∪ {e} を含むレコードから N 本をランダムに選び，それらにおいて d を抑圧
return D
```

具体的な手順をアルゴリズム 1 に示す．ここで，$F(d, e, Q, D_0, D)$ は次のように定義される：

$$F(d, e, Q, D_0, D) = \frac{D(d) \log \frac{D(d)}{D_0(d)}}{N_s(d,\ Q \to \{e\})}. \qquad (5)$$

これは，前項で述べた有用性基準に則り，抑圧するアイテム数を少なくする，D_0，D 間のアイテムの出現頻度分布の差を修正する，という 2 つの目的を組み合わせて設計された関数である．アルゴリズム内では，発見された安全でない攻撃者に対し，この関数が最も大きくなるようなアイテム d を抑圧するアイテムとして決め，抑圧数分だけレコードをランダムに選んで抑圧するという操作を繰り返すことで，すべての攻撃者に対して安全となる加工済データセットを出力する．

3. モデルの緩和

本節では，計算量の課題について考える．

3.1 個人適応型 ρ^m-不確実性

個人適応型 ρ-不確実性を保証するためには，各ユーザーのレコードに対して，考えられる任意サイズの部分集合を背景知識として持った攻撃者を想定し，その安全性を調べなければならない．従って，攻撃者として考えられるパターンはデータセットの最大レコード長に対して組み合わせ的に増大するため，大規模データに対してこれを直接適用することは困難である．これは既存の ρ-不確実性モデルにおいても同様の問題であり，先行研究においてはその適用対象を最大レコード長がごく小さいデータに限定せざるを得なかった．

本研究では，これを現実的な大規模データへも適用できるよう条件を緩和したモデルも新たに提案する．まずは単純なアプローチとして，攻撃者の背景知識となる部分集合のサイズを一定以下に制限した次のモデルを定義する．

定義 4. データセット D，機微制約 E は，$|Q| \leq m$ を満たす任意の攻撃者 $adv = (u, Q)$ に対して安全であるとき，個人適応型 ρ^m-不確実性を満たすという．

このモデルは，想定される攻撃者の強さを制限することで条件を緩めている．ただしこれについても，チェックしなければならないパターンは最大レコード長と m に対して組み合わせ的に増大するため，場合によっては適用は困難となる．次節ではさらなる工夫として，モデルを確率的に緩和することで計算コストを軽減する手法について説明する．

3.2 個人適応型 (ε, δ)-ρ^m-不確実性

モデルの確率的緩和の基本的な考え方は，想定される攻撃者がある確率分布に従って現れると想定し，このような攻撃者に対して一定以上の確率でデータが安全であることを，サンプリングを用いた手法によって確率的に保証することである．本節で述べるモデルは，k^m-匿名性を確率的に緩和することを提案した研究 [11] のアイデアを応用している．

いま，ターゲットの持つレコードの大きさ ℓ の部分集合を背景知識として持つ攻撃者の集合を $\mathcal{A}^\ell = \{(u, Q) : u \in \mathcal{U}, Q \subseteq D_u, |Q| = \ell\}$ と書き，α_ℓ を $\mathcal{A}^\ell$ 上に値をとる確率変数とする．また，D が α_ℓ に対して安全でない確率を H_ℓ とする．これらの準備の下で，定義 4 を確率的に緩和したモデルの定義を以下に述べる．

定義 5. データセット D 及び機微制約 E を考える．すべての $\ell \leq m$ について，$Pr[H_\ell < \varepsilon] \geq 1 - \delta$ が成り立つとき，D は個人適応型 (ε, δ)-ρ^m-不確実性を満たすという．

十分小さい $\varepsilon, \delta > 0$ についてこの定義が満たされるとき，D は高い確率で大半の攻撃者に対して安全であるということが言える．

3.3 サンプリングによるアルゴリズム

データセットが定義 5 の条件を満たすかどうかをサンプリングを用いて確かめたい．次の定理は，このために必要なサンプルについての条件を与える．

定理 1. データセット D，機微制約 E において，α_ℓ が従う確率分布から独立にサンプリングされたサンプル集合を S とする．$\varepsilon, \delta > 0$ について，$|S| \geq \frac{\log(1/\delta)}{2\varepsilon^2}$ が満たされているとする．このとき，すべての $adv = (u, Q) \in S$ について D が安全ならば，$Pr[H_\ell < \varepsilon] \geq 1 - \delta$ が成り立つ．

証明. $n = |S|$ とする．X_i $(i = 1, \ldots, n)$ を，D がサンプル $adv_i = (u_i, Q_i) \in S$ に対して安全でないならば 1，安全ならば 0 をとる確率変数とする．S に含まれるうちで D が安全でないような攻撃者の割合を $\hat{H}_\ell$ とする（$\hat{H}_\ell = \sum_i X_i/n$）と，$S$ は $\mathcal{A}^\ell$ の従う確率分布における独立なサンプル集合であることから，$E[\hat{H}_\ell] = H_\ell$ が成り立ち，Hoeffding の不等式より

$$\forall \varepsilon > 0,\ Pr[H_\ell - \hat{H}_\ell \geq \varepsilon] \leq e^{-2|S|\varepsilon^2} \tag{6}$$

が従う．条件より D はサンプル集合 S に含まれるすべての攻撃者について安全であるから $\hat{H}_\ell = 0$ であり，これらから

$$Pr[H_\ell < \varepsilon] \geq 1 - e^{-2|S|\varepsilon^2} \tag{7}$$

を得る．これより，$\varepsilon, \delta > 0$ に対して

$$|S| \geq \frac{\log(1/\delta)}{2\varepsilon^2} \tag{8}$$

となる S が条件を満たせば，$Pr[H_\ell < \varepsilon] \geq 1 - \delta$ が言えることがわかり，定理が示される． □

この結果からおおよそ，例えば $\varepsilon = \delta = 0.1$ とするときは $|S| \geq 116$，$\varepsilon = \delta = 0.05$ とするときは $|S| \geq 600$，$\varepsilon = \delta = 0.01$ とするときは $|S| \geq 23{,}026$ とすればよいことがわかる．これは m に依存しない値であり，従ってこの条件を用いることによって，計算量がパラメータに対して爆発することを防ぐことができ，大規模データに対しても現実的なコストで処理を行うことが可能となる．

この定理を利用して，データセットが条件を満たすかどうかを確認しながら抑圧を進めていく手順をアルゴリズム 2 に示す．ここでは，各 $\ell = 1, \ldots, m$ についてサンプル内に安全でない攻撃者が存在するかどうかを確認し，存在するならばアルゴリズム 1 と同様の手法でこれを解決し，存在しないならば次の ℓ へと進む．最終的に，すべての ℓ において一度も安全でない攻撃者が現れなければ，定義 5 の条件が満たされていることになるため，その時点での D を

Algorithm 2 SampleSuppressor

Input: データセット D, 機微制約 E, パラメータ ρ, m, ε, δ
Output: 個人適応型 (ε, δ)-ρ^m-不確実性を満たす加工済データセット D

$D_0 \leftarrow D$
$n_S \leftarrow \left\lceil \frac{\log(1/\delta)}{2\varepsilon^2} \right\rceil$
$safe \leftarrow$ **false**
while not $safe$ **do**
  $safe \leftarrow$ **true**
  for $\ell = 1$ to m **do**
    $S \leftarrow \mathcal{A}^\ell$ からサンプリングした n_S 個の攻撃者 α_ℓ のサンプル
    while 安全でない $\alpha_\ell = (u, Q) \in S$ が存在 **do**
      $safe \leftarrow$ **false**
      $C \leftarrow \{e \mid e \in E_u \backslash Q,\ conf(Q \to e) > \rho\}$
      $(d, e) \leftarrow \underset{d \in Q \cup \{e\}, e \in C}{\arg\max} F(d, e, Q, D_0, D)$
      $N \leftarrow N_s(d, Q \to \{e\})$
      $Q \cup \{e\}$ を含むレコードから N 本をランダムに選び，それらにおいて d を抑圧
return D

結果として出力する．

サンプリングを行うためには，α_ℓ が $\mathcal{A}^\ell$ 上のどのような確率分布に従っているかを定める必要がある．文献 [11] では，アイテム集合の出現頻度の違いを考慮し，マルコフ連鎖モンテカルロ法を用いて一様分布からサンプリングを行う手法を提案している．しかしここでは，攻撃者のターゲットがどのユーザーとなるかもサンプリングの対象とする必要があるため，より単純に，まず全ユーザーの中から ℓ 個以上のアイテムを持つユーザー 1 人を一様ランダムに選択し，そのユーザーのレコードから ℓ 個のアイテムの部分集合を一様ランダムに選ぶ，という手法をとることにする．

4. 実験

アルゴリズム 1，2 について，実データを用いた実験によってその性能を評価する．

4.1 設定

実験データとして，データマイニングの分野でベンチマークとして広く使われている，BMS-WebView-1，BMS-WebView-2 を用いた．これらは，e-commerce サイトにおいて数か月に渡って蓄積された click-stream から得られたデータである．データの性質を**表 1** に示す．先述した通り，アルゴリズム 1 はデータの最大レコード長が大きいとき計算量が爆発するため，これらのデータに対して直接適用することは難しい．そこで今回は先行研究に倣い，これらのデータセットから大きさが 5 以下であるレコードのみを抜き出したデータを作成して用いる．この操作を行った結果，BMS-WebView-1 のレコード数は 54,737，BMS-WebView-2 のレコード数は 58,044 となった．アルゴリズム 2 に対しては，オリジナルのデータをそのまま用いた．

これらデータセットにおいては，機微アイテムとそうでないアイテムとの区別が存在していないが，先行研究 [6], [8], [10] では全アイテムのうち一定の割合（40%）をランダムに選択して機微アイテムとし，それ以外を機微でないアイテムと定めている．そこで今回は実験 1 として，一定の割合のアイテムをランダムに選択して一律に全員に対する機微アイテムとする場合（fixed），各個人に対して独立に一定の割合のアイテムをランダムに選択してそのユーザーに対する機微アイテムとする場合（personalized）の 2 通りについて実験を行う．これにより，個人がそれぞれでプライバシー基準を選択することの影響を観察することにする．特に断りのない場合，機微アイテムの割合は 40%とする．

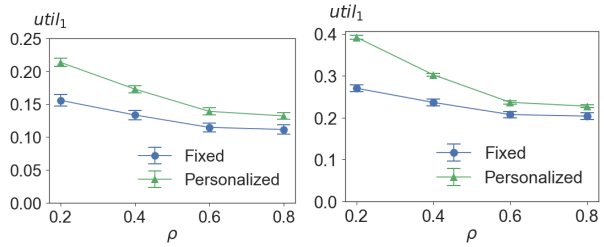
しかし実際には，個人が別々に機微アイテムを選択できるようなモデルでも，一人ひとりがまったく異なるアイテムをランダムに選ぶような状況は現実的とは言い難い．そこで実験 2 として，より現実的な状況を想定し，全アイテムのうちで一定の割合のアイテムを「多くの人にとって機微になりやすい」アイテムとして指定する．その上で，それ以外にも特に一人ひとりが個別に守りたいものを指定しているとして，機微になりやすいアイテムについてもそれぞれ 1%の確率で機微ではなく，また機微になりやすくないアイテムも 1%の確率で機微とする，というようにノイズを加えて現実的な設定を行う（personalized+）．このような状況に対して既存モデルで対処するためには，1 人以上にとって機微であるアイテムは全員にとって機微であると指定するしかない．そのように設定したデータ（fixed+）も用意し比較することで，提案モデルの優位性を確認する．

注意すべき点として，一般に，匿名化の結果はデータセットの性質に大きく依存する．また，今回は機微アイテムをランダムに選んでいるため，結果はこの選び方にも少なからず影響を受けると考えられる．従って，これらの要因によってどのように手法の性能が変動するのかも考慮する必要がある．そのためここでは，まずデータセットの全レコードのうち 10%をランダムに抽出し，さらに機微アイテムをランダムに選択して処理を行う，ということを 10 回繰り返し，結果の平均値と標準偏差を確かめることとする．

有用性の評価は，主に式 (2)，(3) の値によって行う．これに加えて実験 2 では，より現実的な場面でのデータの有用性を調べるため，集合値データの典型的なアプリケーションとして相関ルールマイニングを考え，その結果が匿名化処理によってどれほど変化するかも確かめることにする．

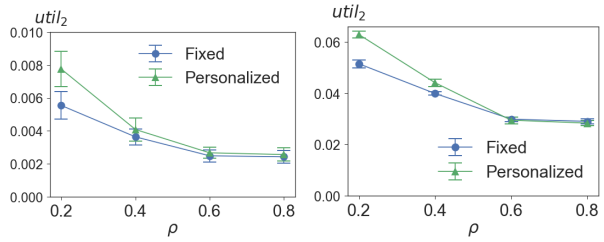
表 1: データセットの性質

データセット	$\|D\|$	$\|\mathcal{I}\|$	最大レコード長	平均レコード長
BMS-WebView-1	59,602	497	267	2.5
BMS-WebView-2	77,512	3,340	161	5.0



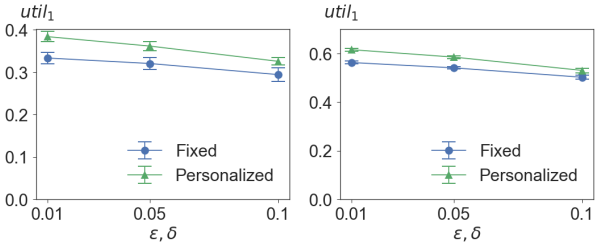
(a) BMS-WebView-1 (b) BMS-WebView-2

図 3: ρ の影響（$util_1$）



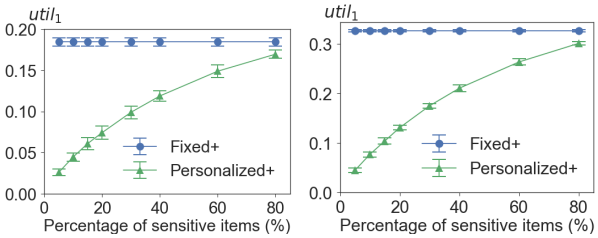
(a) BMS-WebView-1 (b) BMS-WebView-2

図 4: ρ の影響（$util_2$）



(a) BMS-WebView-1 (b) BMS-WebView-2

図 5: 確率的緩和モデルにおける ε, δ の影響



(a) BMS-WebView-1 (b) BMS-WebView-2

図 6: 現実的な設定における比較

4.2 結果

4.2.1 実験 1

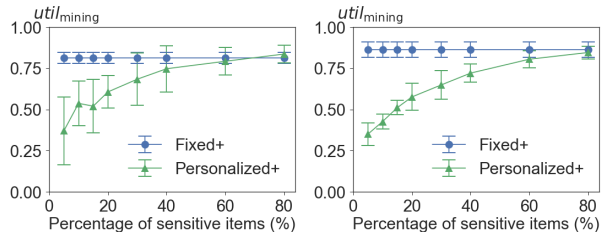
図 3，**図 4** に，ρ の値を変えながらアルゴリズム 1 を最大レコード長 5 のデータに対して適用した結果を示す．機微アイテムが個人によって異なる場合（personalized）の方がやはりプライバシー保証の基準が複雑であるため，抑圧されたアイテムの割合については一貫して有用性の損失が大きくなっているが，多くの場合は 7〜9 割のアイテムが抑圧されずに残されており，この場合も十分実用に足るものと考えられる．KL-ダイバージェンスに関してはその差は僅かであり，この傾向は他の実験結果においても同様であった．BMS-WebView-2 の結果は全体的に BMS-WebView-1 の結果よりも悪くなっているが，これはよりアイテムの種類が増加し，また平均レコード長が長くなることで，発生しうるプライバシー漏洩のリスクも増し，より多くのアイテムを抑圧しなければならなくなっていることによると考えられる．

次に，アルゴリズム 2 をオリジナルデータに対して適用した結果において，抑圧されるアイテムの割合を調べる．**図 5** に，$\rho = 0.5$，$m = 5$ とし，さらに $\varepsilon = \delta$ としてその値を 0.01，0.05，0.1 と変えながら実験を行った結果を示す．全体として，最大レコード長を 5 に制限した場合よりも多くのアイテムが抑圧される結果となっているが，これは長大なレコードによって考えるべきアイテムの組み合わせの数が大幅に増加したことによると考えられる．ε, δ については小さくとるほど，つまり式 (8) に従ってサンプル数を増加させるほど，僅かに抑圧される数が増えているが，これはプライバシー保証の条件が厳しくなることから自然な結果であると言える．KL-ダイバージェンスについての結果もほぼ同様の傾向を持っており，ここでは省略している．

4.2.2 実験 2

より現実的な状況として，personalized+，fixed+の設定において，機微になりやすいアイテムの割合を 5%から 80%まで変えながら抑圧されるアイテムの割合を調べた結果を**図 6** に示す．ここでは再び，最大レコード長を 5 に限定したデータセットを用いる．この状況では，fixed+の場合は実質的にほぼすべてのアイテムを機微と指定して守らなければならないということになっており，必要以上の抑圧によって情報量が大きく落ちてしまう．提案モデルとの差は特に機微になりやすいアイテムの割合が小さいときほど顕著であり，提案モデルの方が有用性損失は非常に少なくなっている．現実には，多くの人が機微であると見なすようなアイテムの割合はそれほど高くないと考えられるため，そのような状況においては特に提案モデルによって匿名化を行うことの優位性が強く表れると言える．

最後に，データを利用するタスクを具体的に想定し，現実的な意味での有用性を調べる．最も典型的なタスクと



(a) BMS-WebView-1　(b) BMS-WebView-2

図 7: 相関ルールマイニング結果の比較

して，ここでは相関ルールマイニングを考える．これは，データセット内で一定の support と確信度の閾値を超えるような相関ルールを見つけるタスクである．そのアルゴリズムとして，Agrawal らによる Apriori アルゴリズム [12] を，最大レコード長 5 のデータに対して適用することを考える．探す相関ルールの support の閾値は全レコード数の 0.05%，確信度の閾値は 30%に設定する．有用性は以下のように測る．まず，元データと加工データのそれぞれに対して Apriori アルゴリズムを適用し，見つかった中から最大で 100 個の相関ルールを，support が大きい順に選ぶ．こうして得られた 2 つの相関ルール集合がどれほど異なるものになっているかを，Jaccard 距離によって測る．すなわち，元データにおいて見つかった相関ルール集合を R_{D_0}，加工データにおいて見つかった相関ルール集合を R_D とするとき，有用性の損失は

$$util_{\text{mining}}(D_0, D) = 1 - \frac{|R_{D_0} \cap R_D|}{|R_{D_0} \cup R_D|} \qquad (9)$$

と定義される．

図 7 に，機微になりやすいアイテムの割合を変えながらこの実験を行った結果を示す．ほぼすべてのアイテムが機微であると見なされている fixed+では，Jaccard 距離はほぼ 0.8 以上と非常に大きい．これは，元データにおいて存在していた相関ルールのうちの多くが失われ，また本来存在しなかったはずの相関ルールが匿名化加工によって多く発生しているということを意味している．一方で personalized+の場合では，特に機微になりやすいアイテムの割合が低いときには多くのルールが加工によって失われることなく残されている．つまり，このような現実的なタスクを考慮した場合でも，提案モデルを用いて匿名化を行うことによって有用性を維持できるということが確認できる．

5. おわりに

本稿ではまず，集合値データのプライバシー保護に関連する既存の匿名化モデルについて解説し，これらにおいては機微アイテムを一律に定めなければならないという課題について論じた．これに対し，各個人が異なるアイテムを機微アイテムとして指定し，それぞれに合ったプライバシー保護基準を与えることのできる新たなモデル，個人適応型 ρ-不確実性及び，さらに計算量的課題に対応するためのモデル，個人適応型 ρ^m-不確実性，個人適応型 (ε, δ)-ρ^m-不確実性を提案した．そしてこれらを実現するためのアルゴリズムを設計し，その性能の実験的評価を行った．

今後の課題として，より性能の良いアルゴリズムの開発及びアルゴリズムの性能についての理論的理解が挙げられる．

謝辞　この研究は科学研究費基盤 (B)「情報検索システムにおけるプライバシー保護に関する研究」（課題番号：15H02700）の助成を受けました．

参考文献

[1] Benjamin C. M. Fung, Ke Wang, Rui Chen, and Philip S. Yu. Privacy-preserving data publishing. *ACM Computing Surveys*, 42(4):1–53, 2010.

[2] Latanya Sweeney. k-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570, 2002.

[3] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. l-diversity: privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, 1(1), 2007.

[4] Manolis Terrovitis, Nikos Mamoulis, and Panos Kalnis. Privacy-preserving anonymization of set-valued data. In *Proceedings of the VLDB Endowment*, volume 1, pages 115–125, 2008.

[5] Manolis Terrovitis, Nikos Mamoulis, and Panos Kalnis. Local and global recoding methods for anonymizing set-valued data. *The VLDB Journal*, 20(1):83–106, 2011.

[6] Yabu Xu, Ke Wang, Ada Wai-Chee Fu, and Philip S. Yu. Anonymizing transaction databases for publication. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 767–775, 2008.

[7] Gabriel Ghinita, Yufei Tao, and Panos Kalnis. On the anonymization of sparse high-dimensional data. In *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*, pages 715–724, 2008.

[8] Jianneng Cao, Panagiotis Karras, Chedy Raïssi, and Kian-Lee Tan. ρ-uncertainty: inference-proof transaction anonymization. In *Proceedings of the VLDB Endowment*, volume 3, pages 1033–1044, 2010.

[9] Grigorios Loukides, Aris Gkoulalas-Divanis, and Jianhua Shao. Efficient and flexible anonymization of transaction data. *Knowledge and Information Systems*, 36(1):153–210, 2013.

[10] Xiao Jia, Chao Pan, Xinhui Xu, Kenny Q. Zhu, and Eric Lo. ρ-uncertainty anonymization by partial suppression. In *Database Systems for Advanced Applications*, pages 188–202, 2014.

[11] Gergely Acs, Jagdish Prasad Achara, and Claude Castelluccia. Probabilistic km-anonymity efficient anonymization of large set-valued datasets. In *Proceedings of the 2015 IEEE International Conference on Big Data*, pages 1164–1173, 2015.

[12] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference on Very Large Data Bases*, pages 487–499, 1994.