# ランダム化データベース上の $k$-匿名性の一般的算出法

五十嵐　大 †　　千田　浩司 †　　高橋　克巳 †

† 日本電信電話株式会社 NTT 情報流通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, chida.koji, takahashi.katsumi }@lab.ntt.co.jp

**あらまし** 近年, プライバシーを保護しつつデータを活用する研究が活発に行われており, その中でも $k$-匿名性およびこれを実現する k-匿名法は特に注目され, 実用に向けた議論がされている. 一方統計やデータベースの分野で, 確定的アルゴリズムを用いる k-匿名化ではなく, データにランダム成分を加えるランダム化による匿名化手法も有効とされ研究されているが, 安全性の評価は不十分であった. これに対し CSS2009 で筆者らはランダム化されたデータベース上の $k$-匿名性である $Pk$-匿名性を定義し, 維持-置換撹乱と呼ばれるランダム化で $Pk$-匿名性を満たす方法を示した. 本稿ではこれをさらに拡張し, 一般のランダム化で $Pk$-匿名性を満たすための理論を示す.

# Generalized Evaluation of $k$-Anonymity on Randomization

Dai IKARASHI†　　Koji CHIDA†　　Katsumi TAKAHASHI†

†NTT Information Sharing Platform Laboratories, NTT Corporation
3-9-11, Midoricho, Musashino-shi, Tokyo 180-8585, Japan

{ ikarashi.dai, chida.koji, takahashi.katsumi }@lab.ntt.co.jp

**Abstract** Recently, studies on the utilizing private data are widely conducted. Among them, $k$-anonymity is especially paid attention to, and practical use of it is being discussed. On the other hand, in the community of database and statistics, not only $k$-anonymity, which adopt non-probabilistic algorithms, but also anonymization methods using data randomization are expected and have been studied. However, it was the problem that there had been no enough security evaluation. In response, we proposed $Pk$-anonymity, which is $k$-anonymity on a randomized database, in CSS2009, and showed a method for satisfying $Pk$-anonymity using a randomization called retention-replacement perturbation. In this paper, we extend it, and show a theory for satisfying $Pk$-anonymity in general randomizations.

## 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] では以下が示された.

- “$k$-匿名 $\Rightarrow$ $Pk$-匿名” が成り立つこと
- ランダム化手法のひとつ維持-置換撹乱 [4][5] での $Pk$-匿名性

本稿では, [3] では離散かつ有限の空間上で議論されていた $Pk$-匿名性を, 連続かつ無限空間に拡張することで, $Pk$-匿名性を一般の数値データや,

[6] のような数値を扱うランダム化にも柔軟に適用可能なより一般的なプライバシーとする．特に本稿では重要な定理として以下を述べる．

- 非確率的匿名化アルゴリズムでは “$k$-匿名 $\Leftrightarrow$ $Pk$-匿名が成り立つこと”
- 任意の属性とランダム化に適用可能な, $Pk$-匿名性の $k$ の値の公式

前者は $Pk$-匿名性が $k$-匿名性の数学的な拡張であることの主張である. 後者は, 匿名化の強度である $k$ の値の算出が自明でなかった $Pk$-匿名性において, この値を算出するための定理であり, ランダム化を用いた匿名化に $k$-匿名性を適用する上で最大の定理となるであろう.

なお本稿では紙面の都合から, 省略する事項が多いことをご容赦願いたい.

## 2 関連研究

本稿のモチベーションは [3] と同じであるため, そちらを参照されたい. 差分としてはランダム化に対する安全性指標として, [7] で情報量理論に基づいた指標が提案されている．これは属性推定に対する安全性であり, 本稿で扱う匿名性の議論とは異なる.

## 3 連続空間における $Pk$-匿名性

[3] では攻撃者によるテーブル型データベースの事前推定 $f_T$, 公開テーブル $\tau'$, 秘密テーブル上のレコード $r$, 公開テーブル上のレコード $r'$ とするとき攻撃者の識別推定を $\mathcal{E}(f_T, \tau', r, r')$, すなわち $\Pr[\Pi(r) = r' | T' = \tau']$) とした. そして任意の上記 4 パラメータに対して $\mathcal{E}(f_T, \tau', r, r') \leq 1/k$ が成り立つことを $Pk$-匿名性とした.

正確な定義は本稿では紙面の都合上再掲しない. 記法は [3] と同じくするため, 詳細な定義は [3] を参照されたい.

連続空間も含む一般の確率空間の場合, すなわちテーブルが連続属性も含む場合, 秘密テーブルの集合 $\mathcal{T}$, 公開テーブルの集合 $\mathcal{T}'$ が連続確率空間となる．このとき, $Pk$-匿名性の定義として以下がより正確である.

任意の $r \in \mathcal{R}, r' \in \mathcal{R}'$, $f_T \in \mathbb{P}^{\mathcal{T}}$ に対して,

$$\operatorname*{ess\,sup}_{\tau' \in \mathcal{T}'} \mathcal{E}(f_T, \tau', r, r') \leq 1/k \tag{1}$$

ただし $\mathbb{P}^{\mathcal{T}}$ は $\mathcal{T}$ 上確率密度関数の集合である. また ess sup, ess inf は本質的上限 (下限) と呼ばれる概念で, 測度空間 (または確率空間) を $\mathcal{X}$, 可測関数 (または確率密度関数) を $f$ とすれば以下のように定義される．積分値を左右しない部分を除いた部分の上下限と思えばよいであろう.

$$\operatorname*{ess\,sup}_{x \in \mathcal{X}} f(x) = \inf \{a \in \mathbb{R} \mid \mu(\{f > a\}) = 0\}$$
$$\operatorname*{ess\,inf}_{x \in \mathcal{X}} f(x) = \sup \{a \in \mathbb{R} \mid \mu(\{f < a\}) = 0\}$$

$\mu(\{f > a\})$ は, 関数 $f$ の関数値が $a$ を超えるような定義域の部分の測度 ($\mathcal{X}$ が 1 次元なら長さ, 2 次元なら面積など) である.

## 4 $Pk$-匿名性が $k$-匿名性の拡張であること

[3] では “$k$ 匿名なら $Pk$-匿名” を示した．これは連続属性を含んでも成り立つほか, 実は逆向きの命題である, “$Pk$ 匿名なら $k$-匿名” も, 匿名化アルゴリズムが非確率的である場合には成り立つ.

**定理 4.1** 任意の正整数 $k$ と, 任意の匿名化システム $\mathcal{P}(= (\mathcal{R}, \mathcal{V}, \mathcal{R}', \mathcal{V}', \Pi, \Delta))$ 及び公開テーブル $\tau'$ に対して, もし $\Delta$ が非確率的アルゴリズムであるならば, 以下の関係が成り立つ.

$$\tau' \text{ が } k\text{-匿名} \Leftrightarrow (\mathcal{P}, \tau') \text{ が } Pk\text{-匿名}$$

すなわち, 非確率的な匿名化を考える上では $k$-匿名性と $Pk$-匿名性は同値であり, $Pk$-匿名性が数学的に $k$-匿名性の拡張になっていることが分かる．残念ながら紙面の都合上本稿では証明を記すことができないため, また別の機会に示すこととする.

## 5 $Pk$-匿名性の評価定理

[3] では維持-置換撹乱と呼ばれるランダム化について, ある $k$ の値で $Pk$-匿名性を満たすことを証明した. 本稿ではより一般, すなわち任意の確率空間を定義域とする属性に対する, レコー

ド毎に独立にランダム化を行う任意のランダム化に対する $Pk$-匿名性の解析を行う. ランダム化に対する上記制限は, それほど厳しい制限ではない. 実際 [6], [4] などの, 再構築法と呼ばれるランダム化は全てこの条件を満たす.

$Pk$-匿名性はデータベースの確率密度関数全体, ランダム化後のデータベース全体, 全レコードの置換全体という, それぞれをとっても非常に高次元な空間の直積確率空間の中での確率の上限で定義されている. 特にレコード数の増加によりこの次元は無限大へと発散する. 本定理はこの超高次元空間で定義された $Pk$-匿名性が, 実は任意のランダム化において, 4 次元という定数次元空間の関数値の下限に帰着させることができるという, 驚くべき定理である. 本定理はランダム化における $k$-匿名性を議論する上で非常に有用な定理であり, 我々は [8] で直ちに用いる.

**定理 5.1** 匿名化システム $(\Delta, \mathcal{R}, \mathcal{R}', \mathcal{V}, \mathcal{V}')$ をとる. $\Delta$ をレコード毎に独立にランダム化を行う匿名化とし, 値 $v \in \mathcal{V}$ のレコードが値 $v' \in \mathcal{V}'$ にランダム化される確率を $A_{v,v'}$ と書く. このとき $\Delta$ は $k$ を以下の値として $Pk$-匿名化である.

$$k = 1 + (|\mathcal{R}| - 1)\operatorname*{ess\,inf}_{\substack{u,v \in \mathcal{V} \\ u',v' \in \mathcal{V}'}} \frac{A_{v,u'}A_{u,v'}}{A_{u,u'}A_{v,v'}}$$

**証明** 本証明では *3* つの補題を用いている. 紙面の都合上補題の証明は, 本証明において最も重要な, 補題 *5.2* のみ掲載する.

$k$ の値は攻撃者の識別推定の逆数であるからこれを評価すると, 任意の $f_T \in \mathbb{P}^{\mathcal{T}}, \tau' \in \mathcal{T}', r \in \mathcal{R}, r' \in \mathcal{R}'$ に対して, まず以下のように変形できる.

$$\begin{aligned}
\frac{1}{\mathcal{E}(f_T, \tau', r, r')} &= \frac{1}{\Pr[\Pi(r) = r' | T' = \tau']} \\
&= \frac{\Pr[T' = \tau']}{\Pr[\Pi(r) = r' \wedge T' = \tau']} \\
&= \frac{\Pr[\Delta(T) = \tau' \circ \Pi]}{\Pr[\Pi(r) = r' \wedge \Delta(T) = \tau' \circ \Pi]} \\
&\qquad (\Delta \text{ と } \Pi \text{ の定義より}) \\
&= \frac{\sum_{\pi:\text{全単射}} \Pr[\Delta(T) = \tau' \circ \pi]}{\sum_{\pi:\text{全単射}} \Pr[\pi(r) = r' \wedge \Delta(T) = \tau' \circ \pi]} \\
&\qquad (\Pi \text{ が } \Delta \text{ と独立より}) \\
&= \frac{\sum_{\pi:\text{全単射}} \Pr[\Delta(T) = \tau' \circ \pi]}{\sum_{\pi(r)=r'} \Pr[\Delta(T) = \tau' \circ \pi]}
\end{aligned}$$

ここで $\Delta$ がレコード毎に独立なランダム化であることから, 任意の $\pi : \mathcal{R} \to \mathcal{R}'$(全単射) で以下が成り立つ.

$$\Pr[\Delta(T) = \tau' \circ \pi] = \prod_{s \in \mathcal{R}} A_{T(s),\tau'(\pi(s))}$$

よって

$$\frac{1}{\mathcal{E}(f_T, \tau', r, r')} = \frac{\sum_{\pi:\text{全単射}} \prod_{s \in \mathcal{R}} A_{T(s),\tau'(\pi(s))}}{\sum_{\pi(r)=r'} \prod_{s \in \mathcal{R}} A_{T(s),\tau'(\pi(s))}}$$

が言える. ここで以下の補題が成立する.

**補題 *5.1*** $\mathcal{X}$ 上可測関数 $f$ に対して, 以下のような関数を $\overline{f}$ と書く.

$$\overline{f}(x) = \max(\operatorname*{ess\,inf}_{y \in \mathcal{X}} f(y), f(x))$$

すると任意の $\mathcal{X}$ 上可測集合 $X$ で

$$\int_{x \in X} f(x)dx = \int_{x \in X} \overline{f}(x)dx$$

$$\operatorname*{ess\,inf}_{x \in \mathcal{X}} f(x) = \operatorname*{ess\,inf}_{x \in \mathcal{X}} \overline{f}(x) = \inf_{x \in \mathcal{X}} \overline{f}(x)$$

$$\overline{f+g} = \overline{f} + \overline{g}, \quad \overline{fg} = \overline{f}\,\overline{g}, \quad \overline{\left(\frac{f}{g}\right)} = \frac{\overline{f}}{\overline{g}}$$

が成り立つ.

**証明** 紙面の都合上省略する.

すなわち, ess inf を inf に置き換えて四則演算を行うことが可能となる. すると以下が言える.

$$\begin{aligned}
&\inf_{f_T \in \mathbb{P}^{\mathcal{T}}} \operatorname*{ess\,inf}_{\tau' \in \mathcal{T}'} \frac{1}{\mathcal{E}(f_T, \tau', r, r')} \\
&= \inf_{\substack{f_T \in \mathbb{P}^{\mathcal{T}} \\ \tau' \in \mathcal{T}'}} \frac{\sum_{\pi:\text{全単射}} \prod_{s \in \mathcal{R}} \overline{A}_{T(s),\tau'(\pi(s))}}{\sum_{\pi(r)=r'} \prod_{s \in \mathcal{R}} \overline{A}_{T(s),\tau'(\pi(s))}} \\
&\qquad (\text{補題 } 5.2 \text{ より}) \\
&= \inf_{\substack{f_T \in \mathbb{P}^{\mathcal{T}} \\ \tau' \in \mathcal{T}'}} \frac{\int_{\tau \in \mathcal{T}} \sum_{\pi:\text{全単射}} \prod_{s \in \mathcal{R}} \overline{A}_{T(s),\tau'(\pi(s))} d\tau}{\int_{\tau \in \mathcal{T}} \sum_{\pi(r)=r'} \prod_{s \in \mathcal{R}} \overline{A}_{T(s),\tau'(\pi(s))} d\tau} \qquad (2)
\end{aligned}$$

次に以下の補題を示す. 本補題は上の式の積分を除去する, 本証明でも最も大きく次元を削減する部分である.

**補題 *5.2*** $\mathcal{X}$ を測度空間とし, $\mu$ をその測度とする. $\mathbb{R}^{\mathcal{X}+}$ を, 積分値が正で有限値である $\mathcal{X}$ 上非負可測関数の集合とする. このとき任意の $a, b \in \mathbb{R}^{\mathcal{X}+}$ に対して,

$$\sup_{f \in \mathbb{R}^{\mathcal{X}+}} \frac{\int_{x \in \mathcal{X}} bf dx}{\int_{x \in \mathcal{X}} af dx} = \operatorname*{ess\,sup}_{x \in \mathcal{X}|_{b/a}} \frac{b(x)}{a(x)}$$

である. 下限についても同様の性質が成り立つ.

ただし $\mathcal{X}|_{b/a}$ は, $\mathcal{X}$ から $b/a$ が 0/0 となる点を除いた集合とする.

**証明** sup の方のみ示す.

$$(1)\ \sup_{f\in\mathbb{R}^{\mathcal{X}+}}\frac{\displaystyle\int_{x\in\mathcal{X}} bfdx}{\displaystyle\int_{x\in\mathcal{X}} afdx} \le \operatorname*{ess\,sup}_{x\in\mathcal{X}|_{b/a}}\frac{b(x)}{a(x)}$$

および

$$(2)\ \sup_{f\in\mathbb{R}^{\mathcal{X}+}}\frac{\displaystyle\int_{x\in\mathcal{X}} bfdx}{\displaystyle\int_{x\in\mathcal{X}} afdx} \ge \operatorname*{ess\,sup}_{x\in\mathcal{X}|_{b/a}}\frac{b(x)}{a(x)}$$

を示す.

[(1) の証明] $\operatorname*{ess\,sup}_{x\in\mathcal{X}|_{b/a}}\frac{b(x)}{a(x)} = z$ とおき, 可測関数列 $\{b'_i \in \mathbb{R}^{\mathcal{X}}\}_{i\in\mathbb{N}\setminus\{0\}}$ を,

$$b'_j(x) = \begin{cases} b(x) & z < \dfrac{b(x)}{a(x)} \le z + \dfrac{1}{j} \text{のとき} \\ za(x) & \text{そうでないとき} \end{cases}$$

と定義する. この $b_j$ が, 任意の $f \in \mathbb{R}^{\mathcal{X}+}$ に対して

$$(i)\lim_{j\to\infty}\int_{x\in\mathcal{X}} b'_j fdx = z\int_{x\in\mathcal{X}} afdx$$

および

$$(ii)\int_{x\in\mathcal{X}} bfdx \le \lim_{j\to\infty}\int_{x\in\mathcal{X}} b'_j fdx$$

を満たすことを示す. そうすれば

$$\int_{x\in\mathcal{X}} bfdx \le \lim_{j\to\infty}\int_{x\in\mathcal{X}} b'_j fdx = z\int_{x\in\mathcal{X}} afdx$$

となり,

$$\frac{\displaystyle\int_{x\in\mathcal{X}} bfdx}{\displaystyle\int_{x\in\mathcal{X}} afdx} \le \frac{z\displaystyle\int_{x\in\mathcal{X}} afdx}{\displaystyle\int_{x\in\mathcal{X}} afdx} = z = \operatorname*{ess\,sup}_{x\in\mathcal{X}|_{b/a}}\frac{b(x)}{a(x)}$$

が分かることになる. $f$ が任意だから,

$$\sup_{f\in\mathbb{R}^{\mathcal{X}+}}\frac{\displaystyle\int_{x\in\mathcal{X}} bfdx}{\displaystyle\int_{x\in\mathcal{X}} afdx} \le \operatorname*{ess\,sup}_{x\in\mathcal{X}|_{b/a}}\frac{b(x)}{a(x)}$$

となる.

[($i$) の証明] 区間 $(z, z+\frac{1}{j}]$ は $j\to\infty$ で空集合に収束するから, $b'_j$ の定義より $\lim_{j\to\infty} b'_j = za$, さらに $\lim_{j\to\infty} b'_j f = zaf$ である. ($a = b = 0$ となる点でも成り立つ. ) また $b'_j$ は単調減少列になっているから任意の $j$ で $b'_j \le b'_1$ である. よって *Lebesgue* の収束定理[1] を適用でき,

$$\lim_{j\to\infty}\int_{x\in\mathcal{X}} b'_j fdx = \int_{x\in\mathcal{X}}\lim_{j\to\infty} b'_j fdx = z\int_{x\in\mathcal{X}} afdx$$

となる.

[($ii$) の証明] ess sup の定義より,

$$\mu\left(\left\{\frac{b}{a} > z + \frac{1}{j}\right\}\right) = 0,$$

すなわち $\left\{\frac{b}{a} > z + \frac{1}{j}\right\}$ は零集合であり, 積分に影響しない. そして $\mathcal{X}$ のその他の領域では $b'_j$ の定義より $b'_j > b$ が成り立つから, 積分の単調性より

$$\int_{x\in\mathcal{X}} bfdx \le \int_{x\in\mathcal{X}} b'_j fdx$$

である. 任意の $j$ で上記が成り立つから,

$$\int_{x\in\mathcal{X}} bfdx \le \lim_{j\to\infty}\int_{x\in\mathcal{X}} b'_j fdx$$

である.

以上より

$$\sup_{f\in\mathbb{R}^{\mathcal{X}+}}\frac{\displaystyle\int_{x\in\mathcal{X}} bfdx}{\displaystyle\int_{x\in\mathcal{X}} afdx} \le \operatorname*{ess\,sup}_{x\in\mathcal{X}|_{b/a}}\frac{b(x)}{a(x)}$$

である.

[(2) の証明] 次に

$$\sup_{f\in\mathbb{R}^{\mathcal{X}+}}\frac{\displaystyle\int_{x\in\mathcal{X}} bfdx}{\displaystyle\int_{x\in\mathcal{X}} afdx} \ge z = \operatorname*{ess\,sup}_{x\in\mathcal{X}|_{b/a}}\frac{b(x)}{a(x)}$$

を示す. 可測関数列 $\{f_j\}_{j\in\mathbb{N}\setminus\{0\}}$ を,

$$f_j(x) = \begin{cases} 1 & x \in X_j \text{のとき} \\ 0 & \text{そうでないとき} \end{cases}$$

と定義する. ただし

$$X_j = \left\{x \in \mathcal{X} \,\middle|\, \frac{b(x)}{a(x)} > z\left(1 - \frac{1}{j}\right)\right\}$$

とする. すると ess sup の定義より, 任意の $j$ で $X_j$ は $b/a$ の零集合ではなく, 関数 $f_j$ の積分は正である. よって $\int_{x\in\mathcal{X}} af_j dx \ne 0$ であり, $\dfrac{\displaystyle\int_{x\in\mathcal{X}} bf_j dx}{\displaystyle\int_{x\in\mathcal{X}} af_j dx}$ が定義できて

$$\frac{\displaystyle\int_{x\in\mathcal{X}} bf_j dx}{\displaystyle\int_{x\in\mathcal{X}} af_j dx} = \frac{\mu(X_j)\displaystyle\int_{x\in X_j} bf_j dx}{\mu(X_j)\displaystyle\int_{x\in X_j} af_j dx}$$

$$\ge \frac{\displaystyle\int_{x\in X_j} za\left(1 - \frac{1}{j}\right) f_j dx}{\displaystyle\int_{x\in X_j} af_j dx} = z\left(1 - \frac{1}{j}\right)$$

となる. よって $\lim_{j\to\infty} z\left(1 - \frac{1}{j}\right) = z$ より,

[1]可測関数列 $\{f_j\}_{j\in\mathbb{N}}$ が, (1) 任意の $j \in \mathbb{N}$ で $|f_j| < f$ を満たす可測関数 $f$ が存在 (2) 極限が可測関数, の 2 条件を満たすなら, 以下が成り立つ.

$$\lim_{j\to\infty}\int_{x\in\mathcal{X}} f(x)dx = \int_{x\in\mathcal{X}}\lim_{j\to\infty} f(x)dx$$

$$\lim_{j\to\infty}\frac{\int_{x\in\mathcal{X}} bf_j dx}{\int_{x\in\mathcal{X}} af_j dx} \geq z$$

であり,

$$\sup_{f\in\mathbb{R}^{\mathcal{X}+}}\frac{\int_{x\in\mathcal{X}} bfxdx}{\int_{x\in\mathcal{X}} afdx} \geq z = \operatorname*{ess\,sup}_{x\in\mathcal{X}|_{b/a}}\frac{b(x)}{a(x)}$$

であることが分かる.

(1), (2) で結局両方向の不等式が示されたから,

$$\sup_{f\in\mathbb{R}^{\mathcal{X}+}}\frac{\int_{x\in\mathcal{X}} bfdx}{\int_{x\in\mathcal{X}} afdx} = z = \operatorname*{ess\,sup}_{x\in\mathcal{X}|_{b/a}}\frac{b(x)}{a(x)}$$

である.

□ (補題 5.2)

以後, $\mathcal{X}_{b/a}$ のように書かなくても sup, ess sup, inf, ess inf は定義される値の中での上下限とする. 補題 *5.2* より

$$\text{式 (2)} = \inf_{\tau'\in\mathcal{T}'}\operatorname*{ess\,inf}_{\tau\in\mathcal{T}}\frac{\sum_{\pi:\text{全単射}}\prod_{s\in\mathcal{R}}\overline{A}_{\tau(s),\tau'(\pi(s))}}{\sum_{\pi(r)=r'}\prod_{s\in\mathcal{R}}\overline{A}_{\tau(s),\tau'(\pi(s))}}$$

$$= \inf_{\substack{\tau\in\mathcal{T}\\ \tau'\in\mathcal{T}'}}\frac{\sum_{\pi:\text{全単射}}\prod_{s\in\mathcal{R}}\overline{A}_{\tau(s),\tau'(\pi(s))}}{\sum_{\pi(r)=r'}\prod_{s\in\mathcal{R}}\overline{A}_{\tau(s),\tau'(\pi(s))}}$$

$$= \operatorname*{ess\,inf}_{\substack{\tau\in\mathcal{T}\\ \tau'\in\mathcal{T}'}}\left(1+\frac{\sum_{\pi(r)\neq r'}\prod_{s\in\mathcal{R}}A_{\tau(s),\tau'(\pi(s))}}{\sum_{\pi(r)=r'}\prod_{s\in\mathcal{R}}A_{\tau(s),\tau'(\pi(s))}}\right)$$

この ess inf の内側を $K(\tau,\tau')$ とおく.

ここで, $\pi(r)=r'$ である置換 $\pi:\mathcal{R}\to\mathcal{R}'$ と $t\neq r$ であるレコードの組 $(\pi,t)$ からは, 以下のようにして $\hat{\pi}_t(r)\neq r'$ なる置換 $\hat{\pi}_t:\mathcal{R}\to\mathcal{R}'$ への全単射が作れる.

$$\hat{\pi}_t(s)=\begin{cases}\pi(t) & (s=r\text{ のとき})\\ \pi(r) & (s=t\text{ のとき})\\ \pi(s) & (\text{それ以外の場合})\end{cases}$$

すなわち $\pi(r)$ と $\pi(s)$ を入れ替える. この入れ替わりに注意すると, 上式の分子は以下のように変形される.

$$\sum_{\pi(r)\neq r'}\prod_{s\in\mathcal{R}}A_{\tau(s),\tau'(\pi(s))} = \sum_{\substack{\pi(r)=r'\\ t\neq r}}\prod_{s\in\mathcal{R}}A_{\tau(s),\tau'(\hat{\pi}_t(s))}$$

$$= \sum_{\substack{\pi(r)=r'\\ t\neq r}}A_{\tau(r),\tau'(\pi(t))}A_{\tau(t),\tau'(\pi(r))}\prod_{\substack{s\neq r\\ s\neq t}}A_{\tau(s),\tau'(\pi(s))}$$

$$= \sum_{\pi(r)=r'}\sum_{t\neq r}\frac{A_{\tau(t),\tau'(\pi(r))}A_{\tau(r),\tau'(\pi(t))}}{A_{\tau(t),\tau'(\pi(t))}}\prod_{s\neq r}A_{\tau(s),\tau'(\pi(s))}$$

ここで $\pi(r)=r'$ なる任意の置換 $\pi:\mathcal{R}\to\mathcal{R}'$ に対して $g_\pi, h_\pi:\mathcal{T}\times\mathcal{T}'\to\mathbb{R}$ を以下のように定義する.

$$g_\pi(\tau,\tau') = A_{\tau(r),\tau'(\pi(r))}\prod_{s\neq r}A_{\tau(s),\tau'(\pi(s))}$$

$$h_\pi(\tau,\tau') = \sum_{t\neq r}\frac{A_{\tau(t),\tau'(\pi(r))}A_{\tau(r),\tau'(\pi(t))}}{A_{\tau(t),\tau'(\pi(t))}}\prod_{s\neq r}A_{\tau(s),\tau'(\pi(s))}$$

すると上記の変形より以下が成り立つ.

$$K(\tau,\tau') = 1+\frac{\sum_{\pi(r)=r'}h_\pi(\tau,\tau')}{\sum_{\pi(r)=r'}g_\pi(\tau,\tau')}$$

ここで以下の補題を用いる.

**補題 *5.3*** $\mathcal{I}$ を可算の添数集合, $(\mathcal{X},\mu)$ を測度空間, $g_i, h_i:\mathcal{X}\to\mathbb{R}$ を, 有限の積分値をもつ非負可測関数とする. もしある可測集合列 $\{X_j\subseteq\mathcal{X}\}_{j\in\mathbb{N}}$ と $z\in\mathbb{R}$ が任意の $i\in\mathcal{I}$ に対して,

$$\lim_{j\to\infty}\frac{\int_{x\in X_j}h_i dx}{\int_{x\in X_j}g_i dx} = \operatorname*{ess\,inf}_{x\in\mathcal{X}}\frac{h_i(x)}{g_i(x)} = z$$

を満たすならば, 以下が成り立つ.

$$\operatorname*{ess\,inf}_{x\in\mathcal{X}}\frac{\sum_{i\in\mathcal{I}}h_i}{\sum_{i\in\mathcal{I}}g_i} = \lim_{j\to\infty}\frac{\int_{x\in X_j}\sum_{i\in\mathcal{I}}h_i dx}{\int_{x\in X_j}\sum_{i\in\mathcal{I}}g_i dx} = z$$

すなわち, 補題 *5.3* は条件を満たせば分子および分母の $\sum$ を除いてよいことを言っている.

**証明** 紙面の都合上省略する.

$\mathcal{X}=\mathcal{T}\times\mathcal{T}'$ とおく. 補題より, もしある $\mathcal{X}$ 上可測集合の列 $\{X_j\}_{j\in\mathbb{N}}$ と, $z\in\mathbb{R}$ が存在して, $\pi(r)=r'$ なる任意の置換 $\pi:\mathcal{R}\to\mathcal{R}'$ に対して

$$\lim_{j\to\infty}\frac{\int_{(\tau,\tau')\in X_j}h_\pi(\tau,\tau')d(\tau,\tau')}{\int_{(\tau,\tau')\in X_j}g_\pi(\tau,\tau')d(\tau,\tau')}$$

$$= \operatorname*{ess\,inf}_{(\tau,\tau')\in\mathcal{X}}\frac{h_\pi(\tau,\tau')}{g_\pi(\tau,\tau')}$$

となるなら,

$$\operatorname*{ess\,inf}_{\substack{\tau\in\mathcal{T}\\ \tau'\in\mathcal{T}'}}K(\tau,\tau') = 1+\operatorname*{ess\,inf}_{(\tau,\tau')\in\mathcal{X}}\frac{h_\pi(\tau,\tau')}{g_\pi(\tau,\tau')}$$

となる.

ここで補題 *5.3* を適用するため, $\operatorname*{ess\,inf}_{(\tau,\tau')\in\mathcal{X}}\frac{h_\pi(\tau,\tau')}{g_\pi(\tau,\tau')}$ についてさらに変形する.

任意の $\tau\in\mathcal{T},\tau'\in\mathcal{T}'$ を考える. すると $\mathcal{R}$ が有限集合であることから, $\frac{A_{\tau(t),\tau'(r')}A_{\tau(r),\tau'(\pi(t))}}{A_{\tau(t),\tau'(\pi(t))}}$ の最小値を与える $t\in\mathcal{R}$ が存在する. この $t$ と, $\pi(t)$ をそれぞれ $t_{\min}$, $t'_{\min}$ とおく. そして *2* つの写像 $\mathscr{T}:\mathcal{V}^2\to\mathcal{T}$, $\mathscr{T}':\mathcal{V}'^2\to\mathcal{T}'$ を, 任意の $(u,v)\in\mathcal{V}^2$, $(u',v')\in\mathcal{V}'^2$, $t\in\mathcal{R}$, $t'\in\mathcal{R}'$ に対して以下のように定義する.

$$(\mathscr{T}(u,v))(t) = \begin{cases} v & (t = r \text{ のとき}) \\ u & (\text{そうでないとき}) \end{cases}$$

$$(\mathscr{T}'(u',v'))(t') = \begin{cases} v' & (t' = r' \text{ のとき}) \\ u' & (\text{そうでないとき}) \end{cases}$$

そして

$$\hat{\tau} = \mathscr{T}(\tau(t_{\min}), \tau(r)), \quad \hat{\tau}' = \mathscr{T}'(\tau'(t'_{\min}), \tau'(r'))$$

とおけば, $t \neq r$ として

$$\frac{A_{\hat{\tau}(t),\hat{\tau}'(r')} A_{\hat{\tau}(r),\hat{\tau}'(\pi(t))}}{A_{\hat{\tau}(t),\hat{\tau}'(\pi(t))}} = \frac{A_{\tau(t_{\min}),\tau'(r')} A_{\tau(r),\tau'(t'_{\min})}}{A_{\tau(t_{\min}),\tau'(t'_{\min})}}$$

となる. この右辺は最小であったから, このとき

$$u = \tau(t_{min}), \quad u' = \tau'(t'_{min}), \quad v = \tau(r), \quad v' = \tau'(r')$$

とすれば

$$\begin{aligned} \frac{h_\pi(\mathscr{T}(u,v), \mathscr{T}'(u',v'))}{g_\pi(\mathscr{T}(u,v), \mathscr{T}'(u',v'))} &= \frac{h_\pi(\hat{\tau}, \hat{\tau}')}{g_\pi(\hat{\tau}, \hat{\tau}')} \\ &= \frac{1}{A_{\hat{\tau}(r),\hat{\tau}'(r')}} \sum_{t \neq r} \frac{A_{\hat{\tau}(t),\hat{\tau}'(r')} A_{\hat{\tau}(r),\hat{\tau}'(\pi(t))}}{A_{\hat{\tau}(t),\hat{\tau}'(\pi(t))}} \\ &\leq \frac{1}{A_{\tau(r),\tau'(r')}} \sum_{t \neq r} \frac{A_{\tau(t),\tau'(r')} A_{\tau(r),\tau'(\pi(t))}}{A_{\tau(t),\tau'(\pi(t))}} \\ &= \frac{h_\pi(\tau, \tau')}{g_\pi(\tau, \tau')} \end{aligned}$$

が成り立つ. すなわち任意の $(\tau, \tau')$ に対して,

$$\frac{h_\pi(\mathscr{T}(u,v), \mathscr{T}'(u',v'))}{g_\pi(\mathscr{T}(u,v), \mathscr{T}'(u',v'))} \leq \frac{h_\pi(\tau, \tau')}{g_\pi(\tau, \tau')}$$

である. よって

$$\begin{aligned} &\operatorname*{ess\,inf}_{u,v \in \mathcal{V}, u',v' \in \mathcal{V}'} (|\mathcal{R}| - 1) \frac{A_{u,v'} A_{v,u'}}{A_{u,u'} A_{v,v'}} \\ &= \operatorname*{ess\,inf}_{(\tau,\tau') \in \mathrm{Im}(\mathscr{T},\mathscr{T}')} \frac{h_\pi(\mathscr{T}(u,v), \mathscr{T}'(u',v'))}{g_\pi(\mathscr{T}(u,v), \mathscr{T}'(u',v'))} \\ &= \operatorname*{ess\,inf}_{\substack{\tau \in \mathcal{T} \\ \tau' \in \mathcal{T}'}} \frac{h_\pi(\tau, \tau')}{g_\pi(\tau, \tau')} \end{aligned}$$

となる.

補題 *5.3* の適用に話題を戻す. $\mathscr{T}'(u',v')$ は $r'$ 以外のレコードについては, 全て同じ値 $u'$ をもつようなテーブルである. よって $g_\pi$ と $h_\pi$ の式の形と, $\pi$ が $\pi(r) = r'$ を満たす置換に制限されていることに注意すれば, $g_\pi(\mathscr{T}(u,v), \mathscr{T}'(u',v'))$, $h_\pi(\mathscr{T}(u,v), \mathscr{T}'(u',v'))$ が $\pi$ によらないことが分かる.

また補題 *5.2* の *(2)* の証明で, 可測関数 $a$ を $g_\pi$, $b$ を $h_\pi$ とおき,

$$z = \operatorname*{ess\,inf}_{(\tau,\tau') \in \mathrm{Im}(\mathscr{T},\mathscr{T}')} \frac{h_\pi(\tau, \tau')}{g_\pi(\tau, \tau')}$$

とすれば, $X_j = \{b/a < c(1 + 1/j)\}$ とするような可測関数列 $\{X_j\}_{j \in \mathbb{N} \setminus \{0\}}$ が補題 *5.3* の収束条件を満たすことが分かる. よって補題 *5.3* より,

$$\begin{aligned} &\operatorname*{ess\,inf}_{\substack{\tau \in \mathcal{T} \\ \tau' \in \mathcal{T}'}} K(\tau, \tau') \\ &= 1 + (|\mathcal{R}| - 1) \operatorname*{ess\,inf}_{u,v \in \mathcal{V}, u',v' \in \mathcal{V}'} \frac{A_{u,v'} A_{v,u'}}{A_{u,u'} A_{v,v'}} \end{aligned}$$

となる.

右辺が $r, r'$ に依らないことから, これは任意の $r \in \mathcal{R}$, $r' \in \mathcal{R}'$ で成り立つ.

□ (定理 5.1)

定理 5.1 を用いることで, 例えば維持-置換撹乱の *Pk*-匿名性は容易に評価することができる. 維持-置換撹乱で $A$ が最大となるのは属性値が全て維持されるときであり, 最小は属性値が全て置換されるときである. よって [3] と同じ式が即座に導かれる.

他の例は [8] で与える.

## 6 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, 匿名性の評価値である $k$ の値を算出する一般的な定理を与えた. [3] では単一のランダム化手法に対する *Pk*-匿名性の評価のみが与えられたのに対して, 本稿の定理を用いることで容易に他のランダム化手法の *Pk*-匿名性を評価できることが期待される. 実際 [8] では定理を用いて加法ノイズを用いた手法 [6][9] の評価を行う.

## 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. 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.

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

[7] D. Robello-Monedero, J. Ferne, J. Domingo-Ferrer. From t-Closeness-Like Privacy to Postrandomization via Infomation Theory. *TKDE2010*, 2010.

[8] 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.

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