ユビキタスコンピューティングにおける1-エントロピーを満たす匿名データ収集

ユビキタスコンピューティングにおける l-エントロピーを満たす匿名データ収集

清　雄一[†a)]　　大須賀昭彦[†b)]

Anonymized Data Collection Satisfying l-Entropy in Ubiquitous Computing

Yuichi SEI[†a)] and Akihiko OHSUGA[†b)]

あらまし　ユビキタスコンピューティング環境において多くのユーザからセンシングしたデータを収集し，その分布を把握することによって，国の政策や企業における意思決定に役立てることができる．しかし，これらのデータには個人を特定できる情報が含まれることがあり，ユーザのプライバシー情報が漏洩するリスクがある．このような問題に対応し，全てのユーザが必ず正しくない情報を提供することで，プライバシーを保護しつつ，サーバ側で真のデータ分布を推測する Negative Survey という手法が提案されている．従来の Negative Survey では多数のユーザ情報を収集しなければ分布を高精度に推測できないという欠点があった．近年，少ないユーザ数から真の分布を推測することができる手法が複数提案されているが，いずれもプライバシー保護レベルが低いという課題がある．本研究では，プライバシー保護レベルを一定レベルに保ち，従来手法よりも真の分布に近い情報を得られる手法を提案する．近年提案されている手法と比較して平均 2 乗誤差を約 1/2 から 1/30 程度にまで削減できることを数学的解析及びシミュレーションによって示す．

キーワード　ユビキタスコンピューティング，プライバシー，匿名化

1. まえがき

ユビキタスコンピューティング技術やセンシング技術の発展により，ユーザに関する様々な情報を収集する研究が盛んに行われている [1], [2]．各ユーザに対して個別最適なサービスを提供するためには，それぞれのユーザ情報を正確に収集する必要がある．しかし，これらのデータには個人を特定できる情報が含まれることがあり，ユーザのプライバシー情報が漏洩するリスクがある．一方で，データマイニングの目的が各ユーザの属性分布を統計的に把握することである場合には，各ユーザの情報を詳細に収集する必要はない．そのための手法として，ユーザの属性データをカテゴリー化し，ユーザが属さないカテゴリーからランダムに選択されたカテゴリー情報を収集する Negative Survey という手法が提案されている [3]〜[6]．

† 電気通信大学，調布市
The University of Electro-Communications, Chofu-shi, 182–8585 Japan
a) E-mail: sei@is.euc.ac.jp
b) E-mail: ohsuga@euc.ac.jp

Negative Survey では，取り得るユーザの属性値の範囲が 0 から 99 であるとしたとき，0 から 9 までをカテゴリー C_1，10 から 19 までをカテゴリー C_2 のように表現する．例えば，あるユーザの属性値が 37 である場合，カテゴリーは C_4 である．このとき，Negative Survey では C_4 以外の任意のカテゴリーをサーバへ報告する．サーバへ報告された情報からは，ユーザが属すカテゴリーを正確に判定できないため，プライバシー情報を一定レベルで保護することができる．

一方で Negative Survey では，多くのユーザから情報を収集することで，各カテゴリーに属すユーザ数を統計的に推測することができる．そのため，プライバシー保護レベルと，各カテゴリーに属すユーザ数の真の値と推測値との**平均 2 乗誤差**（以下では，単に「平均 2 乗誤差」と記述する）との間に，トレードオフの関係がある（プライバシー指標は **4. 2** に，平均 2 乗誤差は **4. 4. 3** において定義する）．

従来の Negative Survey（以降では **Straight 手法**と呼ぶ）では，カテゴリー数の増加や，ユーザ情報の項目数の増加に従って平均 2 乗誤差が急速に増大するという問題がある．近年，サーバへ送信するカテゴ

リーを選択する際の確率分布を正規分布にする手法 [3] や，カテゴリーを多次元化する手法 [6] が提案され，平均 2 乗誤差の減少が図られている．しかし，これらの手法はプライバシー保護レベルが大きく低下するといった課題がある．

本研究では，l-多様性 [7] と呼ばれるプライバシー情報の匿名化指標を Negative Survey で利用できるように拡張した指標に基づいた匿名化手法を提案する．本提案手法においても，Straight 手法と比較してプライバシー保護レベルは低下するが，データの平均 2 乗誤差は大きく減少する．例えば，カテゴリー数が 48 であるとする．l の理論上の最大値は 47 であるが，これを 46 に緩和した場合，提案手法は従来手法の平均 2 乗誤差を約 47%削減する．

本論文の構成を示す．**2.** では，本論文が想定しているサービスモデル，攻撃モデル及びプライバシーモデルを定義する．**3.** では既存研究について述べる．**4.** において，本論文が提案する手法を記述する．**5.** では，提案手法と既存手法の比較を，数学的解析及びシミュレーションによって実施する．**6.** において，l の設定値や，提案手法の課題とその対策について考察を述べる．**7.** で本論文のまとめを記す．

2. 想定環境

2.1 サービスモデル

公共施設等に設置されたセンサネットワークが，付近を通るユーザに関するデータを取得し，取得情報をサーバへ送信する．これらの情報を基に，あらかじめ設定された各カテゴリーに属するユーザの人数を把握することを目的としたサービスを想定する．取得するユーザの情報は，Public Health [8] におけるユーザの年齢，性別，人種，体重や病名，匿名交通モニタリングにおける自動車の速度 [9] 等が考えられる．

2.2 攻撃モデル

サーバは，semi-honest であることを想定する．semi-honest とは，サーバはプロトコルから逸脱したことは行わないが，受信したデータから各ユーザの属性を推測しようとするという攻撃モデルである．また，センサノード自体への攻撃による情報抽出等の攻撃も想定される [10]．これらの問題は本論文のスコープ外であるが，暗号鍵の配備手法 [11]，悪意をもつノードの検知手法 [12] 等の手法を用いて対応することが考えられる．

2.3 プライバシーモデル

ここでは，「ユーザが Negative Survey に参加しているかどうかにかかわらず攻撃者が当該ユーザに関して推測できる情報をプライバシー情報とみなすかどうか」をプライバシーモデルと捉えて議論する．本論文では，各ユーザが自分の情報を開示することによって攻撃者に与える情報をプライバシー情報として定義する．言い換えると，ユーザが Negative Survey に参加しているかどうかにかかわらず攻撃者が当該ユーザに関して推測できる情報はプライバシー情報とみなさない．本論文が採用するプライバシーモデルに基づき，具体的な数値として，プライバシーの保護の度合いを算出する指標を「プライバシー指標」と捉え，既存のプライバシー指標は **3. 2** に，提案するプライバシー指標は **4. 2** に記述する．

本論文が採用するプライバシーモデルは，数学的には，Evfimievski らの論文に基づき，以下のように定義される [13]．

各ユーザ u の真のカテゴリー Cu は，全てのユーザで共通の確率分布から独立にランダムに抽出されたものとみなす．この確率分布を p_c とおくと，この p_c 自体はプライバシー情報ではなく，サーバが p_c を知ることをユーザは許容する．言い換えると，ユーザ u を除く全てのユーザについて真のカテゴリーの情報が得られたとしても，その情報とユーザ u の真のカテゴリー Cu とは独立しているため，Cu に対する推測には何の影響も与えない．

一般的なアンケート調査を例に挙げて説明する．C_1 を 0〜1000 万円，C_2 を 1000 万円〜2000 万円のようにカテゴリーを定義して行う給料についてのアンケート調査を考える．あるユーザ A の回答が「C_1 か C_2 のいずれか」であり，ユーザ B は未回答であったとする．ユーザ A としては「C_1 か C_2 のいずれか」という情報しか開示おらず，またユーザ B は何の情報も開示していないが，その他ほぼ全てのユーザが「C_1 である」と回答した場合，ユーザ A や B についてのカテゴリーも「ほぼ確率 1 で C_1 である」と推測される．

しかしアンケートに回答していないユーザの情報が，その他多くのユーザの回答結果から推測されたとしても，通常はプライバシー情報の漏洩とはみなされないと考えられ，本論文ではそのようなプライバシーモデルを想定する．したがって本論文では上述の例の場合にユーザ A については「C_1 か C_2 のいずれか」という情報のみがプライバシー情報として開示されており，

ユーザ B については何のプライバシー情報も開示されていないと考える．これは，[13]～[15] 等において想定されているプライバシーモデルと同じである．

3. 関 連 研 究

ユーザが自身のデータを改変してサーバに送信し，サーバは得た情報から解析を行う，というプライバシー保護モデルはローカルモデルと呼ばれる [15]．本章では，ローカルモデルの匿名化手法や，利用されているプライバシー指標を記述する．

3.1 匿名化手法

3.1.1 Negative Survey

カテゴリー数を F とし，それぞれ $C_1, C_2, \ldots, C_F$ と表す．ユーザはいずれかのカテゴリーに属し，これを **True Category (TC)** と呼ぶ．あるユーザの True Category が C_i であるとき，C_i 以外のカテゴリーを選択し，サーバへ報告する．サーバへ報告するカテゴリーを **Negative Category (NC)** と呼ぶ．TC が C_i であるとき，C_j を NC として選択する確率を $p_{j,i}$ とし，次のような確率行列をあらかじめ設定しておく．

$$M = \begin{pmatrix} 0 & p_{1,2} & p_{1,3} & \cdots \\ p_{2,1} & 0 & p_{2,3} & \cdots \\ p_{3,1} & p_{3,2} & 0 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \quad (1)$$

$$\text{where} \quad 0 \le p_{j,i} \le 1 \text{ and } \sum_{j=1}^{F} p_{j,i} = 1$$

また，必須の条件ではないが，既存研究においては，確率行列として巡回行列が用いられていることが多い．この場合，

$$\sum_{i=1}^{F} p_{j,i} = 1 \quad (2)$$

も成り立つ．

サーバ側において，C_i が NC として報告された回数を Y_i とする．各 TC に属すユーザ数の推測値を A_i とすると，各 A_i は次のように計算することができる．

$$A = M^{-1}Y \quad (3)$$

ここで，A=$(A_1, A_2, \ldots, A_F)^\tau$，$Y$=$(Y_1, Y_2, \ldots, Y_F)^\tau$ である．

Straight 手法では，$i = j$ のとき $p_{j,i} = 0$，$i \neq j$ のとき $p_{j,i} = 1/(F-1)$ が設定される [5]．しかし Straight 手法は平均 2 乗誤差が大きいという問題がある．

Straight 手法よりも平均 2 乗誤差を小さくすることができる手法が，Xie らによって提案された [3]．以降で，この手法を **GSN 手法**と呼ぶ．GSN 手法では，式 (1) で表される確率行列の各行について，$1/(F-1)$ ではなく正規分布に基づいて値を設定する．したがって，例えばユーザが NC として C_5 をサーバに通知したとき，当該ユーザの TC は，C_4 または C_6 である可能性が最も高く，C_1 や C_{10} である可能性はほとんどないことが判明する．

Groat らは，カテゴリーを多次元化させる手法を提案している [6]．以降で，この手法を **MDA 手法**と呼ぶ．MDA では，$i \neq j$ における幾つかの $p_{j,i}$ の値も 0 に設定する．そのため，サーバに報告された NC を基に，ユーザの TC 候補から，これらのカテゴリーを完全に除外できる．

本論文では，「ユーザの NC から当該ユーザの TC を推測する際における不確かさ」に基づくプライバシー指標（l-エントロピー）を **4.2** において提案するが，この指標の下では，GSN 手法及び MDA 手法における，プライバシーを保護する度合いが大きく減少すると推測される（l-エントロピーで計算される値が小さいほど，プライバシーを保護する度合いも小さい）．その理由は以下のとおりである．後述するが，l-エントロピーは，確率行列の各行を確率分布とみなしたとき，情報理論におけるエントロピー（シャノンの情報量）の計算式と同じ形をしており，シャノンの情報量は一般に，要素のばらつきが大きいほど，エントロピー値が減少することが知られている．GSN 手法や MDA 手法のように，確率行列の要素に 0 が多く含まれると，確率行列の各行を確率分布とみなしたときの，その確率分布のばらつきが大きくなるため，l-エントロピーの値が減少すると推測される．実際に，これらの l-エントロピーの値が，提案手法における確率行列と比べて，大きく減少するかどうかは，**5.1** の数学的解析で確認する．

3.1.2 ランダムアプローチ

Negative Survey と近い手法として，ランダムアプローチ [13], [14] がある．ユーザの属性値にランダムな値を加えてサーバへ報告し，サーバ側で真の値を推測する．しかし各ユーザは真の値から大きく外れることのない値を加えることになるため，攻撃者にとって各

ユーザにおける真の値の取り得る範囲が明らかになるという問題が指摘されている [16].

3.1.3 その他のプライバシー保護手法

盛んに研究されているプライバシー保護手法として，k-匿名性 [17]，l-多様性 [7]，差分プライバシー [18] 等に基づく手法がある．これらのプライバシー保護手法は，ユーザの真のデータを「完全に信頼できる」サーバに集め，それを信頼できない第三者に開示する際に適用する匿名化手法である．したがって，ユーザがサーバを完全には信頼できない場合，つまり，サーバに真のデータを保存することを許諾しない場合は，このようなプライバシー保護手法を利用することができない．ユーザ間で真のデータをやりとりして，k-匿名化等を行う手法もあるが，この場合は，見知らぬ他ユーザを信頼する必要がある．一方 Negative Survey というプライバシー保護手法は，サーバが semi-honest である場合でも利用することができるというメリットがある．

3.2 ローカルモデルにおけるプライバシー指標

本節ではローカルモデルに対して利用できる既存のプライバシー指標について述べ，**2.3** で定義したプライバシーモデルにおいては，これら既存のプライバシー指標で Negative Survey という手法を評価する際には問題が生じることを示す．

Evfimievski ら [13] は大きく分けて二つの指標を提案している．一つ目の指標では，各ユーザの「真の値」を「誤差を含む値」に変換する操作のみにプライバシー指標が依存するとしている．[15] や [14] 等においても同様である．本論文の用語で述べると，確率行列 M がどの程度 TC をかく乱するか，ということをプライバシー指標としている．これらの研究では，**2.3** で定義したプライバシーモデルが想定されている．

Evfimievski ら [13] で定義されている γ-amplifying や Kasiviswanathan ら [15] で定義されている ϵ-local は，本論文の用語で述べると，確率行列 M の各項の値における，最小値と最大値の比をプライバシー指標としている．したがって M に一つでも 0 が含まれている場合はどのような確率行列を用いたとしても，プライバシーを保護する度合いは等しいことになる．本論文では，M に 0 が一つ以上含まれている場合でも，確率行列によってはプライバシーを保護する度合いに大小をつけることが望ましいと考え，γ-amplifying や ϵ-local とは異なったプライバシー指標を提案する．そのように考える理由を以下に述べる．

例えば，カテゴリー数が 100 である状況を考える．確率行列 M_A として，$p_{i,(i+1) \mod F} = 1$，それ以外の $p_{j,i} = 0$ と定義する．また，確率行列 M_B として，$i = j$ のとき $p_{j,i} = 0$，それ以外の $p_{j,i} = 1/99$ と定義する．

確率行列 M_A を利用した場合，例えばユーザの NC が C_1 のとき，当該ユーザの TC は必ず C_{100} である．したがって，ユーザの TC は全く保護されていない．一方，確率行列 M_B を利用した場合，例えばユーザの NC が C_1 であるとき，当該ユーザの TC は C_2 から C_{100} までいずれの可能性もある．

γ-amplifying や ϵ-local の指標を用いると，確率行列 M_A 及び M_B のいずれを利用した場合でも，プライバシーを保護する度合いは等しくなってしまう．

本論文では，ユーザの TC を保護するという目標を考えているため，上記の例のような場合においては，確率行列 M_B を用いたほうがプライバシーを保護する度合いが大きくなるプライバシー指標を用いることが望ましいと考える．後述するが，本論文では，情報理論のエントロピーと同じ考え方に基づき，「ユーザの NC から当該ユーザの TC を推測する際における不確かさ」をプライバシー指標として採用する．プライバシーを保護する度合いを測る指標として情報理論のエントロピーを利用する考えは，Boutsis らの研究等においても採用されている [19].

4. 提案手法

本章では，提案手法の流れを述べた後，利用するプライバシー指標を定義し，提案手法の実施に必要な確率行列の提案を行う．利用する変数やパラメータ名を表 1 に定義する．

4.1 提案手法に基づく匿名データ収集の流れ

4.1.1 前準備

後述する l-エントロピーの定義に基づき，満たすべき l の値を設定する．次にユーザ属性のカテゴリーを設定し，**4.3** で提案する確率行列を構築する．

表 1 Notation
Table 1 Notation.

F	ユーザ属性のカテゴリー数
N	ユーザ数
C_i	ID が i であるカテゴリー
X_i	TC が C_i である実際のユーザ数
Y_i	NC が C_i であるユーザ数
A_i	TC が C_i であると推測されたユーザ数

4.1.2 データ収集及び分析

ユビキタスコンピューティング環境において対象ユーザの属性値を測定し，TC を決定する．次に，確率行列に基づき NC を決定してサーバへ報告する．

多数のユーザから NC を受け取ったサーバは，既存の Negative Survey と同様，式 (3) を利用して各カテゴリーに属す真のユーザ数を推測する．

4.2 プライバシー指標

3.2 で述べたように，ローカルモデルで利用されている [13] や [15] のプライバシー指標を，Negative Survey においてそのまま利用することはできない．本論文では近年，匿名化の分野で広く利用されている l-多様性 [7] を，Negative Survey で利用できるよう拡張した指標（**l-エントロピー**）を提案する．

l-多様性には幾つかのバリエーションがあるが，最もプライバシーを保護する度合いが強いとされている [20]Entropy l-多様性をベースにする．通常の Entropy l-多様性は，直観的には「公開データから推測される，ユーザの非公開情報の属性値として考えられる値の確率分布の分散が十分大きいこと」が要求されている．

ここで，以下の定理が成り立つ．

Theorem 1. Negative Survey において，ユーザが NC として C_j をサーバに報告するということは，TC が $C_i\,(i=1,\ldots,F)$ である確率は $p_{j,i}\,(i=1,\ldots,F)$ である，と宣言していることと同義である．

証明は付録 2. に記載する．

したがって Theorem 1. より，l-エントロピーは直観的には,「NC から推測される，ユーザの TC として考えられるカテゴリーの確率分布（$p_{j,i}\,(i=1,\ldots,F)$）の分散が十分大きいこと」を要求する．

l-エントロピーの厳密な定義は以下のとおりである．ユーザの NC が C_j であるときに，真のカテゴリーが C_i である確率を $p_{j,i}$ とすると，全ての j に対して，

$$H=-\sum_{1\le i\le F,\, i\neq j} p_{j,i}\ln(p_{j,i})\ge \ln(l) \tag{4}$$

が満たされているとき，l-エントロピーが満たされる．また，任意の j について計算される H の最小値を **l-エントロピー値**と呼ぶ．

4.3 確率行列の提案

ユーザの TC が C_i であるとき，NC として C_j を選択する確率を $p_{j,i}$ とし，その確率行列を作成する．本手法では，

$$\begin{cases} 2\le z\le F-1 \\ 1\le m\le z-1 \end{cases} \tag{5}$$

を満たす実数 z 及び自然数 m を用意し，$p_{i+1,i}$ から $p_{i+m,i}$ までの確率を $1/z$ と比較的高い値に設定する．また，$i=j$ の場合は 0，それ以外の場合は，和が 1 になるよう確率を均一に設定する．

結果，得られる確率行列は以下のようになる．

$$p_{j,i}=\begin{cases} 0 & i=j \\ 1/z & j\ge i+1 \& j\le i+m, \\ & i+m\ge F \& j\le i+m-F \\ \frac{1-m/z}{F-m-1} & otherwise \end{cases} \tag{6}$$

ここで，Theorem 1. より，あるユーザの NC が C_j である場合，当該ユーザの TC が C_i である確率は $p_{j,i}$ である．したがって，式 (6) より，あるユーザの NC が C_j である場合，C_j が当該ユーザの TC である確率は 0 である．また，$C_{j-1},\ldots,C_{j-m}$ が TC である確率はそれぞれ $1/z$ であり，それ以外のカテゴリーである確率は $(1-m/z)/(F-m-1)$ となる．

具体例を以下に示す．$F=5$，$z=2$，$m=1$ の場合，提案する確率行列は以下のように表される．

$$M=\begin{pmatrix} 0 & 1/6 & 1/6 & 1/6 & 1/2 \\ 1/2 & 0 & 1/6 & 1/6 & 1/6 \\ 1/6 & 1/2 & 0 & 1/6 & 1/6 \\ 1/6 & 1/6 & 1/2 & 0 & 1/6 \\ 1/6 & 1/6 & 1/6 & 1/2 & 0 \end{pmatrix}$$

ここで，例えば NC として C_3 が選択されたとする．このとき，カテゴリー C_i が TC である確率は $p_{3,i}$ である．したがって，C_3 が TC である確率は 0，C_2 が TC である確率は $1/2$（$=1/z$），それ以外のカテゴリーが TC である確率はそれぞれ $1/6$（$=(1-m/z)/(F-m-1)$）である．

Groat ら [6] は，確率行列と平均 2 乗誤差との関係について考察を行っている．この考察によると，カテゴリー数を F としたとき，確率行列の各要素が $1/F$ より外れるほど平均 2 乗誤差が減少する．本論文においてもこの考察に基づき，$1/F$ より外れた値を設定することで，平均 2 乗誤差が減少することを期待する．Straight 手法においては，確率行列の F^2 成

分中，$(F-1)^2$ 個の成分の値を $1/(F-1)$ の固定値に設定している．提案手法では，多くの要素の値が $(1-m/z)/(F-m-1)$ と設定されるため，m 及び z の値を適切に設定することにより，Straight 手法よりも平均 2 乗誤差が減少することが期待される（m 及び z の値の設定方法は **4.4** において述べる）．この期待が正しいことは，**5.1** の数学的解析において確認する．

一方，同様にパラメータを適切に設定することにより，確率行列の各要素に対角成分を除いて 0 に近い値を設定することなく，大きなばらつきを与えないようにすることができるため，GSN や MDA よりも l-エントロピー値が大きくなると推測される．**5.1** における数学的解析において，提案手法のパラメータを適切に設定することにより，既存手法よりも l-エントロピー値が大きくなることを確認する．

4.4 確率行列のパラメータ最適化

本節では，確率行列で利用するパラメータである z 及び m について，最適な値の設定方法を述べる．

4.4.1 パラメータ最適化の概要

l-エントロピーの指標における l 及び，カテゴリー数 F が与えられるものとする．

式 (5) の範囲内において m 及び z を変動させ，l-エントロピーを満たす m と z の組合せを算出する．m と z からエントロピー値を算出する方法を **4.4.2** において述べる．

次に，上記で得られた範囲内で m 及び z の値を変動させ，それぞれの組合せにおいてサーバにおける平均 2 乗誤差の期待値を算出する．平均 2 乗誤差の定義及び期待値の算出方法を **4.4.3** において述べる．最も平均 2 乗誤差の期待値が小さい組合せを採用する．

4.4.2 プライバシー

式 (4) 及び式 (6) より，提案手法における H は下記の式で表される．

$$H = -\frac{m}{z}\ln(\frac{1}{z}) - (1-\frac{m}{z})\ln(\frac{1-m/z}{F-m-1}) \quad (7)$$

この H が，設定された l において $\ln(l)$ 以上になる必要がある．

また，ここで次の定理が成り立つ．したがって，$H \geq \ln(l)$ を満たす m 及び z の範囲を算出する際に，m 及び z の取り得る可能性のある範囲を狭めることができる．

Theorem 2.

H は，z の値に対して単調増加し，m の値に対して単調減少する．

Proof. 式 (7) を z について微分すると，

$$\frac{\partial H}{\partial z} = -\frac{m}{z^2}\ln(\frac{z-m}{F-m-1}) \quad (8)$$

となる．ここで，$1 \leq m < z \leq F-1$ であるため $\ln(\frac{z-m}{F-m-1})$ は常に 0 以下の値を取る．したがって式 (8) は常に 0 以上であるから，H は z の値に対して単調増加する．

また，式 (7) を m について微分すると，

$$\frac{\partial H}{\partial m} = \frac{1}{z}(\frac{F-z-1}{F-m-1} + \ln(\frac{z-m}{F-m-1})) \quad (9)$$

である．式 (9) が常に 0 以下の値を取ることを示す．

式 (9) を m で微分すると，

$$\frac{\partial H}{\partial m^2} = \frac{(F-z-1)^2}{(F-m-1)^2(m-z)z} \quad (10)$$

であり，式 (10) は常に負の値を取るため，$m=1$ のとき，式 (9) は最大値を取る．

式 (9) に $m=1$ を代入し，それを z で微分すると，

$$\frac{1}{z^2}\left(\frac{1}{z-1} - \frac{1}{F-2} + \ln(\frac{F-2}{z-1})\right) \quad (11)$$

であり，式 (11) は常に 0 以上の値を取るため，z が取り得る最大値である $F-1$ のとき，式 (9) は最大値を取る．

これらより $m=1$，$z=F-1$ のとき式 (9) は最大値を取り，このときの値は 0 である．したがって，式 (9) は常に 0 以下であるため，H は m の値に対して単調減少する． □

4.4.3 平均 2 乗誤差

TC が C_i である実際のユーザ数を X_i，TC が C_i であると推測されたユーザ数を A_i とし，平均 2 乗誤差を次式の MSE（Mean Squared Error）として定義する．

$$MSE = \frac{1}{F}\sum_i (\frac{A_i}{N} - \frac{X_i}{N})^2 \quad (12)$$

ここでは，MSE の期待値を導出する．以下で述べる MSE の期待値はおおむね [21] と同じであるが，[21] には一部誤りがある．[21] では，式 (13) における右辺の第 2 項において $1 \leq j < k \leq F$ とすべきところ，$1 \leq j \leq F, 1 \leq k \leq F, j \neq k$ としており，この第 2 項

を本来の値の 2 倍にしている．

Theorem 3.

ユーザから NC として C_i が報告された総数を Y_i とおき，ユーザ総数を N とする．このとき，$P_{y_i} = Y_i/N$ と定義する(注1)．また，行列 M の逆行列における s 行 t 列目の値を $M^{-1}_{s,t}$ とする．MSE の期待値を $E[MSE]$ とおくと，

$$E[MSE] = \frac{1}{F \cdot N} \sum_i \Big\{ \sum_j (M^{-1}_{i,j})^2 P_{y_j}(1 - P_{y_j}) - 2 \sum_{1 \le j < k \le F} M^{-1}_{i,j} M^{-1}_{i,k} P_{y_j} P_{y_k} \Big\} \tag{13}$$

と表すことができる．

Proof. 期待値を E で表すと，式 (12) より，

$$E[MSE] = \frac{1}{F} \sum_i E[(\frac{A_i}{N} - \frac{X_i}{N})^2] \tag{14}$$

と表すことができる．本論文のように式 (3) を用いて A_i の計算を行った場合，A_i は X_i の不偏推定量となることが示されている [21]〜[23]．不偏推定量の定義から，$E[A_i] = X_i$ である．

また，$E[(A_i/N - E[A_i/N])^2]$ は A_i/N を確率変数とみなしたときの A_i/N の分散を表している．したがって確率変数 A_i/N の分散を $Var(A_i/N)$ と表現すると，

$$E[MSE] = \frac{1}{F} \sum_i Var(\frac{A_i}{N}) \tag{15}$$

である．式 (3) より $A_i/N = \sum_j M^{-1}_{i,j} \cdot Y_j/N$ である．ここで，一般に確率変数 S 及び T に対して，共分散を $Cov(S,T)$ とおくと，$Var(aS + bT) = a^2 Var(S) + b^2 Var(T) + 2abCov(S,T)$ が成り立つ (a, b は定数)．したがって，

$$Var(\frac{A_i}{N}) = \sum_j (M^{-1}_{i,j})^2 Var(\frac{Y_j}{N}) + 2 \sum_{1 \le j < k \le F} M^{-1}_{i,j} M^{-1}_{i,k} Cov(\frac{Y_j}{N}, \frac{Y_k}{N}) \tag{16}$$

である．ここで，NC が C_i である割合を P_{y_i} とすると以下が成り立つ．

(注1)：P_{y_i} は Y_i から計算されるものであるため，y の添字を用いている．

$$\begin{aligned} Var(\frac{Y_j}{N}) &= E[(\frac{Y_j}{N})^2] - (E[\frac{Y_j}{N}])^2 \\ &= \frac{1}{N^2} \sum_{s=0}^{N} P^s_{y_j} (1 - P_{y_j})^{N-s} {}_N C_s \cdot s^2 - P^2_{y_j} \\ &= \frac{1}{N^2} \cdot N \cdot P_{y_j} \cdot (1 + P_{y_j}(N-1)) - P^2_{y_j} \\ &= \frac{1}{N} \cdot P_{y_j}(1 - P_{y_j}) \end{aligned} \tag{17}$$

$$\begin{aligned} Cov(\frac{Y_j}{N}, \frac{Y_k}{N}) &= E[\frac{Y_j}{N}\frac{Y_k}{N}] - E[\frac{Y_j}{N}]E[\frac{Y_k}{N}] \\ &= \frac{1}{N^2} \sum_{s=0}^{N} \sum_{t=0}^{N-s} \Big\{ P^s_{y_j} P^t_{y_k} (1 - P_{y_j} - P_{y_k})^{N-s-t} \\ &\quad \cdot {}_N C_s \cdot {}_{N-s} C_t \cdot s \cdot t \Big\} - P_{y_j} P_{y_k} \\ &= \frac{1}{N^2}(-N + N^2) P_{y_j} P_{y_k} - P_{y_j} P_{y_k} \\ &= -\frac{1}{N} \cdot P_{y_j} P_{y_k} \end{aligned} \tag{18}$$

式 (15)〜(18) より，式 (13) が導出される． □

TC が C_i であるユーザ数の分布が事前にある程度予測できる場合は，その分布に応じて，利用予定の確率行列を用いて P_{y_i} を計算し，MSE の期待値を算出する．ユーザ総数を N とし，TC が C_k であるユーザ数の事前の予測値を X'_k とおくと，

$$P_{y_i} = \frac{1}{N} \sum_{k=1}^{F} p_{i,k} X'_k \tag{19}$$

と設定することができる．なお，TC が C_i であるユーザ数の分布が事前に分からない場合は，一様分布に設定する．つまり，全ての k について $X'_j = 1/F$ として P_{y_i} を計算することになるが，この場合，式 (19) を計算した結果として $P_{y_i} = 1/F \sum_{k=1}^{F} p_{i,k}$ が得られる．式 (6) より，提案手法の確率行列は巡回行列であることが分かる．したがって，式 (2) より $\sum_{k=1}^{F} p_{i,k} = 1$ であるため，$P_{y_i} = 1/F$ となる．

5. 評　　価

本章では，各パラメータが，提案手法に与える影響を評価する．また，l-エントロピーにおける l の値と平均 2 乗誤差（MSE）のトレードオフの関係について，既存の Negative Survey 手法と比較する．

5.1 MSE の期待値の分析

提案手法について，$N = 10{,}000$，$F = 48$ としたと

きの MSE を計算した結果を図 1 に示す．提案手法では，大きな傾向として z の値が大きくなるほど MSE も増大していることが分かる．また，一部分（$m=1$ の場合 $z=24$ 等）において大きく MSE が増加している．この理由は次のように考えられる．$m=1$ かつ $z=24$ の場合，確率行列のある行に注目すると，ある 1 成分の値は 0，$m(=1)$ 成分の値は 1/24，残り 46 成分の値は $(1-1/24)/46=1/48$ となり，ほとんどの値が $1/F$ と一致している．例えば，NC として C_5 が報告された例を考える．このときの TC は，C_5 である可能性は 0，C_4 である確率は 1/24，その他のカテゴリーについては $1/F$ である．したがってこの NC からは，C_5 と C_4 以外のカテゴリーの出現頻度に関する情報は全く得られない．このように $(1-m/z)/(F-m-1)$ の値が限りなく $1/F$ に近い場合は平均 2 乗誤差が大きい（$m=2$ のときの $z=32$ 等）．

このように一部分は Straight 手法よりも MSE が大きく上回っているが，m や z をうまく調整することによって，常に Straight 手法よりも MSE の期待値を削減することが可能である．

次に，既存手法と比較した結果を図 2 (a) に示す．図では，エントロピー値 $\ln(l)$ を変化させ，そのエントロピー値を満たすように各手法のパラメータを設定している．どのエントロピー値を取っても，提案手法の MSE が最も小さい値を実現していることが分かる．$l=30$ のときエントロピー値は約 3.4 であるが，このとき Straight，GSN，MDA，提案手法の MSE はそれぞれ約 4.5×10^{-3}，1.1×10^{-3}，1.9×10^{-3}，3.8×10^{-5} であり，提案手法の MSE が 1/30 以下になっていることが分かる．また，Straight 手法のエントロピー値は $\ln(47)=3.85$ である．このエントロピー値を満たす必要がある場合，Straight，GSN，MDA，提案手法の MSE は同一の値になる．しかし，$\ln(46)=3.82$ を満たすよう設定する場合，GSN 及び MDA 手法における MSE はほとんど変わらないが，提案手法の MSE は約 1.9×10^{-3} であった．これは既存手法の約 42%程度の値である．また，MSE が同じ値であるとき，l-エントロピー値は，提案手法が最も大きいことが分かる．

次に，ユーザ数 N を変化させたときの結果を図 2 (b) に示す．エントロピー値が $\ln(30)$ を超えるよう，各手法におけるパラメータを設定している．MSE が N に比例して減少していることが分かる．

次に，カテゴリー数 F を変化させたときの結果を図 2 (c) に示す．ここでも，エントロピー値が $\ln(30)$ を超えるよう，各手法におけるパラメータを設定した．Straight 手法はカテゴリー数が増加するほど MSE も増加しているが，その他の手法はカテゴリー数の増加に応じて MSE が減少している．これらの中でも，提案手法における MSE が最も小さいことが分かる．

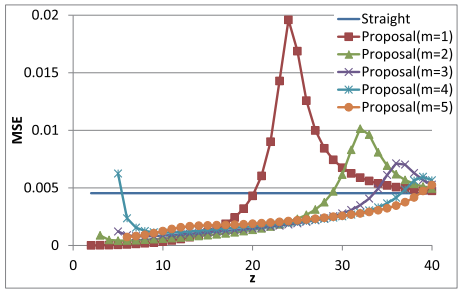
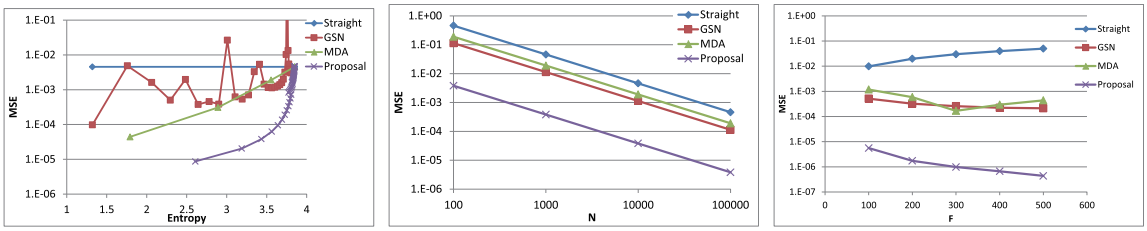


図 1　$N=10,000$, $F=48$ における提案手法の MSE
Fig. 1　MSE where $N=10,000$, $F=48$ of proposal.

5.2　シミュレーション

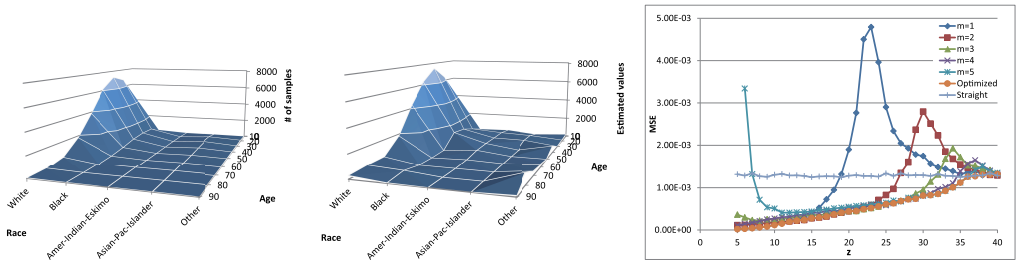
5.2.1　実データを利用した MSE の計測

匿名化手法における研究分野で広く利用されている UCI の Adult データセット [24] を用いて評価を行った．Adult データセットの属性のうち，カメラセンサネットワーク [25] 等で推測可能だと考えられる年齢，人種の 2 属性を利用した．年齢は 10 代から 90 代ま



(a) $N=10,000$, $F=48$　(b) entropy $>\ln(30)$, $F=48$　(c) entropy $>\ln(30)$, $N=10,000$

図 2　MSE の比較
Fig. 2　Comparison of MSE.



(a) Original data set　(b) Estimated result　(c) MSE of the estimated result

図 3　Adult データセットの分布と推測結果
Fig. 3　Distribution of Adult data set and estimated results.

で存在しており，それを 10 代，20 代のようにカテゴリー化して利用した．また，人種は White, Black, Amer-Indian-Eskimo, Asian-Pac-Islander, Other の 5 種類であった．したがって合計で 45 の組み合わせが存在する．データ総数は 32,561 件であった．

年齢及び人種の組み合わせにおけるデータ分布を，図 3 (a) に示す．また $z = 10$ に設定した際の提案手法における推測結果を図 3 (b) に示す．図より，データ分布の傾向が十分に復元できていることが分かる．

z を変化させたときの MSE を計測した結果を図 3 (c) に示す．NC の生成と，サーバによる TC の分布推測までを 100 回繰り返した結果の平均値を示している．あらかじめ **4. 4** で提案した手法を用いて m 及び z の最適値を求め，その値を利用した結果を Optimized として示している(注2)．実際に測定された各 m における最小値と，Optimized の値がほぼ一致しており，最適な m が導出されていると言える．

5. 2. 2　計 算 時 間

提案手法において，収集された NC のデータから，真のデータ分布を求めるために必要な時間を計測した．実験は，OS が Windows 7 Professional 64 bit, CPU が Intel Core i7-3712QM CPU @ (2 CPUs), RAM が 8GB である PC を利用して行った．

上述した Adult データセットにおいて，収集された NC のデータから，各 TC に属すユーザ数を求めるために要した時間は，1,017ms であった．

次に，ランダムにデータを生成し，カテゴリー数を 100 から 1,000 まで変化させて計測を行った結果を図 4 に示す．$N = 10,000$，$z = 10$，$m = 1$ に設定した．また，N，z，m をそれぞれ変動させた場合の計算時間を計測したが，結果は図 4 とほぼ同一となった．

(注2)：m 及び z の最適な値を事前に計算する際は，TC が C_i であるユーザ数の分布は事前に判明していないと想定し，全ての i に対して，$P_{y_i} = 1/F$ と設定して MSE の期待値を算出した．

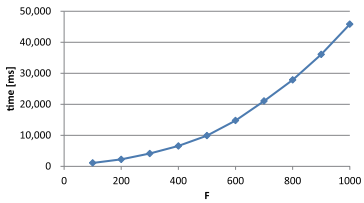


図 4　シミュレーション時間
Fig. 4　Simulation time.

6. 考　　察

6. 1　l の設定値

l-多様性を提案している [7] では，シミュレーションのデフォルト値として $l = 6$ を用いている．また，l-多様性よりも匿名化の度合いが緩いとされている k-匿名性の指標において，k の値は多くの既存研究において 10 程度以下に設定されていることを考慮すると，l の値を本論文のように理論上の最大値未満に設定することは妥当であると考えられる．また，l を Negative Survey における理論上の最大値よりも 1 だけ小さい値を設定しても，カテゴリー数が 48 の場合に MSE を 1/2 程度に削減できている．適切な l の値の決定方法は，将来課題として今後検討する必要がある．

6. 2　手法の限界

手法の限界として，ID が近いカテゴリーが意味的にも近い場合，ユーザの真の属性が推測されるリスクが高まることが挙げられる．ここで，「ある二つのカテゴリーが意味的に近い」かどうかの判断はアプリケーションに依存する．数値をそのままカテゴリー化したような例では，ID が近いカテゴリー，つまり C_i と C_{i+1} が表現する内容は近いと考えられる．また，病名を調査するような場合においても，C_1 が A 型肝炎，

C_2 がB型肝炎，C_3 がA型インフルエンザ，C_4 がB型インフルエンザ，のように設定されている場合，C_1，C_2 のカテゴリー群及び C_3，C_4 のカテゴリー群はそれぞれ意味的に近いと考えられる．

このような場合，カテゴリー設定を工夫する必要が生じる．例えば，あるユーザ属性値の取り得る範囲が1から80までであり，C_i $(i = 1, \ldots, 8)$ がそれぞれ1から10，11から20，21から30，31から40，41から50，51から60，61から70，71から80を表すとする．$m = 2$，$z = 3$ の場合，ユーザのNCが C_5 だとすると，ユーザのTCが C_i $(i = 1, \ldots, 8)$ である確率はそれぞれ，1/15, 1/15, 1/3, 1/3, 0, 1/15, 1/15, 1/15 である．したがって，2/3の確率でユーザ属性値が21から40であることが判明する．

一方，C_i $(i = 1, \ldots, 8)$ がそれぞれ1から10，41から50，11から20，51から60，21から30，61から70，31から40，71から80を表すようにカテゴリー設定を行うことを考える．$m = 2$，$z = 3$ の場合，ユーザのNCが C_5 だとすると，ユーザの真の値は，1/3の確率で11から20，同じく1/3の確率で51から60となる．

この二つの例を考えたとき，「TCを推測されない」という意味では，同一レベルでプライバシーは保護されている．しかし，「ユーザの真の値を推測されない」という意味では，前者のほうがプライバシー情報を保護できていない．この課題に対する対策として，次節で述べる**カテゴリーランダム化**を行うことが考えられる．

6.3 カテゴリーランダム化

6.3.1 問題の定義

カテゴリー C_i が表す**定義内容**を λ_i とおく．例えばユーザ属性のデータ値の範囲が0から39であり，0から9までを C_1，10から19までを C_2 のように表現すると，$\lambda_1 = [0, 9]$，$\lambda_2 = [10, 19]$ のようになる．

このように，カテゴリーランダム化を行う前は，λ_i は C_i の定義内容を表している．以下で v_i を算出し，カテゴリーランダム化を行った後は，λ_i が C_{v_i} の定義内容を表すように変更する．例えば $v_1 = 4, v_2 = 3, v_3 = 1, v_4 = 2$ である場合，λ_i $(i = 1, ...4)$ はそれぞれ C_4，C_3，C_1，C_2 の定義内容を表すようになる．逆に言うと，C_1 は λ_3 $(= [20, 29])$，C_2 は λ_4 $(= [30, 39])$，C_3 は λ_2 $(= [10, 19])$，C_4 は λ_1 $(= [0, 9])$ を表すことになる．

添字 i が隣接する m 個の C_i $(i = 1 \mod F, \ldots (i + m - 1) \mod F)$ を考える．これらの各 C_i の定義内容を表す m 個の λ_j に対し，添字である j の値の最大値と最小値の差をできるだけ大きくすることを目標とする．これは次の問題を解くことと同じである．

Problem 1. 1から F まで番号の付いた F 個の球を円形に並べ，連続する m 個の球に書かれている番号を読み取る．その番号の最大値と最小値の差を d とおく．連続する m 個の球の選択の仕方は F 通りある．この F 通りについてそれぞれ d を求め，その最小値を d_{min} とする．この d_{min} を最大化するように，F 個の球を円形に並べるにはどうすれば良いか？

6.3.2 カテゴリーランダム化アルゴリズム

ここでは，インデクスが連続しているカテゴリー（例えば C_i と C_{i+1}）の定義内容が意味的に遠くなるようにカテゴリーの定義内容を設定すると，上記「手法の限界」が軽減されると仮定し，更に，この「意味的に遠くなるようにカテゴリーの定義内容を設定する問題」を Problem 1 に置き換えられると仮定する．このとき，カテゴリーの定義内容を入れ替えることによって上記「手法の限界」を軽減しようとする方針においては，提案するカテゴリーランダム化を行うことで，手法の限界を理論上最も軽減させることができる．ここで述べた仮定を取り除いた場合に，提案するカテゴリーランダム化が手法の限界に対して具体的に何を保証するのかについて明らかにすることは将来課題とする．

ここでは並び替えアルゴリズムを提案し，得られる d_{min} が理論上の最大値を満たすことを付録1.において証明する．

まず以下を満たす変数 n を用意する．

$$n = \begin{cases} F/m & F \bmod m = 0 \\ \lfloor F/m \rfloor & F \bmod m = 1 \\ \lceil F/m \rceil & otherwise \end{cases} \tag{20}$$

$(1 \le i \le n)$ or $(F - n + 1 \le i \le F)$ or $(m \ne 2$ and $i = F - n)$ を満たす i に対して，次式で v_i を導出する．ここで，$\min(a, b)$ は，a と b のうち小さいほうを返す関数である．

$$v_i = \begin{cases} (i-1)m+1 & 1 \le i \le n \\ \min((i-F+n)m, F) & F-n+1 \le i \le F \\ 2 & i = F-n \ \&\ m \ne 2 \end{cases} \tag{21}$$

上記式で得られた v_i の集合を V とおく．この導出範囲に含まれない i に対しては，V に含まれない値から，1 以上 F 以下の値の範囲で重複を許さずランダムに v_i を設定する．

任意の i（$i = 1, \ldots, F$）に対し，v_i と $v_{(i+1) \bmod F}$ の値の差の最小値は d は以下を満たす．

$$d \geq F - n - 1 \tag{22}$$

これは理論上の最大値である．証明は付録 1. に記す．

7. むすび

本論文では，ユーザが属す真のカテゴリーを隠してサーバへ通知し，サーバ側で各カテゴリーに属すユーザ数を推測する手法を提案した．まず，匿名化に関する研究分野で広く利用されている l-多様性の概念を Negative Survey の分野に適用させた．次に，真のカテゴリーを隠蔽し，どのカテゴリーを報告するかを決定するための新しい確率行列を構築した．また，近接したカテゴリーが類似の意味をもつことがないように，カテゴリーランダム化手法の提案を行った．これらの提案要素を用いることにより，プライバシー保護レベル及び平均 2 乗誤差のトレードオフを従来手法より高いレベルで取れることを，数学的解析及び実データを用いたシミュレーション実験で示した．設定によっては，従来手法よりも平均 2 乗誤差を約 1/2 から 1/30 程度までに削減することができた．

謝辞　本研究は JSPS 研究費 24300005，23500039，25730038 の助成を受けたものです．

本研究を遂行するにあたり，研究の機会と議論・研鑽の場を提供して頂き，ご指導頂いた国立情報学研究所/東京大学 本位田 真一 教授をはじめ，活発な議論と貴重なご意見を頂いた研究グループの皆様に感謝致します．

文　献

[1] J. Froehlich, E. Larson, S. Gupta, G. Cohn, M. Reynolds, and S. Patel, “Disaggregated end-use energy sensing for the smart grid,” IEEE Pervasive Computing, vol.10, no.1, pp.28–39, 2011.

[2] Z. Li, M. Li, J. Wang, and Z. Cao, “Ubiquitous data collection for mobile users in wireless sensor networks,” Proc. IEEE INFOCOM, pp.2246–2254, 2011.

[3] H. Xie, L. Kulik, and E. Tanin, “Privacy-aware collection of aggregate spatial data,” Data & Knowledge Engineering, vol.70, no.6, pp.576–595, 2011.

[4] S. Forrest and M. Groat, “Reconstructing spatial distributions from anonymized locations,” Proc. IEEE ICDEW, pp.243–250, 2012.

[5] J. Horey, M.M. Groat, S. Forrest, and F. Esponda, “Anonymous data collection in sensor networks,” Proc. MobiQuitous, pp.1–8, IEEE, 2007.

[6] M.M. Groat, B. Edwards, J. Horey, W. He, and S. Forrest, “Enhancing privacy in participatory sensing applications with multidimensional data,” Proc. IEEE PerCom, pp.144–152, 2012.

[7] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, “l-diversity: Privacy beyond k-anonymity,” ACM TKDD, vol.1, no.1, Article No.3, 2007.

[8] J. Corburn, “Confronting the challenges in reconnecting urban planning and public health,” American journal of public health, vol.94, no.4, pp.541–546, 2004.

[9] C. Sharp, S. Schaffert, A. Woo, N. Sastry, C. Karlof, S. Sastry, and D. Culler, “Design and implementation of a sensor network system for vehicle tracking and autonomous interception,” Proc. EWSN, pp.93–107, IEEE, 2005.

[10] E. Platon and Y. Sei, “Security software engineering in wireless sensor networks,” Progress in Informatics, vol.5, no.1, pp.49–64, 2008.

[11] L. Eschenauer and V.D. Gligor, “A key-management scheme for distributed sensor networks,” Proc. ACM CCS, pp.41–47, 2002.

[12] F. Ye, H. Yang, and Z. Liu, “Catching “moles” in sensor networks,” Proc. IEEE ICDCS, p.69, 2007.

[13] A. Evfimievski, J. Gehrke, and R. Srikant, “Limiting privacy breaches in privacy preserving data mining,” Proc. ACM PODS, pp.211–222, 2003.

[14] R. Agrawal and R. Srikant, “Privacy-preserving data mining,” Proc. ACM SIGMOD, pp.439–450, 2000.

[15] S.P. Kasiviswanathan, H.K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith, “What can we learn privately ?,” SIAM Journal on Computing, vol.40, no.3, pp.793–826, 2013.

[16] R. Chen and A. Reznichenko, “Towards statistical queries over distributed private user data,” Proc. USENIX NSDI, pp.169–182, 2012.

[17] P. Samarati, “Protecting respondents identities in microdata release,” IEEE Trans. Knowl. Data Eng., vol.13, no.6, pp.1010–1027, 2001.

[18] C. Dwork, “Differential privacy,” Automata, Languages and Programming, vol.4052, pp.1–12, Lecture Notes in Computer Science, Springer, 2006.

[19] I. Boutsis and V. Kalogeraki, “Privacy preservation for participatory sensing data,” Proc. IEEE PerCom, pp.103–113, 2013.

[20] N. Li, T. Li, and S. Venkatasubramanian, “t-closeness: Privacy beyond k-anonymity and ldiversity,” Proc. IEEE ICDE, pp.106–115, 2007.

[21] Z. Huang and W. Du, “OptRR: Optimizing randomized response schemes for privacy-preserving data

mining," Proc. IEEE ICDE, pp.705–714, 2008.
[22] F. Esponda, "Negative surveys," Technical report, arXiv.org, 2006.
[23] M.M. Groat, B. Edwards, J. Horey, W. He, and S. Forrest, "Application and analysis of multidimensional negative surveys in participatory sensing applications," Pervasive and Mobile Computing, vol.9, no.9, pp.372–391, Jan. 2013.
[24] "UCI Machine Learning Repository: Adult Data Set," http://archive.ics.uci.edu/ml/datasets/Adult
[25] H. Ma, D. Li, Y. Hong, and W. Chen, "Minimum camera barrier coverage in wireless camera sensor networks," Proc. IEEE INFOCOM, pp.217–225, 2012.

付　　録

1. カテゴリーランダム化アルゴリズム

Problem1 の問題に対し，d_{min} の理論上の最大値を導出する．また，本文中で提案したアルゴリズムがこの理論上の最大値を実現することを証明する．

以降では λ_i に対応する球を，球 $[i]$ と記述する．

1.1 d_{min} の理論上の最大値

まず $F \bmod m = 0$ の場合を考える．

円形に並べられた F 個の球を重複無しで，連続する m 個ずつグループ分けする（図 A·1 (a)）．このときグループは n 個作成される（n の定義は式 (20)）．各グループに対し，$G_1, G_2, \ldots$ のように ID を振る．

G_1 に球 $[1]$，G_2 に球 $[2]$，のように各グループに対し球の ID が小さいものから順に配置する（ここでは理論上の最大値を導出するため，各グループ内における球の配置の順番は問わない）．また，G_n に球 $[F]$，G_{n-1} に球 $[F-1]$，のように逆順に球の ID が大きいものから配置する．結果，G_i には球 $[i]$ 及び球 $[F-n+i]$ が配置される．明らかに，注目している n 個の各グループにおいて d を求めると，この配置のときのみ d の最小値が最大化される．このとき各グループにおいて以下が成り立っている．

$$d = F - n \tag{A·1}$$

図 A·1 (a) にはグループが n 個ある．残り $F-n$ 個のグループ分けを考えると，必ず，いずれかのグループ分けについては，「G_j には球 $[j]$ 及び球 $[F-n+j]$ が配置される」ことを実現できない．したがって，d_{min} の理論上の最大値は式 (A·1) より 1 だけ小さくなる．

次に $F \bmod m = 1$ の場合を考える．F 個の球を重複無しで，連続する m 個ずつグループ分けする．余った一つの球はどのグループにも属さない（図 A·1 (b)）．このときグループは $n(=\lfloor F/m \rfloor)$ 個作成される．上述と同様の議論により，d_{min} の理論上の最大値は，式 (A·1) より 1 だけ小さくなる．

最後に $F \bmod m > 1$ の場合を考える．F 個の球を重複無しで，連続する m 個ずつグループ分けする．余った $F \bmod m$ 個の球も 1 グループにカウントする（図 A·1 (d)）．このときグループは $n(=\lceil F/m \rceil)$ 個作成される．上述と同様の議論の結果，d_{min} の理論上の最大値は，式 (A·1) より 1 だけ小さくなる．

以上より，以下の式が満たされる．

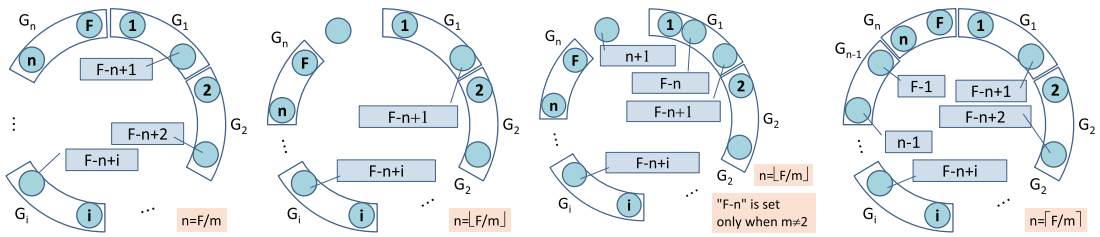
$$d_{min} = F - n - 1 \tag{A·2}$$

1.2 提案アルゴリズムの証明

提案アルゴリズムが d_{min} の理論上の最大値 $F-n-1$ を満たすことの証明を行う．

$F \bmod m \neq 1$ の場合

式 (21) より，v_i が $1, m+1, \ldots, (n-1)m+1$ となる球 ID は $1, 2, \ldots, n$ であり，v_i が m，$2m$，…，$\min(nm, F)$ となる球 ID は $F-n+1, F-n+2, \ldots, F$ である．結果，球を v_i に応じて時計回りに並べると，$F \bmod m = 0$ の場合は図 A·1 (a)，$F \bmod m > 1$ の場合は図 A·1 (d) のようになる．図中の各グループ内



(a) $F \bmod m = 0$　(b) $F \bmod m = 1$ (Ideal)　(c) $F \bmod m = 1$ (Proposal)　(d) $F \bmod m > 1$

図 A·1　球の配置
Fig. A·1　Location of balls.

において $d = F - n$ である．また，G_i 内の m 番目の球 $[F - n + i]$ と G_{i+1} 内の 1 番目の球 $[i + 1]$ における ID の差は $F - n - 1$ である．また，G_n の m 番目の球 $[F]$ と G_1 の 1 番目の球 [1] における ID の差は $F - 1$ である．したがって，任意の m 個の連続する球に対して $d \geq F - n - 1$ が成り立つ．

$F \bmod m = 1$ の場合

式 (21) より，v_i が $1, m + 1, \ldots, nm + 1$ となる球 ID は $1, 2, \ldots, n + 1$ であり，v_i が $m, 2m, \ldots, nm$ となる球 ID は $F - n + 1, F - n + 2, \ldots, F$ である．$m \neq 2$ の場合は，v_i が 2 となる球 ID は $F - n$ である．結果，球を v_i に基づいて時計回りに並べると図 A·1 (c) のようになる．図中の各グループ内において $d = F - n$ である．また，G_i 内の m 番目の球と G_{i+1} 内の 1 番目の球における ID の差は $F - n - 1$ である．

次に，G_n と G_1 の間にある余りの球 $[n + 1]$ に注目する．

m の値によらず，球 $[F]$，球 [1]，球 $[n + 1]$ を含めて連続する m 個の球を選択する場合，$d \geq F - n - 1$ が満たされる．

$m \neq 2$ の場合，球 $[F]$ を含まず，球 $[n + 1]$ を含めて連続する m 個の球を選択する場合，必ず球 [1] と球 $[F - n]$ が選択される．したがって，$d \geq F - n - 1$ が満たされている．球 [1] を含まず，球 $[n + 1]$ を含めて連続する m 個の球を選択する場合，必ず球 $[F]$ が含まれる．したがって $d \geq F - n - 1$ が満たされる．

$m = 2$ の場合，球 $[F]$ を含まず，球 $[n + 1]$ を含めて連続する $m(= 2)$ 個の球を選択する場合，球 $[n + 1]$ と球 [1] が選択される．この二つの球における ID の差は，$n = \lfloor F/2 \rfloor = F - \lfloor F/2 \rfloor - 1$ であるから，$d \geq F - n - 1$ が満たされている．球 [1] を含まず，球 $[n + 1]$ を含めて連続する $m(= 2)$ 個の球を選択する場合，球 $[n + 1]$ と球 $[F]$ が選択される．この二つの球における ID の差は $F - n - 1$ である．

以上より，常に $d \geq F - n - 1$ が満たされる．したがって，式 (A·2) が成り立つ．

2. Theorem 1. の証明

カテゴリー数を F とする．ユーザの TC が C_i であるとき，NC として C_j を選択する確率を $p_{j,i}$ とおく．ここでは，ユーザが NC として C_j をサーバに報告した場合において，この NC のみから，サーバが当該ユーザの TC を推測する場合，TC が $C_i\,(i = 1, \ldots, F)$ である確率は $p_{j,i}\,(i = 1, \ldots, F)$ となることを証明する．

サーバは，NC として C_j を受け取ったとき，当該ユーザの TC が C_i であった条件付き確率 $P(TC = C_i | NC = C_j)$ を次のように求めることができる．ここで，$P(NC = C_j)$ は NC が C_j である確率を表し，$P(TC = Ci)$ は TC が C_i である確率を表す．ベイズの定理より，

$$P(TC = C_i | NC = C_j) = \frac{P(NC = C_j | TC = C_i) \cdot P(TC = C_i)}{P(NC = C_j)} \quad \text{(A·3)}$$

となる．ここで，$p_{j,i} = P(NC = C_j | TC = C_i)$ である．

また，当該ユーザの TC が C_i である確率は未知であるため，全ての i について $P(TC = C_i)$ が同一の値をもつと仮定する．つまり，以下が成り立つ．

$$P(TC = C_i) = \frac{1}{F} \quad \text{for all } i \qquad \text{(A·4)}$$

また，ユーザの TC が C_i であるとき，NC として C_j を選択する確率が $p_{j,i}$ であるため，

$$P(NC = C_j) = \sum_{i=1}^{F} P(TC = C_i) \cdot p_{j,i} \qquad \text{(A·5)}$$

と表される．したがって，式 (A·3) より，

$$P(NC = C_j) = \frac{1}{F} \sum_{i=1}^{F} p_{j,i} \qquad \text{(A·6)}$$

である．式 (A·3)，(A·4)，(A·6) より，

$$P(TC = C_i | NC = C_j) = p_{j,i} / \sum_{i=1}^{F} p_{j,i} \qquad \text{(A·7)}$$

が成り立つ．ここで，式 (6) より，提案する確率行列は巡回行列であることが分かる．したがって，式 (2) 及び式 (A·7) より，以下が導かれる．

$$P(TC = C_i | NC = C_j) = p_{j,i} \qquad \text{(A·8)}$$

（平成 25 年 7 月 1 日受付，10 月 21 日再受付）

清　　雄一　（正員）

1981 年生．2009 年東京大学大学院情報理工学系研究科博士後期課程修了．同年（株）三菱総合研究所入社．同社情報技術研究センター，金融ソリューション本部等に所属．2013 年より電気通信大学助教，現在に至る．分散コンピューティング，セキュリティ，プライバシー保護技術等の研究に従事．情報処理学会，電子情報通信学会，IEEE Computer Society 各会員．

大須賀昭彦　（正員）

1958 年生．1981 年上智大学理工学部数学科卒．同年 (株) 東芝入社．同社研究開発センター，ソフトウェア技術センター等に所属．1985〜1989 年 (財) 新世代コンピュータ技術開発機構 (ICOT) 出向．2007 年より，電気通信大学大学院情報システム学研究科教授．2012 年より，国立情報学研究所客員教授兼任．工学博士 (早稲田大学)．主としてソフトウェアのためのフォーマルメソッド，エージェント技術の研究に従事．1986 年度情報処理学会論文賞受賞．IEEE Computer Society Japan Chapter Chair，人工知能学会理事，日本ソフトウェア科学会理事を歴任．情報処理学会，電子情報通信学会，人工知能学会，日本ソフトウェア科学会，IEEE Computer Society 各会員．