Vol.2014-DPS-158 No.4
Vol.2014-CSEC-64 No.4
2014/3/6

実サービスのデータを用いた k-匿名状態の推移調査と，合理的な匿名状態評価指標の検討

小栗 秀暢[†1]　　曽根原 登[†2]

現代において，BigData 分析における個人のプライバシーの扱いは，非常に大きな関心事になっている．そのため情報を匿名化することによって漏洩のリスクを減少させる手段が必要とされている．k-匿名化技術は個人情報の属性値を k-1 個以上に区分することで個人識別性を減少させる技術だが，合理的な匿名状態を規定する指標が存在しないため，実サービスでの適用が難しい．そこで本稿では，実際のサービスに登録したユーザ群に対して一律のユーザ分類を行い，k 匿名状態がどのように推移するかを調査した．サービスを利用するユーザ数と k 匿名状態の関係を明らかにすることで，実サービスで合理的な匿名化を行う際の基準となる評価指標を提案する．

The change investigation of k-anonymity level applying the real services, and examination of the rational anonymity level evaluation index.

Hidenobu Oguri[†1]　　Noboru Sonehara[†2]

Privacy is a major concern in management of Big Data, especially for datasets that contain sensitive personal information.
A model that is widely used to protect privacy is k-anonymity, which can be generally defined as a clustering method in which any record in a dataset is indistinguishable from at least (k-1) other records in the same dataset. The k-anonymity is used as a technique to decrease personal discrimination. But in the present, rational anonymity evaluation index does not exist. Therefore it is difficult applying anonymity technique to real services. We performed a various user classification in the user who enrolled in many services, and investigated how k-anonymity level changed. We calculated and prove the relations of the k-anonymity level and the number of the real service users. We suggest an evaluation index of k-anonymity for applying rational anonymity technique.

1. はじめに

近年の個人情報保護意識の高まりによって，個人情報を保持する事業者が，その維持のために多額のデータ保持費用と漏えいリスクを負うようになった．個人情報の価値は主に拡散した際の被害額から算出されることが多く，日本において 2011 年に発生した個人情報被害総額は 1899 億円と算出されている．[1]

漏えい人数	628万4363人
インシデント件数	1551件
想定損害賠償総額	1899億7379万円
一件あたりの漏えい人数	4238人
一件あたり平均想定損害賠償額	1億2810万円
一人あたり平均想定損害賠償額	4万8533円

図 1　2011 年度のセキュリティ被害について[1]

企業や団体はこれらの漏えいリスクに備えるため，常にセキュリティ設備の保護に追われており，SOC3(米国内部統制レポート基準)，PCIDSS(クレジットカード業界のセキュリティ基準)，ISO27001(情報セキュリティマネジメントシステム)のような厳しいセキュリティ基準を満たすため，多額の費用を投じている．　**図 2** はN社のクラウド価格表から抽出した，各セキュリティオプションの価格を示した例である．これらのコストや安全規定はインターネット上のインシデントが発生するたびに増加していく．

○ 主なセキュリティ維持にかかるコスト		
SSL証明書	¥78,750/年/複数台	N社
システム監視	¥5250/月/台	クラウド
遠隔バックアップ	¥42000/月/TB	価格表より
追加容量	¥11550/TB/月	(2013/12)
PCIDSS 更新	年間1度の更新業務	

図 2　主なセキュリティコスト例

これらの情報の多くは，セキュリティの要件を決定する技術者の側が管理する基準であるため，データ内に含まれるプライバシー情報の漏洩リスクとのバランスによって要件が決められることになる．

このようなセキュリティ技術の側面から考えると，k-匿名化技術などの匿名化技術は，個人が特定，又は識別されるリスクを低減させることで，個人情報とそれ以外の情報を区分し，運用やセキュリティ保持コストを削減するために行うものである．

プライバシー情報の漏洩が問題となる反面，企業が新しい顧客獲得や事業を拡大するためには，自社が保持する情報だけでは分析する範囲が狭く，有為な結果を出すことが難しいという問題がある．そのため，他の事業者の保持するマーケティングデータなどと結合する仕組みとして”Linked Open Data”のような公開可能なデータを相互に結合する仕組みが求められている．

このような分析と必要とするのは，事業を推進する企画

†1 総合研究大学院大学 複合化学研究科 情報学専攻/ニフティ株式会社
The Graduate University for Advanced Studies, School of Multidisciplinary, Informatics Department, Tokyo, Japan. NIFTY Corporation

†2 国立情報学研究所
National Institute of Informatics

Vol.2014-DPS-158 No.4
Vol.2014-CSEC-64 No.4
2014/3/6

者やマーケティング分析者である．彼らは匿名化技術を，情報を公開し他の情報と接続するための技術と考え，そのセキュリティ上のリスクを考慮せず，なるべく有益な情報のみを扱いたいと考える．

セキュリティのリスクとコストを削減することを目的とする集団と，情報接続と分析を目的とする集団の間には，匿名化に対する認識のギャップが存在する．

ビッグデータを適切に匿名化し，顧客の活動状況を分析する活動は，協調フィルタリングによるレコメンドエンジン[2]などに活用できる重要な技術と考えられているが，現状では安全性と有益性の両方を満たすアルゴリズムは存在せず，トレードオフの関係性[9]と考えられている．

本稿では，匿名化処理の妥当性を検討するため，匿名化の安全性を計測する上で一般的な技術である k-匿名化処理を，実データに対して一律で実施し，その匿名状態の変化を計測し，実際のサービス等で活用可能なレベルの匿名化処理の指標を検討する．

2. 過去研究について

まず，匿名化とは，個人情報やプライバシー情報などのパーソナル情報を加工して，他の情報との容易照合性を減少させる処理のことである．

ここでパーソナル情報とは「属性」と「属性値」として表現されるユーザに関する情報であり，あるユーザのパーソナル情報をテーブルのレコードとして表現する．そして，単一の属性ではユーザを特定できないが，複数組み合わせるとユーザを特定できる可能性のある属性の組合せを準識別子(quasi-identifier，QID) と呼ぶ．

また，ユーザを特定された状態で開示されることが望ましくない属性をセンシティブ属性(sensitive attribute : SA)と呼ぶ．この時，もし攻撃者があるユーザの QID の属性値を知っていたとすると，そのユーザのレコードを特定できてしまい，SA の属性値を知られてしまう．これを防ぐために，QID の属性値を一般化して，より抽象的な値にする方法が知られている．そして，QID の属性値によって識別されるレコードが少なくとも k 個以上ある場合，そのテーブルは k-匿名性を満たすという[3]．

k-匿名化を実現するための手法として， Datafly 方式[3]や μ -Argus 方式[4][5][6]などのアルゴリズムが主に使われており，公共データや医療データの匿名化アプリケーションとして提供されている．

それらの k-匿名化手法は，データの出現数に合わせて切り落としや抽象化を行い，データの出現数を k 値以下にする．多くの匿名化アルゴリズムは，上記のような情報の変更の組み合わせにデータを Re-coding し抽象化を行うことで利用者を特定するデータの組み合わせ出現数を k 値以下にすることで成立する．

これらの方式は階層化された選択肢を用いて情報量の減少のみを判断基準とした処理を行う．そのために，実際の匿名化処理を行う際に必要となる情報要素などが失われる可能性がある．Tian と Zhang は，匿名化処理された情報に対して，分析の目的達成に必要な情報が残されており，かつ，匿名化処理が求められたレベルに達しているかを確認する手法を提案している．[11]

図 3 は[11]において情報の有益性と匿名性を確認するための区分表である．ある匿名化されていない情報 *An* を匿名化処理された情報 *E* に変換した場合，*E* が最低匿名レベル *r* を満たしているか(yes/no)と，*E* の一要素 *Ei* が元情報 *Ai* の中の有益性 *S* を満たしているか(yes/no)によって 4 区分 (*gv,gnv,bv,bnv*)に評価し，それぞれに対して処理が変更する仕組みを取り入れることを提案している．

例えば，20 才(*Ai*)の情報を抽象化して大学生(*Ei*)に変更し，最低匿名レベル(*r*)を満たしたが，成年か未成年かをフラグで分類する必要(*S*)がある場合，有益性 *S* を満たしていないため *gnv* となる．この場合，区分の方式を変更して再計算することを推奨している．この方式は k-匿名性だけでなく l-多様性などにも汎用的に適用できる点が優れている．

An EC E		Does $E[i]$ contain values of A_i in S?	
		Yes	No
Does E satisfy r?	Yes	gv	gnv
	No	bv	bnv

図 3　匿名化後のデータに含まれる情報の判定表

これらの手法は，分析の目的が明確である医療データなどの匿名化処理のためには有益だが，探索的な顧客の分析を行うマーケティング分析などには不向きである．

また，一般的な指標として適切な匿名化レベルを設定することができないためにサービス実装が難しい．

3. k-匿名化処理と合理的な匿名化の関係

2013 年 12 月に発表された内閣府の”パーソナルデータに関する検討会”[12]では，非特定情報と非識別情報という定義を新たに作成し，用途に応じた***”合理的な匿名化”***を行うことを推奨している．

だが，非特定や非識別などの概念を全てのプライバシー保護技術に適用することは難しい．例えば，l-多様化技術は定性的な情報の区分方法であることから，これらの議論とは同一視できない．また，個人情報(属性値などのスプレッドシート型のマスタデータ)とパーソナルデータ(特に行動や位置情報などのフロー型トランザクションデータ)を区分して考えておらず，ビッグデータが再識別可能である点を区分して議論されていない．

フロー型のデータの匿名化処理が難しいことは多く論じられているが，スプレッドシート型データの匿名化処理技術は，欠測化・ノイズ付加・観測値の交換・ランダム化等の手法が提案され，それぞれの個人情報保護の効果について研究が進んでいる．本稿では，その中からスプレッド

Vol.2014-DPS-158 No.4
Vol.2014-CSEC-64 No.4
2014/3/6

シート型のプライバシー保護技術の中で，一番基本的な概念である k 匿名化レベルに閉じて問題を整理し，合理的な匿名化状態を満たす条件を検討する．

まず，合理的な匿名化を検討する上で用語と概念の定義を行う．**図 4** は典型的な個人情報とパーソナルデータの組み合わせである．多くはサービス入会時に登録した自分の属性情報と，その後のアクションを記述したフロー情報に分割されて保持されている．

ID	姓	年齢	性別
A001	山田	17	男
A002	岩鬼	17	男
A003	里中	16	男
A004	山田	8	女
A005	殿馬	18	男
A006	夏川	16	女
A007	土井垣	20	男
A008	犬飼	20	男

ID	日付	店舗	内容
A001	11/1	ツタヤ-中野	2001年宇宙の旅
A001	11/1	ガスト-新宿	お得ランチ
A001	11/2	バーミヤン-新宿	ラーメン
A001	11/2	DHC-ネット	精力剤
A001	11/3	ファミマ-新中野	アダルト雑誌
A001	11/3	ニッセン-ネット	スカート
A001	11/4	シダックス-新宿	カラオケ2時間
A001	11/4	マルエツ-中野	弁当
A001	11/4	牛角-新宿	食べ放題

図 4　特定可能・識別可能な情報の例

この状態では個人の名称などが存在するために，個人が特定できる状態にある．そこで特定できる氏名と ID を消去したものが**図 5** である．

ID	姓	年齢	性別
-	-	17	男
-	-	17	男
-	-	16	男
-	-	8	女
-	-	18	男
-	-	16	女
-	-	20	男
-	-	20	男

ID	日付	店舗	内容
Z001	11/1	ツタヤ-中野	2001年宇宙の旅
Z001	11/1	ガスト-新宿	お得ランチ
Z001	11/2	バーミヤン-新宿	ラーメン
Z001	11/2	DHC-ネット	精力剤
Z001	11/3	ファミマ-新中野	アダルト雑誌
Z001	11/3	ニッセン-ネット	スカート
Z001	11/4	シダックス-新宿	カラオケ2時間
Z001	11/4	マルエツ-中野	弁当
Z001	11/4	牛角-新宿	食べ放題

図 5　特定不可能だが，識別可能な情報の例

だが，この**図 5** のスプレッドシート型データには，準識別子の組み合わせによって再識別可能となる情報が存在する．そのため，現状の識別性を減少させるために k 匿名性を確保する処理を行う必要がある．

出現数に応じて属性値を抽象化したものが**図 6** だが，この状態では，フロー情報との接続が可能なため，再識別のリスクが存在している．これを連結可能匿名状態，と定義する場合もある．

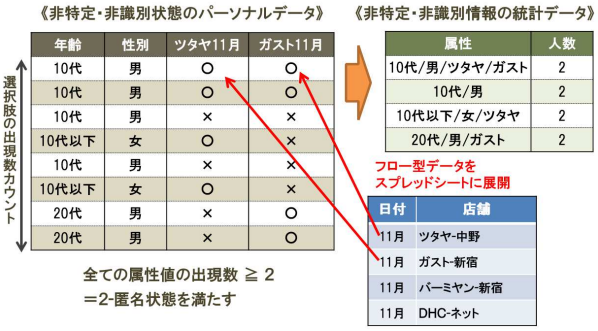
ID2	ID1
B008	A001
B007	A002
B006	A003
B005	A004
B004	A005
B003	A006
B002	A007
B001	A008

ID1	日付	店舗
A001	11月	ツタヤ-中野
A001	11月	ガスト-新宿
A001	11月	バーミヤン-新宿
A001	11月	DHC-ネット
A001	11月	ファミマ-新中野
A001	11月	ニッセン-ネット
A001	11月	シダックス-新宿

ID2	年齢	性別
B008	10代	男
B007	10代	男
B006	10代	男
B005	10代以下	女
B004	10代	男
B003	10代以下	女
B002	20代	男
B001	20代	男

図 6　再識別不可能情報と連結された情報

フロー型データの再識別可能性は，ある程度の量の情報が含まれていた場合でも可能な場合が多い．例として，[13] では JR 東日本の SUICA データを元に再識別可能性を計算した場合，2.237 駅の利用があれば再識別可能性があるとしている．そこで，フロー情報との接続性を消去し，かつ k 匿名性を確保するためには，**図 7** のような形で，出現した情報に対する適切なフラグ形式に落とし込むなどの情報の変換が必要となる．



年齢	性別	ツタヤ11月	ガスト11月
10代	男	○	○
10代	男	○	○
10代	男	×	×
10代以下	女	○	×
10代	男	×	×
10代以下	女	○	×
20代	男	×	○
20代	男	×	○

属性	人数
10代/男/ツタヤ/ガスト	2
10代/男	2
10代以下/女/ツタヤ	2
20代/男/ガスト	2

日付	店舗
11月	ツタヤ-中野
11月	ガスト-新宿
11月	バーミヤン-新宿
11月	DHC-ネット

図 7　特定不可能・識別不可能な情報への変換例

これによって各選択肢の出現数と属性などの関係性を分析することが可能となる．だが，現実的にこのような処理を行う際に，k 値を幾らに設定するか，またはフラグをどのように抽象化するか，などの問題が発生する．安全度を高めるために k 値を高く設定した場合，利用性が失われる場合がある．そのため通常のサービスにおいて可能な匿名化処理レベルを確認し，安全性も確保できる最適な値を検討する必要がある．

本稿では，実際のサービスを用いて属性値の種類と k 値の変化量を調査し，合理的な k 値を導き出すための判断材料を提供する．

4.　k の値と経済指標との関係性

まず，k-匿名化で利用される k の値は，多くの匿名化処理の中で利用されているが，それらの値がどのような数値的な特徴を持つかの定義は余り行われていない．

ある個人情報全体の数量を S とし，それらの情報を k-匿名化処理した際の最小値を k とする．

まず，準識別子が 1 種類しか存在しない場合 (k＝S，k/S＝1.0)がある．それ以外の場合，選択肢の分類は 2 種類以上存在するため，S*1/2 が実質的な最大値となり，かつ S は 1 以上存在する．(1≦k＜S/2，k/S＜0.5)　その中で，2 を最低ラインとして識別確率を 1/k まで低下させる処理が k-匿名化処理である．

この数値を利用性という観点から考えるため，一般的な分析の目的である「マーケティング分析と利用」という点に当てはめる．マーケティングの定義はアメリカマーケティング協会によると”***顧客，依頼人，パートナー，社会全体にとって価値のある提供物を創造・伝達・配達・交換するための活動であり，一連の制度，そしてプロセス***”となるが，ここでは，“顧客の求める商品を提供するために必要なプロセス”と定義する．

Vol.2014-DPS-158 No.4
Vol.2014-CSEC-64 No.4
2014/3/6

k 値とは，マーケティングに用いるならば，顧客をクラスタリング分析した結果，標準分布の両端にあたる．事業者の狙いと比較するならば，対象顧客と比較すると一番イレギュラーな分類である．**(図 8)**

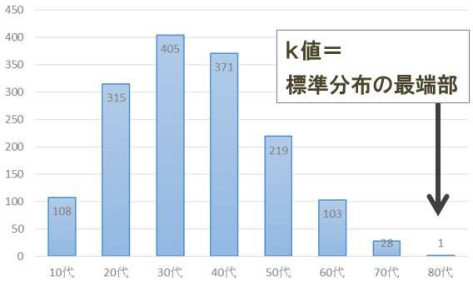


図 8　通常の分析で扱う場合の k 値

例えば，図は 30 代をターゲットとしたあるサービスの顧客分布例である．このような分布の場合，全員が 30 代に集中しているならば対応コストは少なくて済む．だが，ターゲット以外の顧客が多く存在する場合，例えば老年代に向けた文字の大きな表記の追加や，若年層に向けたサービスの拡充などが必要となる．

マーケティングに匿名情報を用いる目的は，個人情報をクラスタリングし，その顧客の求める商品を理解し，該当商品を提供する活動と考えると，k 値とコストと保護レベルで構成される**図 9** のような関係性が成立する．

k 値が低いということは，その顧客に対して別途商品を検討し，提供するコストが増加すると考えられる．

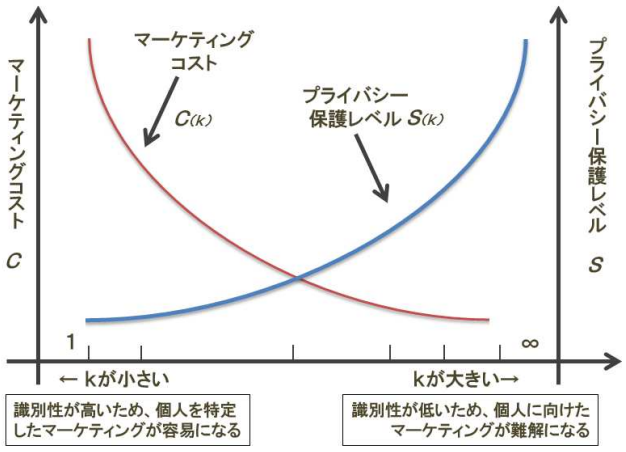


図 9　顧客対応コストと k 値の関係性

逆に，k 値が高いということは，その顧客群が多いため，より効率の良い活動が出来ることからコストが軽減する．

これらのコストの大小は，顧客層の偏りや，全体の数量によっても変化する．より詳細にこれらの値による変化を検討する場合は，k 値以外の要素も含めた指数とし，実験を行う必要がある．

本稿の実験では，セキュリティの標準的な値は k 値であるため，まずは k 値とサービスの関係性を確認した．

5.　実験概要

実験は，N 社で展開する 1635 のサービスに対して，2013 年 10 月に 1 度以上課金決済を行った顧客群に対して行った．顧客群は，対象顧客の中で，性別/年齢/都道府県の 3 つの要素を全て入力している顧客を抜き出し，全体の 40%程度である．それらの顧客属性を一律に分類し，分類方法によって k 匿名レベルがどのように変化するかを検証した．

対　象	数　量
サービス数	1635サービス
ユーザ総数	4,344,922人
平均ユーザ数	2645人
ユーザ数の標準偏差	16291.45
最大ユーザ数	350,527人
最小ユーザ数	1人

表 1　対象となったサービスと顧客群

サービスの区分は，サービスの内容によって区分するのではなく，k 値の特性を確認するため人数による階級を作成した．個人情報の分類方法として，以下の属性に対して抽象化判定樹を作成し，それぞれの値での分類を行った．これら全サービスが全国に提供されているのではなく，特定の地域にしか提供しないものなど，k 値が低くなる要素が含まれるものも多く存在している．

これらのサービス群の顧客全体に対して，上記の区分を実施し，それぞれの組み合わせ(性別 2 区分×年代 3 区分等)も含めて実施した．

実施総パターンは　単体[1+3+3＝7 種]，2 種組み合わせ[(1*6+3*4+3*4)/2)=15 種]，3 種組み合わせ[1*3*3=9 種]の合計 31 種類の情報区分を行い，サービスのユーザ総数と区分数の関係性などを調査した．

登録人数	サービス数	人数	平均ユーザ数
100001人以上	10	1,669,482	166,948
50001～100000人	16	1,147,872	71,742
10001～50000人	36	870,965	24,193
5001～10000人	34	241,927	7,116
1001～5000人	124	266,613	2,150
1000人以下	1415	148,063	105
合計	1635	4,344,922	2,657

表 2　対象となったサービスと顧客群

属性	区分数	分類1	分類2	分類3	分類4	分類5	分類6	分類7	分類8	分類9
性別	2区分	男性	女性							
年代	3区分	未成年	成人	老人						
	5区分	20代以下	30代	40代	50代	60代以上				
	9区分	0代	10代	20代	30代	40代	50代	60代	70代	80代以上
地域	2区分	西日本	東日本							
	9区分	北海道	東北	関東	中部	近畿	中国	四国	九州	沖縄
	47区分	北海道	青森	岩手	・・・	鹿児島	沖縄			

表 3　顧客群に対する区分方式の種類

Vol.2014-DPS-158 No.4
Vol.2014-CSEC-64 No.4
2014/3/6

6. 実験 1：区分数と k 値の関係性

まず，k 値の推移を調査するため，各区分数と k 値の推移を調査した．縦軸は平均 k 値，横軸は区分数となっている．区分数とは，その属性が持つ選択肢の区分数を合計したものである．例えば[性別＝2 区分], [都道府県＝47 区分], [男女 2 区分*年代 9 区分*都道府県 47 区分の組合せ＝846 区分(グラフ内の最端値)]となる．

区分数が 2 までは非常に大きな値が出るのだが，5～9 区分になると極端に値が減少し，そこから同じような数値が続く．特徴が解り難い．そのため k 値という意味は薄れるが，対数グラフにしたものが**図 10** となる．

これを見ると，値に対する大小はあるが，区分数が多くなるほど k 値が小さくなる傾向が見られる．これは想定通りである．だが，10 万人超のサービスと 1～5 万人のサービスでも，10 区分以下になると大きく値が変化しないことは想定外な結果である．

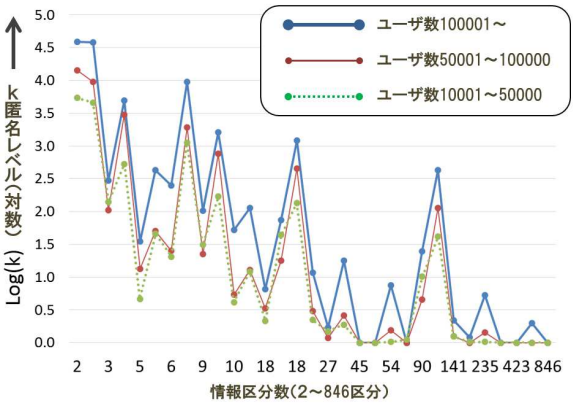


図 10　属性の区分数と k 値の推移(対数)

各増減の状況を見ると特定の区分の際に数値が増減している傾向が見えるため，数学的な区分の大小よりも，区分の方法による増減の差の方が影響が大きいと考えられる．グラフの中で急激に上がっている箇所は，多くが“地域”の属性が含まれている箇所である．”地域”は 47 区分を行った場合でも万遍なく存在し，標準偏差が小さいことが判明している．

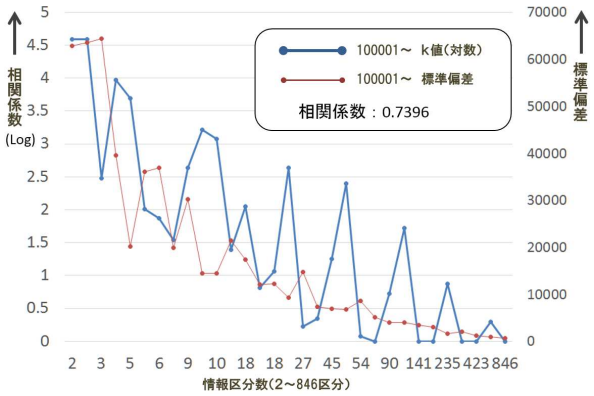


図 11　属性の区分数(対数)と標準偏差平均の比較

そのため，標準偏差との比較を行ったところ，相関係数 0.74 と比較的高い数値が出た．標準偏差と k 値の増減には一定の関係性があると考えられる．**(図 11)**

各サービス数の階級ごとの相関関係を求めたものが**図 12** である．k 値の整数で求めた相関係数は，サービスの登録者数が少なくなるほど相関が少なくなる．対数の相関でみると，その関係性がより顕著となる．

また，対数の比較として，各サービス群を対象に近似直線を引いたところ，ユーザ登録数が多いサービスの方が k 値が急激に減少する傾向にあることが解った．だが，k の対数値の減少率は，各サービス群を比較しても[-0.10±0.02]に収まっており，サービスの登録人数の多寡と k 値の減少率には関連性が薄く，人数よりも区分の方法，特に標準偏差が大きな意味を持つことが解った．

▽k値(整数)の相対関係

	100001−	50001−	10001−	5001−	1001−	−1000
100001−	1					
50001−	0.974	1				
10001−	0.997	0.986	1			
5001−	0.989	0.995	0.997	1		
1001−	0.988	0.961	0.987	0.980	1	
−1000	0.916	0.856	0.905	0.888	0.960	1

▽k値(対数)の相対関係

	100001−	50001−	10001−	5001−	1001−	−1000
100001−	1					
50001−	0.977	1				
10001−	0.971	0.984	1			
5001−	0.941	0.981	0.983	1		
1001−	0.873	0.905	0.948	0.957	1	
−1000	0.774	0.797	0.866	0.880	0.946	1

表 4　選択肢の区分数と k 値の推移

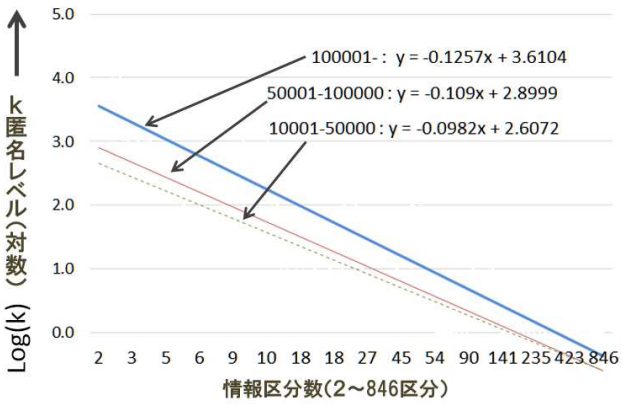


図 12　k 値(対数)の線形近似値の比較

7. 実験 2：k 値と区分方式の関係性

前項の実験は，属性を区分する数に対しての関係性を確認したが，本項は属性値を判定樹で抽象化した際に変化する値について検証した．性別と年齢の情報を抽象化していき，区分数が変化した場合に発生する変化の確認を行った．

〇k-匿名レベル比較パターン
-6 分類：(男,女)*(未成年,成人,老人)
-10 分類：(男,女)*(20 代以下,30 代,40 代,50 代,60 代以上)
-18 分類:(男,女)*(0 代,10 代,20 代,30 代,40 代,50 代,60 代,70 代,80 代以上)

18 分類は，元となった情報と同じものであり，10 分類は，情報を一度分析し，両端値をすそ切りして最適と思われる区分を人力で探索した分類である．

18 分類はデータの詳細さが一番高い群である．これは，サービスの要件定義を行う際に，ある程度不必要と考えられる階層も確認できるように区分されている．

この 18 分類を行った場合の 10 万人超のユーザを持つサービス群での k 値は 6.6 と小さな値となった(**図 13**)．10 万人超サービスの登録ユーザ数平均値は 166,948 人であるため，18 区分した場合の理論的最大値は 9274 人(1/18=5.6%)であり，6.6 はその 0.07%の値である．

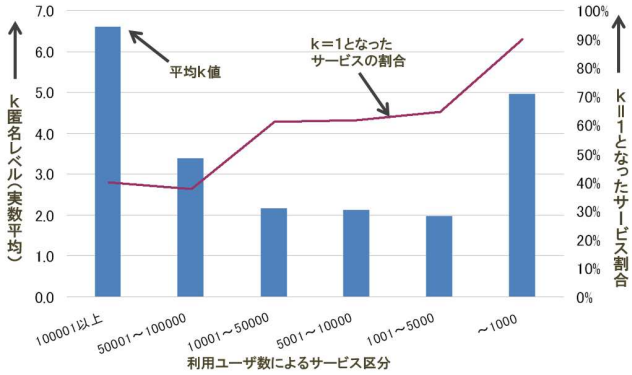


図 13　18 分類における k 値平均と，k=1 サービス率

また，10 万人超のサービスの場合でも 40%のサービスで k=1(匿名化できない)状態となった．1000 人以下のサービスにおいては 90%以上のサービスが k=1 となった．つまり通常のサービスで利用される顧客区分をそのまま用いた場合，k 値を高く維持することは難しいことが解る．

同様の分析を 6 分類/10 分類で行った結果が**図 14/図 15** となる．6 分類は年代情報をほぼ最大限に抽象化し，簡単な分析にしか利用できない形としたものであるため k 値は 18 分類と比較すると大きいが，人間が最適化した分類よりも k の値は低い．多くのサービス郡において，k 値平均は分類数の多い 10 分類の方が，6 分類の 10 倍以上となっている．安全性という面においても，情報の詳細さという面においても 10 分類は 6 分類よりも優れていると考えられる．

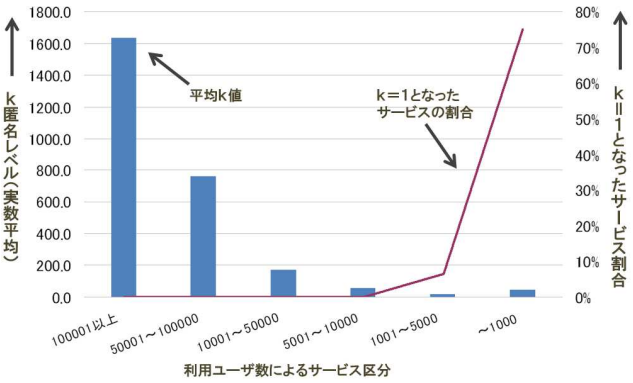


図 14　10 分類における k 値平均と，k=1 サービス率

10 分類の特徴として，1001-5000 ユーザのサービスまでの全てのサービスで k≧2 が維持されており，非常にユーザ数の少ないサービスにおいても匿名状態が維持できていることが解る．

セキュリティを重視する技術者ならば，この情報は情報の詳細性も維持されている上に安全性も格段に高い．そのためにこの方針で匿名化処理を進めることは妥当と考える．だが，この 10 区分の問題点は，k 値を高めるために”**20 代以下**”という属性を使用している点にある．サービス登録者として少数である 10 代を排除するために行った措置のため，安全性の観点から作成した区分であり，利用性の減少については配慮されていない．

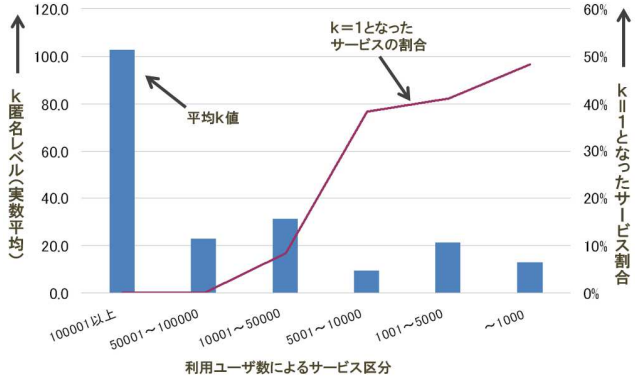


図 15　6 分類における k 値平均と，k=1 サービス率

そこで，k 値だけの指標に加えて我々が過去に行った SEM(検索エンジン広告)の価格による k 匿名化の評価[14]という観点を追加して比較した(表 5)．2013 年 8 月における SEM 価格によって評価した属性値の価格表によって，最高値をつけた概念に収束させた区分”12 分類”を作成し，他の区分と比較した．

12 分類 : (男，女)*(未成年，成人，30's，40 代，50's，老人)

元データ	元単価	変更データ	変更単価	単価の変化率
10代	¥27	10代→未成年	¥195	722%
20代	¥46	20代→成人	¥87	189%
30代	¥76	30代→30's	¥100	132%
40代	¥203	40代→40代	¥203	100%
50代	¥124	50代→50's	¥1,855	1496%
60代	¥266	60代→老人	¥520	195%
70代	¥51	70代→老人	¥520	1020%
80代	¥0	80代→老人	¥520	-

6区分に収束
未成年
成人
30's
40代
50's
老人

表 5　各属性値の変更パターン図

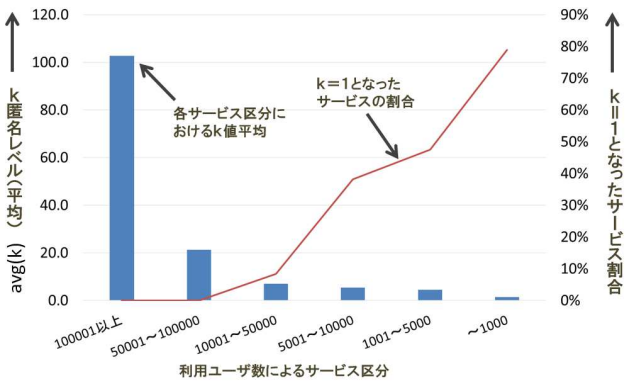


図 16　12 分類(SEM) k 値平均と，k=1 サービス率

12 分類は 50001-100000 人のサービス群までは，6 分類とほぼ同じ k 値を維持しているが，それ以下のサービス群で

Vol.2014-DPS-158 No.4
Vol.2014-CSEC-64 No.4
2014/3/6

はｋ値，及びk=1 サービス率が高くなっている．**(図 16)**

この４種類の分類方法について，100001 人超のサービスについてのｋ値と，その SEM 価格による指標(各属性の人数*SEM 単価の総額)を比較したものが図となる．

この比較によって，当初は最適値と考えられていた 10 分類については，ｋ値は高いが SEM 価格総額による指標は低いと考えることができ，安全性と組み合わせて評価することが可能となる．**(図 17)**

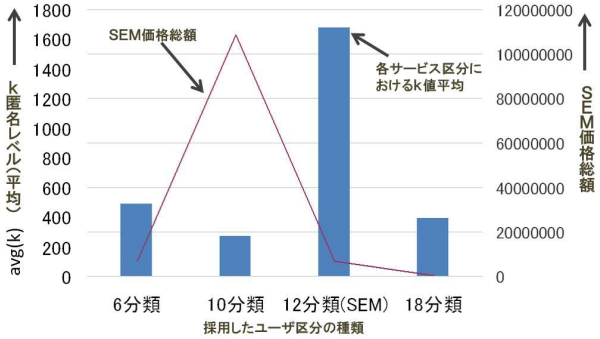


図 17　10 万人超サービスにおける分類種類ごとのｋ値と SEM 価格総額比較

	SEM総額	k 値平均	詳細性（順位）	安全性（順位）	SEM価格（順位）
6分類	32388746	102.8	4	2	2
10分類	18060198	1631.0	3	1	4
12分類(SEM)	112009307	102.8	2	2	1
18分類	25964414	6.6	1	4	3

表 6　10 万人超サービスの各分類の値と相対的な順位

8. 実験 3：国勢調査の相関とｋ値の分布調査

今回対象とした情報群の中で，地域分布については日本の国勢調査(2010 年)との相関関係が比較的高いことが判明した．そこで，基準値として日本国内の人口分布を用いて，その値との相関係数を計測．相関関係とｋ値に関連性があるかについて調査した．

○国勢調査の相関係数とｋ値の調査対象
-地域 9 分類：(北海道,東北,関東,中部,近畿,中国,四国,九州,沖縄)
国勢調査との相関 0.934

対象サービスのｋ値を国勢調査との相関係数で並べたところ，相関係数 0.93 までのサービス群はｋ値が 200 以上と高い値で推移したのに対し，それ以下の相関係数になるとｋ値が多く 1 で推移する．これはある程度サンプルを多く入れ込んだ場合でも傾向は同じである．**(図 18)**

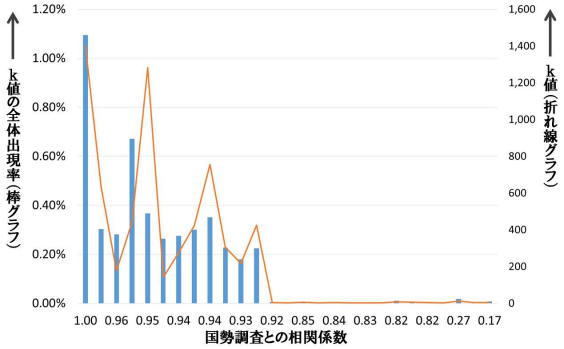


図 18　地域 9 分類のｋ値と国勢調査との相関係数- 50001 人以上の 26 サービスでの調査

50001 人以上の大規模サービスでの数字ではなく，上位 100 サービスでの分布状態を調査したものが下記の図である．相関係数が 0．9 以上のサービスにｋ値が高いものが集中しており，相関係数が低いものについては殆どｋ値を維持することができなくなっている．**(図 19)**

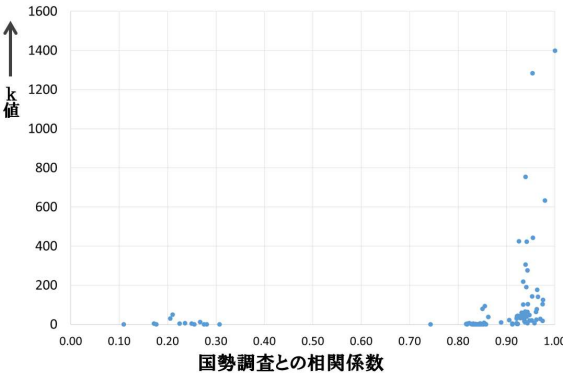


図 19　地域 9 分類のｋ値と国勢調査との相関係数- 上位 100 サービスでの調査

上記２つの調査は，元が同じ顧客群から生成されている情報であるため，似た数値となることはある程度想定できる．本情報の中には”東日本”のみで展開されているサービスなど国勢調査と反する結果を持つものが多く存在する．

匿名化するための区分を決定するアプローチとしては，国勢調査やオープンデータと情報の区分の方法を同一にして相関係数を取得することで，情報を初期の段階でスクリーニングすることが可能であると考えられる．

9. まとめと今後の課題

実データを用いた実験で判明した知見は以下の通り．

○実情報を区分する際に，年代/地域などの一律的な区分を用いると，ｋ値は低くなる．10 万人超のサービスであっても 18 区分を行うと平均ｋ値は 6.6 となり，理論的最大値との比較では 0.07%の値であった．
○サービス利用者が多いほうがｋ値が大きくなる傾向があることは明白だが，サービス利用者で区分した中でも傾向の相関は高い．サービス利用者区分のｋ値平均値の相関係数は 0.856-0.997 の間にあり，ｋ値平均値の対数の場合は 0.774-0.983 に収まる．

○情報の詳細性と k 値には緩やかな相関が見られるが，単純な区分数で並べると数値のばらつきが大きい．むしろ，標準偏差との相関が 0.74 と強い．
○出現率の低い情報のすそ切り/丸めなどの処理を行うことで k 値を高く維持することができる．だが，出現率だけに着目した分類では利用性が低くなる場合がある．その場合, SEM 広告費などの指標を用いることで利用と安全性のバランスを確保することができる．
○比較するべき基礎情報がある場合は，その情報との相関係数によって k 値を高く維持できる区分かの判断が簡便にできるようになる．

上記の内容を実施したうえで，k 値のや安全性の指標について検討した意見としては，サービスの値に関わらず，k 値は 10 以上に維持するには，ある程度の顧客属性への配慮/データ処理作業が伴うと考えられる．

サービス登録者数の多寡による違いはあるが，データを扱う技術者がある程度のプライバシーへの配慮を行ったかどうかの基準として，k＝10 程度を基準に考えることは有意義であると考える．また k 匿名レベルはマーケティングの観点で換算すると両端の外れ値であることから，全体の傾向などを分析する情報としては不要である．そのため，分析初期の段階で裾きりなどを行って処理することは安全性と利用性の両面の観点から望ましい．これ以上の詳細な情報が必要な場合は，目的を明確化して適応的に必要な数値だけを抽出する必要があるため，悪用を避けるために利用者履歴等を管理する方法を採用すべきと考える．

だが，今回の実験は主にひとつの顧客群をベースにして作られた群であるため，本当に多様なサービスとの関連性は判明していない．より多くの事業者の情報とこのような匿名化情報を比較検討するプラットフォームが必要とされるだろう．

また，今回の情報は個人情報の数ある属性の中から，年齢・性別・地域だけを利用している．実際には多様な属性値の中から，有意義な抽象化レベルを探し，売り上げ情報などとの相関性を調査することで，利用性と安全性の両立が可能となる．

だが，今回行った属性区分でも 31 種類だけである．実際に存在する属性は 47 種類*抽象化パターンだけ，指数的に増加する．さらに，実際には今回実験で利用したパターン以外にも様々な分析観点が存在する．どのような値の分析を行うかまで遡って検討する必要がある．

属性ごとの抽象化パターンは無数に存在する上に，その価値評価のための指標も現在は確立されていない．各数値の評価を行いながら最適な匿名化を提供する，仕組みを提供することが今後の課題である．

我々は，種々の属性の変化パターンとその SEM 価格や検索数などの量に応じた適応的匿名化サービスの開発を進めている．このような仕組みを通じてユーザ属性の区分を簡便に決定することができることで，技術者や分析者にとって有益な情報利用が推進されるだろう．

参考文献

1) 2011 年情報セキュリティインシデントに関する調査報告書, NPO Japan Network Security Association(JNSA)日本ネットワークセキュリティ協会 セキュリティ被害調査ワーキンググループ
2) 本多 克宏,個人情報のクラスタリングによる匿名化と安心･安全な推薦システム (特集 安全社会における情報科学の役割), ケミカルエンジニヤリング 58(3), 188-192, 2013-03
3) Sweeney, L, k-anonymity: a model for protecting privacy, Int. J. Uncertainty, Fuzziness and Knowledge-Based Systems, Vol. 10, pp. 557. 570 (2002)
4) Mitsubishi Research Institute, Inc. 情報技術研究センター 松崎和賢，データ匿名化の現状に関する一考察. 医療・統計分野を中心とした国内外の動向．2011-7-8
5) 日本情報経済社会推進協会(JIPDEC)，パーソナル情報の利用のための調査研究報告書，2011-3
6) Anco J. Hundepool and Leon C. R. J. Willenborg, Statistics,m-and t-ARGUS: Software for Statistical Disclosure Control,Record Linkage Techniques,1997
7) Josep Domingo-Ferrer, Francesc Sebe and Agusti Solanas, A polynomial-time approximation to optimal multivariate microaggregation. Comput. Math. Appl. , 55(4):714-732, 2008.
8) El Emam K and Dankar FK and Issa R and Jonker E and Amyot D and Cogo E and Corriveau JP and Walker M and Chowdhury S and Vaillancourt R and Roffey T and Bottomley J, A globally optimal k-anonymity method for the de-identification of health data, September--October 2009
9) Daniel C. Barth-Jones, How should we understand re-identification risks under HIPAA?, 2011
10) Kristen LeFevre David J. DeWitt Raghu Ramakrishnan, Incognito: Efficient Full-Domain K-Anonymity, SIGMOD '05 Proceedings of the 2005 ACM SIGMOD international conference on Management of data, Pages 49-60, 2005
11) Hongwei Tian and Weining Zhang,Privacy-Preserving Data Publishing Based on Utility Specification ,Social Computing (SocialCom), 2013 International Conference,114-121,8-14 Sept.2013
12) 首相官邸 高度情報通信ネットワーク社会推進戦略本部(ＩＴ総合戦略本部)パーソナルデータに関する検討会,技術検討ワーキンググループ 報告書,2013-12-10
13) 首相官邸 高度情報通信ネットワーク社会推進戦略本部(ＩＴ総合戦略本部)パーソナルデータに関する検討会 菊池浩明,匿名化レベルの分類について,2013-10-17
14) H.Oguri and N.Sonehara,A K-Anonymity Method Based on SEM (Search Engine Marketing) Price of Personal Information, Social Computing (SocialCom), 2013 International Conference, 1011-1015, Sept.2013