サンプリングによる安全性指標についての考察

三本 知明[1]　清本 晋作[1]　宮地 充子[2]

概要：2015 年 9 月の個人情報保護法の改正により，様々な組織でパーソナルデータの利活用を目指した取り組みが進められている．これまで k-匿名性や l-多様性，差分プライバシなど様々な安全性指標が提案されてきたが，いずれも単体の匿名化技術だけでは実際のデータに対してその安全性を満たすのが困難であったり，加えるノイズが大きすぎて有用性が担保できないなどの課題が指摘されている．現在では差分プライバシを考慮した k-匿名化アルゴリズムなど，それぞれの手法を組み合わせたものが注目されている．本研究ではサンプリングと k-匿名化に着目し，サンプリングによる安全性を個人が一意に識別されうる確率としてどのようにあらわすかについて考察する．サンプリングによる安全性を確率であらわすことで，k-匿名化とサンプリングを組み合わせた際に，同じ確率という指標での評価が可能となる．

キーワード：プライバシ，k-匿名性，サンプリング

A study on a safety index of sampling

Tomoaki Mimoto[1]　Shinsaku Kiyomoto[1]　Atsuko Miyaji[2]

Abstract:
The low concerning the protection of personal information was promulgated on Sep. 9, 2015 and many organizations have started to prepare for utilizing personal data. Some safety indexes for personal data such as k-anonymity, l-diversity and differential privacy are proposed so far, but they are pointed out that it is difficult to fill safety standards by using only one approach. Therefore, some researchers proposed combined methods such as k-anonymization algorithm toward to differential privacy. In this study, we focus on combining sampling and k-anonymization, and consider the relation between the safety index of sampling and k-anonymity. Specifically, we constructed a sample dataset, and evaluated the relation between a probability of identification disclosure of any individual using the posterior probability of population uniqueness and $(1, \epsilon, \delta)$-private.

Keywords: privacy, k-anonymity, sampling

1. はじめに

情報通信技術の発展により，多種多様かつ大量のデータの収集・分析が可能となってきており，それにより新たなサービスが生み出されてきている．とりわけパーソナルデータは位置情報や購買履歴など，個人の行動・状態等に関する情報であり，利用価値が高いデータであると期待されている．しかしパーソナルデータの利活用に注目が集まる一方で，プライバシに関する懸念も同様に広がってきている．そこでプライバシを守りつつ，高い有用性を保持するための匿名化手法の確立が求められる．このトレードオフの関係について，これまでも多くの研究が行われてきており，様々な安全性の指標や分析手法が提案されてきたが，扱うデータの種類や想定する攻撃者，利用シーンなど状況によってどのような匿名化が最適なのかという基準は依然として決まっていない．

匿名化の手法としては，k-匿名化 [1] が代表的な手法である．k-匿名化とは，あるデータセットを，k 人以上が同じ準識別子 (性別，年齢など複数の属性を組み合わせるこ

[1] 株式会社 KDDI 総合研究所

[2] 大阪大学大学院工学研究科

とで個人を特定し得る属性) を持つようなデータセットへと変換する手法のことである．データセットを変換するにあたり，レコードの追加や削除，一般化というような処理が行われる．一般化とは，例えば東京や大阪という地名を日本というより大きな区分へ置き換えたり，26 歳を 21-30 歳というような範囲や，クラスタの平均値等に置き換える手法であるが，当然一般化の度合いを強くするに従い，情報量の損失が増えてしまうという問題がある．このような問題を解決するため，一般化の手法についても数多くの研究が行われているが，年々個人に関する情報へのアクセスが容易となってきており，他のデータセットと組み合わせることで個人に関する情報が漏洩するといった課題が依然として残っている．2006 年からは Dwork が提唱したプライバシの定義である差分プライバシ [2] に注目が集まり，データベース、機械学習，暗号理論などの理論計算機科学の様々な分野で研究が行われている．差分プライバシとは，何らかのクエリに対する答えに対してノイズを加えることで，ある特定のレコードの存在に関わらず出力結果に差がほとんどない，すなわちそのレコードについての情報が漏れないことを指標としている．しかし差分プライバシについても，研究者であっても理解が困難であることや，クエリによっては加えるノイズが大きくなり，実利用に乏しくなるというような問題点が挙げられる．

そのような背景から，近年では複数の匿名化技術を組み合わせる手法が提案されている．例えば [3] では，Soria-Comas らは差分プライバシを満たすような k-匿名化を提案しており，あらかじめ insensitive な MDAV を用いて k-匿名化を施しておくことで，比較的小さなノイズの追加で単純にノイズを加える場合と同等の安全性を満たすことを示している．匿名化技術の組合せについては，[4] などのガイドラインでは，k-匿名化について述べられており，また欧米諸国のセンサス局では実際にデータセットをサンプリングした後に加工を施すことが多いことから，実際の利用ではサンプリングと k-匿名化は有効な組み合わせであることが伺える．

本研究の目的は，一つのデータセットに対して k-匿名化とサンプリングを行う場合を考え，サンプリングによる安全性を k-匿名性と比較しやすい確率で表現する点にある．我々は擬似データセットに対して，提案されている安全性指標である ϵ, δ を入力として与え，得られたサンプリング率をもとに生成したサンプルデータセットにおいて一意であるレコードの母集団一意確率をパラメータを変更しながら求める．サンプリングによる安全性を確率であらわすことで，サンプリングと k-匿名化アルゴリズムを組み合せた際に，同じ確率で安全性をあらわすことが可能となると考える．

2. 準備

以下では，k-匿名性とサンプリングによる安全性を比較するにあたり，それぞれで用いられている指標について述べる．

2.1 k-匿名性

匿名化処理にあたり，データセットの属性は，識別子，準識別子，機密属性に分類することができる．識別子とは，名前や国民 ID などそれだけで個人を特定できる属性であり，準識別子とは，それだけでは個人を特定できるとは限らないが，他のデータセットにあらわれる可能性があり，それらを紐付けると個人を特定できる可能性がある属性である．また，機密属性とは病名や購買履歴などユーザが個人と紐付けられた状態で公開されたくないような属性のことである．k-匿名性とは，任意の一つ以上の準識別子を選択した時に，同じ準識別子の組合せを持つレコードがデータセットに k 個以上存在することを保証する指標である．k-匿名性を概念を拡張し，機密情報についても考慮した，l-多様性 [5] や t-近似性 [6] なども知られている．

2.2 母集団一意性

統計分野における代表的な識別リスク指標であり，データセット全体，すなわち母集団においてある準識別子の組合せが一意に定まるレコードが存在するとき，そのレコードを母集団一意であるという．また母集団からサンプリングによって得られた標本において，ある準識別子の組合せが一意に定まるレコードが存在するとき，そのレコードを標本一意という．標本一意であるレコードの存在は，そのすべてが必ずしもプライバシ侵害の脅威とはならないが，それが母集団でも一意である場合について考える必要がある．ランダムサンプリングを行った際，事前分布を考慮することで，標本一意であるレコードが母集団一意でもある確率を求めることが可能である [7]．具体的な計算方法は以下のとおりである．

まず，母集団は既知ではない限り確率的に変動すると考え，ある確率関数で定義された母集団の上の超母集団から抽出されていると考える．また，標本は母集団から非復元抽出されるものとする．母集団には N 個のレコードが存在し，準識別子の組合せが K 通りある，すなわち K 個のクラスタがあるとする．また i 番目のクラスタに含まれるレコード数を F_i とする．確率 n/N で非復元抽出によるランダムサンプリングを行った際に，i 番目のクラスタに含まれるレコード数を f_i とする．このとき $f = (f_1, ..., f_K)'$ は次のような確率関数を持つ多変量幾何分布に従う（$(n; f) \sim MH(H; F)$ と表記する）．

$$\Pr(f|F) = \prod_{i=1}^{K} \frac{\begin{pmatrix} F_i \\ f_i \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}}$$

$$\left(\text{ただし}, N = \sum_{i=1}^{K} F_i, n = \sum_{i=1}^{K} f_i, F = (F_1, ..., F_K)'\right)$$

超母集団における i 番目のクラスタの相対頻度 π_i が既知であり，母集団が単純復元抽出によって超母集団から抽出されたと考えれば F は以下の様な確率関数を持つ多項分布に従い（$F \sim MH(N;\pi)$ と表記する），これは母数 F の事前分布として解釈することができる．

$$\Pr(F|\pi) = \frac{N!}{F_1! \cdots F_K!} \pi_1^{F_1} \cdots \pi_K^{F_K}$$
$$\left(\text{ただし}, \pi = (\pi_1, ..., \pi_K)'\right)$$

以上から母数 F の事後分布は次の式であらわされる．

$$\Pr(F|\pi, f) = \frac{(N-n)!}{(F_1 - f_1)! \cdots (F_K - f_k)!} \pi_1^{F_1 - f_1} \cdots \pi_K^{F_K - f_k}$$

次に標本一意であるレコードが m 個であったとして，更にそれが母集団一意でもある場合を考える．$\pi_i = \pi_0 (i = 1, ..., m)$ とし，r を標本一意かつ母集団一意であるレコード数とすると，標本一意であるレコードが母集団一意でもある確率は

$$\Pr(F_{i_1} = \cdots = F_{i_r} = 1 | f_1 = \cdots = f_m = 1) = (1 - r\pi_0)^{N-n}$$

ここから以下の定理を得ることができる（証明は [7] を参照）．

定理 (母集団一意確率): $(n;f) \sim MH(H;F), F \sim MH(N;\pi)$ であるとき，$\pi_1 = \cdots = \pi_m = \pi_0$ とする．また，$\{i_1, ..., i_k\} \subseteq \{1, ..., m\}$ とすると，少なくとも一組の $\{i_1, ..., i_k\}$ に対して，$f_1 = \cdots = f_m = 1$ の時に $F_{i_1} = \cdots = F_{i_k} = 1 (1 \leq k \leq m)$ となる確率 α_k は

$$\alpha_k = \sum_{r=k}^{m} \begin{pmatrix} m \\ r \end{pmatrix} \cdot \sum_{j=0}^{m-r} (-1)^j \begin{pmatrix} m-r \\ j \end{pmatrix} (1 - (r+j)\pi_0)^{N-n}$$

したがって $k = 1$ として超母集団におけるクラスタの相対頻度 π_0 を決めることで，標本一意であるレコードの中に，母集団一意であるものが少なくとも一つある確率を得ることができる．相対頻度については事前情報が全く無い場合は $\pi_0 = 1/n$，すなわち標本にレコードが 1 個存在する場合，母集団にはそのレコードは平均して N/n 個存在すると仮定するのが自然であるが，実際は相対頻度の値を変えてリスク評価をすることが必要である．

2.3 $(1, \epsilon, \delta)$-private

サンプリングによる安全性の評価として，母集団一意性とは別に $(1, \epsilon, \delta)$-private を指標とするものがある [8].

定義 ($(1, \epsilon, \delta)$-private): i 番目のレコードの値が v であるようなデータセットを $T_{\{i \to v\}}$ とする．i 番目のレコードのみが異なる任意のデータセットを $T_{\{i \to v\}}, T_{\{i \to v'\}}$ とし，あるサニタイゼーションメカニズムによってサニタイズされたデータセットを S とする．以下の式を満たすとき，このサニタイゼーションメカニズムは $(1, \epsilon, \delta)$-private であるという．

$$\Pr\left[\frac{\Pr[S|T_{\{i \to v\}}]}{\Pr[S|T_{\{i \to v'\}}]} < 1 + \epsilon\right] < 1 - \delta$$

この定義を満たすようなサンプリングを行うため，更に以下を定義する．

定義 (レコード種別):あるデータセットにおいて，少なくとも $\frac{12 \log(\frac{k}{\alpha})}{p}$ 回出現するレコードを common value とする．また，高々 $\frac{2 \log(\frac{k}{\alpha})}{\epsilon}$ 回出現するレコードを rare value とする．common value でも rare value でもないレコードを infrequent value とする．ただし，k はデータセットにおけるクラスタの数であり，$\alpha = \frac{\delta}{2}$，また p はサンプリング率である．

定義 (Good sample):サンプルデータセットにおいて，rare value が出現せず，infrequent value v の出現回数が高々 $n_{v_i} p + 2 \log(\frac{k}{\alpha})$ 回であり，common value v の出現回数が高々 $n_{v_c} p + \sqrt{3 n_{v_c} p \log(\frac{k}{\alpha})}$ 回であるようなデータセットを good sample とする．ただし，n_{v_i}, n_{v_c} はそれぞれもとのデータセットにおける infrequent value, rare value の出現回数である．

このとき，以下の定理を導くことができる．

定理 ($(1, \epsilon, \delta)$-private):データセット T において，$\alpha = \frac{\delta}{2}, p + \epsilon < 1 - \frac{2}{3 n_v^1}, p < \frac{2}{3}$ とする．ただし，n_v^1 は $T \backslash \{i\}$ におけるレコード v の数をあらわす．また K をデータセット T におけるクラスタの数，そして t を rare value のレコード数とする．$t > 0$ の場合，T に対して $p < \frac{\epsilon \log(\frac{1}{1-\alpha})}{4t \log(\frac{K}{\alpha})}$ のサンプリング，また $t = 0$ の場合，T に対して $p < \epsilon$ のサンプリングは $\left(1, \max\left(\frac{p+\epsilon}{1-(p+\epsilon)}, \frac{\frac{3}{2}p}{1-\frac{3}{2}p}\right), \delta\right)$-private である．

証明は基本的に [8] と同じだが，lemma 1 において v が infrequent value のとき，$n_v^1(p+\epsilon) + p < n_v^1, \frac{3}{2}(n_v^1 + 1)p < n_v^1 + 1$ を満たす範囲として，$p < \frac{2}{3}, p + \epsilon < 1 - \frac{2}{3 n_v^1}$ としている．

3. 実験方法

以下では，本実験における攻撃者モデル，および比較方法について述べる．

3.1 攻撃者モデル

ある個人のレコード，あるいは出力されたデータセットを持つ攻撃者を想定する．このとき母集団一意確率は，攻撃者がある個人のレコードを特定できる確率の最大値であると考えることができる．この値は，k-匿名性を持つデータセットにおいて $\frac{1}{k}$ とすることができるため，サンプリングと k-匿名化を行った場合の安全性は，これらの確率を掛けあわせたものと考えることができる．

3.2 比較方法

サンプリングによる安全性は，母集団一意確率および ϵ, δ であらわすことができる．本稿では，まずクラスタ数および rare value の数 k, t を設定した擬似データを生成し，安全性 ϵ, δ を決定，サンプリング率 p を計算する．その後擬似データを確率 p でサンプリングし，標本一意であるレコード数 m を求め，相対頻度 π_0 を変えながら母集団一意確率を計算する．これにより，ϵ, δ と母集団一意確率との関係を確認することが可能であり，上記攻撃者モデルのもとサンプリングと k-匿名化との比較が可能となる．

4. 実験結果

$(1, \epsilon, \delta)$-private なサンプリングを行う際，rare value 数が多いと good sample を得ることが困難になるため，データセットはあらかじめ汎化を行う必要がある．本稿ではレコード数 $N = 10000$，クラスタ数 $k = 100$，rare value 数 $t = 1$ として評価を行った．ϵ, δ とサンプリング率との関係を表 1 に示す．

表 1 ϵ, δ とサンプリング率の関係

$\{\epsilon, \delta\}$	サンプリング率
$\{0.5, 0.5\}$	0.60%
$\{0.5, 0.9\}$	1.38%
$\{0.9, 0.5\}$	1.08%
$\{0.9, 0.9\}$	2.49%
$\{0.97, 0.96\}$	3.00%

上記のように事前に汎化した場合であっても rare value が存在する場合，$(1, \epsilon, \delta)$-private の安全性指標のもとでは高々 3% のサンプリングしか許されないことが分かる．

続いてこのサンプリング率を元に擬似データをサンプリングし，母集団一意確率を求めた．この時，超母集団におけるセルの相対頻度 π_0 を $1/n$（標本に 1 個存在する場合，母集団には平均して N/n 個存在するという想定）から $1/N$（標本に 1 個存在する場合，母集団には平均して 1 個存在するという想定）の範囲で評価を行った．サンプリングは複数回行い，標本一意であるレコード数 m をその平均値としている．この時の π_0，母集団一意確率の関係を表 2-6 に示す．

表 2 サンプリング率 0.60%，標本一意のレコード数 14.2

π_0	母集団一意確率
$1/n$	$4.0E-72$
$1/3n$	$1.3E-23$
$1/5n$	$5.5E-14$
$1/10n$	$8.9E-7$
$1/50n$	0.41
$1/100n$	0.95
$1/N$	0.99

表 3 サンプリング率 1.38%，標本一意のレコード数 11.5

π_0	母集団一意確率
$1/n$	$7.9E-31$
$1/3n$	$5.0E-10$
$1/5n$	$7.1E-6$
$1/10n$	0.0088
$1/50n$	0.95
$1/100n$	1.0
$1/N$	1.0

表 4 サンプリング率 1.08%，標本一意のレコード数 11.8

π_0	母集団一意確率
$1/n$	$1.3E-39$
$1/3n$	$6.2E-13$
$1/5n$	$1.3E-7$
$1/10n$	0.0012
$1/50n$	0.87
$1/100n$	1.0
$1/N$	1.0

4.1 考察

実験結果をみると，母集団一意確率は π_0 が $1/n$ では過小評価であり，おおよそ $1/50n$ 以下では過大評価であることが分かる．今回のようなデータの場合，実際の事後確率を考える際は，$1/5n \leq \pi_0 \leq 1/10n$ 程度が適当だと考えられる．また rare value は 1 個だけであるから，サンプリング率を上げるにしたがって rare value 以外のレコードは

表 5　サンプリング率 2.49%，標本一意のレコード数 8.7

π_0	母集団一意確率
$1/n$	$7.9E-17$
$1/3n$	$1.9E-5$
$1/5n$	0.0034
$1/10n$	0.16
$1/50n$	1.0
$1/100n$	1.0
$1/N$	1.0

表 6　サンプリング率 3.00%，標本一意のレコード数 5.2

π_0	母集団一意確率
$1/n$	$4.5E-14$
$1/3n$	$1.1E-4$
$1/5n$	0.008
$1/10n$	0.19
$1/50n$	1.0
$1/100n$	1.0
$1/N$	1.0

重複がみられるようになり m は減少するが，それが rare value である可能性が高まるため母集団一意確率は上昇していると考えられるため，自然な結果といえる．

本実験を通して課題も幾つか挙げられる．まず，擬似データセットは rare value の数が 1 としており，ある程度匿名化されたデータに外れ値が含まれているというような想定である．これは $(1,\epsilon,\delta)$-private が満たす条件が，rare value が 0 である必要があるなど非常に厳しい制約があるためで，また ϵ,δ は上限値が決まっているためサンプリング率は高々 3% までの評価しかできなかった．また結果としても，母集団一意確率に対して，ϵ,δ の値は高いものとなっている．このような厳しい制約は，$(1,\epsilon,\delta)$-private がサンプリングのみに対する安全性指標であるというのが理由の一つと思われる．したがって，その後の k-匿名化を見据えたサンプリングの安全性指標が必要である．

また，今回の擬似データセットのように rare value の数が 1 であるような場合，k-匿名性の安全性指標を $\frac{1}{k}$ とすると，そのレコードの有無でリスクが大きく変化する．したがって k-匿名性を考える場合，サンプリングによってクラスタサイズが最小のレコードがサンプルに含まれるかどうかが大きな問題となる．そのためサンプリングと組み合わせる際は，単純に $\frac{1}{k}$ をリスクとするのではなく，それとは別にクラスタサイズの平均なども考慮すべきだろう．

4.2　おわりに

本稿ではサンプリングによる安全性を二通りの安全性指標で比較した場合のパラメータの取り方について考察した．実験結果から，擬似データセットにおいて $(1,\epsilon,\delta)$-private を満たすとき，超母集団における相対頻度は $1/10n$ 程度，すなわち標本一意であるデータは母集団において平均 $N/10n$ 個のレコードを含むと考えるのが適当であることが分かった．

しかし，$(1,\epsilon,\delta)$-private を満たす条件としてサンプリング率が $O(\frac{\epsilon\delta}{t})$ である必要があり，本研究で用いたようなある程度汎化されたデータセットに外れ値が含まれているような状況でないと，サンプリング率が非常に小さくなってしまうため，k-匿名化を組み合わせる場合は，それを考慮した安全性指標が必要となる．

謝辞

本研究の一部は JSPS 科研費基盤 C (JP15K00183) と (JP15K00189) 及び科学技術振興機構（JST）の CREST と国際科学技術協力基盤整備事業の助成を受けています．

参考文献

[1] L. Sweeney. "k-anonymity: a model for protecting privacy" *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10 (5), pp. 557-570, 2002.

[2] C. Dwork. "Differential privacy" *ICALP 2006*, LNCS 4052, pp.1-12, Springer, 2006.

[3] J. Soria-Comas, J. Domingo-Ferrer, D. Sanchez and S. Martinez. "Enhancing data Utility in differential privacy via microaggregation-based k-anonymity" The VLDB Journal, vol.23(5), pp.771-794, 2014.

[4] M. Oswald. "Anonymisation Standard for Publishing Health and Social Care Data Specification (Process Standard)", ver. 1.0, 2013.

[5] A. Machanavajjhala, D. Kifer, J. Gehrke and M. Venkitasubramaniam. "l-Diversity: Privacy Beyond k-Anonymity" *ACM Transactions on Knowledge Discovery from Data(TKDD)*, vol.1, no.1, p.3, 2007

[6] N. Li, T. Li, and S. Venkatasubramanian. "t-closeness: Privacy beyond k-anonymity and l-diversity" *ICDE 2007*, pp.106-115, 2007.

[7] 大森 裕浩. "ミクロデータにおける母集団一意性の事後確率" 統計数理, 51(2), pp.223-239, 2003.

[8] K. Chaudhuri and N. Mishra. "When Random Sampling Preserves Privacy" *CRYPTO 2006*, pp.198-213, 2006.