再識別リスク評価と匿名化の展望

南 和宏[1]　千田 浩司[2]

概要：ビッグデータの活用が進む中，パーソナルデータを適切に保護するための匿名化に関する基準が国内外で整備されつつある．特に再識別に対するリスク評価やリスク低減のための匿名化技術の確立は急務である．しかし汎用的なリスク評価や匿名化は困難であるため，状況に応じた適切な対策が重要といえる．本稿では，代表的な再識別リスク指標である k-匿名性に着目し，攻撃者の知識，およびパーソナルデータの提供形態や加工方法に応じた適切な対策について再考する．そして k-匿名性を満たしていても元のデータを再構築できる場合がある問題を取り上げ，その対策について論じる．

キーワード：再識別，匿名化，k-匿名性

Research Challenges for Re-identification Risk Assessments and De-identification

Kazuhiro Minami[1]　Koji Chida[2]

Keywords: re-identification, de-identification, k-anonymity

1. はじめに

ビッグデータの活用が進む中，パーソナルデータを適切に保護するための匿名化 (De-identification)[*1]に関する基準が国内外で整備されつつある．特に匿名化したパーソナルデータから特定の個人のデータを識別する**再識別** (Re-identification) の問題が深刻化しており [3], [4], [5]，再識別に対するリスク評価や，リスク低減のための匿名化技術の確立が求められている．我が国においても，昨年 9 月に成立した個人情報保護法改正に伴い，本人の同意無くパーソナルデータの流通を認める「匿名加工情報」の基準が検討されているところである [6].

代表的な再識別リスク指標として，k-匿名性 [4] が挙げられる．k-匿名性は特定の個人のデータを k 個未満に識別できる場合にリスクが高いと考える指標である．そのため，k-匿名性を満たさずリスクが高いと判断されたパーソナルデータは，k-匿名性を満たすように加工する必要がある．

匿名化の基準に関して先進的な欧米の公的文書では，k-匿名性の利用について具体的に触れているものもいくつか存在する [7]．しかし実際には，攻撃者の知識，およびパーソナルデータの提供形態や加工方法によっては，k-匿名性が適用できなかったり，適切な指標でない場合がある．そこで本稿では，各種状況下での再識別リスクについて整理し，現状の指標との対応付けや課題，そして対策について考察する．特に k-匿名性を満たしていても元のデータを再構築できる場合がある問題を取り上げ，その対策について論じる．

2. 準備

2.1 再識別リスク

氏名等の削除は古典的な匿名化としてよく知られるが，

[1] 統計数理研究所
Institute of Statistical Mathematics
[2] NTT セキュアプラットフォーム研究所
NTT Secure Platform Laboratories

[*1] ISO/TS 25237(Health informatics — Pseudonymization)[1] では，匿名化 (Anonymization) を「身元が分かるデータの集合 (Identifying Data Set) とデータ主体 (Data Subject) の間の関連を取り除くプロセス」と定義しており，匿名化手法には Masking と De-identification の 2 種類がある [2].

それだけでは不十分な場合がある．例えば時刻と詳細な位置情報（緯度経度情報）を含むパーソナルデータは，夜間には自宅にいる可能性が高いことから自宅の場所が推測されやすい．また，氏名を削除した患者の医療情報（診断結果，投薬情報等に加え性別，生年月日，ZIP コードが含まれていた）を米国マサチューセッツ州が公開したところ，マサチューセッツ州知事の医療情報が特定されてしまったという事例も知られる [4]．同じく公開されている投票者名簿（氏名，性別，生年月日，ZIP コードが含まれていた）から，州知事と同じ性別，生年月日，ZIP コードの人がいないことが明らかになったためである．さらには，1990 年の米国国勢調査の回答データは性別，生年月日，および 5 桁の ZIP コードだけで全体の 87%が一意の情報になっているという報告もある [3]．以上の例から，情報の一意性や他のパーソナルデータとの突合等を踏まえた再識別リスクの評価が重要であることが分かる．

2.2 匿名化

表 1 元のパーソナルデータ

name (正識別子)	sex (準識別子)	age (準識別子)	income (センシティブ属性)	item 1 (非センシティブ属性)
Alice	F	24	$46K	coffee
Bob	M	25	$52K	beer
Chris	M	30	$57K	cola
Dan	M	30	$81K	milk
Eve	F	32	$50K	cola
Flora	F	32	$104K	whiskey

本稿で対象とするパーソナルデータは，表 1 に例示するように各個人のデータが 1 レコードに記載されたテーブルとする．そして各列の属性は以下の何れかに分類できるものとする．

- **正識別子**： 個人を一意に識別できる属性，または当該属性の組．例えば氏名，住所のような属性の組み合わせは無視できない確率で正識別子となる [8].
- **準識別子**：間接的に個人を識別できる属性．性別や年齢のような属性は間接的に個人の識別に用いることができる [9].
- **センシティブ属性**：正識別子，準識別子以外で，個人のプライバシーに関するもの等，他人にむやみに知られたくない属性．センシティブ属性の値をセンシティブデータとよぶ．
- **非センシティブ属性**：上記以外の属性．非センシティブ属性の値を非センシティブデータとよぶ．

表 1 に例示するようなテーブルを匿名化するために，公的統計分野等において様々な技法（加工方法）が知られている．表 2 は，先行文献 [10] の表 4 に基づき代表的な加工方法をまとめたものである．詳細を記した先行文献がいくつかあるので参照されたい [11], [12], [13]．なおノイズ付加や PRAM 等は属性値が確率的に変化する．このような加工方法は撹乱的（perturbative）とよばれる．一般化や曖昧化のように確率的な要素を伴わない加工方法は非撹乱的（non-perturbative）とよばれる．

2.3 k-匿名性

あるパーソナルデータのテーブルについて，全ての準識別子の値が等しいレコードの集合を準識別クラスとよび，全ての準識別クラスが k 個以上のレコードをもつとき，そのテーブルは k-匿名性を満たすという [4]．k-匿名性を満たす加工方法を k-匿名化とよび，準識別子に対して非撹乱的な手法が一般に用いられる．表 3 は local recoding およびセル削除を用いて 2-匿名化を行ったテーブルの例である．

匿名化の重要な点は，再識別リスクを低減させつつ，本来のパーソナルデータが持つ有用性 (utility) もなるべく損ねないことである．情報損失を最小とする k-匿名化は NP-困難であることが知られ [16]，有用性を可能な限り損ねない k-匿名化のアルゴリズムが数多く提案されている [17]．しかし本質的には匿名性と有用性はトレードオフの関係にあり，用途に応じて適切な匿名化データを作成することが望ましい．

表 3 2-匿名化されたテーブルの例

sex (準識別子)	age (準識別子)	income (センシティブ属性)	item 1 (非センシティブ属性)
×	[24,25]	$46K	coffee
×	[24,25]	$52K	beer
M	30	$57K	cola
M	30	$81K	milk
F	32	$50K	cola
F	32	$104K	whiskey

ところで Sweeney は k-匿名性について以下の問題を指摘している [4].

(1) 同一個人のデータが複数のレコードに含まれている：もしある準識別クラスが同一個人のデータで占めており，攻撃者がその事実を知っていれば，当該個人のデータを推定しやすくなる．

(2) レコードの並び替えをしない：元のデータが辞書順に並んでいれば，特定の個人のデータを推定しやすくなる．

(3) 同一の（非）センシティブ属性を複数回開示する：元のパーソナルデータの準識別子に対して異なる加工を施し，複数の匿名化データを作成して開示した場合，（非）センシティブ属性の値をキーにして，k-匿名性を

表 2 匿名化の代表的な加工方法（先行文献 [10] の表 4 に基づき作成）

分類	技法	概要
属性情報の削除	属性（列）削除	正識別子等，開示すべきでない属性を削除する
	仮名化	開示すべきでない属性を符号や番号等に置き換える
属性情報の置換え	一般化	属性値を上位の値や概念に置き換える（例：「年齢」→「年代」，「キュウリ」→「野菜」） データ全体に行うものを Global Recoding，局所的に行うものを Local Recoding とよぶ 四捨五入や最も近い定数の倍数への変換等を丸め法（Rounding）とよぶ
	曖昧化	特に大きい，もしくは小さい値をまとめる （例：100 歳以上の人を「100 歳以上」とする） 大きい値のまとめを Top Coding，小さい値のまとめを Bottom Coding とよぶ
	ミクロアグリゲーション	複数のレコードをグループ化し，同じグループの値を代表値（例：平均値）に置き換える
	ノイズ付加	一定の分布に従った乱数的なノイズを加える
	スワッピング	レコード間で属性値を（確率的に）入れ替える
	PRAM	マルコフ推移確率行列に基づき，確率的に属性値を置き換える [14], [15] Post RAndomization Method の略
その他技法	レコード（行）削除	特殊な属性値を持つレコードを削除する （例：120 歳以上のレコードを削除）
	セル削除	開示すべきでない属性値を削除する
	疑似データ作成	元データと統計的に疑似させる人工的な合成データを作成する
	サンプリング	元データ全体から一定の割合・個数でランダムに抽出する
	並び替え	レコードの順序をランダムまたは値の大小順やアルファベット順等に置き換える

満たさないデータを再構築できる場合がある．

(4) 時系列に変化するデータを逐次開示する：複数の開示データの差分から，k-匿名性を満たさないデータを再構築できる場合がある．

上記について，(1) および (2) は容易に回避できる問題だが，(3) および (4) は複数回のデータ開示において生じる問題であり，k-匿名性は指標として適切でない．

3. k-匿名性の課題

2.3 節では，Sweeney が指摘している k-匿名性の問題について触れた．本節では，攻撃者の知識，およびパーソナルデータの提供形態や加工方法の視点から，k-匿名性の課題を再考する．

3.1 攻撃者の知識

k-匿名性は攻撃者に対して以下の知識を (暗黙に) 仮定していると考えられる．

(1) ターゲット（攻撃対象者）の準識別子を知っている．

(2) ターゲットのデータが元のパーソナルデータに含まれていることを知っている．

(3) ターゲットのデータが匿名化データに含まれていることを知っている．

(4) 元のパーソナルデータの識別子と準識別子を知っている．

これらの条件を弱めることで，比較的弱い匿名化で済み，より有用性の高い匿名化データの作成が期待できる．逆に，匿名化データを作成する k-匿名化のアルゴリズムは，攻撃者の知識として特に考慮されていなかったが，最近では k-匿名化のアルゴリズムの最小化原理を利用して，元のデータを再構築する攻撃が提案されている [18]．以降，この種の攻撃を既知アルゴリズム攻撃と呼ぶ．

3.2 パーソナルデータの提供形態

k-匿名性は元のパーソナルデータについて唯一の匿名化データを作成・開示することを暗黙に仮定していると考えられる．しかし元のパーソナルデータが多数の属性，特に多数の準識別子を含むとき，k-匿名化によって有用性が著しく低下する次元の呪いの問題がある [19], [20]．次元の呪いの問題を回避するため，分析の都度，必要な属性のみ抽出して匿名化データを作成・開示する方法が提唱されている [20]．すなわち有用性を高めるために依存関係のある複数の匿名化データを開示することになる．ただし Sweeney が指摘した問題 (3) のように，(非) センシティブ属性が複数回開示されるとそれがキーとなり得るため，センシティブ属性は唯一の開示，またはセンシティブ属性も加工対象とする必要がある．

一方，元のパーソナルデータが時系列に変化する場合も想定される．一つは時が経つにつれ属性が増える場合であり，もう一つはレコード数や属性値が変化する場合である．何れも元のパーソナルデータの変化前と変化後それぞれの匿名化データを作成・開示することが想定される．

まとめると，パーソナルデータの提供形態は以下の 4 つに分類できる．

(1) 元のパーソナルデータが変化しない場合

(a) 唯一の匿名化データを作成・開示
(b) 依存関係のある複数の匿名化データを作成・開示

(2) 元のパーソナルデータが変化する場合
(a) 属性を追加して匿名化データを逐次的に作成・開示
(b) レコード数や属性値が変化した匿名化データを逐次的に作成・開示

本稿では以降，(1) の提供形態に絞って検討を行う．

3.3 匿名化の加工方法

k-匿名性は一般化や削除等の非撹乱的な加工により匿名化データを作成することを前提とする．しかしノイズ付加，スワッピング，PRAM のような撹乱的な加工には対応しておらず，匿名化に制限を与えてしまうという課題がある．

3.4 攻撃の具体例

3.4.1 既知アルゴリズム攻撃

表 3 は，表 1 について 2-匿名性を満たすよう匿名化したデータだが，例えば上位 2 件のレコードの年齢が 24 歳以上 25 歳以下であることに着目すると，アルゴリズムが最小化原理に基づいていれば，片方が 24 歳，もう片方が 25 歳と推定できる．その理由は，もし 2 件とも 24 歳または 25 歳であれば，24 歳以上 25 歳以下のように一般化しないためである．したがって，24 歳と 25 歳が 1 人ずついることが分かり，表 4 のような 2-匿名性を満たさないテーブルを再構築できる．

表 4 2-匿名性を満たさないテーブルの再構築

sex	age
×	24
×	25
M	30
M	30
F	32
F	32

3.4.2 差分攻撃

依存関係のある複数の k-匿名化データを作成・開示する場合，複数の k-匿名化データの差分から，k-匿名性を満たさないテーブルを再構築できる可能性がある．これを本稿では差分攻撃と呼ぶ．

表 5 のパーソナルデータに対し，表 6 と表 7 の 2 つの 2-匿名化データを作成・開示したとする．このとき，表 6 と表 7 の差分により，(k-匿名化のアルゴリズムを知らなくても) 表 5 の正識別子を除いたテーブルを一意に再構築できる．

表 5 元のパーソナルデータ

name (正識別子)	sex (準識別子)	age (準識別子)	blood (準識別子)	income (センシティブ属性)	item 1 (非センシティブ属性)
Alice	F	24	O	$40K	coffee
Eve	F	24	O	$40K	coffee
Flora	F	25	A	$40K	cola
Chris	M	26	A	$50K	cola
Dan	M	26	A	$50K	cola

表 6 2-匿名化されたテーブルの例 1

sex (準識別子)	age (準識別子)	income (センシティブ属性)
F	[24,25]	$40K
F	[24,25]	$40K
F	[24,25]	$40K
M	26	$50K
M	26	$50K

表 7 2-匿名化されたテーブルの例 2

age (準識別子)	blood (準識別子)	item 1 (非センシティブ属性)
24	O	coffee
24	O	coffee
[25,26]	A	cola
[25,26]	A	cola
[25,26]	A	cola

4. 対策の考察

4.1 攻撃者の知識に応じた再識別リスク指標

攻撃者の知識が以下の状況の場合の再識別リスク指標について考察する．

(1) ターゲットのデータが匿名化データに含まれているかどうか分からない．
(2) 元のパーソナルデータの識別子と準識別子を完全には分からない．

(1) については，サンプリングが有効と考えられる．すなわち，元のパーソナルデータを知っていたとしても，サンプリングされたデータとの照合は自明でない．ただし元のパーソナルデータのうち一意のレコードについては，サンプリングされたデータにも含まれていれば照合されてしまう．そのため，サンプリングされたデータをさらに加工する等の対策が必要となる．サンプリングと一般化により加工したデータの再識別リスク指標として，δ-存在性が提案されている [21]．δ-存在性を満たせば，任意の個人のレコードについて，匿名化データに含まれることの確信度が δ 以下であることが保証される．

(2) については，(1) を援用する仮定といえる．その理

由は，元のパーソナルデータの中で一意のレコードであっても，その事実が分からないため，サンプリングされたデータとの照合が自明でなくなるためである．このような仮定に基づくリスク指標は，菊池らが考察を与えている [22].

4.2 既知アルゴリズム攻撃の対策

従来，匿名化に関する安全性の検証は，匿名化プログラムが出力する匿名化データの中身の確認で十分と信じられていた．実際，k-匿名化の要件は匿名化データの準識別グループのサイズが k 以上であるかを確認するだけである．しかし近年，匿名化アルゴリズムの知識を利用したアルゴリズムベースの情報漏洩の問題が指摘されている [18]．特に多くの k-匿名化アルゴリズムは匿名化のためのデータ加工による情報損失を最小に抑える最小化原理を採用しており，匿名化アルゴリズムの最小化原理を利用した推論攻撃が可能であることが分かってきた．

既知アルゴリズム攻撃は，k-匿名化において元のテーブルを準識別クラスを決定する際，準識別子の値の等価性に加え，センシティブ属性の値の多様性に関する追加要件を考慮した匿名化アルゴリズムで顕著な問題になる．このような追加要件は ℓ-多様性として提案されており，準識別クラスのレコード群は ℓ 個以上の異なるセンシティブ属性の値を取ることが要求される．

表 8–11 で既知アルゴリズム攻撃の事例を説明する．表 8 の医療データから氏名（正識別子）の属性を削除すると，表 9 のテーブルになる．表 9 のレコードを性別の男女で 2 つの準識別クラスに分けると，数が少ない男性グループに 2 つのレコードが含まれるので，2-匿名化の要件を満足している．したがって，匿名化のアルゴリズムが最小化原理に従うならば，性別の属性をこれ以上一般化する必要はない．

しかし準識別クラスのセンシティブ属性の多様性に着目すると，表 9 の 2 人の男性は病名がエイズであり，2-多様性の要件を満足していない．この場合，男性のグループに別の病名の女性を追加した準識別クラスを作る必要がある．表 10 はそのような 2-多様性を満足するテーブルの一例である．癌と糖尿病をもつ女性を男性のグループに加え，性別の識別子の値は‘∗’に一般化している．

ただし，攻撃者が外部知識として表 11 の準識別子の情報を持ち，テーブルに含まれる男性の数が 2 人と知っていた場合はどうであろうか？この場合，エイズの患者数は 2 人なので，男性のエイズ患者が何人いるかで 3 つの場合に分けられる．男性のエイズ患者がいない場合，表 10 より 2 人の男性の病名は必ず異なるものであることが分かる．また女性のグループには既に胃炎と心臓病の患者が含まれるので，エイズ患者が 2 名加わっても，2-多様性は満足される．したがってこの場合は男女のレコードを混合した準識別クラスを作成するために性別を一般化する必要はない．男性のエイズ患者が 1 名の場合も同様に性別属性の一般化の必要性は生じない．3 つめの男性エイズ患者が 2 名いる場合のみ，表 9 のテーブルから表 10 への一般化が必要であり，男性 2 名がエイズ患者であることが判明する．

このように準識別クラスの決定の際にセンシティブ属性を参照する行為は非常に危険であり，匿名化データのグループ化の情報から間接的に機微な情報が漏洩する危険性がある．

このような既知アルゴリズム攻撃の対策として，Xin ら [23] はアルゴリズムセーフな匿名化アルゴリズムを定義しており，下記の 2 つの要件を提示している．

(1) 匿名化データの準識別クラスを（非）センシティブ属性に依存せずに決定する．
(2) （非）センシティブ属性の値を変更しない．

一方，表 4 の例では上記の要件を満たしている．これは最小化原理の採用によるものと考えられ，対策として確率的な処理の追加が挙げられる．すなわち，多少の有用性を犠牲にし，表 3 において例え上位 2 件のレコードの年齢がともに 24 歳だとしても，確率的に 24 歳以上 25 歳以下と一般化することで，表 4 の再構築を防ぐことができる．

表 8 医療データ

名前 (正識別子)	性別 (準識別子)	病名 (センシティブ属性)
鈴木	女	胃炎
佐藤	女	心臓病
木村	女	癌
高橋	女	糖尿病
田中	男	エイズ
本田	男	エイズ

表 9 2-匿名化した医療データ

性別 (準識別子)	病名 (センシティブ属性)
女	胃炎
女	心臓病
女	癌
女	糖尿病
男	エイズ
男	エイズ

4.3 差分攻撃の対策

依存関係のある複数の匿名化データを開示する場合，k-匿名性を満たさないテーブルを再構築されないようにする必要がある．Kifer らは，匿名化データの開示と一部の属性の集計値を組み合わせたときの再識別リスクについて評価している [24]．すなわち，複数のデータから矛盾無く再構築できるテーブルについて k-匿名性を検証する．特に表 6

表 10 2-多様性の医療データ

性別 (準識別子)	病名 (センシティブ属性)
女	胃炎
女	心臓病
*	癌
*	糖尿病
*	エイズ
*	エイズ

表 11 攻撃者の外部知識

名前 (正識別子)	性別 (準識別子)
鈴木	女
佐藤	女
木村	女
高橋	女
田中	男
本田	男

および表 7 のように，重複開示された属性（age）の値が異なる粒度で一般化されている場合には注意が必要である．

4.4 匿名化の加工方法に応じた再識別リスク指標

最近では撹乱的な加工に対応した再識別リスク指標が提案されている．例えばノイズ付加や PRAM に対して，Pk-匿名性と呼ばれる指標が提案されている [25], [26]．Pk-匿名性は，特定の個人のデータを $1/k$ を超える確率で識別できればリスクが高いと考える指標であり，k-匿名性と等価な匿名性を持つことが証明されている．また，差分プライバシーと呼ばれる指標 [27] と Pk-匿名性との関係についても解析されてきており [25]，撹乱的な加工に対する再識別リスク評価も技術的に可能になりつつある．

5. おわりに

代表的な再識別リスク指標である k-匿名性について，攻撃者の知識，およびパーソナルデータの提供形態や加工方法に応じた再考を行った．攻撃者の知識を抑制することで，より有用性の高い匿名化データの作成可能性を示した．逆に攻撃者が匿名化アルゴリズムを知っている場合や，依存関係のある複数の匿名化データを得る場合は，k-匿名性では検出できない攻撃が存在することを明らかにした．特に k-匿名性を満たしていても元のデータを再構築できる場合がある問題を具体的に示し，その対策方針を提案した．

謝辞 本稿は，2016 年 7 月 28 日に電気通信大学にて開催された 2016 年度第 3 回 PWS 勉強会 (タイトル：「良い k-匿名化と悪い k-匿名化，差分開示の問題」，発表者：南，千田) の発表内容に基づき執筆しました．PWS 勉強会で熱心にご議論頂いた参加者の皆様に感謝いたします．

参考文献

[1] ISO/TS 25237:2008 (en), Health informatics — Pseudonymization, https://www.iso.org/obp/ui/#iso:std:iso:ts:25237:ed-1:v1:en
[2] K. El Emam and L. Arbuckle (木村映善，魔狸，笹井崇司 訳), データ匿名化手法, オライリー・ジャパン, 2015.
[3] L. Sweeney, “Uniqueness of simple demographics in the U.S. population,” LIDAP-WP4, Carnegie Mellon University, Laboratory for International Data Privacy, 2000.
[4] L. Sweeney, “k-anonymity: a model for protecting privacy,” International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, **10(5)**, pp. 557–570, 2002.
[5] A. Narayanan and V. Shmatikov, “How to break anonymity of the Netflix Prize Dataset,” arXive.org. Retrieved 19 January 2014.
[6] 個人情報保護委員会，匿名加工情報に関する委員会規則等の方向性について （平成 28 年 6 月 3 日) http://www.ppc.go.jp/files/pdf/280603_siryou2.pdf
[7] 千田 浩司，吉浦 裕，島岡 政基，「匿名化基準に関する欧米公的文書 7 選の考察」，CSS2016.
[8] 独立行政法人 統計センター，「統計データ開示抑制に関する用語集 改訂版（対訳）」，2005 年 8 月， http://www.nstac.go.jp/services/pdf/skk-yogosyu2.pdf
[9] 竹村 彰通，「個票開示問題の研究の現状と課題」統計数理，**51(2)**，pp. 241–260，2003.
[10] 内閣官房，「技術検討ワーキンググループ報告書（第 5 回パーソナルデータに関する検討会配布資料」，2013 年 12 月 10 日，https://www.kantei.go.jp/jp/singi/it2/pd/dai5/siryou2-1.pdf
[11] L. Willenborg and T. de Waal, “Statistical disclosure control in practice,” Lecture Notes in Statistics, **111**, Springer, 1996.
[12] L. Willenborg and T. de Waal, “Elements of statistical disclosure control,” Lecture Notes in Statistics, **155**, Springer, 2001.
[13] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, R. Lenz, J. Naylor, E. S. Nordholt, G. Seri, and P.-P. De Wolf, “Handbook on statistical disclosure control (version 1.2),” http://neon.vb.cbs.nl/casc/handbook.htm, Jan. 2010.
[14] P. Kooiman, L.C.R.J. Willenborg, and J.M. Gouweleeuw, “PRAM: A method for disclosure limitation of microdata,” Report, Department of Statistical Methods, Statistics Netherlands, Voorburg/Heerlen, 1997.
[15] 藤野友和，垂水共之，「PRAM の理論とその実用上の諸問題」，統計数理，**51(2)**，pp. 321–335，2003.
[16] A. Meyerson and R. Williams, “On the complexity of optimal k-anonymity,” Proc. of PODS 2004, pp. 223–228, ACM, 2004.
[17] C. Aggarwal and P. Yu, “Privacy-preserving data mining: Models and algorithms,” Springer, 2008.
[18] R. C.-W. Wong, A. W.-C. Fu, K. Wang, and J. Pei, “Minimality attack in privacy preserving data publishing,” Proc. of VLDB ’07, pp. 543–554, 2007.
[19] C. Aggarwal, “On k-anonymity and the curse of dimentionality,” Proc. of VLDB 2005, pp. 901–909, ACM, 2005.
[20] 千田浩司，五十嵐大，高橋克巳，濱田浩気，菊池亮，冨士仁，「集合匿名化クラウドの課題と対策」，電子情報通信学会論文誌，**J96-A(4)**，pp. 149–156，2013.
[21] M. E. Nergiz, M. Atzori, and C. Clifton, “Hiding the presence of individuals from shared databases,” Proc. of

ACM SIGMOD, pp. 665–676, 2007.
[22] 菊池 亮,「サンプリングを用いた際の個人識別リスクの評価」, CSS2016.
[23] X. Jin, N. Zhang, and G. Das, "Algorithm-safe privacy-preserving data publishing," Proc. of EDBT '10, pp. 633–644, 2010.
[24] D. Kifer and J. Gehrke, "Injecting utility into anonymized datasets," Proc. of SIGMOD 2006, pp. 217–228, ACM, 2006.
[25] D. Ikarashi, R. Kikuchi, K. Chida, and K. Takahashi, "k-anonymous microdata release via post randomisation method," IWSEC 2015, Lecture Notes in Computer Science, **9241**, pp. 225–241, 2015.
[26] 五十嵐大，長谷川聡，納竜也，菊池亮，千田浩司,「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」，コンピュータセキュリティシンポジウム 2012 (CSS2012)，情報処理学会，2012.
[27] C. Dwork, "Differential privacy," ICALP 2006, Lecture Notes in Computer Science, **4052**, pp.1–12, 2006.