データ通信協会　情報法制研究会
第6回シンポジウム
2017年11月25日(土)
一橋講堂

Innovative R&D by NTT

パネルディスカッション＜匿名加工情報の利活用に向けて＞

k-匿名性に関する疑問に答える

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
高橋克巳

もくじ

- k-匿名性の基本知識
  - k-匿名性とは何か？
  - k-匿名化とは何か？
  - k-匿名化の方法は？
- 匿名加工情報との関係
  - k-匿名性は匿名加工情報に必須？
  - 識別子／準識別子って何？
  - 識別子／準識別子は誰かが決めているのか？
  - k-匿名化の製法を表示するなら？
- k-匿名性の実体
  - 「k=1データ」って何？
  - ところでkの値はいくつがいいの？
  - k-匿名化の対象にしない項目はどうすればいいの？
  - データの性質に応じたリスクと対応
  - k-匿名化したデータの有用性
- 提案

k−匿名性とは何か？

- データの中のどの人も、データ中の他の「最低 k-1人」と区別がつかない

2-匿名性の例

住所	年齢	性別	
千代田区	20	男	
千代田区	20	男	3
千代田区	20	男	
渋谷区	10	女	
渋谷区	10	女	2

k−匿名化とは何か？

- 個人データを「k-匿名性」を満たすように加工すること
- 方法はたくさんある

k-匿名化の方法は？

（雑だが実用的な一例）

① 氏名を削除
② 住所・年齢・性別をほどよく一般化
③ カウント
④ k を満たさないレコードがあれば削除

氏名	住所	年齢	性別
田中	千代田区一ツ橋	22	男
鈴木	千代田区大手町	23	男
山田	千代田区霞ヶ関	29	男
佐藤	渋谷区神南	18	女
山本	渋谷区恵比寿	17	女
山田	豊島区池袋	21	男

①↓ ②↓ ②↓ ②↓

氏名	住所	年齢	性別
*	千代田区	2x	男
*	千代田区	2x	男
*	千代田区	2x	男
*	渋谷区	1x	女
*	渋谷区	1x	女
*	豊島区	2x	男

③

④→

k−匿名性は匿名加工情報に必須？

- 必須ではないが、以下の理由で望ましい
  - データに個人識別性がないと根拠をもって主張できる
  - 他に、使いやすいツールがない

- 無論、k-匿名性によらない匿名加工情報作成もある
  - サンプリング、擬似データ作成、等

識別子／準識別子って何？

- 定義
  - 識別子：項目単独で個人識別性があるデータ項目（氏名等）
  - 準識別子：項目組み合わせで個人識別性があるデータ項目（住所と生年月日等）
- k-匿名性の保護範囲である
  - k-匿名化は、識別子は削除し、準識別子は一般化等をして「k人」いるようにする
    - 上記以外はk-匿名性は何も保証しない
- 「規則19条1号」の対象とほぼ同じ概念
- どの項目を選定するかが匿名化の心臓

識別子／準識別子は誰かが決めているのか？

- 誰も決めていません（–__–;）
- 作成者が決める必要がある

- 『NIIレポート』では限定的な列挙を試み「特定対象項目」と呼んでいる
  a. 氏名以外の基本4情報（住所、生年月日、性別）
  b. 所属する／した会社、学校等名称
  c. 本人到達性のあるメールアドレス、SNSのID
  d. 本人到達性のある電話番号
  e. クレジットカード番号
  f. 単体で特定の個人を識別することができるもの（氏名、顔写真）

k−匿名化の製法を表示するなら？

- k-匿名化の対象項目の明記が重要

表示例の提案　その１

本匿名加工情報は、氏名を削除し、住所・年齢・性別に対するk-匿名化を行いました。

表示例の提案　その２

本匿名加工情報は、氏名を削除し、住所・年齢・性別に対する非識別化を行いました。

識別子　　準　識　別　子

例として
住所・年齢・性別を準識別子に設定した場合

氏名	住所	年齢	性別	購入品名	購入日時	数量	単価
田中	千代田区一ツ橋	22	男	コラコーラ	2017/11/23 15:01	1	130円
鈴木	千代田区大手町	23	男	爆笑ミルク	2017/11/10 17:23	1	189円
山田	千代田区霞ヶ関	29	男	ランチパッコ	2017/11/09 12:34	1	152円
佐藤	渋谷区神南	18	女	クールガム	2017/11/01 10:10	1	141円
山本	渋谷区恵比寿	17	女	西の月	2017/10/30 04:01	1	234円
山田	豊島区池袋	21	男	虹鯛焼	2017/10/29 11:38	1	181円

「k=1データ」って何？

（まぎらわしいので整理）

- 下記のデータの各レコードは、
  - 住所・年齢・性別に対してk=2を満たす
  - データの中で一意（識別可能）である

氏名	住所	年齢	性別	購入品名	購入日時	数量	単価
*	千代田区	2x	男	コラコーラ	2017/11/23 15:01	1	130円
*	千代田区	2x	男	爆笑ミルク	2017/11/10 17:23	1	189円
*	千代田区	2x	男	ランチパッコ	2017/11/09 12:34	1	152円
*	渋谷区	1x	女	クールガム	2017/11/01 10:10	1	141円
*	渋谷区	1x	女	西の月	2017/10/30 04:01	1	234円
*	*	*	*	*	*	*	*

3

2

ところで kの値はいくつがいいの？

- 共通汎用的な答えはありません

- ２、または３でよいという説があります
- ５、または１０がよいという説があります
- ２０以上だとデータが破壊されてしまうという説があります

k-匿名化の対象にしない項目はどうすればいいの？（その１）

（極端な記述がある場合）

- ３行目のダイヤ5億円はどうみても特異なので削除する（規則19条4号）
- ４行目のガム100個はこのデータベースでは目立つので削除する（規則19条5号）

氏名	住所	年齢	性別	購入品名	購入日時	数量	単価
*	千代田区	2x	男	コラコーラ	2017/11/23 15:01	1	130円
*	千代田区	2x	男	爆笑ミルク	2017/11/10 17:23	1	189円
*	千代田区	2x	男	ダイヤモンド	2017/11/09 12:34	1	5億円
*	渋谷区	1x	女	クールガム	2017/11/01 10:10	100	141円
*	渋谷区	1x	女	西の月	2017/10/30 04:01	1	234円
*	*	*	*	*	*	*	*

k-匿名化の対象にしない項目はどうすればいいの？（その2）

（その他の留意）

- データの性質に応じたリスクを考えて対応する

氏名	住所	年齢	性別	購入品名	購入日時	数量	単価
*	千代田区	2x	男	コラコーラ	2017/11/23 15:01	1	130円
*	千代田区	2x	男	爆笑ミルク	2017/11/10 17:23	1	189円
*	千代田区	2x	男	ランチパッコ	2017/11/09 12:34	1	152円
*	渋谷区	1x	女	クールガム	2017/11/01 10:10	1	141円
*	渋谷区	1x	女	西の月	2017/10/30 04:01	1	234円
*	*	*	*	*	*	*	*

データの性質に応じたリスクと対応

リスクの例	具体例	対応の例
(購入品名が)明らかに個人識別に貢献する	野球選手のデータで「バットか？グローブか？ミットか？」を調べると職務がわかる	《k-匿名化対応》 購入品名もk-匿名化の対象とする
参照情報があってマッチングが可能である	「コラコーラ 2017/ 11/ 23 15:01 購入」が記録されたデータが流通している	《ノイズ》 タイムスタンプにノイズを入れて、マッチングできないようにする
潜在的な個人識別性がある	位置情報の履歴を分析すると自宅がわかる	《部分削除》 自宅の位置情報を削除する

k−匿名化したデータの有用性(1)

- 人口データで模擬してみる
  - 総務省　住民基本台帳に基づく人口、人口動態及び世帯数（平成27年1月1日現在）
    - 【総計】平成27年住民基本台帳年齢階級別人口（市区町村別）

1902　全国の市区町村数

405　人口10万人以上の市区町村数

k−匿名化したデータの有用性(2)

- 人口データを元にk-匿名化を模擬
  - 市区町村・性別・年代に対するk-匿名性の評価
  - あるパーソナルデータが人口見合いで偏りなく構成されている仮定で実験
  - 下表のセルがk-匿名性の単位
- 結果（全人口データからNレコードをランダム抽出してk-匿名化を模擬）
  - N=1万レコード → k=2を満たすレコード数12%、k=10は無理
  - N=10万レコード → k=2を満たすレコード数84%、k=10は35%程度
- 考察
  - 少ないレコード数でk-匿名性を実現するのは困難で、日本住民の場合、10万レコード程度確保しないと、満足なk-匿名化ができない

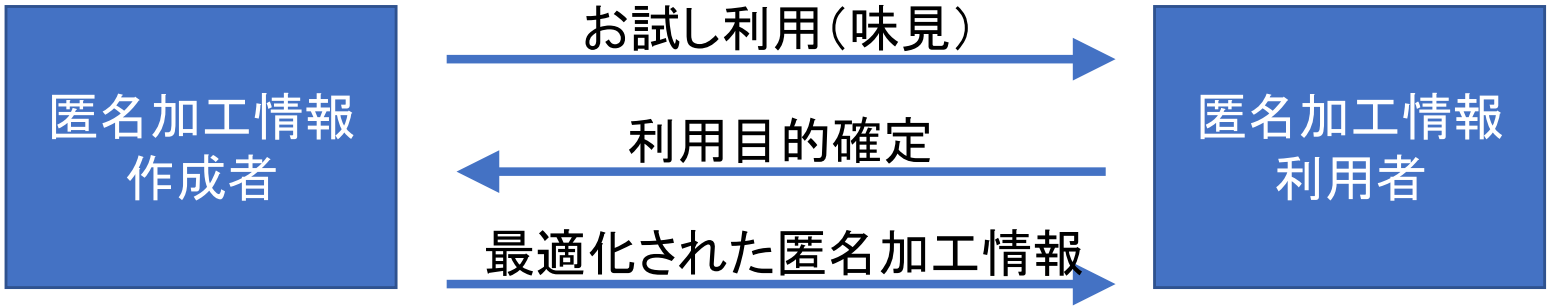
都道府県	市区町村名	性別	0-9歳	10−19歳	・・・	90−99歳	100歳以上
北海道	札幌市中央区	男	4317	3909		87	13
北海道	札幌市中央区	女	4058	3845		419	95
北海道	函館市	男	4302	4781		149	15
沖縄県	八重山郡与那国町	女	40	40		2	0

提案1：匿名加工情報の利用目的

- 匿名加工情報には汎用目的なビッグデータ提供への期待がある
- 他方、匿名加工情報の規律は、データから個人識別性をなくすことである
- 個人識別性をなくすためには、レコードを削除や、個々のレコードの項目の情報量抑制をせざるを得ない（データ破壊）
- それでもデータの有用性を確保するには「不要なレコードや項目は入れない」に行き着く（匿名加工情報の最適化）
- 匿名加工情報の利用目的を決めて作成することが有用なデータ作成に貢献する

提案2：匿名加工情報のお試し利用

- 元々ビッグデータは発見的な情報分析であるという側面がある
  - やってみないとわからない
- 発見的な営みと、目的の事前確定は相入れない
- 「お試し利用」の仕組みが必要である
  - 技術的に解決するならば、リアルな擬似データを作成する技術の研究開発
  - 制度的に解決するならば、各種制限をつけた「お試し」利用の運用可能性検討
  - 業界的には AI/IoT を含んだデータエコシステム設計構築の一環



まとめ

- k-匿名性の基本知識
- 匿名加工情報との関係
- k-匿名性の実体
- 提案
  - 匿名加工情報作成は利用目的を決めてから作成
  - 「お試し利用」の仕組みが必要