情報処理学会論文誌 Vol.55 No.9 2120–2133 (Sep. 2014)

多次元属性のための匿名データ収集アルゴリズムの提案

清 雄一[1,a)] 大須賀 昭彦[1,b)]

受付日 2013年11月26日, 採録日 2014年6月17日

概要：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，マーケティング等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情報が漏洩するリスクがある．このような問題に対応し，ユーザが一定の確率で真実でない情報をサーバに送信するよう制約を設けることで，プライバシを保護しつつ，サーバ側で真のデータ分布を再構築する Randomized Response（RR）という手法が提案されている．再構築された結果と真のデータ分布との間には誤差があるが，収集対象となるユーザ属性が複数ある場合，従来手法ではどの程度の誤差が発生するか知ることができず，再構築した結果から有効な分析ができないという実用上の課題があった．また，要求されるプライバシ保護レベルを満たしたうえで，この誤差を最小化できるような RR のパラメータの設定方法も提案されていない．さらに，ユーザ属性の数が増加するほど，再構築に要する計算時間が膨大になるという課題もある．本論文ではこれら実用上の課題を解決する手法を提案する．数学的解析および実データを利用したシミュレーション結果により，提案手法の有効性を示す．

キーワード：プライバシ，データマイニング

Anonymized Data Collection for Multi-dimensional Attributes

Yuichi Sei[1,a)] Akihiko Ohsuga[1,b)]

Received: November 26, 2013, Accepted: June 17, 2014

Abstract: Ubiquitous computing environment can collect sensing data of users. These data can be used for several purposes such as decision-making of companies. However, collecting user data may include personally identifiable information and violate their privacy. Randomized response scheme which collect disguised data of each user and can assume true data distributions of users have been proposed. However, existing studies do not provide a calculation method of estimated errors between the true data distributions and the reconstructed data distributions when multiple attributes are needed to be anonymized and collected. Also, they do not provide a method of setting an RR's parameter that will minimize erros and ensures a required privacy level. Moreover, existing studies need a lot of calculation time for reconstructing data distributions if the number of user attributes is large. We prove out proposed method is effective by mathematical analysis and simulations.

Keywords: privacy, data mining

1. はじめに

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さかんに行われている [9], [17]．ユーザの属性データを直接収集することはプライバシ情報の漏洩につながる場合もあるため，ユーザの属性データをカテゴリ化し，一定の確率で真実でないカテゴリをサーバへ送信することでプライバシを保護する，Randomized Response（RR）[2], [4], [13], [23] という手法が提案されている．

3 章に示すとおり，RR には，同一カテゴリ選択確率というパラメータを，要求されるプライバシ保護レベルに応じて変更させることができる．このパラメータを 0 に固定した RR を特に，Negative Survey（NS）と呼び，NS に特

[1] 電気通信大学大学院情報システム学研究科
Graduate School of Information Systems, The University of Electro-Communications, Chofu, Tokyo 182–8585, Japan
[a)] sei@is.uec.ac.jp
[b)] ohsuga@uec.ac.jp

化した手法も複数提案されている [8], [10], [11], [12]．その中でも Groat ら [10], [11] は，ユーザ属性が複数存在するときに特に有効な NS の手法を提案している．また Aoki ら [3] は Groat らの手法を拡張し，複数のユーザ属性に対応する NS に，1 つのユーザ属性に対応する RR を組合せた手法を提案している．

複数属性に対応した RR も提案されているが [1], [26], [27]，以下のような課題がある．

再構築された結果と真のデータ分布との間には誤差があるが，収集対象となるユーザ属性が複数ある場合，既存手法ではどの程度の誤差が発生するか知ることができない．もし誤差について何も分からなければ，データ分布を再構築できたとしても，真のデータ分布とほとんど同じであるのか，大きく異なっている可能性が高いのか，まったく分からないということになる．これは本来の目的であるデータマイニングを行うにあたって，重大な問題となりうる．また，要求されるプライバシ保護レベルを満たしたうえで，この誤差を最小化できるような RR のパラメータの設定方法も提案されていない．本論文では，データ分布を再構築する計算式から，誤差の期待値を計算する式を導出することによってこの課題を解決する．

また，RR に関する既存手法は，本論文も採用している逆行列手法と，反復ベイズ法とに大きく分けることができる．反復ベイズ法のほうが経験的に誤差が小さいことが分かっているが [1], [26], [27]，特に収集対象のユーザ属性数や各属性におけるカテゴリ数が多い場合，再構築にかかる時間が長い．ユーザがセンシングした結果を収集する参加型環境センシングの分野では広く逆行列手法が利用されており [3], [10], [11]，本論文においても，計算時間が早い逆行列手法を利用し，かつ，従来手法では必要な逆行列の計算を必要としないシンプルな計算式で再構築を可能とする．

本論文の構成を示す．2 章では，本論文が想定しているモデルについて述べ，3 章では既存研究について述べる．4 章ではプライバシ指標および有効性指標を定義する．5 章において，本論文が提案する手法を記述し，数学的な解析を 6 章で行う．7 章では，提案手法と既存手法の比較を，数学的解析およびシミュレーションによって実施する．8 章において考察を述べ，9 章で本論文のまとめを記す．

2. 想定モデル

2.1 アプリケーションモデル

ユーザがスマートフォン等のデバイスを用いて周囲の環境をセンシングし，自分の属性情報（年齢，性別，位置情報等）とともにその結果をサーバへ通知する．これらの情報を基に，あらかじめ設定された各カテゴリに属するユーザの人数を把握するアプリケーションモデルを想定する（**図 1**）．取得するユーザの情報は，Public Health [5] におけるユーザの年齢，性別，人種や病名，匿名交通モニタリング [21] における自動車の速度や運転者の年齢等が考えられる．このアプリケーションモデルは，既存の RR や NS の研究が想定しているモデルと同様のものである [3], [10], [11]．

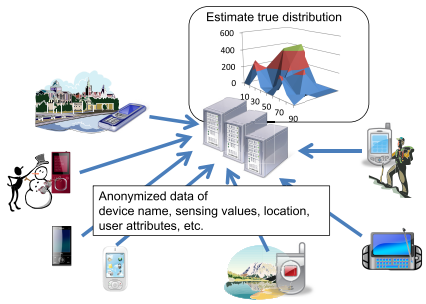


図 1 アプリケーションモデル
Fig. 1 Application model.

たとえば，年齢を属性 0，位置情報を属性 1 とする．サーバは，どのような属性（年齢および位置情報）を持った人が何人ずつ存在しているかを知りたい．ここで，たとえば年齢について，0 歳代 ,...,90 歳代のようにカテゴリ分けをし，それぞれのカテゴリ ID を $0,\ldots,9$ とする．同様に位置情報について，東京都，神奈川県，千葉県のようにカテゴリ分けをし，それぞれのカテゴリ ID を 0，1，2 とする．あるユーザの年齢が 25 歳，現在地が東京都であるとき，当該ユーザの属性 0 におけるカテゴリ ID は 2，属性 1 におけるカテゴリ ID は 0 である．

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攻撃モデルである．また，スマートフォン等のセンシングデバイス自体への攻撃による情報抽出等も想定される [16]．この問題は本論文のスコープ外であるが，マルウェア検知手法等を用いて対処することが考えられる [20]．

2.3 プライバシモデル

本論文で想定するプライバシモデルを述べる．各ユーザが自分の情報を開示することによってサーバに与える情報をプライバシ情報と定義する．いい換えると，ユーザが RR に参加しているかどうかにかかわらず，サーバが当該ユーザに関して推測できるような情報はプライバシ情報とはみなさない．

一般的なアンケート調査を例にあげて説明する．C_1 を 0〜1,000 万円，C_2 を 1,000 万円〜2,000 万円，C_3 を 2,000 万円〜3,000 万円，のようにカテゴリを定義して行う給料についてのアンケートを考える．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．このとき，そのほかほぼすべてのユーザが「C_1 である」と回答した場合，ユーザ A やユーザ B について

のカテゴリも「高い確率で C_1 である」と推測することができる．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，通常はプライバシ情報の漏洩とはみなされないと考えられ，本論文においてはこのようなプライバシモデルを想定する．

このプライバシモデルのフォーマルな定義は次のとおりである．これは，Evfimievski ら [7] や Kasiviswanathan ら [14] が想定するモデルと同一である．

各ユーザ u の真のカテゴリ C_u は，すべてのユーザで共通の確率分布から独立にランダムに選択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．いい換えると，ユーザ u を除くすべてのユーザについて真のカテゴリ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 C_u とは独立しているため，C_u に対する推測には何の影響も与えない．

サーバが，ユーザ u の属性がある性質 Q を持つ確率を事前に求める状況を考える．このとき，その求められる確率を事前確率と呼ぶ．また，サーバが当該ユーザから RR に基づく情報を受け取った状態において，ユーザ u の属性が性質 Q を持つ確率を求める状況を考える．このとき，その求められる確率を事後確率と呼ぶ．本論文では，事後確率と事前確率の差が大きくなりすぎるとプライバシが保護されていないと考える．このモデルに基づいて，プライバシ保護レベルを具体的な数値として表すプライバシ指標は，4.1 節において述べる．

3. 関連研究

3.1 Randomized Response と Negative Survey

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，というプライバシ保護モデルはローカルモデルと呼ばれる [14]．本節では，ローカルモデルの代表的手法である Randomized Response（RR）および Negative Survey（NS）について述べる．

3.1.1 Randomized Response（RR）

まず，収集対象のユーザ属性が 1 つだけであると想定して説明する．当該属性のカテゴリ数を F とし，それぞれ $C_0, \ldots, C_{F-1}$ と表す．ユーザはいずれかのカテゴリに属し，これを **True Category（TC）**と呼ぶ．あるユーザの TC が C_i であるとき，ある確率で C_i 以外のカテゴリを選択し，サーバへ報告する．サーバへ報告するカテゴリを **Disguised Category（DC）**と呼ぶ．TC が C_i であるとき，C_j を DC として選択する確率を $p_{j,i}$ とし，確率行列をあらかじめ設定しておく．確率行列は，対角成分を $p_{i,i} = p$，対角成分以外の全 $i \neq j$ における各成分を同一の値に設定する Uniform Perturbation が広く利用されている [2], [4], [13], [23]．本論文でも Uniform Perturbation を利用する．また，このときの対角成分の値を**同一カテゴリ選択確率**と呼ぶ．この確率行列は以下のように表すことができる．

$$M = \begin{pmatrix} p & \frac{1-p}{F-1} & \frac{1-p}{F-1} & \cdots \\ \frac{1-p}{F-1} & p & \frac{1-p}{F-1} & \cdots \\ \frac{1-p}{F-1} & \frac{1-p}{F-1} & p & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \tag{1}$$

RR に参加したユーザ総数を N とおく．このユーザ集合の中においてあるユーザに着目した際に，当該ユーザの TC が C_i である確率を $P(X = C_i)$ とおく．同様にユーザ集合の中においてあるユーザに着目した際に，当該ユーザの DC が C_i である確率を $P(Y = C_i)$ とおくと，

$$\begin{aligned} &\overrightarrow{Y} = M \cdot \overrightarrow{X} \\ \text{where } &\overrightarrow{X} = (P(X = C_0), \ldots, P(X = C_{F-1}))^\tau, \\ &\overrightarrow{Y} = (P(Y = C_0), \ldots, P(Y = C_{F-1}))^\tau \end{aligned} \tag{2}$$

という関係がある [13]．ここで，τ は転置行列を表す．サーバは $\overrightarrow{X}$ の値を推測することを目的としている．

実際に DC として C_i を報告したユーザ数を N_i とおく．このとき $\overrightarrow{Y}$ の最尤推定量が $\widehat{\overrightarrow{Y}} = \{N_0/N, \ldots, N_{F-1}/N\}$ と求まる．したがって，$\overrightarrow{X}$ の最尤推定量を $\overrightarrow{A}$ とおくと，

$$\begin{aligned} &\overrightarrow{A} = M^{-1} \cdot \widehat{\overrightarrow{Y}} \\ \text{where } &\overrightarrow{A} = (P(A = C_0), \ldots, P(A = C_{F-1}))^\tau \end{aligned} \tag{3}$$

と求めることができる [13]．

次に，収集対象のユーザ属性が複数個あると想定して説明する．

ユーザの各属性を個別に RR を用いてユーザからサーバに報告し，サーバ側でも各属性を個別に再構築することを考える．この場合，各属性が独立しているため，たとえば属性 0 のカテゴリ ID が 2 であり，かつ，属性 1 のカテゴリ ID が 3 であるユーザ数を導出する，ということはできない．したがって，属性の組合せに対してもサーバ側でデータマイニングしたい場合は，各属性の関係を保ったまま，サーバ側で再構築を行う必要がある．

ここで，ユーザ属性を単純に一次元化する手法として，指定されているプライバシ保護レベルから，ただ 1 つの同一カテゴリ選択確率 p を導出し，式 (1) で表される確率行列を利用することを考える．本来であれば，属性ごとに異なる値を設定すべき同一カテゴリ選択確率 p を，同一の値に設定して確率行列を作成していることから，特に各属性のカテゴリ数に偏りが大きいときに，要求されているプライバシ保護レベル以上に過剰にユーザの属性を保護してしまう．このため，サーバ側において再構築した際の平均二乗誤差の値が大きくなってしまう，という課題がある．

各ユーザの属性の組合せ情報を維持したままサーバ側で

再構築を行うために，全属性の組合せを 1 つの属性とみなして再構築を行うことができる．五十嵐ら [26] は，属性ごとに最適な同一カテゴリ選択確率 p_i を設定し，その組合せを考慮することによって確率行列を生成している．具体的には，ある属性 i に注目した場合，その属性の真のカテゴリと，サーバに報告する確率が同一である確率を p_i，異なる確率を $(1-p_i)/(F_i-1)$ に設定し，全属性についてそれらの組合せを導出する．これにより，プライバシ保護レベルが指定された場合，各属性に対して厳密にその保護レベルを満たす同一カテゴリ選択確率 p_i を設定することができているため，要求されているプライバシ保護レベルに対して過不足のない匿名化が可能となる．

文献 [26] は，RR を多次元属性に拡張した手法を 2 通り提案している．1 つは本論文が提案する逆行列手法であり，もう 1 つは反復ベイズ法である．逆行列手法は本論文が提案するアルゴリズムと本質的に同じものであるが，以下の違いがある．

- MSE の期待値を算出する手法が提案されていない
  サーバ側で再構築したデータが，真のデータとどのくらい異なるものであるかが分からなければ，再構築したデータを分析したとしても，どの程度の確信度を持って分析結果を支持できるかが分からない．文献 [26] では，各属性について確率行列を作成し，その逆行列のクロネッカー積を計算し，その結果を利用してサーバ側で再構築を行っている．この計算結果がどうなるかについて議論されていないため，このままだと MSE の期待値を求めることはできない．本論文のアルゴリズムでは，再構築に必要な行列の各項の値を定式化しているため，この式を利用して，MSE の期待値を算出することができている．
- MSE の期待値を最小化する同一カテゴリ選択確率の決定方法が述べられていない．
  Uniform Perturbation である RR を行うためには，同一カテゴリ選択確率を決定する必要がある．文献 [26] では，MSE の期待値を導出することができていないため，要求されるプライバシ保護レベル γ を満たしたうえで，MSE の期待値を最小化する同一カテゴリ選択確率を導出できない．

反復ベイズ法については，同一カテゴリ選択確率 p が特に小さい場合は，反復ベイズ法のほうが，逆行列手法よりも MSE の値が小さくなることがシミュレーション結果により示されている．しかし，7 章で示すように，本論文においてはほとんどの設定の下で，MSE の値はほぼ等しい．また，反復ベイズ法は 7 章に示すように，データによっては計算時間が大きくかかるという課題がある．たとえば，属性数 4，各属性におけるカテゴリ数が 20 である場合，1 回の再構築に約 5.8×10^4 秒が必要であった．再構築を 1 回のみ行う場合はこれでも大きな問題にならないかもしれないが，収集対象のユーザ属性数がより多い場合はさらに問題が大きくなる．たとえば，ユーザ属性数が 20 ある場合を考える．20 のうち 4 個の属性の組合せについて分析を行おうとすると，全部で ${}_{20}C_4 = 4,845$ 通りの再構築を行うことができる．したがって，4 個の属性の組合せ 1 つに対して 5.8×10^4 秒必要である場合，仮にすべての組合せについて再構築を行おうとすると 2.8×10^8 秒必要となり，膨大な時間を要する．もし，最初にすべての組合せについて再構築するのではなく，必要に応じて再構築する場合，必要が生じるたびに 5.8×10^4 秒を要するのは，データ分析に支障をきたすと考えられる．提案手法は約 100 分の 1 の時間で再構築が可能である．

文献 [1] は，各カテゴリに属すユーザ数の推測だけでなく，決定木を作るというアプリケーションにおける工夫がなされている．各カテゴリに属すユーザ数の推測については，文献 [26] と同じアルゴリズムが採用されているため，上記と同じ課題を持つ．

文献 [27] は文献 [1] を拡張し，決定木を作るというアプリケーションにおいて，目的属性が 3 値以上の決定木を作る際における効率的な手法を提案している．決定木を作る際には特に有効な手法であるが，本論文が対象とするような，各カテゴリに属すユーザ数の推測，という目的に関しては，文献 [26] や文献 [1] と同じ性質を持っている．

3.1.2 Negative Survey（NS）

RR において，同一カテゴリ選択確率を 0 に設定する手法は特に Negative Survey（NS）と呼ばれ，複数の研究が提案されている [8], [10], [11], [12].

Groat らは，ユーザ属性が複数存在するときに特に有効な NS の手法を提案している [10], [11]．通常，ユーザ属性が複数存在するときは，それを単純に一次元化して NS を行うことになる（たとえば属性 0 として 10 カテゴリ，属性 1 として 10 カテゴリがある場合，それらを単純に組合せ，100 カテゴリの属性が 1 つだけあるとみなす）が，Groat らはさらに確率行列を工夫することで推測精度を向上させることができることを示した．Aoki ら [3] は Groat らの手法を拡張し，複数のユーザ属性に対応する NS に，1 つのユーザ属性に対応する RR を組合せた手法を提案している．

NS は，RR では調整可能な同一カテゴリ選択確率を 0 で固定することから，実装が簡単であることや，プライバシおよび有効性の解析を簡単に行えるという利点がある．しかしながら，4.1 節において記述するプライバシ保護レベルを柔軟に変更できない等のデメリットも存在する．本論文では 4.1 節に述べるように，プライバシ指標として，値が小さいほどプライバシ保護レベルが高い γ を利用する．NS では確率行列に 0 が含まれているため，γ の値は必ず無限大となる[*1]．したがって本論文で採用するプライバシ

[*1] 正確にはゼロ除算が発生するため計算不可能であるが，$\lim_{\gamma \to 0+}$ を考えることで無限大と算出される．

指標の下では，NS が RR よりもプライバシ保護レベルが高くなることはない．一方，4.2 節で述べる有効性指標においても，RR のほうが NS よりも優れている場合がある（7 章を参照）．本論文とは異なる指標を利用する場合においても，RR の同一カテゴリ選択確率を最適な値に設定することができれば，RR は NS と必ず同等以上の性能を出すことができる．

3.2 その他のプライバシ保護手法

さかんに研究されているプライバシ保護手法として，k-匿名性 [19], l-多様性 [18], 差分プライバシ [6] 等に基づく手法がある．これらのプライバシ保護手法は一般的に，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 RR や NS というプライバシ保護手法は，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．

4. 指標

本章では，本論文で利用するプライバシ指標および有効性指標について述べる．

4.1 プライバシ指標

本論文で考えるプライバシモデルのように，サーバにおける事前知識と事後知識の差に着目した指標として，ρ_1-to-ρ_2 プライバシ [7] が提案されている．これはローカルモデルで広く用いられており，ローカルモデルの 1 つである RR の分野でも利用されている [2], [4], [24].

事前確率が ρ_1 以下である場合に，事後確率が ρ_2 以上となるとき，upward ρ_1-to-ρ_2 プライバシが保護されていない，と定義される．逆に，事前確率が ρ_2 以上である場合に，事後確率が ρ_1 以下となるとき，downward ρ_2-to-ρ_1 プライバシが保護されていないと定義され，upward ρ_1-to-ρ_2 プライバシおよび downward ρ_2-to-ρ_1 プライバシの両方が保護されているとき，ρ_1-to-ρ_2 プライバシが保護されていると定義される．

サーバの事前知識が既知である場合は，ρ_1-to-ρ_2 プライバシはそれぞれ ρ_1 と ρ_2 を明示してプライバシを保護することが可能であるが，一般に，サーバの事前知識が既知である状況は少ない．そこで，0 より大きい値を持つパラメータ γ を用意し，サーバの事前知識が不明な場合においても，事前知識と事後知識の差が大きくなりすぎることがないよう制約を設けることができる．

属性が 1 つしかない RR においては，

$$\frac{p_{j,i}}{p_{j,k}} \leq \gamma \ \text{forall}\ i,j,k \tag{4}$$

かつ

$$\gamma \leq \frac{\rho_2}{\rho_1} \times \frac{1-\rho_1}{1-\rho_2} \tag{5}$$

の場合，サーバの事前知識が不明な場合においても，ρ_1-to-ρ_2 プライバシが満たされることが証明されている（γ-amplification という）．

たとえば，あるユーザが性質 Q を持つ確率が，サーバの事前知識として 5%であったとする．この場合，$\gamma = 10$ であると，サーバの事後知識において，当該ユーザが性質 Q を持つ確率は最大約 35%に抑えられる．一方，事前知識が 1%であり，$\gamma = 10$ であると，当該ユーザが性質 Q を持つ確率は最大約 10%に抑えられる．

このように，サーバにおける事前確率と事後確率の差が大きくなるとプライバシが侵害されているというプライバシモデルにおいては，ρ_1-to-ρ_2 プライバシという指標が，そのモデルを表現することができていると考えられ，また，サーバの事前知識が不明な状況においても，パラメータ γ の値を設定することで，サーバにおける事後確率が事前確率より大きくなりすぎることのないよう制約を設けることができるため，本論文ではこのパラメータ γ を**プライバシ保護レベル**として利用する．この γ の値が小さいほど，プライバシ保護レベルが高い，つまり，サーバにおける事前確率と事後確率との差が小さくなる．

本論文のように Uniform Perturbation を利用する場合，式 (4) は具体的には，同一カテゴリ選択確率 p とカテゴリ数 F を使って，

$$\frac{p}{(1-p)/(F-1)} \leq \gamma \ \text{and}\ \frac{(1-p)/(F-1)}{p} \tag{6}$$

と表すことができる．なぜなら，式 (4) において $p_{j,i}$ および $p_{j,k}$ が取り得る値は，p または $(1-p)/(F-1)$ の 2 値しかないからである．

これを多次元属性へ拡張すると次のようになる．属性 i におけるカテゴリ数を F_i，同一カテゴリ選択確率を p_i とし，属性数が D であるとすると，

$$\frac{p_i}{(1-p_i)/(F_i-1)} \leq \gamma \ \text{and}\ \frac{(1-p_i)/(F_i-1)}{p_i} \leq \gamma \quad \text{for all}\ i = \{0,\ldots,D-1\} \tag{7}$$

となる．直観的には，あるユーザの DC から推測される当該ユーザの TC として，最も可能性の高いカテゴリと最も可能性の低いカテゴリにおける可能性の比がプライバシ指標となっている．

なお本論文では，プライバシ保護レベルは全ユーザ共通であると想定するが，8 章で述べるように，各ユーザで異なるプライバシ保護レベルが設定される状況にも応用可能である．

4.2 有効性指標

有効性指標として，真のデータ分布と推測されたデータ分布の間における，平均二乗誤差（MSE: Mean Squared Error）を利用する．収集対象の属性数を D，各属性 i におけるカテゴリ数を F_i とおく．たとえば，収集対象の属性として年齢および位置情報があり，年齢を「0 歳代」,...,「90 歳代」までに分け，位置情報を「東京都」「神奈川県」「千葉県」の 3 つに分けた場合，$D=2$，$F_0=10$，$F_1=3$ となる．また，全属性における全カテゴリの組合せを表現するベクトルとして $\vec{V}$ を用意する．この例では，10 カテゴリと 3 カテゴリの組合せを表現するため，$10\cdot 3=30$ の要素を持つ．ベクトル $\vec{V}$ の j 番目の要素を $\vec{V}_j$ と表記する．具体的に $\vec{V}$ をどのように設定すれば良いかは，5.3 節において述べる．

このとき MSE を σ^2 で表すと，以下の式で定義する．

$$\sigma^2 = \frac{1}{\sum_{i=0}^{D-1} F_i} \sum_{\vec{V}_j} (P(A=\vec{V}_j) - P(X=\vec{V}_j))^2 \quad (8)$$

ここで，$P(X=\vec{V}_j)$ は，ユーザ集合においてあるユーザに着目した場合に，当該ユーザの各 TC が $\vec{V}_j$ で表現される組合せである確率を表している．たとえば，ユーザ集合全体の数が 1,000 人，「20 歳代」および「東京都」であるユーザ数が 100 人であり，$\vec{V}_6$ が「20 歳代」および「東京都」の組合せを表現している場合，$P(X=\vec{V}_6)=1/10$ である．また，$P(A=\vec{V}_j)$ は，サーバ側で $P(X=\vec{V}_j)$ を推測した結果を表している．

この指標は，RR や NS において広く利用されている [11], [12], [13], [25] [*2]．

5. 提案手法

5.1 事前準備

ここでは，要求されるプライバシ保護レベルを満たし，その範囲において，MSE の期待値を最小化することができる最適な同一カテゴリ選択確率を導出する．

要求されるプライバシ保護レベルを γ とすると，式 (7) より，属性 i においてこの γ を満たす同一カテゴリ選択確率 p_i の範囲は以下のように表される．

$$\frac{1}{1+\gamma(F_i-1)} \le p_i \le \frac{\gamma}{\gamma+F_i-1} \quad (9)$$

この範囲内で，MSE の期待値を最小化できる p_i の値を導出する．ある属性 l の同一カテゴリ選択確率 p_l に注目し，$i\neq l$ の各 p_i を定数とみなしたとき，p_l をどのように設定すると MSE の期待値が最小化されるかを導出する．

後述の式 (29) で表される MSE の期待値 $E[\sigma^2]$ を p_l について偏微分すると以下のようになる．

$$\frac{\partial E[\sigma^2]}{\partial p_l} = \frac{1+\Delta}{N\Delta^2}\mathcal{S}(l)\mathcal{T}(l) \quad (10)$$

where

$$\mathcal{S}(l) = \prod_{\substack{i\neq l,\\ 0\le i\le D-1}} \frac{3-2p_k+F_k(F_k+p_k^2-3)}{(p_kF_k-1)^2}, \quad (11)$$

$$\mathcal{T}(l) = -\frac{2(F_l-1)^3}{(p_lF_l-1)^3} \quad (12)$$

ここで，$\Delta=\prod_i F_i$ であり，また，式 (11) は p_l を含まない式であることに注意する．

式 (12) は，$p_l<1/F_l$ のとき正の値を取り，$1/F_l<p_l$ のとき負の値を取る．式 (11) はつねに正の値を取ることから，式 (10) より，$p_l<1/F_l$ のとき p_l の値が増加するほど MSE の期待値は増加し，$1/F_l<p_l$ のとき p_l の値が増加するほど MSE の期待値が減少することが分かる．

また，式 (29) に対して，$p_l=1/F_l+\delta$ と設定した場合も，$p_l=1/F_l-\delta$ と設定した場合も等しい値を取るため[*3]，MSE の期待値は $p_l=1/F_l$ で対称な関数となる．以上より，式 (9) を満たす範囲の中で，$1/F_l$ との差の絶対値が最も大きい値を選ぶことになる．これは，$1\le\gamma$ かつ $2\le F_l$ の場合は，

$$p_l = \frac{\gamma}{\gamma+F_l-1} \quad (13)$$

である．したがって，各属性 l について，式 (13) を満たす p_l を同一選択確率に設定することで，与えられたプライバシ指標を満たし，MSE の期待値を最小化することができる．

5.2 ノードプロトコル

あるユーザについて，各属性 i の TC を x_i と表す．全属性の TC をまとめて表現したものを **TC セット**と呼び，$\{x_0,\ldots,x_{D-1}\}$ と表す．属性 i については，p_i の確率で DC は x_i となり，$(1-p_i)/(F_i-1)$ の確率で x_i 以外のカテゴリからランダムに選択して DC とし，これを y_i とおく．すべての属性 i について DC を選択した結果，得られたデータ $\{y_0,\ldots,y_{D-1}\}$ を **DC セット**呼び，この情報をサーバへ報告する．

4.2 節で述べた例で考える．年齢を属性 0，位置情報を属性 1 とする．年齢について，0 歳代 ,...,90 歳代のカテゴリ ID をそれぞれ $0,\ldots,9$ とし，位置情報について，東京都，神奈川県，千葉県のカテゴリ ID をそれぞれ 0，1，2 とする．あるユーザの年齢が 25 歳，現在地が東京都であるとき，当該ユーザの属性 0 における TC は 2，属性 1 における TC は 0 であり，TC セットは $\{2,0\}$ である．属性 0 について，DC として TC と同じ 2 を選択する確率は，同一カテゴリ選択確率 p_0 で表される．同様に，属性 1 について，

[*2] ここであげた文献の中には MSE の平方根を取るもの等もあるが，本質的には同じ指標であると考える．

[*3] 具体的には $\frac{1+\Delta}{N\Delta^2}\mathcal{S}(l)\frac{-1+F_l(3+F_l(-3+\delta^2+F_l))}{\delta^2F_l^3}$ となる．

DC として TC と同じ 0 を選択する確率は，同一カテゴリ選択確率 p_1 で表される．したがって，たとえば DC セットとして $\{5,0\}$ が選ばれる確率は，$(1-p_0)/(F_0-1)\cdot p_1$ となる．

このように，ノードは単純な計算により，DC セットを導出することができる．

5.3 サーバプロトコル

式 (2) の関係を満たす $\overrightarrow{X}$，$\overrightarrow{Y}$，M が導出できれば，式 (3) と同様に，TC セットの真のユーザ分布を以下のように推測することができる．

$$\overrightarrow{A} = M^{-1}\cdot\overrightarrow{Y} \tag{14}$$

5.3.1 $\vec{X}$，$\vec{Y}$，$\vec{A}$ の定義

$\overrightarrow{X}$，$\overrightarrow{Y}$，$\overrightarrow{A}$ はそれぞれ，すべての属性のすべてのカテゴリ ID の組合せを表現できる必要がある．したがって，

$$\Delta = \prod_{i=0}^{D-1} F_i \tag{15}$$

とおくと，$\overrightarrow{X}$，$\overrightarrow{Y}$，$\overrightarrow{A}$ は，要素数 Δ の 1 次元ベクトルとして表現することができる．

これをふまえ，$\overrightarrow{X}$，$\overrightarrow{Y}$，$\overrightarrow{A}$ がそれぞれ，$\overrightarrow{X}=(X_0,\ldots,X_{\Delta-1})^\tau$，$\overrightarrow{Y}=(Y_0,\ldots,Y_{\Delta-1})^\tau$，$\overrightarrow{A}=(A_0,\ldots,A_{\Delta-1})^\tau$ であり，各 X_j，Y_j，A_j はそれぞれ，TC セットが $\overrightarrow{V}_j$ である真のユーザ数，DC セットが $\overrightarrow{V}_j$ であるユーザ数，TC セットが $\overrightarrow{V}_j$ であると推測されるユーザ数を表すと定義する．ここで，各 $\overrightarrow{V}_j$ は以下のように定義することができる．

$\overrightarrow{V}_j$ は，D 個の各属性がそれぞれ $\{h(j)_0,\ldots,h(j)_{D-1}\}$ である組合せを表す．たとえば，$\{2,1,2\}$ は，属性 0，属性 1，属性 2 における各カテゴリ ID が，2，1，2 であることを表している．

ここで，$h(j)_i$ は次のように定義することができる．

$$h(j)_i = \left\lfloor \frac{j}{\prod_{k=i+1}^{D-1} F_k} \right\rfloor \mod F_i \tag{16}$$

たとえば，属性数 3，各属性のカテゴリ数がそれぞれ $F_0=5$，$F_1=3$，$F_2=2$ である例を考える．この場合たとえば，$\overrightarrow{V}_7$ について考えると，

$$h(7)_0 = \left\lfloor \frac{7}{F_1F_2} \right\rfloor \mod F_0 = 1$$
$$h(7)_1 = \left\lfloor \frac{7}{F_2} \right\rfloor \mod F_1 = 0$$
$$h(7)_2 = \left\lfloor \frac{7}{1} \right\rfloor \mod F_2 = 1$$

となり，$\overrightarrow{V}_7$ は各属性のカテゴリ ID が $\{1,0,1\}$ である組合せを表していることになる．このようにして計算した，$\overrightarrow{V}_j$ と $\{h(j)_0,h(j)_1,h(j)_2\}$ との関係は**表 1** のようになる．

表 1 $F_0=5, F_1=3, F_2=2$ における $\overrightarrow{V}_j$ と $\{h(j)_0,h(j)_1,h(j)_2\}$ との関係

Table 1 Relationship between $\overrightarrow{V}_j$ and $\{h(j)_0,h(j)_1,h(j)_2\}$ when $F_0=5, F_1=3, F_2=2$.

$\overrightarrow{V}_j$	$\{h(j)_0,h(j)_1,h(j)_2\}$
$\overrightarrow{V}_0$	{0, 0, 0}
$\overrightarrow{V}_1$	{0, 0, 1}
$\overrightarrow{V}_2$	{0, 1, 0}
$\overrightarrow{V}_3$	{0, 1, 1}
$\overrightarrow{V}_4$	{0, 2, 0}
$\overrightarrow{V}_5$	{0, 2, 1}
$\overrightarrow{V}_6$	{1, 0, 0}
$\overrightarrow{V}_7$	{1, 0, 1}
⋮	⋮
$\overrightarrow{V}_{29}$	{4, 2, 1}

したがって，X_j，Y_j，A_j はそれぞれ，TC セットが $\{h(j)_0,\ldots,h(j)_{D-1}\}$ である真のユーザ数，DC セットが $\{h(j)_0,\ldots,h(j)_{D-1}\}$ であるユーザ数，TC セットが $\{h(j)_0,\ldots,h(j)_{D-1}\}$ であると推測されるユーザ数を表すことになる．

5.3.2 確率行列 M の定義

属性 i における同一カテゴリ選択確率を p_i としたとき，式 (2) の関係が満たされるような確率行列 M を導出する．ある Y_j に注目すると，式 (2) より，

$$Y_j = \sum_{k=0}^{\Delta-1} M_{j,k}X_k \tag{17}$$

である．あるユーザの TC セットが $\overrightarrow{V}_k$ $(k=0,\ldots,\Delta-1)$ であるとき，属性 i $(i=0,\ldots,D-1)$ の TC は $h(k)_i$ で表される．同様に DC セットが $\overrightarrow{V}_j$ $(j=0,\ldots,\Delta-1)$ であるとき，属性 i の DC は $h(j)_i$ で表される．$h(k)_i=h(j)_i$ のとき，属性 i について，TC と DC が同一である確率は p_i である．したがって，あるユーザの TC セットが $\overrightarrow{V}_k$ であるときに，当該ユーザの DC セットが $\overrightarrow{V}_j$ となる確率を $m_{j,k}$ とおくと，各属性 i について，$h(k)_i=h(j)_i$ のとき TC と DC が同一である確率が p_i となるような各 $m_{j,k}$ を算出し，確率行列 M を構成すればよい．

この M は以下の式として定義することができる．

$$M = \begin{pmatrix} m_{0,0} & m_{0,1} & \cdots & m_{0,\Delta-1} \\ m_{1,0} & m_{1,1} & \cdots & m_{1,\Delta-1} \\ \vdots & \vdots & \vdots & \vdots \\ m_{\Delta-1,0} & m_{\Delta-1,1} & \cdots & m_{\Delta-1,\Delta-1} \end{pmatrix} \tag{18}$$

where $m_{j,k} = \prod_{i=0}^{D-1} n_{i,j,k}$,

$$n_{i,j,k} = \begin{cases} p_i & \left(\left\lfloor \frac{j}{\prod_{l=i+1}^{D-1} F_l} \right\rfloor \equiv \left\lfloor \frac{k}{\prod_{l=i+1}^{D-1} F_l} \right\rfloor \pmod{F_i}\right) \\ \frac{1-p_i}{F_i-1} & \text{(otherwise)} \end{cases}$$

ここで，2 つの整数 a と b が n を法として合同であるとき，$a \equiv b \pmod{n}$ と表している．

以下に例を示す．式 (2) より，Y_7 について考えると，

$$Y_7 = \sum_{k=0}^{\Delta-1} m_{7,k} \cdot X_k$$

と計算される．例として，属性数 3，各属性のカテゴリ数がそれぞれ $F_0 = 5$，$F_1 = 3$，$F_2 = 2$ である場合を考える．以下では，$m_{7,0}$ の値を算出する．式 (18) に当てはめると，$m_{7,0} = \prod_{i=0}^{2} n_{i,7,0}$ となり，各 $n_{i,7,0}$（$i = 0, 1, 2$）は次のように計算される．

$$n_{0,7,0} = \frac{1-p_0}{F_0-1} \quad \text{because} \left\lfloor \frac{7}{F_1F_2} \right\rfloor \not\equiv \left\lfloor \frac{0}{F_1F_2} \right\rfloor \pmod{F_0}$$

$$n_{1,7,0} = p_1 \quad \text{because} \left\lfloor \frac{7}{F_2} \right\rfloor \equiv \left\lfloor \frac{0}{F_2} \right\rfloor \pmod{F_1}$$

$$n_{2,7,0} = \frac{1-p_2}{F_2-1} \quad \text{because} \left\lfloor \frac{7}{1} \right\rfloor \not\equiv \left\lfloor \frac{0}{1} \right\rfloor \pmod{F_2}$$

したがって，

$$m_{7,0} = \frac{1-p_0}{F_0-1} \cdot p_1 \cdot \frac{1-p_2}{F_2-1} \tag{19}$$

となる．これは，あるユーザの TC セットが $\vec{V}_0$ である場合，式 (19) で表される $m_{7,0}$ の確率で，当該ユーザの DC セットが $\vec{V}_7$ となることを表している．式 (16) より，$\vec{V}_0 = \{0, 0, 0\}$，$\vec{V}_7 = \{1, 0, 1\}$ であり，属性 0 と属性 2 については TC と DC が異なっており，属性 1 については TC と DC が同一である．属性 i について TC と DC が同一である確率は p_i であり，TC と異なる各カテゴリが DC となる確率は $(1-p_i)/(F_i-1)$ であることからも，式 (19) が正しいことを確認できる．

5.3.3 確率行列 M の逆行列の導出

式 (18) で表される M の逆行列 M^{-1} を導出することができれば，式 (14) に基づいて，収集した $\vec{Y}$ と M^{-1} から，TC セットの推測値 $\vec{A}$ を計算することができる．

式 (18) より，この逆行列 M^{-1} は以下の式で表される．

$$M^{-1} = \frac{1}{\prod_{i=0}^{D-1}(p_iF_i-1)} \begin{pmatrix} m'_{0,0} & m'_{0,1} & \cdots & m'_{0,\Delta-1} \\ m'_{1,0} & m'_{1,1} & \cdots & m'_{1,\Delta-1} \\ \vdots & \vdots & \vdots & \vdots \\ m'_{\Delta-1,0} & m'_{\Delta-1,1} & \cdots & m'_{\Delta-1,\Delta-1} \end{pmatrix}$$

where $m'_{j,k} = \prod_{i=0}^{D-1} n'_{i,j,k}$,

$$n'_{i,j,k} = \begin{cases} F_i + p_i - 2 & \left(\left\lfloor \frac{j}{\prod_{l=i+1}^{D-1} F_l} \right\rfloor \equiv \left\lfloor \frac{k}{\prod_{l=i+1}^{D-1} F_l} \right\rfloor \pmod{F_i} \right) \\ p_i - 1 & \text{(otherwise)} \end{cases} \tag{20}$$

ここでも例として，属性数 3，各属性のカテゴリ数がそれぞれ $F_0 = 5$，$F_1 = 3$，$F_2 = 2$ である場合を考える．このときたとえば A_7（TC セットが $\vec{V}_7$ であるユーザ数の推測値）は，

$$A_7 = \sum_{k=0}^{\Delta-1} m'_{7,k} \cdot Y_k$$

と計算される．以下では例として，$m'_{7,0}$ の値を算出する．式 (20) に当てはめると，$m'_{7,0} = \prod_{i=0}^{2} n'_{i,7,0}$ となり，$m_{7,0}$ を計算したときと同様に計算を行うことで，

$$m'_{7,0} = (p_0 - 1) \cdot (F_1 + p_1 - 2) \cdot (p_2 - 1)$$

が得られる．

6. 解析

6.1 MSE の期待値

本節では MSE の期待値を導出する．MSE の期待値は 5.1 節において述べたように，満たすべきプライバシ保護レベル γ が与えられたとき，このプライバシ保護レベルを満たしつつ，MSE の期待値を最小化するための式 (13) を導出する際にも利用している．

属性数が 1 つだけある場合，MSE の期待値は以下のとおり求められる [13]．属性のカテゴリ数を F とし，各カテゴリを C_i で表す．ユーザ集合の中においてあるユーザに着目した場合に，当該ユーザの TC が C_i である確率を $P(X = C_i)$ とおく．同様にユーザ集合の中においてあるユーザに着目した際に，当該ユーザの DC が C_i である確率を $P(Y = C_i)$ とおく．また，$P(X = C_i)$ の最尤推定量を $P(A = C_i)$ とおく．このとき，MSE の期待値は以下の式で表される．

$$\begin{aligned} E[\sigma^2] &= \frac{1}{F} \sum_{i=0}^{F-1} E[P(A = C_i) - P(X = C_i)]^2 \\ &= \frac{1}{F} \sum_{i=0}^{F-1} \left[\sum_{j=0}^{F-1} \left(M_{i,j}^{-1}\right)^2 Var\left(\frac{N_j}{N}\right) \right. \\ &\quad \left. + \sum_{j,k,\ j\neq k} 2 \cdot M_{i,j}^{-1} M_{i,k}^{-1} Cov\left(\frac{N_j}{N}, \frac{N_k}{N}\right) \right], \end{aligned} \tag{21}$$

where

$$Var\left(\frac{N_j}{N}\right) = \frac{1}{N} \cdot P(Y = C_j)(1 - P(Y = C_j)),$$

$$Cov\left(\frac{N_j}{N}, \frac{N_k}{N}\right) = -\frac{1}{N} \cdot P(Y = C_j)P(Y = C_k)$$

ここで，N はユーザ総数を表し，$M_{i,j}^{-1}$ は確率行列 M の逆行列における i 行 j 列目を表している．

式 (21) をユーザ属性が複数ある状況に適用できるよう拡張する．ユーザ集合の中においてあるユーザに着目した際に，当該ユーザの TC セットが $\vec{V}_i$ である確率を

$P(X=\overrightarrow{V}_i)$ とおく．同様にユーザ集合の中においてあるユーザに着目した際に，当該ユーザの DC が $\overrightarrow{V}_i$ である確率を $P(Y=\overrightarrow{V}_i)$ とおく．また，$P(X=\overrightarrow{V}_i)$ の最尤推定量を $P(A=\overrightarrow{V}_i)$ とおく．

式 (21) は，$P(A=C_i)$ と $P(X=C_i)$ の二乗誤差の期待値を各 i $(i=0,\ldots,F-1)$ について算出し，その平均値を求めている．本論文においては，$P(A=\overrightarrow{V}_i)$ と $P(X=\overrightarrow{V}_i)$ の二乗誤差の期待値を各 i $(i=0,\ldots,\Delta-1)$ について算出し，その平均値を求める必要がある．したがって，式 (21) において $P(A=C_i)$ と $P(X=C_i)$ をそれぞれ $P(A=\overrightarrow{V}_i)$ と $P(X=\overrightarrow{V}_i)$ に置き換え，F で除算して平均値を取っている部分を Δ で除算して平均値を取るよう置き換えることで，属性が複数ある場合における，MSE の期待値を算出することができる．

結果，MSE の期待値 $E[\sigma^2]$ は以下の式で表すことができる．

$$\begin{aligned}E[\sigma^2] &= \frac{1}{\Delta}\sum_{i=0}^{\Delta-1} E[P(A=\overrightarrow{V}_i)-P(X=\overrightarrow{V}_i)]^2 \\ &= \frac{1}{\Delta}\sum_{i=0}^{\Delta-1}\left[\sum_{j=0}^{\Delta-1}\left(M_{i,j}^{-1}\right)^2 Var\left(\frac{N_j}{N}\right)\right. \\ &\quad \left.+\sum_{j,k,\ j\neq k}^{\Delta-1} 2\cdot M_{i,j}^{-1}M_{i,k}^{-1}Cov\left(\frac{N_j}{N},\frac{N_k}{N}\right)\right],\end{aligned} \tag{22}$$

where

$$Var\left(\frac{N_j}{N}\right)=\frac{1}{N}\cdot P(Y=\overrightarrow{V}_j)(1-P(Y=\overrightarrow{V}_j)),$$
$$Cov\left(\frac{N_j}{N},\frac{N_k}{N}\right)=-\frac{1}{N}\cdot P(Y=\overrightarrow{V}_j)P(Y=\overrightarrow{V}_k)$$

$P(Y=\overrightarrow{V}_j)$ は事前には不明であるため，すべての j について同確率で $P(Y=\overrightarrow{V}_j)$ が設定されると仮定すると，$P(Y=\overrightarrow{V}_j)=1/\Delta$ とおくことができる．7 章では，このような仮定の下で算出した MSE の期待値と，偏りの大きい実データを基にシミュレーションを行った結果の MSE とを比較し，おおむね一致していることを示す．

また，式 (20) より

$$\sum_{j=0}^{\Delta-1}(M_{i,j}^{-1})^2=\Phi(0)\Big/\prod_{k=0}^{D-1}(p_kF_k-1)^2 \tag{23}$$

where $\Phi(k)=((F_k-2+p_k)^2+(p_i-1)^2(F_k-1))\Phi(k+1)$,
$\Phi(D)=1$

と計算され，この数式を整理することで次式が得られる．

$$\sum_{j=0}^{\Delta-1}(M_{i,j}^{-1})^2=\prod_{k=0}^{D-1}\frac{3-2p_k+F_k(F_k+p_k^2-3)}{(p_kF_k-1)^2} \tag{24}$$

また，

$$\sum_{j,k,j\neq k}^{\Delta-1}M_{i,j}^{-1}M_{i,k}^{-1}=\sum_{j,k}^{\Delta-1}M_{i,j}^{-1}M_{i,k}^{-1}-\sum_{j=0}^{\Delta-1}(M_{i,j}^{-1})^2 \tag{25}$$

である．さらに，式 (20) より任意の i,j に対して

$$\sum_{j=0}^{\Delta-1}M_{i,j}^{-1}=1 \tag{26}$$

であるから，

$$\sum_{j,k}^{\Delta-1}M_{i,j}^{-1}M_{i,k}^{-1}=1 \tag{27}$$

である．したがって，

$$\sum_{j,k,j\neq k}^{\Delta-1}M_{i,j}^{-1}M_{i,k}^{-1}=1-\prod_{k=0}^{D-1}\frac{3-2p_k+F_k(F_k+p_k^2-3)}{(p_kF_k-1)^2} \tag{28}$$

と計算される．

結果，式 (22)，(24)，(28) より，MSE の期待値は次の式で表される．

$$E[\sigma^2]=\frac{1}{N\Delta^2}\left((1+\Delta)\prod_{k=0}^{D-1}\frac{3-2p_k+F_k(F_k+p_k^2-3)}{(p_kF_k-1)^2}-2\right) \tag{29}$$

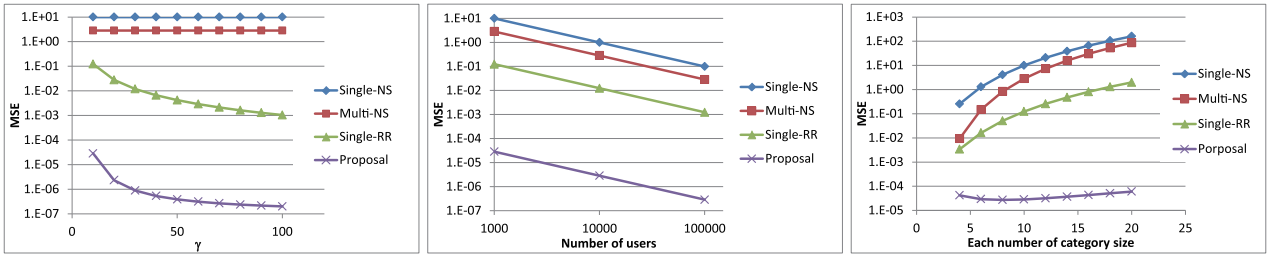
6.2 計算量

サーバ側でユーザの分布を推測する際の計算量は，確率行列の逆行列（式 (20)）を構成するために必要な計算量と，この逆行列を用いて $\overrightarrow{A}$ を求める（式 (14)）ために必要な計算量とに分けて考えることができる．式 (20) を構成するのに必要な計算量は $O(D\cdot\Delta^2)$ である．また，$\overrightarrow{A}$ を求めるのに必要な計算量は $O(\Delta^2)$ である．したがって，最終的な計算量は $O(D\cdot\Delta^2)$ と表すことができる．

7. 評価

式 (29) で定義した MSE の数学的評価および，実際に TC から DC を生成して MSE を計算するシミュレーション評価を行った．

比較対象として次の 4 手法を用意した．1 つは，複数の属性を単純に一次元化して取り扱う NS 手法であり，**Single-NS** と呼ぶ．2 つ目は，複数の属性を複数であることを考慮して推測する NS 手法 [10], [11] であり，ここでは **Multi-NS** と呼ぶ．3 つ目は，複数の属性を単純に一元化して取り扱う RR の手法であり，**Single-RR** と呼ぶ．また，3 章で紹介した反復ベイズ法 [26] も比較対象とし（図中では **Iterative** と呼ぶ），反復を終了させる条件は文献 [26] に従って決定した．反復ベイズ法については，MSE を数学的に導出することができないため，数学的評価は行わず，シミュレーション評価のみを行った．



(a) MSE with different γ (b) MSE with different N (c) MSE with different category sizes

図 2 MSE の比較（数学的解析）
Fig. 2 Comparison of MSE (mathematical analysis).

すべての評価実験においてプライバシ保護レベル γ を設定しているが（γ の値が小さいほどプライバシ保護レベルは高い），3.1.2 項で述べたとおり，γ を無限大に設定しない限り NS に基づく手法はいずれもこのプライバシ保護レベルを**満たさない**ことに注意されたい．したがってここでは，Single-NS および Multi-NS については，γ の値を考慮せず MSE を算出している．

文献 [7] では，サーバにおいてあるユーザがある性質 Q を持つ事前確率が 1%から 10%である状況を想定し，それぞれ RR を行うことで事後確率が 50%を超えないよう，プライバシ保護レベル γ を設定している．事前確率が 1%であり，事後確率が 50%である場合は $\gamma = 99$，事前確率が 10%であり，事後確率が 50%である場合は $\gamma = 9$ である．

本論文では，文献 [7] の設定範囲を包含するよう，γ の値を 5 から 100 まで変動させて実験を行った．上記とは別の例をあげると，あるユーザが性質 Q を持つ確率が，サーバの事前知識として 5%であったとする．この場合，$\gamma = 5$ であると，事後知識において，当該ユーザが性質 Q を持つ確率は最大約 21%に抑えられる．$\gamma = 19$ のときは 50%であり，$\gamma = 100$ のときは約 85%になる．したがってこの場合，$\gamma = 19$ 以下であれば，サーバは当該ユーザが性質 Q を持つかどうかはっきりしたことはいえず，$\gamma = 100$ のときは高い確率でユーザが性質 Q を持つことが分かるが，確信できる程ではないということになる．

また，シミュレーションは，OS が Windows 7 Professional 64 bit，CPU が Intel Xeon E5-2667 v2 (3.30 GHz)，Memory が 128 GB である機器を用いて行った．

7.1 数学的解析

Single-NS，Multi-NS，Single-RR，提案手法の各手法で，設定された γ に応じた MSE を計算した結果を**図 2** に示す．明記しない限り，ユーザ数 $N = 1{,}000$，属性数 4，各属性のカテゴリ数 10 に設定している．ここで，Single-NS および Single-RR は，4 つの属性を 1 次元化し，10,000 カテゴリの属性が 1 つだけあるとみなしている．

図 2 (a) は，プライバシ保護レベル γ の値を 10 から 100 に変動させて MSE を算出した結果を表している．前述のとおり，Single-NS および Multi-NS は γ に依存せず値は一定である．Single-RR および提案手法のいずれも γ の値が大きい（プライバシ保護レベルが低い）ほど，MSE の値が小さくなっていることが分かる．また，いずれの γ においても，提案手法が最も小さい MSE を実現していることが図から分かる．

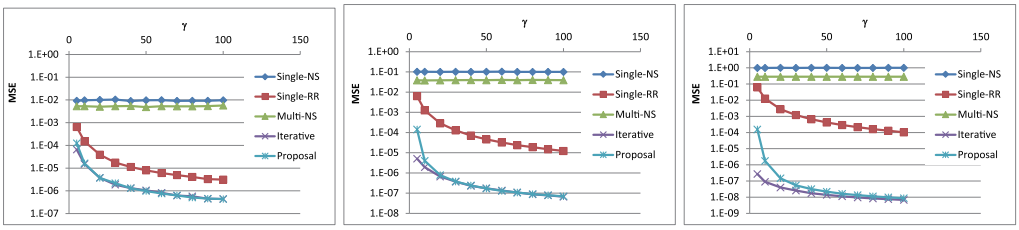
RR や NS に参加するユーザ総数 N を変動させた結果を図 2 (b) に表す．いずれの手法においても，ユーザ数に反比例して MSE が減少していることが分かる．これは式 (29) からも明らかである．

次に，各属性におけるカテゴリ数を変動させて MSE を算出した結果を図 2 (c) に表す．属性数は 4 で固定しているため，カテゴリ数の変動に応じて，全属性の全カテゴリの組合せ総数（Δ）も変動する．いずれの手法においても，カテゴリ数の増加にともなって MSE が増加している．しかし，提案手法における増加率が最も小さくなっていることが図から分かる．

Single-NS, Multi-NS, Single-RR, 提案手法の各手法で, 属性数と MSE との関係を表した結果を**図 3** に示す．図 3 (a) は，全属性の全カテゴリの組合せ総数（Δ）を 3^{12} に設定し，属性数と各属性のカテゴリ数を変化させて MSE を算出した結果を表している．ここで，各属性におけるカテゴリ数は同一になるように設定している．たとえば，属性数（Number of attributes）が 4 であるときは，属性数が 4 であり，各属性におけるカテゴリ数は共通して 27 ($27^4 = 3^{12}$) である．図 3 (b) についても同様である．Single-NS および Single-RR については，属性数にかかわらずそれらを一次元化して 1 つの属性とみなす手法であるため，属性数に依存せずに MSE は一定の値となっている．図より，全属性の全カテゴリの組合せ総数が一定である場合は，属性数が多いほど提案手法の効果が高くなっていることが分かる．

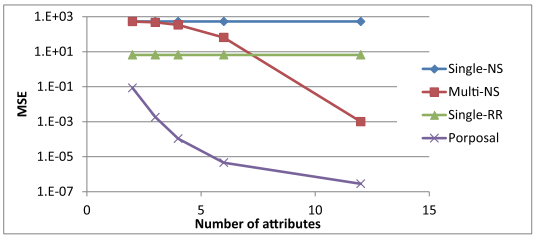
7.2 シミュレーション

まず，人工的に作成したデータでシミュレーション評価を行った．各人工データは，標準正規分布にしたがってデータを作成した．結果を**図 4** に示す．これらの図は収集対象の属性数を 2 から 4 まで変化させ，それぞれの属性に

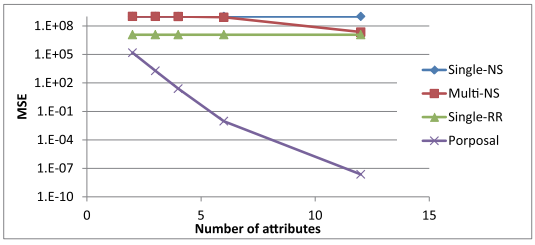


(a) Number of attributes is 2　(b) Number of attributes is 3　(c) Number of attributes is 4

図 4　人工のデータセットにおける MSE
Fig. 4　MSE of synthetic data.

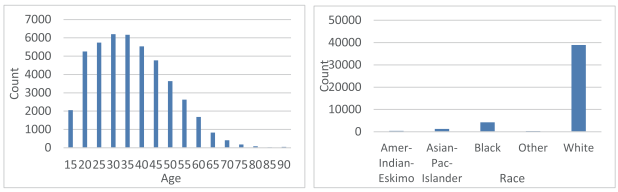


(a) Product of number of categories is 3^{12}



(b) Product of number of categories is 10^{12}

図 3　属性数と MSE の関係（数学的解析）
Fig. 3　Relationship between number of attributes and MSE.



(a) Age disbribution　(b) Race disbribution

図 5　Adult データセットにおける年齢と人種のデータ分布
Fig. 5　Data distributions of age and race in Adult data set.

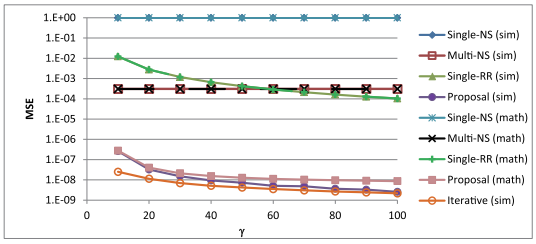


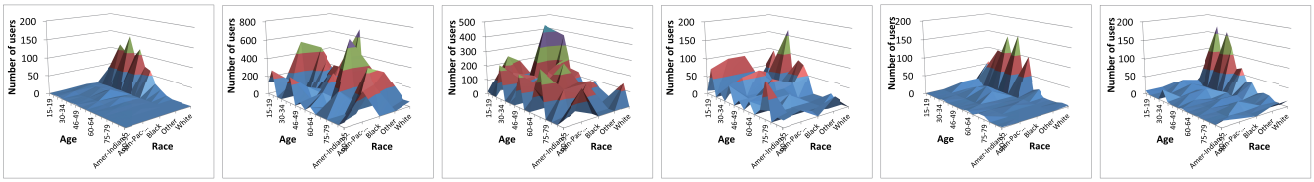
図 6　Nursey データセットにおける推測結果
Fig. 6　Estimated results of Nursey.

おけるカテゴリ数を 10 に設定している．また，一様分布，ポアソン分布についても評価を行ったが，ほとんど同じ結果になった．逆行列手法を用いた RR に関しては，データ分布に依存せずほぼ同じ MSE となることが既存研究においても示されている [11], [13].

図から分かるとおり，γ の値が大きい，つまりプライバシ保護レベルを下げるほど，MSE の値が小さくなっている．また，Single-NS，Single-RR，Multi-NS よりも，Iterative や提案手法の MSE が大きく下回っていることが分かる．γ の値が特に小さいときは Iterative のほうが提案手法よりも小さい MSE を実現できているが，多くの設定においては，ほぼ同じ MSE の値となっている．

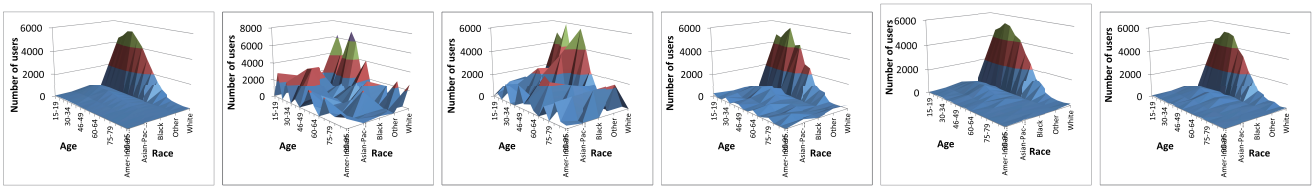
次に，RR や NS にかかわらず，匿名化に関する研究 [13], [15], [18] 等で広く使われている，UCI の Nursery データセットおよび Adult データセット [22] を用いて評価を行った．Nursery データセットは属性数 8，各属性におけるカテゴリ数はそれぞれ 3，5，4，4，3，2，3，3 であり，12,960 のデータが格納されている．本シミュレーションにおいては，偏りが大きいデータにおいて本提案手法の有効性を示すため，Adult データセットについては，偏りが大きい年齢，人種の 2 属性を利用した．年齢は 15–19，20–24 のように 5 歳ずつカテゴリ化して利用した．抽出されたデータには [15–19] から [90–94] の 16 カテゴリまで存在し，人種は White，Black，Amer-Indian-Eskimo，Asian-Pac-Islander，Other の 5 種類であった．また，Adult データセットには欠損値を含まないデータとして 45,222 人分のデータが格納されている．年齢および人種のデータ分布を**図 5** に示す．

まず Nursey データセットに対して評価を行った．プライバシ保護レベル γ の値を 10 から 100 まで変化させ，各 γ に対し，TC セットから DC セットを生成してサーバ側で TC セットを推測するという処理を 100 回繰り返した．推測された TC セットと実際の TC セットの MSE をそれぞれ計測し，その平均値を表したものが**図 6** である．ここでは，数式を基に算出した数学的解析結果も図に載せて



(a) Original data set　(b) Estimated result (Single-NS)　(c) Estimated result (Multi-NS)　(d) Estimated result (Single-RR)　(e) Estimated result (Iterative)　(f) Estimated result (Proposal)

図 7 Adult データセットにおけるオリジナルの分布と推測結果（1,000 ユーザ）
Fig. 7 Distribution of Adult data set and estimated results (Number of users is 1,000).



(a) Original data set　(b) Estimated result (Single-NS)　(c) Estimated result (Multi-NS)　(d) Estimated result (Single-RR)　(e) Estimated result (Iterative)　(f) Estimated result (Proposal)

図 8 Adult データセットにおけるオリジナルの分布と推測結果（45,222 ユーザ）
Fig. 8 Distribution of Adult data set and estimated results (Number of users is 45,222).

いる．（sim）と記載されているものがシミュレーション結果であり，（math）と記載されているものが数学的解析結果である．図から，反復ベイズ法および提案手法における MSE が小さい値を実現していることが分かる．また，数学的解析の結果とシミュレーション結果がよく一致していることも図から分かる．

次に Adult データセットに対して評価を行った．まず，Adult データセットからランダムに 1,000 人分のデータを抽出して実験を行った．抽出された 1,000 人のデータ分布および，各手法における推測結果を**図 7** に示す．ここでは，$\gamma = 10$ に設定した．また，RR および NS では，推測結果が負の値を取り得るという特徴がある．図 7 では，推測結果が負の値になった場合は 0 とみなして作図している．図より，NS に基づく Single-NS や Multi-NS は真のデータ分布をうまく推測できていないことが分かる．Single-RR についても，真のデータ分布との誤差は大きく，提案手法が最も精度良く真のデータ分布を推測できていることが分かる．このときの MSE はそれぞれ，0.063，0.024，0.0012，0.0001，0.00017 であり，反復ベイズ法の MSE が最も小さいが，提案手法もおおむね真のデータ分布を再構築できていることが図から分かる．

次に，Adult データセットの 45,222 件のすべてを利用して評価を行った．45,222 人分のデータ分布および，各手法における推測結果を**図 8** に示す．ここでも，$\gamma = 10$ に設定し，推測結果が負の値になった場合は 0 とみなしている．Single-NS や Multi-NS は，1,000 人分のデータに対して推測を行った結果よりは推測精度が向上していると考えられるが，まだ正しく推測できているとはいいにくい．一方，Single-RR と提案手法は精度良く真のデータ分布を推測できていると考えられるが，反復ベイズ法や提案手法のほうが明らかに真のデータ分布と近く，ほとんど一致していることが分かる．また，シミュレーションを 100 回繰り返したときの MSE の平均値はそれぞれ，1.7×10^{-3}，7.6×10^{-4}，2.8×10^{-5}，2.9×10^{-6}，4.1×10^{-6} であった．式 (29) を用いて算出した MSE の期待値は，反復ベイズ法を除くとそれぞれ 1.7×10^{-3}，7.6×10^{-4}，2.7×10^{-5}，4.2×10^{-6} であり，おおむね一致していることが分かる．

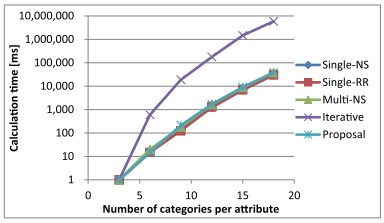


図 9 再構築にかかる計算時間
Fig. 9 Calculation time for reconstruction.

以上より，反復ベイズ法が最も小さい MSE を実現しているが，提案手法も同程度の MSE を実現しているといえる．

次に，推測にかかる時間を計測した．属性を 2 つのみ利用した Adult データセットの再構築には，いずれの手法も 1 秒未満で計算することができた．最後に，属性数を 4 に固定し，各属性におけるカテゴリ数を変動させて人工デー

タを生成し，それぞれにおいて再構築にかかる時間を計測した．結果を図 **9** に示す．

図から分かるように，反復ベイズ法はそれ以外の手法に比べ，最大 100 倍程度，サーバにおける再構築に時間がかかった．

8. 考察

本論文では議論を簡潔にするために，各ユーザおよび各属性について γ は同一の値が設定されると想定した．しかし，ユーザごとおよび属性ごとに γ を変更することも可能である．ユーザごとおよび属性ごとに γ を異なる値に設定した場合は，式 (13) を用いて各ユーザにおける各属性の同一カテゴリ選択確率 p_k を計算するとき，それぞれ設定されている γ の値を利用する．

サーバ側では，属性ごとに異なるプライバシ保護レベルが設定されていたとしても，全ユーザで共通であれば，そのことを考慮することなく DC セットから TC セットを推測することができる．ユーザによって異なるプライバシ保護レベルが設定されていた場合は，式 (14) を用いて，ユーザごとに TC セットの真のユーザ分布 A を推測し，それを全ユーザについて足し合わせることで最終的な推測値を算出することができる．具体的な検証は将来課題とする．

9. おわりに

ユーザ属性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再構築を行うというプライバシ保護モデルにおいて広く利用されている Randomized Response（RR）において，ユーザ属性が複数あるときに有効な手法を提案した．従来手法では，再構築を行った結果の誤差について定量的な評価ができず，再構築後のデータがどの程度信頼できるものなのか判断することができなかった．また，要求されるプライバシ保護レベルを満たしたうえで，誤差を最小化するためのパラメータの設定方法も明確でなかった．本論文では，ユーザ属性が複数存在する場合に有効な RR を提案し，再構築後のデータの期待値をあらかじめ計算できる手法を提案し，その期待値の基いて，誤差を最小化できるパラメータの設定方法を明確化した．数学的解析および実データを用いたシミュレーションによってその有効性を確認した．

謝辞 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いただき，ご指導いただ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いただいた研究グループの皆様に感謝いたします．

参考文献

[1] Agrawal, R., Srikant, R. and Thomas, D.: Privacy preserving OLAP, *Proc. ACM SIGMOD*, pp.251–262 (2005).
[2] Agrawal, S. and Haritsa, J.: A Framework for High-Accuracy Privacy-Preserving Mining, *Proc. IEEE ICDE*, pp.193–204 (2005).
[3] Aoki, S., Iwai, M. and Sezaki, K.: Privacy-Aware Community Sensing Using Randomized Response, *Proc. IEEE International Workshop on Security, Trust, and Privacy* (*STPSA*), pp.127–132 (2013).
[4] Chaytor, R. and Wang, K.: Small domain randomization: Same privacy, more utility, *Proc. VLDB Endow.*, Vol.3, No.1-2, pp.608–618 (2010).
[5] Corburn, J.: Confronting the challenges in reconnecting urban planning and public health, *American Journal of Public Health*, Vol.94, No.4, pp.541–546 (2004).
[6] Dwork, C.: Differential Privacy, *Automata, Languages and Programming*, Lecture Notes in Computer Science, Vol.4052, Springer, pp.1–12 (2006).
[7] Evfimievski, A., Gehrke, J. and Srikant, R.: Limiting privacy breaches in privacy preserving data mining, *Proc. ACM PODS*, pp.211–222 (2003).
[8] Forrest, S. and Groat, M.: Reconstructing Spatial Distributions from Anonymized Locations, *Proc. IEEE ICDEW*, pp.243–250 (2012).
[9] Froehlich, J., Larson, E., Gupta, S., Cohn, G., Reynolds, M. and Patel, S.: Disaggregated End-Use Energy Sensing for the Smart Grid, *IEEE Pervasive Computing*, Vol.10, No.1, pp.28–39 (2011).
[10] Groat, M.M., Edwards, B., Horey, J., He, W. and Forrest, S.: Enhancing privacy in participatory sensing applications with multidimensional data, *Proc. IEEE PerCom*, pp.144–152 (2012).
[11] Groat, M.M., Edwards, B., Horey, J., He, W. and Forrest, S.: Application and analysis of multidimensional negative surveys in participatory sensing applications, *Pervasive and Mobile Computing*, Vol.9, No.9, pp.372–391 (2013).
[12] Horey, J., Groat, M.M., Forrest, S. and Esponda, F.: Anonymous Data Collection in Sensor Networks, *Proc. MobiQuitous*, pp.1–8 (2007).
[13] Huang, Z. and Du, W.: OptRR: Optimizing Randomized Response Schemes for Privacy-Preserving Data Mining, *Proc. IEEE ICDE*, pp.705–714 (2008).
[14] Kasiviswanathan, S.P., Lee, H.K., Nissim, K., Raskhodnikova, S. and Smith, A.: What Can We Learn Privately ?, *SIAM Journal on Computing*, Vol.40, No.3, pp.793–826 (2013).
[15] Kenig, B. and Tassa, T.: A practical approximation algorithm for optimal k-anonymity, *Data Mining and Knowledge Discovery*, Vol.25, No.1, pp.134–168 (2011).
[16] La Polla, M., Martinelli, F. and Sgandurra, D.: A Survey on Security for Mobile Devices, *IEEE Communications Surveys & Tutorials*, Vol.15, No.1, pp.446–471 (2013).
[17] Li, Z., Li, M., Wang, J. and Cao, Z.: Ubiquitous data collection for mobile users in wireless sensor networks, *Proc. IEEE INFOCOM*, pp.2246–2254 (2011).
[18] Machanavajjhala, A., Kifer, D., Gehrke, J. and Venkitasubramaniam, M.: l-diversity: Privacy beyond k-anonymity, *ACM TKDD*, Vol.1, No.1, pp.3–es (2007).
[19] Samarati, P.: Protecting respondents identities in microdata release, *IEEE Trans. Knowledge and Data Engineering*, Vol.13, No.6, pp.1010–1027 (2001).
[20] Shabtai, A., Kanonov, U., Elovici, Y., Glezer, C. and Weiss, Y.: "Andromaly": A behavioral malware detection framework for android devices, *Journal of Intelligent Information Systems*, Vol.38, No.1, pp.161–190

(2011).
[21] Sharp, C., Schaffert, S., Woo, A., Sastry, N., Karlof, C., Sastry, S. and Culler, D.: Design and implementation of a sensor network system for vehicle tracking and autonomous interception, *Proc. EWSN*, pp.93–107, IEEE (2005).
[22] UCI Machine Learning Repository, available from ⟨http://archive.ics.uci.edu/ml/datasets⟩.
[23] Warner, S.L.: Randomized response: A survey technique for eliminating evasive answer bias, *American Statistical Association*, Vol.60, No.309, pp.63–69 (1965).
[24] Xiao, X., Tao, Y. and Chen, M.: Optimal Random Perturbation at Multiple Privacy Levels, *Proc. VLDB Endow.*, Vol.2, No.1, pp.814–825 (2009).
[25] Xie, H., Kulik, L. and Tanin, E.: Privacy-aware collection of aggregate spatial data, *Data & Knowledge Engineering*, Vol.70, No.6, pp.576–595 (2011).
[26] 五十嵐大，千田浩司，高橋克巳：多値属性に適用可能な効率的プライバシー保護クロス集計，コンピュータセキュリティシンポジウム（CCS），pp.497–502 (2008).
[27] 高見澤秀久，有次正義：プライバシーを保護するカウント演算の多値属性分類への適用について，日本データベース学会 Letters，Vol.6, No.1, pp.33–36 (2007).

清 雄一 （正会員）

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，プライバシ保護技術等の研究に従事．電子情報通信学会，IEEE Computer Society 各会員．

大須賀 昭彦 （正会員）

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．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．