Computer Security Symposium 2020
26 - 29 October 2020

匿名化による差分プライベートな購入履歴プロファイリング

菊池浩明[1,a)]

概要：匿名加工技術には，リスクを正確に評価して，セキュリティと有用性のバランスを求めることが肝要である．本稿では，購買履歴データをプロファイリングする新たなモデルを提案する．提案モデルは過去に購買した経験のある商品の集合を表す固定長のベクトルである．全顧客の履歴を統合し，プロファイルとする．このプロファイルを入手した攻撃者がどれくらいのリスクになるかを算出することを試みる．いくつかの仮定の元で，データがあるプロファイルを持つ確率を算出し，差分プライバシーにおけるプライバシー費を評価する．このプロファイル確率は，プロファイルのプライバシーを評価するだけではなく，平均絶対誤差の形でデータの有用性を評価することを可能にしている．
このモデルを用いて，トップコーディング，サンプリング，一般化，および，k-匿名性などの代表的な加工がもたらすプライバシー利得を評価する．加工は単純なパラメータでモデル化されているので，プロファイルの変化を独立同分布の確率変数の和と仮定することで，プライバシー損失と平均絶対誤差の算出を可能にしている．

キーワード：匿名加工，差分プライバシー

Differential-Private Profiling Customer Payment Records with Anonymization

Hiroaki Kikuchi[1,a)]

Abstract: To assess accurately the risk of being compromised, anonymized data requires a balance between utility and security. In this paper, we propose a new model of profiling customer purchase records. Our model is a fixed-size vector that indicates the set of all goods that the customer has purchased before. Aggregated all records for n customers, we define the profile for customers. Our interest is how much risk of re-identification against an adversary who has the profile matrix as an adversarial knowledge. To evaluate a privacy budget of the differential privacy, we show the probability of dataset to have a profile under some reasonable assumptions. The probability of profile allows us to estimate not only the privacy of the profile, but also the utility as the form of mean absolute error of profiles.
We have examined the privacy gain expected by performing representative anonymization including a top/bottom coding, a sampling/suppression, and a generalization (fundamental elements in k-anonymity). These anonymizations are modeled by means of simple factors, which allow us to estimate the privacy loss and the mean absolute error under assumption that the profile's bit errors occur as a sum of independent and identically distributed random variables characterized by the number of records.

Keywords: anonymization, differential privacy

1. Introduction

ビッグデータによる電子化された我々の行動記録は，その規模ではなく，その観測期間の長期化により真の効果をもたらす．購買履歴，決済記録，クレジットカード履歴な

[1] 明治大学総合数理学部
School of Interdisciplinary Mathematical Sciences, Meiji University
[a)] kikn@meiji.ac.jp

© 2020 Information Processing Society of Japan

どのパーソナルデータを集約することで，特定個人の行動パターン，趣味範囲，意思決定などを予測する顧客プロファイリングが実現できるからである．解析されたデータは，マーケット分析を行う企業，政策へ活用する政府機関，自然科学分野の振舞いを分析する科学者らにとても有益である反面，個人の権利を侵害する新たなプライバシーの脅威となる潜在性を持つ．Mavriki と Karyda は，個人やある集団，社会についてのプロファイリングが引き起こすであろう様々な脅威について議論している [1]. Pyrgelis らは，時系列の集約された位置情報から推測されるある集団へのメンバーシップリスクを定量化している [14]. 位置情報だけではなく，視聴した映画履歴 [12], 評価した商品履歴 [6], 治療や投薬などの医療記録 [5]，購買した注文履歴 [4] などによって個人に関する多くのことが推論できることが示されている．

プロファイリングされた個人の行動に関するデータは個人情報とみなされるのだろうか？ GDPR[2] では 4 条 (4) において，プロファイリングを「自然人と関連する一定の個人的側面を評価するために、特に，当該自然人の業務遂行能力，経済状態，健康，個人的嗜好，興味関心，信頼性，行動，位置及び移動に関する側面を分析又は予測するために，個人データの利用によって構成される，あらゆる形式の，個人データの自動的な取扱い[*1]」と定め，事業者側に透明性と説明責任を求め，特定の個人的側面を評価する場合は，個人情報収集時にプロファイリングで利用することを利用目的に明記しておくことを求めている [3]．一方，ビッグデータを特定の個人と紐付けないで集約する技術として，匿名化 (Anonymization, or de-identification) がある．ISO/IEC 20889[13] では，統計処理，暗号学的処理，削除，仮名化，一般化，摂動化などの様々な匿名化のカテゴリーを定め，k-anonymity [7], ℓ-diversity [8], t-closeness [9], β-likeness [10], ϵ-differential privacy [11] などの多くのプライバシーモデルの上で評価することが求められている．

そこで，本研究では，購買履歴データからの顧客プロファイリングが引き起こすプライバシー損失と，各種匿名化技術によってもたらされるプライバシー利得について研究する．プロファイルを特定の顧客が購買した商品の集合を表す固定長ベクトルでモデル化し，全顧客について集約したデータから，差分プライバシーにおけるプライバシー費を定量化する．ここに匿名化を施すによって，プライバシーの度合いは向上するが，その効用は低下する．そこで，このトレードオフを評価するために，代表的な匿名化の技術をシンプルなパラメーターでモデル化し，提案モデルの上で元のデータに対するプロファイルの平均絶対誤差 MAE を評価する．すなわち，MAE により効用損失を，差分プライバシーによってプライバシー利得 (ϵ) をそれぞれ定量化する．

本稿における新規性は次のとおりである．

- 年間で累積された時系列の購買記録を，購買商品と顧客の関係を表す $\ell \times n$ 行列で表すプロファイルのモデルを提案する．
- 購買履歴に関する顧客プロファイルが満足する差分プライバシーにおけるプライバシー費 (budget) を証明し，代表的な匿名化によってもたらされる利得を定量化する．
- 顧客プロファイルと匿名化されたプロファイルとの間の矛盾するビットの確率を評価し，その期待値を用いて MAE，匿名化による効用損失を定量化する．
- 購買履歴のオープンデータとして，UCI Machine Learning Repository の Online Retail data ($m = 45,047$) の匿名化によるプライバシー利得と効用損失の分析結果を報告する．

2. 差分プライバシー

定義 2.1 (ϵ-DP) 任意の $S \in \mathrm{Range}(\mathcal{A})$ と任意の隣接するデータセット D と D' について，ランダムアルゴリズム $\mathcal{A}$ が，

$$e^{-\epsilon} \leq \frac{\Pr[\mathcal{A}(D) = S]}{\Pr\mathcal{A}(D') = S]} \leq e^{\epsilon} \tag{1}$$

である時，ϵ-differential private である．
ここで，隣接する D と D' とは，D に属して，D' には属さない様なあるレコード t が存在するデータセットである．

定義 2.2 ((ϵ, δ)-DP) 任意の $S \in \mathrm{Range}(\mathcal{A})$ と任意の隣接する D と D' について，ランダムアルゴリズム $\mathcal{A}$ が

$$\Pr[\mathcal{A}(D) = S] \leq e^{\epsilon} Pr[\mathcal{A}(D') = S] + \delta$$

である時，(ϵ, δ)-private である．

3. 購買履歴のプライバシー

3.1 購買履歴

表 1 に本研究で対象とする購買履歴 D の例 (the Online Retail dataset in the UCI Machine Learning Repository) を示す．購買のレコード（行）は，顧客 ID, 購買日，購入商品，価格，購入数の $m = 9$ 列から成っている．複数の行に同じ顧客 ID が振られており，例えば，顧客 12362 は 4 つの購買注文を行っていることを示している．

3.2 顧客のプロファイリング

プロファイリングとは，購買履歴，決済履歴，クレジットカード履歴などのパーソナルデータを集約して特定個人の行動や興味，意思決定を予測する特徴を構成する行為である．購買履歴行列 D は，顧客 ID について集約された n 個の互いに素な部分集合 $D = D_1 \cup D_2 \cup \cdots \cup D_n$ に分割される．顧客 i の（期間内の）全レコードを D_i, そのレコー

[*1] 個人情報保護委員会訳

表 1　Example transaction table D

$t_{\cdot,1}$ cust. ID	$t_{\cdot,2}$ inv.	$t_{\cdot,3}$ date	$t_{\cdot,4}$ time	$t_{\cdot,5}$ stock ID	$t_{\cdot,6}$ price	$t_{\cdot,7}$ qty
12361	0	2011/2/17	10:30	21913	3.75	4
12361	0	2011/2/17	10:30	22431	1.95	6
12361	0	2011/2/25	13:51	22630	1.95	12
12361	0	2011/2/25	13:51	22555	1.65	12
12362	0	2011/4/28	9:12	21866	1.25	12
12362	0	2011/4/28	9:12	20750	7.95	2
12362	0	2011/4/28	9:12	22908	0.85	12
12360	0	2011/5/23	9:43	21094	0.85	12
12360	0	2011/5/23	9:43	23007	14.95	6

表 2　Example of profiles for $n = 5, m = 13, \ell = 3$

ID i	rank r_i	profile $\boldsymbol{x}_i$	# records m_i	minority
1	1	(1,1,1)	6	0
2	2	(1,0,1)	3	0
3	3	(1,1,0)	2	0
4	4	(0,1,0)	1	1
5	5	(1,0,0)	1	1

ド数を $m_i = |D_i|$ で表す．

購買歴には，購買日，商品の単価や個数などの多くの情報が含まれているが，本稿では，購買商品の集合に着目する．なぜならば，購買商品集合はその顧客の嗜好の特徴を表しており，多くの場合，この情報だけに基づいて Jacccard 距離などの類似度に基づき，精度の高い識別が可能であるからである [12]．

データ D_i から商品集合への射影を $D_i^{(g)}$ と表す．D における全商品の集合を $G = \{g_1, \ldots, g_\ell\}$ とおく．ここに，ℓ 種類の商品が存在する．例えば，表 1 における顧客 12362 の商品集合は，$D_{12362}^{(g)} = (21866, 20750, 22908)$.

定義 3.1 (profile)　D_i を m_i 個のレコードから成る i 番目のユーザのデータセットとする．i 番目ユーザの**プロファイル** は，$j = 1, \ldots, \ell$ について

$$x_j = \begin{cases} 1 & \text{if } g_j \text{ is in } D_i^{(g)}, \\ 0 & \text{otherwise} \end{cases}$$

となる $\boldsymbol{x} = (x_1, \ldots, x_\ell) \in \{0,1\}^\ell$ で定められる ℓ-次元のベクトルである．

i 番目の顧客のプロファイルを $\mathcal{A}(D_i)$ と書く．n 人分のプロファイルの集約は，$\ell \times n$ 行列となり，$\mathcal{A}(D)$ で表す．例えば，表 2 の，$\ell = 3$ 種の商品について $n = 5$ 名が購入した $m = 13$ レコードの履歴データを考えよう．もしも，2 番めの顧客が $D_2 = (g_1, g_3, g_1)$ の購入を行っていたら，そのプロファイルは，$\mathcal{A}(D_2) = (1, 0, 1)$ である．表は顧客の持つレコード数 m_i でソートされており，簡単化の為に，そのランク r_i を仮 ID i としている．顧客 4, 5 は期間内に一度しか購買をしておらず，後述するボトムコーディングの対象となる．

表 3 にシンボル表を示す．

表 3　List of Symbols

シンボル	記述	値域
U	顧客集合	$\lvert U\rvert = n$
G	商品集合	$\lvert G\rvert = \ell$
D	購買履歴	$\lvert D\rvert = m$
m_i	i 番目顧客のレコード数	$m_{min} \le m_i \le m_{max}$
θ	レコーディングしきい値	$\theta \in [1, m_{max}]$
β	ランダムサンプリング率	$\beta \in [0, 1]$
c	カテゴリに c 値を含む一般化	$1 \le c$
D_i	i 番目の顧客の m_i 個の履歴	$m_i = \lvert D_i\rvert$
$\mathcal{A}(D_i)$	i 番目顧客のプロファイル	$\boldsymbol{x} \in \{0,1\}^\ell$
$\mathcal{A}(D)$	全顧客集約プロファイル	$\ell \times n$ matrix

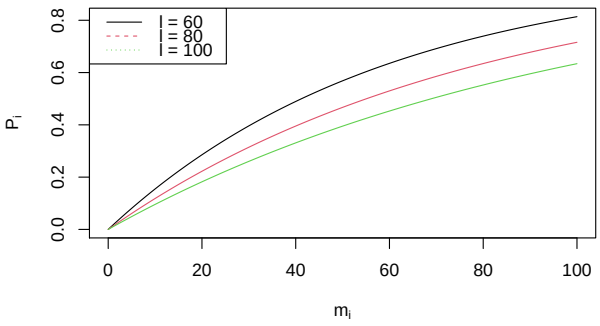


図 1　Probability distribution p_i with parameter m_i (a number of records for i-th customer)

3.3　攻撃者仮定

全顧客の集約プロファイル $\mathcal{A}(D)$ と未知の顧客の履歴 D_i が与えられた時，該当するプロファイル $\mathcal{A}(D_i)$ を識別しようとする攻撃者を考える．本研究では，この攻撃者が正しく識別する確率を評価する．それが正しく推定できれば，そのリスクが許容範囲内に収まるように履歴の累積が長くならないように調整すればいいからである．

4.　評価

4.1　プロファイル確率

定理 4.1 (profile probability)　i 番目の顧客の m_i 個の購買履歴を，$1/\ell$ の確率で G の値を取る独立同分布の確率変数の列 D_i とする．i 番目のプロファイル $\boldsymbol{x}$ は，$\Pr[\mathcal{A}(D_i) = \boldsymbol{x}] = p_i^z (1 - p_i)^{\ell - z}$ の確率で生起する．ただし，ここで，z は $\boldsymbol{x}$ のハミング重みであり，p_i は

$$p_i = 1 - (1 - \frac{1}{\ell})^{m_i}. \qquad (2)$$

で定められる $\boldsymbol{x}$ のあるビットが 1 を取る確率である．

ハミング重み z は二項分布 $B(\ell, p_i)$ に従うので，その期待値は $E(z) = \ell p_i$ で与えられる．

4.2　差分プライベートなプロファイリング

i 番目の顧客が D において m_i 個のレコードを持ち，D'

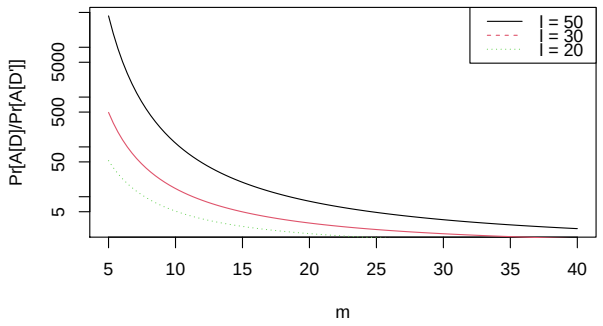


図 2 Distribution of privacy loss with regards to number of records m_i

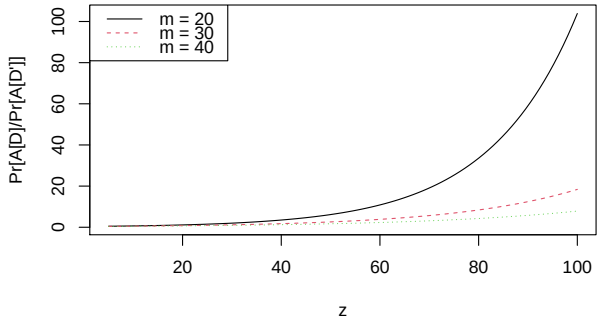


図 3 Distribution of privacy loss with regards to Hamming weight $z = w_H(\boldsymbol{x})$ $(\ell = 100)$

においては $m_i - 1$ 個のレコードを持ち，$t \in D$ で $t \notin D'$ となるようなレコードを含む 2 つの **隣接する** データセット D と D' を考えよう．$j \neq i$ について $D_j = D'_j$ なので，プライバシー損失

$$\frac{Pr[\mathcal{A}(D) = S]}{Pr[\mathcal{A}(D') = S]} = \frac{\prod_i^n Pr[\mathcal{A}(D_i) = S_i]}{\prod_i^n Pr[\mathcal{A}(D'_i) = S_i]} = \frac{Pr[\mathcal{A}(D_i) = \boldsymbol{x}]}{Pr[\mathcal{A}(D'_i) = \boldsymbol{x}]} \tag{3}$$

$$= \left(\frac{p_i}{p'_i}\right)^z \left(\frac{1-p_i}{1-p'_i}\right)^{\ell - z}$$

$$= \left(\frac{1-(1-1/\ell)^{m_i}}{1-(1-1/\ell)^{m_i-1}}\right)^z \left(1-\frac{1}{\ell}\right)^{\ell-z}$$

がある区間に閉じているかどうかが分かれば良い．図 2 と 3 に示される $\ell = 20, 30, 50$ and $m_i = 20, 30, 40$ についての分布から見えるように，$0 \leq z \leq \ell$ について単調増加であり，$1 \leq m_i \leq m$ について単調減少である．しかし，単純な例により，次が言える．

リマーク 4.1 プロファイル $\mathcal{A}(d)$ は（強い）ϵ-差分プライバシーを満たさない．

なぜならば，$m_i = 1$ となる顧客 i がプライバシー損失を最大化するからである．D_i においては，$\boldsymbol{x}_i = (0, \ldots, 1, \ldots, 0)$ となるプロファイルを $1/\ell$ で持つが，D'_i においてはレコードがなく，決して $\boldsymbol{x}_i$ を持つことが出来ず，$Pr[\mathcal{A}(D_i) = \boldsymbol{x}_i] = 0$ により，損失を無限大にするからである．この時，いかなる攻撃者も D と D' を確定的に識別できる．

従って，プロファイリングは強い差分プライバシーを満たさない．そこで，N_1 をその様な一つしか履歴を持たない顧客の集合と置く．幸い，その様な顧客は実際には少数派であることが多く，故障確率 $\delta_1 = N_1/n$ とおくことで，次の弱い保証を得る．

定理 4.2 ((ϵ, δ)-differential private profile) プロファイル $\mathcal{A}(D)$ は $\epsilon_2 = 2\ln(2 - 1/\ell) + (\ell - 2)\ln(1 - 1/\ell)$ and $\delta_1 = N_1/n$. について (ϵ_2, δ_1)-差分プライバシーを満たす．

4.3 匿名化

裸のプロファイルからはある程度のプライバシーが漏れる．そこで，D に匿名加工をした時の有用性と安全性（プライバシーバジェット）を考える．

これまで匿名化には多様なアルゴリズムが提案されている．そこで，ISO/IEC 20889[13] 次のカテゴリーを定めている．統計処理，暗号学的処理，削除，仮名化，一般化，摂動化などのカテゴリーを定めている．本項では，この内，次の代表的な加工について考える．

4.3.1 Sampling/Suppression

定義 4.1 (sampling) データセット D から $[0, 1]$ の β をサンプリング率でレコードを抽出し，$1 - \beta$ で削除をするランダムサンプリングを $f_\beta(D)$ で表す．

ランダムサンプリングにより，i 番目の顧客のレコード数は $m'_i = \beta m_i$ になる．

4.3.2 トップ・ボトムコーディング

定義 4.2 (top coding) θ_{m_+} を顧客一人の持つ最大履歴頻度とする．全ての顧客について，$f_{\theta_{m_+}}(D)$ を超える数のレコードを削除する処理を $f_{\theta_{m_+}}(D)$ と表す．

本稿では，冗長なレコードを選ぶ方法については割愛する．

定義 4.3 (bottom coding) θ_{m_-} を顧客あたりの最小履歴頻度とする．全ての顧客について，θ_{m_-} 未満のレコードを持たないする処理を $f_{\theta_{m_-}}(D)$ で表す．

条件に満たない顧客については，削除するか，ダミーのレコードを追加する方法が考えられる．なお，$\theta_{m_-} > 1$ によるボトムコーディングによって，$m_i =$ となる顧客は存在しなくなるので，強い差分プライバシーにおける損失が有界でない問題を避けることが出来る．

4.3.3 一般化

定義 4.4 (generalization) $1 \leq c \leq \ell$ となる c を一般化係数と呼ぶ．c 個の値を一つのカテゴリーに一般化し，商品集合 G を ℓ/c 個のカテゴリーに分割し，$\ell' = \ell/c$ となるカテゴリー集合を持つ様に処理する一般化を f_c で表す．

c は実数も許されることに注意せよ．

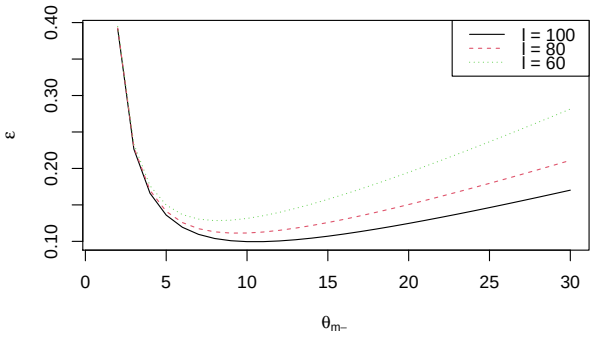


図 4　Differential Privacy for bottom coding θ_{m_-}

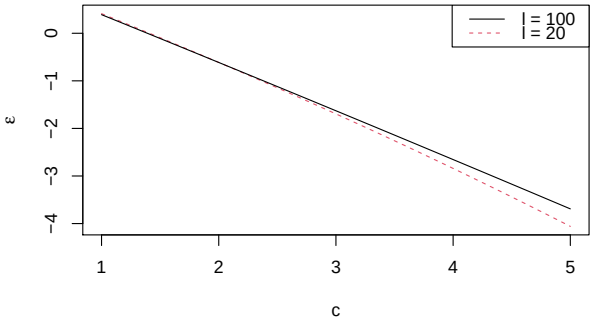


図 5　Differential Privacy for Generalization with c

5. 解析

5.1 匿名化によるプライバシー利得

トップコーディングとサンプリングは，差分プライバシーの保証を変化させない．

リマーク 5.1　(ϵ_2, δ_1)-差分プライバシーであるプロファイル $\mathcal{A}(D)$ がある時，トップコーディング $f_{\theta_{m_+}}$ と サンプリング f_β は (ϵ_2, δ_1)-差分プライバシーを満たす．

定理 5.1　任意の $\theta_{m_-} > 1$ について，ボトムコーディング後のプロファイリング $\mathcal{A}[f_{\theta_{m_-}}(D)]$ は $(\epsilon_{\theta_{m_-}}, 0)$-差分プライバシーを満たす．ここで，$\epsilon_{\theta_{m_-}} = \theta_{m_-} \ln(1-(1-1/\ell)^{\theta_{m-}} - \theta_{m_-} \ln(1-(1-1/\ell)^{\theta_{m-}})$.

しきい値 θ_{m_-} のボトムコーディングによるプライバシー費 ϵ の分布を図 4 に示す．プライバシーの保証は θ_{m_-} に比例して向上（ϵ の減少）しているが，10 以降は緩やかに増加している．これは，プロファイルのハミング重みが最大 θ_{m_-} までしか持ち得ないので，$z = \theta_{m_-}$ で制約されるためである．

定理 5.2　任意の $c > 1$ について，一般化した後のプロファイル $\mathcal{A}[f_c(D)]$ は (ϵ_c, δ_1)-差分プライバシーを満たす．ここで，$\epsilon_c = 2\ln(2-c/\ell) + (\ell-2)\ln(1-c/\ell)$.

一般化係数 $c = 1, \ldots, 5$ についての，プライバシー費 ϵ の節約の分布を図 5 に示す．c が大きくなればプライバシーは向上している．

k-匿名性は通常，一般化，トップ・ボトムコーディングの組合せで実現される．しかし，プライバシー費がサンプリングやトップコーディングで変化しないことが分かったので，本稿では一般化に焦点を絞る．

5.2 匿名化による有用性損失

5.2.1 MAE

匿名化 f によって損失する有用性を，元プロファイル $\mathcal{A}(D)$ と加工後のプロファイル $\mathcal{A}(f(D))$ の平均絶対誤差 MAE で評価する．

定義 5.1 (MAE)　n 人の顧客が関わるデータセット D の匿名化 f による MAE を，

$$w_f(D) = \frac{1}{n}\sum_i^n ||\mathcal{A}(D_i) - \mathcal{A}(f(D_i))|| = \frac{1}{n}\sum_i^n ||\boldsymbol{x}_i - \boldsymbol{x}'_i||,$$

とする．ここで $||\boldsymbol{x} - \boldsymbol{x}'||$ は $\boldsymbol{x}_i \oplus \boldsymbol{x}'_i = (w_1, \ldots, w_\ell)$ のハミング重みである．

プロファイル $\mathcal{A}(D_i)$ と $\mathcal{A}(f(D_i))$ とが異なる時, かつその時に限り，1 となる確率変数 W_j を考えると，ハミング重み $w_1, \ldots, w_\ell$ は独立同分布 W_j の和と考えることが出来る．m_i が与えられた時，$W_j = 1$ の条件付き確率は，$Pr[W_j = 1|m_i]$

$$\begin{aligned} &= Pr[x_j = 1, x'_j = 0 \vee x_j = 0, x'_j = 1|m_i] \\ &= p_i(1-p'_i) + (1-p_i)p'_i - p_i(1-p'_i)p'_i(1-p'_i), \end{aligned}$$

ここで，p_i と p'_i は，それぞれ D_i と $f(D_i)$ において，式 (2) で与えられる，i 番目の商品が購入される確率である．これを，p_i と p'_i を引数として取る関数 $P_F(p_i, p'_i)$ で表す．f による P_F が与えられた時の MAE は，確率 P_F，サイズ ℓ の二項分布に従うので，その期待値は，

$$\begin{aligned} \hat{w}_f(D) &= E[w_f(D)] = E[\sum W_j] \\ &= E[B(P_F(p_i, p'_i), \ell)] = \ell \sum_i^n P_F(p_i, p'_i). \end{aligned}$$

で推定できる．

$P_F(p_i, p_j)$ の確率分布を図 6 に示す．$p_i = p_j$ となる対角線の P_F が必ずしも 0 でなく，$p_i = 0.5$ を最大とする凸になっていることに注意せよ．

5.2.2 データ

解析には， Online Retail dataset ($m = 45,047$, $\ell = 3,090$) を用いる．図 7 は顧客あたりの履歴数 m_i の分布である．履歴頻度で上位 100 のランク r_i に並べている．ここで実線は，非線形最小二乗による近似曲線 $\hat{m}_i = 4250/r_i^{0.908} + 14.06$ である．

5.2.3 Top/bottom coding

しきい値 θ_{m_+} によるトップコーディングは，$i = 1, \ldots, n$

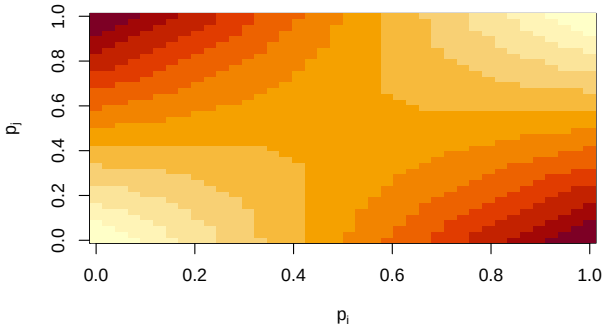


図 6 Probability of an inconsistent bit of w_j to be 1, $P_F(p_i, p_i')$

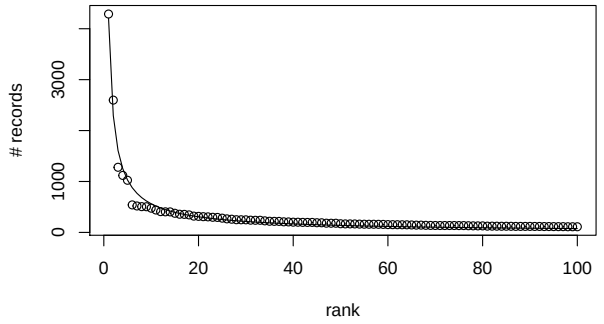


図 7 Number of records in the Online Retail dataset[4] with respect to the rank ranging 1 to 100

について，レコード数を

$$m_i' = \begin{cases} \theta_{m_+} & \text{if } m_i > \theta_{m_+} \\ m_i & \text{otherwise.} \end{cases}$$

にレコードを削除する．そこで，$m_{r_i} > \theta_{m_+}$ となる最小のランクを $r_+ = m_i^{-1}(\theta_{m_+})$ と置くと，MAE は，

$$\begin{aligned} \hat{w}_{\theta_{m_+}} &= \frac{1}{n}\left(\sum_{r=1}^{r_+} \ell P_F(p_{i_r}, p_{i_{r_+}}) + \sum_{r=r_++1}^{n} ||\boldsymbol{x}_{i_r} - \boldsymbol{x}_{i_r}||\right) \\ &\leq \frac{r_+}{n}\ell P_F(p_{max}, p_{i_{r_+}}), \end{aligned}$$

で推定できる．ここで，$i_1, \ldots, i_n$ は，レコード数でソートされた顧客識別子，すなわち，$m_{i_1} \leq \cdots \leq m_{i_n}$，$p_{max} = p_{i_{r_1}}$ である．トップコーディングの閾値 θ_{m_+} が減少するのに伴って，MAE は増加し，ランクの下界 r_+ も上がる．

しきい値 θ_{m_-} によるボトムコーディングは，しきい値 θ_{m_-} 未満の頻度のレコードを削除するか，ダミーレコードを追加することで実施できる．（ここではその方法は問わない）トップコーディングと同様に，$i = 1, \ldots, n$ について，i 番目の顧客のレコード数を

$$m_i' = \begin{cases} \theta_{m_-} & \text{if } m_i < \theta_{m_-} \\ m_i & \text{otherwise} \end{cases}$$

となるように加工を行う．$m_{i_{r_-}} \leq \theta_{m_-}$ を満たす最大のランクを r_- と置くと，それによる MAE は，

$$\begin{aligned} \hat{w}_{\theta_{m_-}}(D) &= \frac{1}{n}\left(\sum_{r=r_-}^{n} \ell Pr(p_{i_r}, p_{i_{r_-}}) + 0\right) \\ &\leq \frac{n - r_-}{n}\ell P_F(p_{min}, p_{i_{r_-}}), \end{aligned}$$

で推定できる．ここで，$p_{min} = p_{i_{r_n}}$ である．ほとんどの n について p_{min} は極小さく，$P_F(p_i, 0) = p_i$ であることとを用いると，よりシンプルな下限を $\hat{w}_{\theta_{m_-}} \leq ((n - r_-)/n)\ell p_{i_{r_-}}$ と得る．

ボトムコーディングのしきい値 θ_{m_-} はプライバシーと有用性の両面に効果を与える．もし大きければ，定理 5.1 で示したように，ϵ を $\epsilon_{\theta_{m_-}}$ まで減少させて，より高いプライバシーの度合いを保証する．同時に，ランクの下限 r_- は上昇し，$p_{i_{r_-}}$ を増加させるので，その MAE は増加する．結局，トレードオフを生じさせる支配的な加工は，トップコーディングではなくてボトムコーディングである．

図 8 と図 9 に，Online Retail データセットに対して評価した，トップコーディングとボトムコーディングのしきい値についての MAE の変化を示す．トップコーディングにおいては，$\theta_{m_+} < 200$ までに大きな MAE の減少が見られ，それ以降は収束してる．ロングテイルの分布をするデータの性質により，数少ないトップのレコードが全体のデータの特性を支配しているためである．

5.2.4 サンプリング

D のレコードをサンプリングレート β で一様にサンプリングするので，i 番目の顧客のレコード数は，$m_i' = \beta m_i$，となり，対応する MAE は，

$$\begin{aligned} \hat{w}_\beta(D) &= \frac{1}{n}\sum_{i=1}^{n} \ell P_F(p_i, \beta p_i) \\ &\geq w(D) + \frac{1}{n}\sum_{i=1}^{n} \ell(1 - (1 - \frac{1}{\ell})^{\beta m_i} - (1 - \frac{1}{\ell})^{(1+\beta)m_i}) \end{aligned}$$

で推定される．ここで，$w(D)$ は加工前の MAE である．すなわち，残りの項がサンプリングによって生じた誤差を表している．

図 10 に β についての MAE の変化を示す．β が小さいほど誤差が大きく，有用性が損なわれている．逆に，$\beta = 1.0$（加工なし）の場合でも，MAE は 0 にならない．これは，節 5.2.1 で述べたように，$P_F(p_i, 1p_i)$ で生じる偽陽性の為である．

5.2.5 一般化

一般化係数 $c \in [1, \ell]$ による一般化は，値域を ℓ/c 個のカテゴリーに変換する．従って，商品数は $\ell' = \ell/c$, と変化するが，顧客の持つレコード数 m_i は変わらない．この加工による MAE は，

$$\hat{w}_c(D) = \frac{1}{n}\sum_{i=1}^{n} \frac{\ell}{c} P_F(p_i, p_i')$$

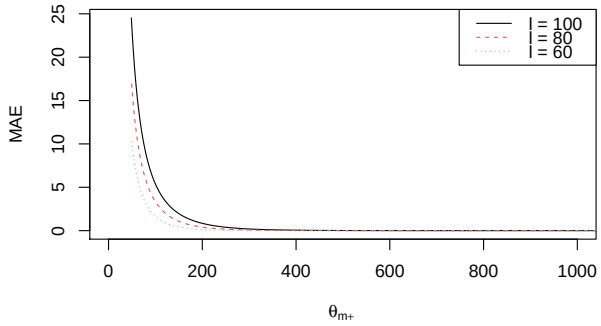


図 8　Mean Absolute Error of top coding

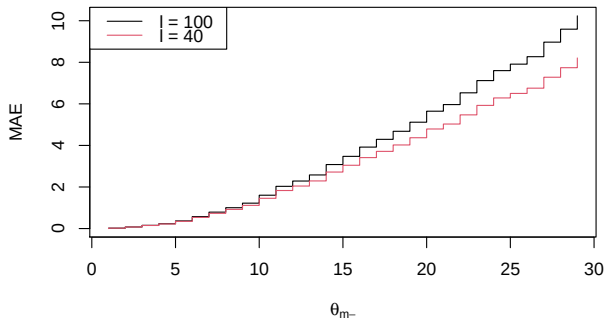


図 9　Mean Absolute Error of bottom coding

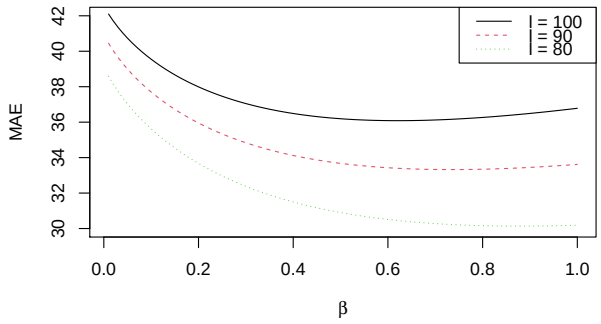


図 10　Mean Absolute Error of sampling

$$= \frac{\ell}{cn}\sum_{i=1}^{n} P_F(1-(1-\frac{1}{\ell})^{m_i}, (1-(1-\frac{c}{\ell})^{m_i})$$

で推定される．ただし，この一般化によって，プロファイルのビット長が ℓ から ℓ/c ビットに変化してしまうので，その誤差評価は注意しなくてはならない．例えば，$\ell = 4$ のあるプロファイル $\mathcal{A}(D_i) = (\underline{0,0},0,1)$ を，$c = 2$ で一般化して，$\mathcal{A}(f_c(D_i)) = (\underline{0},1)$ となったとしよう．この時，下線の 2 ビットには変化がないので，この 2 つのプロファイルは無矛盾であると呼ぶこととする．この様な自然なビットの対応付をして，矛盾の生じるビットの数で MAE を考える．

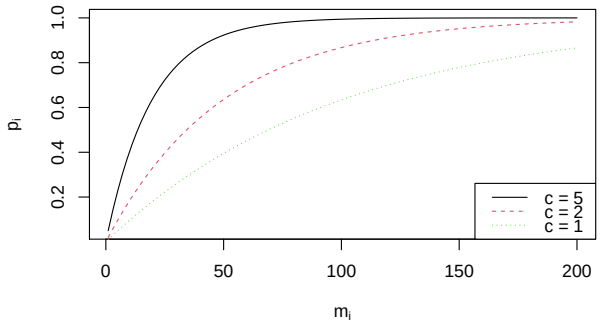


図 11　Probability Distributions p_i after generalization with $c = 1, 2, 5$

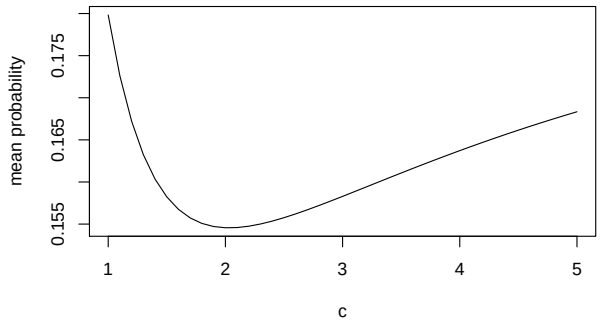


図 12　Mean Probability for inconsistent bit $P_F(p_i, p_i')$ w.r.t. generalization factor c ($\ell = 100$)

一般化係数 c によって，プロファイルのあるビットの確率 p_i は増加し，それゆえ，矛盾ビットの平均確率 P_F も同様に増加する．結局, MAE を上げる．図 11 に，$m_i = 1, \ldots, 200$ についてのこのビット確率 p_i の増加を，図 12 に，c についての MAE の増加をそれぞれ示す．

5.2.6　k-匿名化

k-匿名性は，一般化や削除などの適切な組み合わせによって満たされる．しかし，レコード削除は ϵ には効果がないので，一般化加工に焦点を置く．k は一般化係数 c に対して単調増加し，$c = \ell$ の時 $k = n$ となる．従って，$1 < c_k \leq \ell$ の間にすべての同値類が k 個以上のレコードを持つような c_k が必ず存在し，それについて，次が成立する．

リマーク 5.2　k-匿名性を満たす一般化係数 c_k は，(ϵ_{c_k}, δ)-差分プライバシーを満たし，$\hat{w}_{c_k}$ の MAE を引き起こす．

5.3　まとめ

表 4 に，本研究の評価結果を整理する．購買歴の様な

表 4 Summary of privacy and utility for representative anonymizations techniques

anonymization	parameters			privacy			utility
techniques f	factor	Range(m_i)	ℓ'	z	ϵ	δ	MAE
identity	N/A	$[m_{min}, m_{max}]$	ℓ	m_i	N/A	N/A	$w(D)$
top coding	θ_{m_+}	$[m_{min}, \theta_{m_+}]$	ℓ	2	ϵ_2	δ_0	$\hat{w}_{\theta_{m_+}}(D)$
bottom coding	θ_{m_-}	$[\theta_{m_-}, m_{max}]$	ℓ	θ_{m_-}	$\epsilon_{\theta_{m_-}}$	0	$\hat{w}_{\theta_{m_-}}(D)$
sampling	β	$[\beta m_{min}, \beta m_{max}]$	ℓ	2	ϵ_2	δ_0	$\hat{w}_{\beta}(D)$
generalization	c	$[m_{min}, m_{max}]$	ℓ/c	2	ϵ_c	δ_0	$\hat{w}_{c}(D)$
k-anonymity	c_k	$[m_{min}, m_{max}]$	ℓ/c_k	2	ϵ_{c_k}	δ_0	$\hat{w}_{c_k}(D)$

データセット D に対するいくつかの匿名加工の技術 f を，θ, β, c などのパラメータでモデル化した．それらのいくつかは，プライバシーの観点では，差分プライバシーの ϵ, δ に影響を与え，有用性については，元のデータに対する MAE を増加させる．

6. 結論

本研究では，購入履歴からの顧客プロファイリングにおけるプライバシーと有用性について議論した．顧客プロファイルを期間内に購入したことのある商品の集合を表すベクトルで表すモデルを提案し，顧客が期間内に行った購入記録（レコード）の数が与えられたときのプロファイルの確率分布を明らかにした．しかしながら，このプロファイルは期間内に 1 度しか購入をしない稀な顧客がいる時に，差分プライバシーでいうところのプライバシー損失を発散させ，安全性の下限を算出できない．そこで，δ_1 の小さな確率で不成立することを認めた弱い差分プライバシーを証明した．

一方，識別のリスクを下げるために，トップ・ボトムコーディング，サンプリング，一般化，k-匿名性などの匿名化の技術が知られている．提案プロファイルモデルは，これらの匿名化の加工の安全性の利得と有用性の損失を定量化することを可能にしている．代表的な加工アルゴリズムを単純なパラメータで定式化し，差分プライバシーの ϵ の変化により安全性を，元のデータに対する平均絶対誤差 MAE により有用性を評価する式を示した．更に，Online Retail データセットの約 45,000 レコードによる評価結果を示している．

主要な結論の一つは，差分プライバシーにおけるプライバシー費を抑える最も効果的な方式が，ボトムコーディングであることである．なぜならば，それは，識別しやすい頻度の低い稀な記録を削除するからと考えることが出来るからである．代表的な匿名加工である k-匿名性は，本評価では係数 c_k による一般化で実現できることが示されている．

参考文献

[1] Mavriki, Paola and Karyda, Maria. “Profiling with big data: Identifying privacy implications for individuals, groups and society”, The 12th Mediterranean Conference on Information Systems (MCIS), 2018.

[2] General Data Protection Regulation, Regulation (EU) 2016/679 (`https://gdpr-info.eu` refereed in 2019)

[3] Article 29 Data Protection Working Party, “Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation”, 2018.

[4] D. Chen, S. L. Sain, and K. Guo, “Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining,” Journal of Database Marketing and Customer Strategy Management, Vol. 19, No. 3, pp. 197–208, 2012.

[5] K. El Emam, L. Arbuckle, “Anonymizing health data case studies and methods to get you started”, *O’Reilly*, 2013.

[6] G. Linden, B. Smith, and J. York, “Amazon.com recommendations: Item-to-item collaborative filtering”, IEEE Internet Computing, vol. 7, no.1, pp.76–80, 2003.

[7] P. Samarati and L. Sweeney, “Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression,” SRI Int., Menlo Park, CA, USA, Tech. Rep., 1998.

[8] A. Machanavajjhala, J. Gehrke, D. Kiefer, and M. Venkatasubramanian, “L-diversity: Privacy beyond k-anonymity,” ACM Trans. Knowl. Discovery Data, vol. 1, no. 1, Art. no. 3, 2007.

[9] N. Li, T. Li, and S. Venkatasubramanian, “t-Closeness: Privacy beyond k-anonymity and ℓ-diversity,” in Proc. 23rd IEEE Int. Conf. Data Eng. (ICDE), pp. 106–115, 2007.

[10] J. Cao and P. Karras, “Publishing microdata with a robust privacy guarantee,” Proc. VLDB Endowment, vol. 5, no. 11, pp. 1388–1399, 2012.

[11] C. Dwork, “Differential privacy,” in Proc. 33rd Int. Colloq. Automata, Lang. Program. (ICALP) (Lecture Notes in Computer Science), vol. 4052, Berlin, Germany, Springer, pp. 1–12, 2006.

[12] A. Narayanan and V. Shmatikov, “Robust De-anonymization of Large Sparse Datasets,” 2008 IEEE Symposium on Security and Privacy (S&P 2008), pp. 111-125, 2008.

[13] ISO/IEC 20889, “Privacy enhancing data de-identification terminology and classification of techniques”, 2018.

[14] Apostolos Pyrgelis, Carmela Troncoso, Emiliano De Cristofaro, “Knock Knock, Who’s There? Membership Inference on Aggregate Location Data”, NDSS 2018, 2018.