数値属性に適用可能な, ランダム化により k-匿名性を保証するプライバシー保護クロス集計

五十嵐　大 †　　長谷川　聡 ‡　　納　竜也 ‡　　菊池　亮 †　　千田　浩司 †

† 日本電信電話株式会社 NTT セキュアプラットフォーム研究所
180-8585 東京都武蔵野市緑町 3-9-11
{ ikarashi.dai, kikuchi.ryo, chida.koji }@lab.ntt.co.jp
‡ 筑波大学情報科学類

あらまし プライバシー保護技術において, k-匿名化と呼ばれる代表的な匿名化手法と, これを確率的な手法に拡張した Pk-匿名化と呼ばれる手法がある. CSS2011 において筆者らは, 数値属性を含むデータを Pk-匿名化するための手法, ラプラスノイズ加算を提案した. しかし, ラプラスノイズ加算によりランダム化されたデータを分析する方法は示されていない.

本稿では, ラプラスノイズ加算を用いた場合, クロス集計の再構築という, 代表的なプライバシー保護データ分析処理を施すのが困難であることを指摘し, この問題を解決する**有界ノイズ加算**を提案する. さらに, その場合のクロス集計の再構築に必要な, 遷移確率行列の算出アルゴリズムを示す.

A Privacy Preserving Cross-tabulation which Guarantees k-Anonymity by Randomization for Numeric Attributes

Dai IKARASHI†　　Satoshi HASEGAWA‡　　Tatsuya OSAME‡　　Ryo KIKUCHI†
Koji CHIDA†

†NTT Secure Platform Laboratories, NTT Corporation
3-9-11, Midoricho, Musashino-shi, Tokyo 180-8585, Japan
{ ikarashi.dai, kikuchi.ryo, chida.koji }@lab.ntt.co.jp
‡College of Information Science, University of Tsukuba1

Abstract In the field of privacy protection techniques, there is a method called k-anonymization, which is the most popular anonymization method, and there is Pk-anonymization, which is a probabilistic extension of k-anonymization. In CSS 2011, we proposed a Pk-anonymization method for numeric attributes, Laplase noise addition. However, data analysis method is not proposed yet. In this paper, we point out that Laplase noise addition makes it difficult to conduct reconstruction of cross tabulation, which is a popular privacy-preserving data analysis method, and propose *bounded noise addition*, which solves this problem. Furthermore, we show an algorithm for the calculation of a transition probability matrix to reconstruct cross tabulations on bounded noise addition.

1　はじめに

k-匿名性 [1][2] は “データベース上である人のデータを k 個未満に絞り込むことができない”

という, プライバシー保護データ分析の分野で最もポピュラーなプライバシーの概念である. *k*-匿名性は直感的で分かりやすいため安心感を得られやすいと考えられる一方, データのランダム化を用いた匿名化手法には適用することができなかった. ランダム化は, 統計学及びデータベースの両分野で有望視されている重要な匿名化手法である. しかしランダム化には *k*-匿名性のような直感的なプライバシーの概念が存在せず, 安全性の評価が困難であることが課題であった.

Pk-匿名性 [3] はこの課題に対して *k*-匿名性をランダム化に適用したプライバシーの概念であり, “データベース上である人のデータを $1/k$ 以上の確率で当てることができない” ことを保証する. [3] では維持-置換撹乱 [4] と呼ばれるカテゴリ属性に対するランダム化が *Pk*-匿名性を満たすことが示され, そして [5] ではラプラスノイズ加算と呼ばれる数値属性に対するランダム化も *Pk*-匿名性を満たすことが示された. しかし, [5] では, ランダム化された値に対してデータ分析を行う, 再構築 [6] の方法については論じなかった.

これに対し本稿では, まずラプラスノイズ加算ではクロス集計の再構築 [7] が困難となることを指摘し, これを解決する有界ノイズ加算と呼ぶノイズ変換方法を提案する. また有界ノイズ加算によって Laplase 分布を変換したときの *Pk*-匿名性がラプラスノイズ加算と完全に等しいことを示す. さらに [5] で *Pk*-匿名性を満たさないことが証明された正規分布加算についても, 有界ノイズ加算と組み合わせることで *Pk*-匿名性を満たすことも示す.

そして, 最後に改めてクロス集計の再構築の方法を示す.

2 関連研究

ノイズの加算を用いたプライバシー保護手法に関しては, 統計の分野では 30 年ほど以前より研究されており, [9] などがある. またデータベースの分野では Agrawal らがプライバシー保護データマイニングとして, 各個人のデータにノイズを加算して, データマイニング結果のみを精度良く再構築するという, 再構築法を提案した [6]. 個人のデータにノイズを加えるのではないが, Dwork は対話型の統計データベースにおいて, Laplase 分布と呼ばれる分布をクエリに対する応答に加算してプライバシー保護を実現する手法及び, その場合のプライバシーである Differential Privacy を提案した [8].

ランダム化に対する *k*-匿名性の適用に関しては, [3] で我々が *Pk*-匿名性を提案したものの, 離散値を定義域とする属性を前提としたため, 数値属性に対するランダム化に対しての適用可能性は十分ではなかった. そこで [5] ではラプラスノイズ加算と呼ばれる, 数値属性のための *Pk*-匿名化方法が提案された.

3 準備

3.1 対象とするデータ

本稿で対象とするデータやモデルは, これまでの *Pk*-匿名性に関する文献 ([3] や [5]), また [7] などと同じである. 対象はテーブル型のデータベースであり, これを匿名化する. そして匿名化データベースを用いてデータ分析を行う.

3.2 ラプラスノイズ加算と *Pk*-匿名性

Laplase 分布は両側指数分布とも呼ばれ, 一様分布や正規分布と並び, 一般的な分布である. この Laplase 分布に従うノイズを加算することで, 定理 3.1 に記すように *Pk*-匿名性を実現することができる [5].

なお, [8] では統計値に Laplase 分布を加算することで, Differential Privacy と呼ばれるプライバシーを実現している. Differential Privacy はプライバシー保護の分野で非常に注目されており, ここで用いられている Laplase 分布が *Pk*-匿名性を実現するという事実は, 非常に興味深い事実である.

定理 3.1 $\mathcal{V}$ を任意の $\mathbb{R}^n$ 上の *Lebesgue* 可測集合, Δ を平均 0, 分散 σ^2 の *Laplase* 分布の加算とする. このとき Δ は

$$k = 1 + (|\mathcal{R}| - 1)\exp\left(-2\frac{\sup_{u,v\in\mathcal{V}}(\|u-v\|_1)}{\sigma}\right)$$

として Pk-匿名化である.

ただし, $\|\cdot\|$ は $L1$-ノルムであり, 各成分の絶対値の総和である.

4 再構築における Laplase 分布加算の課題

ラプラスノイズ加算後の属性値は, 任意の実数値をとる. すなわち, 属性の値域を大きく外れた値をとることがあり得るのである. このことは, 再構築の際に問題を引き起こす.

基礎的な再構築方法であるベイズ反復法では, 反復計算により真のデータの分布を推定する. この真のデータの推定分布の初期値として, ランダム化データの分布を代入する. しかしこれは, 真のデータとランダム化データの値域が等しいことを前提としている. つまり, ラプラスノイズ加算後のデータは値域を外れるため, この観点においてベイズ反復法の適用が自明でなくなる.

この課題を解決するには, 以下の 2 つのアプローチが考えられるであろう.

1. 初期値の設定方法を改良する
2. 値域を外れないように加算する分布を改良する

前者の場合, 値域を外れたデータを値域内に振り分ける必要がある. しかしこのプロセスはベイズ反復法の目的である真のデータの推定そのものであり, そのベイズ反復法の開始前に必要である初期値のために高精度の推定を行うことは難しそうである. よって, 本稿では後者のアプローチをとる.

5 提案：有界ノイズ加算

前節で述べたように, 本稿ではノイズを加算しても値域を外れないようにと, ノイズの加算を改良する. なお, 現時点ではノイズの分布を Laplase 分布に限定しない.

5.1 手法 1：反復試行アプローチ

最も正直な以下の手法をまず提案する. (**有界ノイズ加算 (反復試行アプローチ)**)

1. 分布に従うノイズ (**基礎ノイズ**と呼ぶことにする) を生成し, 属性値に加算する.
2. 値域を外れた場合, ノイズの生成からやり直す.

この方法でランダム化された属性値が値域に収まることは明らかである. この方法で生成されたノイズを**有界ノイズ**と呼ぶことにする.

5.2 提案手法の Pk-匿名性

提案手法の匿名性はどのようになっているだろうか？ここでは以下の定理を示す.

補題 5.1 区間 $[a,b] \subseteq \mathbb{R}$*(ただし* $a < b$*)* を値域とする任意の数値属性と, ノイズの分布を表す任意の確率密度関数 $f : \mathbb{R} \to \mathbb{R}$ があるものとする. ただし f は任意の $\varepsilon > 0$ で, $\int_{[0,\varepsilon]} f(x)dx$ も $\int_{[-\varepsilon,0]} f(x)dx$ も正数であるとする.

このとき, 属性の任意の値 $v \in [a,b]$ があるとする. そして, このとき $\alpha_v \in \mathbb{R}$ を以下とおく.

$$\alpha_v = \int_{[a-v,b-v]} f(x)dx$$

すると有界ノイズの確率密度関数 $f'_v : [a,b] \to \mathbb{R}$ は, 基礎ノイズの確率密度関数 f を用いて以下のように表される.

$$f'_v(v') = f(v')/\alpha_v \tag{1}$$

補題 5.1 はすなわち, ランダム化後の値によらず, ランダム化前の属性値 v だけから決まる係数 $1/\alpha_v$ があって, 有界ノイズは基礎ノイズの確率密度が $1/\alpha_v$ 倍されるだけで, ノイズの分布の (区間内での) 形は変化しないということを述べている. α_v は, 基礎ノイズが, 属性値に加算して区間内に収まるような値となる確率である. α_v に関しては, v ごとに値が異なることに注意.

証明 提案手法で起きる乱数の生成のやり直しを表すため, f に従う確率変数列を $\{X_i\}_{i=0,1,\ldots}$ とおく.

このときに, X'_i を, 有界ノイズの生成でやり直しを i 回までに制限した場合のノイズとする. このとき X'_i の確率密度関数が $i \to \infty$ で式 *(1)* に収束することを示せばよい. すると, 帰納的に以下が成り立つ.

$$\begin{cases} \Pr[\![X'_0 = v']\!] = \Pr[\![X_0 = v']\!] \\ \Pr[\![X'_i = v']\!] = \Pr[\![X'_{i-1} = v' \vee (X'_{i-1} \notin [a,b] \wedge X_i = v')]\!] \end{cases}$$

ここで次の 4 つに注意する.

- $\Pr[\![X_0 = v']\!] = f(v)$
- 事象 $X'_{i-1} = v'$ と事象 $X'_{i-1} \notin [a,b] \wedge X_i = v'$ は排反
- ノイズを生成し直すことから $X'_{i-1} \notin [a,b]$ と $X_i = v'$ が独立
- f は任意の $\varepsilon > 0$ で, $\int_{[0,\varepsilon]} f(x)dx$ も $\int_{[-\varepsilon,0]} f(x)dx$ も正数より, $|1-\alpha_v| < 1$

すると以下が成り立つ.

$$\begin{aligned} &\Pr[\![X'_i = v']\!] \\ &= \Pr[\![X'_{i-1} = v']\!] + \Pr[X'_{i-1} \notin [a,b]]\Pr[\![X_i = v']\!] \\ &= \Pr[\![X'_{i-1} = v']\!] + (1 - \textstyle\int_{[a-v,b-v]} f(x)dx)^{i-1}\, f(v') \\ &= \Pr[\![X'_{i-1} = v']\!] + (1-\alpha_v)^{i-1}\, f(v') \\ &= f(v') + (1-\alpha_v)\, f(v') + (1-\alpha_v)^2\, f(v') + \dots \\ &\to (1/\alpha_v) f(v') \ (= f(v')/\alpha_v) \ \text{ as } \ i \to \infty \\ &\qquad\qquad (|1-\alpha_v| < 1 \text{ の等比数列より}) \end{aligned}$$

□

補題 5.1 は, 多次元の基礎ノイズから多次元の有界ノイズを生成する場合も, 区間を多次元長方形としても成り立つ. 証明も同じである.

それでは, 有界ノイズ加算の *Pk*-匿名性を評価する.

定理 5.1 区間 $[a,b] \subseteq \mathbb{R}$*(ただし* $a < b$*)* を値域とする任意の数値属性と, ノイズの分布を表す任意の確率密度関数 $f : \mathbb{R} \to \mathbb{R}$ があるものとする. ただし, f は次の *3* 条件を満たす.

1. f は原点に関して線対称 *(*すなわち $f(x) = -f(-x)$*)*
2. f は $x < 0$ で広義単調増加で $x \geq 0$ で広義単調減少
3. f は有限個の点を除いて連続

すると $\mathcal{V} = \mathcal{V}' = [a,b]$ かつ, また任意の $v, v' \in [a,b]$ に対して $A_{v,v'}$ を値 v が値 v' にランダム化される確率密度 ***(*遷移確率密度*)*** として, 以下が成り立つ.

$$\operatorname*{ess\,inf}_{\substack{u,v\in\mathcal{V} \\ u',v'\in\mathcal{V}'}} \frac{A_{v,u'}A_{u,v'}}{A_{u,u'}A_{v,v'}} = \left(\frac{f(b-a)}{f(0)}\right)^2 \qquad (2)$$

定理 5.1 は, 以下の *Pk*-匿名性の公式 [5] に代入して用いる公式である.

$$k = 1 + (|\mathcal{R}| - 1) \prod_{a:\text{属性}} \operatorname*{ess\,inf}_{\substack{u,v\in\mathcal{V}_a \\ u',v'\in\mathcal{V}'_a}} \frac{(A_a)_{v,u'}(A_a)_{u,v'}}{(A_a)_{u,u'}(A_a)_{v,v'}} \qquad (3)$$

すなわち, k の値を, 基礎ノイズの確率密度関数から容易に算出できることを示している. 本稿では, 属性ごとに決まる $\operatorname*{ess\,inf}_{\substack{u,v\in\mathcal{V} \\ u',v'\in\mathcal{V}'}} \frac{A_{v,u'}A_{u,v'}}{A_{u,u'}A_{v,v'}}$ を, 属性の**匿名率**と呼ぶことにする. 匿名率は, 高い方が匿名性が高くなる.

なお, f の条件は, 3 つあるので強い条件に見えるが, Laplase 分布, 正規分布, 連続一様分布など, 代表的な分布で満たされる条件である. 特に条件 2 は, ランダム化前の値に近い値ほどランダム化によりその値になる確率が高いということを述べており, 直感的にそもそも数値属性のランダム化においては必要な要件である. また条件 3 は, これを満たさない確率密度関数を想像する方が難しいであろう.

証明 補題 *5.1* と同様, 有界ノイズの確率密度関数を $f'_v : \mathbb{R} \to \mathbb{R}$ とする. ただし $v \in [a,b]$ は任意の属性値である. まず, ノイズは属性値に対して加算されるので, 任意の $v, v' \in [a,b]$ に対して, v が v' にランダム化される確率密度 $A_{v,v'}$ は $A_{v,v'} = f'_v(v'-v)$ である. そして, f の条件は補題 *5.1* の仮定を満たすから, ある α_v が存在して $f'_v(v'-v) = f(v'-v)/\alpha_v$ である.

はじめに, f が有限個の点を除いて連続であることから, ess inf は min に直せる.

$$\operatorname*{ess\,inf}_{\substack{u,v\in\mathcal{V}\\u',v'\in\mathcal{V}'}}\frac{A_{v,u'}A_{u,v'}}{A_{u,u'}A_{v,v'}}=\min_{\substack{u,v\in\mathcal{V}\\u',v'\in\mathcal{V}'}}\frac{A_{v,u'}A_{u,v'}}{A_{u,u'}A_{v,v'}}$$

これを踏まえて, $v=v'=a, u=u'=b$ を代入してみる.

$$\frac{A_{v,u'}A_{u,v'}}{A_{u,u'}A_{v,v'}}=\frac{f'_v(u'-v)f'_u(v'-u)}{f'_v(v'-v)f'_u(u'-u)}=\left(\frac{f(b-a)}{f(0)}\right)^2$$

よって最小値の性質から以下が成り立つ.

$$\min_{\substack{u,v\in\mathcal{V}\\u',v'\in\mathcal{V}'}}\frac{A_{v,u'}A_{u,v'}}{A_{u,u'}A_{v,v'}}\le\left(\frac{f(b-a)}{f(0)}\right)^2$$

次に逆向きの不等号が成り立つことを示す.

任意の $v\in[a,b]$ を考える. f の *2* 条件と補題 *5.1* より, v' を動かしたときの $A_{v,v'}$ の最大値は以下である.

$$\max_{v'\in[a,b]}A_{v,v'}=\max_{v'\in[a,b]}f'_v(v'-v)=f'_v(0)=f(0)/\alpha_v$$

また最小値は, また c を a,b のうち v からより遠い方とすれば, 以下となる.

$$\begin{aligned}\max_{v'\in[a,b]}A_{v,v'}&=\min_{v'\in[a,b]}f'_v(v'-v)=f'_v(c-v)\\&=f(c-v)/\alpha_v\end{aligned}$$

よって $\min_{u',v'\in\mathcal{V}'}\frac{A_{v,u'}}{A_{v,v'}}=\frac{f(c-v)}{f(0)}$ であり, f の条件から v も動かしたとき,

$$\min_{\substack{v\in\mathcal{V}\\u',v'\in\mathcal{V}'}}\frac{A_{v,u'}}{A_{v,v'}}=\frac{f(b-a)}{f(0)}$$

となる. min の性質より

$$\begin{aligned}\min_{\substack{u,v\in\mathcal{V}\\u',v'\in\mathcal{V}'}}\frac{A_{v,u'}A_{u,v'}}{A_{u,u'}A_{v,v'}}&\ge\min_{\substack{v\in\mathcal{V}\\u',v'\in\mathcal{V}'}}\frac{A_{v,u'}}{A_{v,v'}}\min_{\substack{u\in\mathcal{V}\\u',v'\in\mathcal{V}'}}\frac{A_{u,v'}}{A_{u,u'}}\\&=\left(\min_{\substack{v\in\mathcal{V}\\u',v'\in\mathcal{V}'}}\frac{A_{v,u'}}{A_{v,v'}}\right)^2=\left(\frac{f(b-a)}{f(0)}\right)^2\end{aligned}$$

□

系 5.1 値域を区間 $[a,b]$ とする数値属性に, 基礎ノイズを分散 $2\sigma^2$ の *Laplase* 分布とする有界ノイズ加算でランダム化したとき, この属性の匿名率は以下となる.

$$\operatorname*{ess\,inf}_{\substack{u,v\in\mathcal{V}\\u',v'\in\mathcal{V}'}}\frac{A_{v,u'}A_{u,v'}}{A_{u,u'}A_{v,v'}}=\exp\left(-2\frac{|b-a|}{\sigma}\right)\qquad(4)$$

系 5.1 で導かれた匿名率は, 面白い性質を持っている. 実は, 基礎ノイズを加算した場合の匿名率と同じ, すなわち [5] の Laplapse ノイズ加算における匿名率と等しいのである. [5] によれば Laplase ノイズ加算における k の値は以下である.

$$k=1+(|\mathcal{R}|-1)\exp\left(-2\frac{\sup_{u,v\in\mathcal{V}}(\|u-v\|_1)}{\sigma}\right)$$

この式の exp 以下に $\mathcal{V}=\mathcal{V}'=[a,b]$ を代入すれば系 5.1 の結果と一致する値であることが確かめられる.

また, [5] で *Pk*-匿名性を満たさないことが示された正規分布の加算なども, 有界ノイズとすれば *Pk*-匿名性を満たす.

系 5.2 値域を区間 $[a,b]$ とする数値属性に, 基礎ノイズを分散 σ^2 の正規分布とする有界ノイズ加算でランダム化したとき, この属性の匿名率は以下となる.

$$\operatorname*{ess\,inf}_{\substack{u,v\in\mathcal{V}\\u',v'\in\mathcal{V}'}}\frac{A_{v,u'}A_{u,v'}}{A_{u,u'}A_{v,v'}}=\exp\left(-\frac{(b-a)^2}{\sigma^2}\right)\qquad(5)$$

一般に, 有界ノイズは基礎ノイズと同等かそれ以上の匿名率をもつ.

定理 5.2 補題 *5.1* と同じ状況を仮定する. このとき, 有界ノイズ加算による匿名率は, 基礎ノイズ加算による匿名率以上である.

証明 基礎ノイズ加算と有界ノイズ加算の遷移確率密度をそれぞれ $A_{v,v'}, B_{v,v'}$ と書く. すると以下のように示される.

$$\begin{aligned}\operatorname*{ess\,inf}_{\substack{u,v\in[a,b]\\u',v'\in\mathbb{R}]}}\frac{A_{v,u'}A_{u,v'}}{A_{u,u'}A_{v,v'}}&=\operatorname*{ess\,inf}_{\substack{u,v\in[a,b]\\u',v'\in\mathbb{R}]}}\frac{f(u'-v)f(v'-u)}{f(u'-u)f(v'-v)}\\&\le\operatorname*{ess\,inf}_{\substack{u,v\in[a,b]\\u',v'\in[a,b]}}\frac{f(u'-v)f(v'-u)}{f(u'-u)f(v'-v)}\\&\quad(u',v'\text{ が限定されたので, ess inf の性質から})\\&=\operatorname*{ess\,inf}_{\substack{u,v\in[a,b]\\u',v'\in[a,b]]}}\frac{B_{v,u'}B_{u,v'}}{B_{u,u'}B_{v,v'}}\end{aligned}$$

□

5.3 処理効率の改善：非反復アプローチ

ここまでで有界ノイズ加算の正直な生成法と，匿名性を見た．しかし，正直な方法は反復回数の期待値が $1/\alpha_v$ であり，α_v が大きい場合に反復回数が大きくなってしまう．このような場合のために，反復なしで同じノイズを生成するアルゴリズムを考える．

一般に，確率密度関数が既知である 1 次元のノイズを発生させる場合，累積分布関数の逆関数を求めることで，一様乱数から所望のノイズを生成することができる．

有界ノイズ加算 (非反復アプローチ)
累積分布関数を $F_v : [a-v, b-v] \to [0,1]$, $F_v(v') = \int_{[a-v,v']} f'_v(x)dx = \int_{[a-v,v']} f_v/\alpha_v dx$ とし，累積分布関数の逆関数を $F_v^{-1} : [0,1] \to [a-v, b-v]$, $[0,1]$ 上一様乱数を X とすれば，$F_v^{-1}(X)$ が f を基礎ノイズとする有界ノイズである．**有界ノイズ加算は，$v + F_v^{-1}(X)$ を行えばよい**．

Laplase 分布を基礎ノイズとして具体例を示す．まず α_v は以下である．

$$\alpha_v = \int_{[a-v,b-v]} f(x)dx = \int_{[a,b]} \frac{f(x)}{2\sigma} \exp\frac{|x-v|}{\sigma} dx$$
$$= 1 - \frac{1}{2}\exp\left(\frac{-(v-a)}{\sigma}\right) - \frac{1}{2}\exp\left(\frac{-(b-v)}{\sigma}\right)$$

次に F_v を計算する．

$$F_v(v') = \begin{cases} \frac{1}{2\alpha_v}\left(\exp\frac{v'-v}{\sigma} - \exp\frac{a-v}{\sigma}\right) & (v' < v) \\ \frac{1}{2\alpha_v}\left(2 - \exp\frac{a-v}{\sigma} - \exp\frac{v-v'}{\sigma}\right) & (v' \geq v) \end{cases}$$

よって逆関数 F_v^{-1} は以下となる．

$$F_v^{-1}(x) = \begin{cases} \sigma\log\left(2\alpha_v x + \exp\frac{a-v}{\sigma}\right) + v \\ \qquad \left(x < \frac{1}{2\alpha_v}\left(1 - \exp\frac{a-v}{\sigma}\right) \text{のとき}\right) \\ -\sigma\log\left(2 - 2\alpha_v x - \exp\frac{a-v}{\sigma}\right) + v \\ \qquad (\text{otherwise}) \end{cases}$$

6 クロス集計の再構築

ここで，前節で提案した有界ノイズ加算によってランダム化された属性を用いてクロス集計を再構築 [7] することを考える．そのためには，遷移確率行列が必要である．以下では，遷移確率行列を導いていく．なお，以降は基礎ノイズを Laplase 分布とする．

遷移確率行列はカテゴリ属性において，値 v が値 v' に変化する確率を行列としたものである．連続属性の場合，直接連続値で再構築を行うことはできない．遷移確率密度関数を，区間 $[a,b]$ を適当な数 n 個の部分区間 $I_0, ..., I_{n-1}$ に分割し，値 v が区間 $I = I_i$ に含まれるときに，区間 $I' = I_j$ に含まれる値に変化する確率，という形に量子化する必要がある．

この確率 $\mathcal{A}_{II'}$ は以下のように表される．ただし，Δ はランダム化アルゴリズムである．

$$\mathcal{A}_{II'} = \frac{1}{d-c}\int_I \Pr[\Delta(v) \in I' | V = v]\Pr[V = v]dv \tag{6}$$

ただし，$I = [c,d]$ とする．

このうち，まずはじめに $\Pr[\Delta(v) \in I' | V = v]$ を算出することから始める．

この式は，

$$\int_{I'} A_{v,v'} dv' \tag{7}$$

と等価である．よって，まずこの v' に関する積分を先に解く．

v' に関する積分　$I' = [c', d']$ とする．ただし $a \leq c' < d' \leq b$ である．

$$\int_{I'} A_{v,v'} dv' = \int_{I'} \frac{1}{\alpha_v}\frac{1}{2\sigma}\exp\frac{-|v'-v|}{\sigma}dv' \tag{8}$$

(1)$c' < v, d' \geq v$ のとき

$$\int_{c'}^{v} \frac{1}{\alpha_v}\frac{1}{2\sigma}\exp\frac{v'-v}{\sigma}dv' \tag{9}$$
$$+ \int_{v}^{d'} \frac{1}{\alpha_v}\frac{1}{2\sigma}\exp\frac{-v'+v}{\sigma}dv'$$
$$= \frac{1}{2\alpha_v}\left(2 - \exp\frac{c'-v}{\sigma} - \exp\frac{-d'+v}{\sigma}\right) \tag{10}$$

(2)$c' < v, d' < v$ のとき

$$\frac{1}{2\alpha_v}\left(\exp\frac{d'-v}{\sigma} - \exp\frac{c'-v}{\sigma}\right) \tag{11}$$

(3)$c' \geq v, d' \geq v$ のとき

$$\frac{1}{2\alpha_v}\left(-\exp\frac{-d'+v}{\sigma} + \exp\frac{-c'+v}{\sigma}\right) \tag{12}$$

v に関する積分　式 (6) に戻ってみよう．いま $\Pr[\Delta(v) \in I' | V = v]$ は導かれた．一方, $\Pr[V = v]$ というのは, ランダム化前の値の分布である．これについては本稿では $[a, b]$ を I の分割 (すなわち n 分割) と同じか, もしくはより細かく mn 個に区間を分割して $J_0, ..., J_{mn-1}$ とし, 各 J_i 内では $\Pr[V = v]$ は定数と見なす方法をとる．すると

$$\mathcal{A}_{II'} = \frac{1}{m}\sum_{J \subseteq I}\mathcal{A}_{JI'} \tag{13}$$

となる．なお, 各 J_i 内での $\Pr[V = v]$ を推定する最も簡単な方法は, 一様分布として推定する方法である．

以下, $\mathcal{A}_{JI'}$ を計算する．

$$\mathcal{A}_{JI'} = \frac{1}{h-g}\int_J \Pr[\Delta(v) \in I' | V = v]\Pr[V = v]dv \tag{14}$$

である．ただし $J = [g, h]$ とする．

$\int_J \Pr[\Delta(v) \in I' | V = v]\Pr[V = v]dv$ の部分に関しては, $I = [c, d]$, $I' = [c', d']$ の関係ごとに, 3 パターンの場合の積分を解く必要がある．

- $c < c', d < d'$ (式 (12) を用いる)
- $c = c', d = d'$ (式 (10) を用いる)
- $c > c', d > d'$ (式 (11) を用いる)

(1)$c > c', d > d'$ の場合

$$\int_g^h \frac{\left(\exp\frac{d'}{\sigma} - \exp\frac{c'}{\sigma}\right)\exp-\frac{v}{\sigma}}{2(1 - \frac{1}{2}\exp\frac{a-v}{\sigma} - \frac{1}{2}\exp\frac{-b+v}{\sigma})}Rdv \tag{15}$$

ここで R は 定数 ($\Pr[V = v]$ のこと) である．

まず変数変換を行う．$\exp\frac{-v}{\sigma} = z$ として変数変換し, 式を整理すると以下のようになる．

$$\int R\sigma\frac{\left(\exp\frac{d'}{\sigma} - \exp\frac{c'}{\sigma}\right)z}{\left(\exp\frac{a}{\sigma}z^2 - 2z + \exp\frac{-b}{\sigma}\right)}dz \tag{16}$$

次に分母の $\left(\exp\frac{a}{\sigma}z^2 - 2z + \exp\frac{-b}{\sigma}\right)$ を因数分解する．判別式より解を 2 つ持つことがわかるから, その解を β, β' とおくことにする．

$$\beta = \frac{1 + \sqrt{1 - \exp\frac{a-b}{\sigma}}}{\exp\frac{a}{\sigma}} \tag{17}$$

$$\beta' = \frac{1 - \sqrt{1 - \exp\frac{a-b}{\sigma}}}{\exp\frac{a}{\sigma}} \tag{18}$$

となる．これを用いて, 変形すると,

$$\int R\sigma\frac{\left(\exp\frac{d'}{\sigma} - \exp\frac{c'}{\sigma}\right)z}{\exp\frac{a}{\sigma}(z-\beta)(a-\beta')}dz \tag{19}$$

となる．これを部分分数分解すると,

$$\int R\sigma\frac{\left(\exp\frac{d'}{\sigma} - \exp\frac{c'}{\sigma}\right)}{\exp\frac{a}{\sigma}}\left(\frac{\beta}{(z-\beta)(\beta-\beta')} - \frac{\beta'}{(z-\beta')(\beta-\beta')}\right)dz \tag{20}$$

となり, 結局

$$R\sigma\frac{\left(\exp\frac{d'}{\sigma} - \exp\frac{c'}{\sigma}\right)}{\exp\frac{a}{\sigma}}\frac{1}{\beta-\beta'}\left[\beta\log|z-\beta| - \beta'\log|z-\beta'|\right]_{\exp\frac{-g}{\sigma}}^{\exp\frac{-h}{\sigma}} \tag{21}$$

となる．

(2)$c = c', d = d'$ の場合

$$\int_g^h \frac{\left(2 - \exp\frac{c'-v}{\sigma} - \exp\frac{-d'+v}{\sigma}\right)}{2(1 - \frac{1}{2}\exp\frac{a-v}{\sigma} - \frac{1}{2}\exp\frac{-b+v}{\sigma})}Rdv \tag{22}$$

これを解く．同様にして解いていくと, 以下となる．

$$R\frac{\sigma}{\exp\frac{a}{\sigma}}\left[2\left(\frac{\log|z-\beta|}{\beta-\beta'} - \frac{\log|z-\beta'|}{\beta-\beta'}\right)\right. \tag{23}$$

$$- \exp\frac{c'}{\sigma}\left(\frac{\beta}{\beta-\beta'}\log|z-\beta| - \frac{\beta'}{\beta-\beta'}\log|z-\beta'|\right) \tag{24}$$

$$\left. - \exp\frac{-d'}{\sigma}\left(\frac{\log|z|}{\beta\beta'} + \frac{\log|z-\beta|}{\beta(\beta-\beta')} - \frac{\log|z-\beta'|}{\beta'(\beta-\beta')}\right)\right]_{\exp\frac{-g}{\sigma}}^{\exp\frac{-h}{\sigma}} \tag{25}$$

(3)$c < c', d < d'$ の場合
同様にして解く．

$$-R\sigma\frac{\left(\exp\frac{-d'}{\sigma} - \exp\frac{-c'}{\sigma}\right)}{(\gamma-\gamma')\exp\frac{-b}{\sigma}}\left[\gamma\log|z-\gamma| - \gamma'\log|z-\gamma'|\right]_{\exp\frac{g}{\sigma}}^{\exp\frac{h}{\sigma}} \tag{26}$$

ここで γ, γ' は,

$$\gamma = \frac{1 + \sqrt{1 - \exp\frac{a-b}{\sigma}}}{\exp\frac{-b}{\sigma}} \tag{27}$$

$$\gamma' = \frac{1 - \sqrt{1 - \exp\frac{a-b}{\sigma}}}{\exp\frac{-b}{\sigma}} \tag{28}$$

である.

最終的に遷移確率行列を求めるアルゴリズムとしては, 各 I, I' について, 式 (23), (25), (26), (27), (28) を用いて式 (16) を計算し, さらにそれを用いて式 (15) を計算すればよい. クロス集計の再構築全体としては, あとは [7] で紹介されている方法をそのまま用いることができる.

7 おわりに

本稿では [3] で提案されたランダム化を用いる匿名化のための匿名性である, *Pk*-匿名性について, [5] のラプラスノイズ加算の, クロス集計の再構築における課題を指摘し, これを解決するノイズ変換方法である, 有界ノイズ加算を提案した. そして有界ノイズ加算の *Pk*-匿名性について以下を示した.

- 有界ノイズ加算の *Pk*-匿名性に関する公式.
- Laplase 分布を基礎ノイズとする有界ノイズ加算が, 基礎ノイズと比較して匿名性を全く低下させないこと.
- [5] で *Pk*-匿名性を満たさないことが証明された正規分布も, 有界ノイズ加算を用いれば *Pk*-匿名性を満たすこと.
- 有界ノイズ加算は一般的に基礎ノイズ加算より匿名性を低下させない.

さらに, Laplase 分布を基礎ノイズとする有界ノイズ加算がされたデータにおける, クロス集計の再構築の方法, 特に, そのために必要な遷移確率行列の算出法を示した.

今後の課題として, 提案したアルゴリズムの評価実験が挙げられる.

また, ここで証明はしないが, 一般的に有界ノイズは基礎ノイズよりもエントロピーが小さくなる. 一方, 本稿で示したように有界ノイズ加算は基礎ノイズ加算より匿名性を低下させない. このことから, 基礎ノイズと比較して匿名性は低下させずに, ランダム化前データとの相互情報量などの有用性指標は高めることができる場合があるのではないかと期待しており, 実験的評価, 理論的評価ともに行っていきたい.

参考文献

[1] P. Samarati and L. Sweeney. Generalizing data to provide anonymity when disclosing information (abstract). *Proc. of the 17th ACM-SIGMOD-SIGACT-SIGART Symposium on the Principles of Database Systems, p. 188, Seattle, WA, 1998.*

[2] L. Sweeney. *k*-anonymity: a model for protecting privacy. *Int'l Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol.10, Issue 5, pp.557-570, 2002.

[3] 五十嵐 大, 千田 浩司, 高橋 克巳. *k*-匿名性の確率的指標への拡張とその応用例. CSS2009, 2009.

[4] R. Agrawal, R. Srikant and D. Thomas. Privacy Preserving OLAP. *SIGMOD Conference ACM*, pp. 251-262, 2005.

[5] 五十嵐 大, 千田 浩司, 高橋 克巳. 数値属性における, *k*-匿名性を満たすランダム化手法. CSS2011, 2011.

[6] R. Agrawal and R. Srikant. Privacy-preserving data mining. *Proc. of the 2000 ACM SIGMOD Intl. Conf. on Management of Data*, 2000.

[7] 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. CSS2008, 2008.

[8] C. Dwork. Differential Privacy. *ICALP (2) 2006*, 2006.

[9] J. J. Kim. A method for limiting discloure in microdata based on random noise and transformation. In *Proceedings of the Section on Survey Research Methods*, 1986.

[10] 五十嵐 大, 千田 浩司, 高橋 克巳. *Pℓ*-多様性:属性推定に対する再構築法のプライバシーの定量化. CSS2010, 2010.