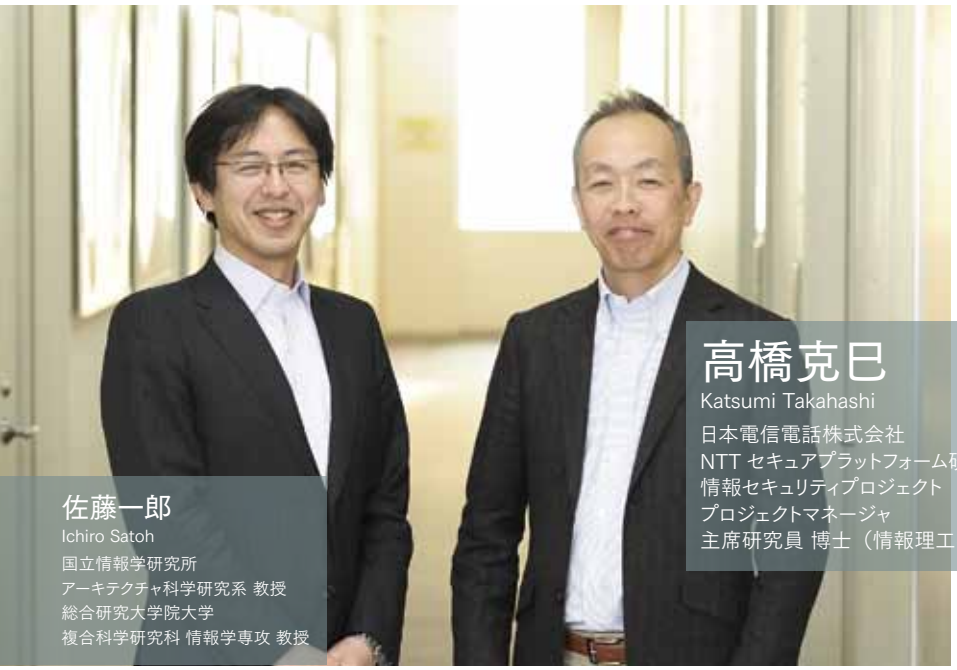
データ活用の利便性と安全性を両立させる

匿名化技術の最新動向とその課題

That's **Collaboration** 2

パーソナルデータの活用に際して、プライバシー保護を実現していくために不可欠となる匿名化技術。現在、匿名化技術はどこまで進化しているのか。そして、実運用にあたり、どのような課題が浮上しているのか。本稿では、匿名化技術の最新動向と、実用化に向けた展望と課題について、政府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（以降、技術WG）のメンバーであるNTTセキュアプラットフォーム研究所の高橋克巳氏に、NIIの佐藤一郎教授（技術WG主査）が話を聞いた。



佐藤一郎
Ichiro Satoh
国立情報学研究所
アーキテクチャ科学研究系 教授
総合研究大学院大学
複合科学研究科 情報学専攻 教授

高橋克巳
Katsumi Takahashi
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
情報セキュリティプロジェクト
プロジェクトマネージャ
主席研究員 博士（情報理工学）

ビッグデータ時代の匿名化とは

佐藤　まず、今後のビッグデータの利活用の鍵となる「匿名化」についてお聞かせください。

高橋　匿名化の基本的な定義ですが、データに含まれる名前や生年月日、住所といった情報を削除したり、変更を加えたりすることで、個人を特定できないようにすることを言います。つまり、データの「匿名性」を高めるための技術や手法、あるいはそれらの組み合わせを匿名化と呼ぶわけですね。しかしビッグデータ時代を迎え、世に多くの情報が蓄積されるようになったなか、単に名前や住所を削除しても、他のデータとの突き合せを行うことで個人が特定されてしまうといった問題が指摘されています。

ビッグデータ活用で匿名化が必要とされるのは法制度解釈上の事情からで、その背景には、さまざまな制約を伴う個人情報を匿名データにして自由に使いたい、というニーズがあります。しかし、個人情報を個人情報でない状態に加工することは技術的には容易ではないケースが多々ある。匿名化技術を使えばいいじゃないかと言われるが、そこには匿名化技術への理解不足があります。データに含まれる名前を削除するなど、「単純な」匿名化は比較的簡単にできる。これが世の中でよく言われている「匿名化」ですが、それだけではデータに含まれる個人が誰であるかわかってしまうリスクは排除できないのです。

佐藤　匿名化というのは単純かつ一律にできるわけではない、ということですね。

高橋　ええ。一方、データに含まれる個人性をほぼ無くしたような状態に加工することは可能です。統計データがこれに相当しますが、これだとデータの抽象度が高くなりすぎて、分析対象データにはならない場合がある。これが、パーソナルデータの匿名性と有用性のジレンマです。

データの利用価値を損なうことなくプライバシーを確保する技術

佐藤　ジレンマの克服のためには、何を目指せばいいのでしょう？

高橋　分析目的に応じて、必要最小限の情報だけを残すことが理論上の目標になります。「技術WG」では、匿名化の基本的な考え方として、「汎用的な匿名化は存在しない。種類・特性・利用目的等に応じケースバイケース」と表現しました。

このような状況においては、個人情報であるか否かということは、必ずしも根源的なことではなくなるでしょう。むしろパーソナルデータ

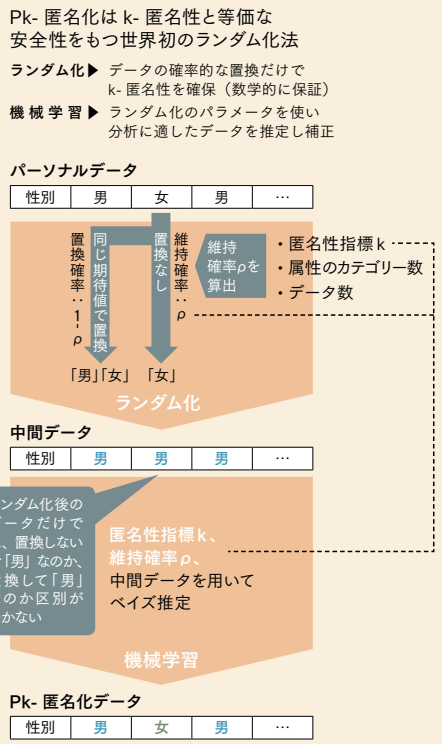
［図1］k-匿名化の例

会員番号	生年月日	住所	年齢	購買品
1001	1979.04.01	東京都中央区A町	34	パン
1002	1986.12.10	神奈川県横浜市A町	26	漫画
1003	1974.10.10	東京都渋谷区B町	38	アイス
1004	1991.05.05	神奈川県鎌倉市B町	22	文庫
1005	2006.11.10	埼玉県川越市A町	17	コーラ
1006	1990.02.06	神奈川県厚木市C町	23	時刻表
1007	2003.08.15	埼玉県浦和市B町	19	牛乳
1008	2000.09.30	埼玉県大宮市C町	9	お茶
1009	1983.01.01	東京都練馬区C町	30	弁当
1010	1994.07.07	埼玉県与野市D町	18	水

k-匿名性（k=3）を満たした状態

会員番号	生年月日	住所	年齢	購買品	
1001	1979.04.01	東京都	30代	食品	3
1003	1974.10.10	東京都	30代	食品	
1009	1983.01.01	東京都	30代	食品	
1002	1986.12.10	神奈川県	20代	書籍	3
1004	削除	神奈川県	20代	書籍	
1006	1990.02.06	神奈川県	20代	書籍	
1005	2006.11.10	埼玉県	未成年	飲料	4
1007	2003.08.15	埼玉県	未成年	飲料	
1008	2000.09.30	埼玉県	未成年	飲料	
1010	1994.07.07	埼玉県	未成年	飲料	

［図2］Pk-匿名化の概念図



の性質と行いたい分析の目的に応じて、最善のデータ処理方法をとることが重要になります。そしてプライバシー上のリスクを情報の提供者に正確に表現して、活用側への制度的な担保も含め、安全なデータ処理ができるようにする必要があります。

佐藤　技術としては、どのような方法があるのでしょうか？

高橋　例えば、分析の目的として個人情報のまま取り扱いたいのであれば、暗号が役に立ちます。我々は、プライバシーの保護と安全管理措置を実現できるよう、暗号化したままデータ処理を行う「秘密分散・秘密計算技術」に関する研究開発を進めてきました。これらの技術群はデータの正確さを損なうことなく、その秘匿性を極限まで追求するものです。

一方、匿名化をとるのであれば、プライバシーのリスクの表現パラメーターである「k-匿名性」があります。k-匿名性とはデータの匿名性を評価する指標で、「同じような属性の人が、必ずk人以上いる状態」のこと。例えば、年齢が20代でも、30代でも、対象とする属性に含まれる人が少なくとも10人以上いる場合、このデータの匿名性は「k＝10」と表現されます。つまり、kの数値が大きいほどプライバシーリスクは小さくなります。なお、k-匿名性を実現するパーソナルデータの加工技術がk-匿名化です。同じ属性の組み合わせをもつレコードを少なくともk個存在するよう、属性の値を粗くする一般化や、希少な人のデータの削除によって「k-匿名化」が実現できます（**図1**）。

佐藤　NTTセキュアプラットフォーム研究所では、さらに進んだ取り組みをされていますね。

高橋　我々のチームが新しく開発したのが、「Pk－匿名化」という手法です。Pk-匿名化は、個々のデータを確率的に変化させる処理である「ランダム化」※1を行って、データが誰のものであるのかをわからなくします。本手法のランダム化は誰のレコードであるか1/k以上の確率で当てることができないように制御するというものです。この性質をPk－匿名性（確率的なk-匿名性）と呼んでいます。その後、「ベイズ推定」※2と呼ばれる機械学習の手法を用いることで、データの元の状態を推定する処理を行います（**図2**）。このことで、実用的な分析が可能な有用な匿名データが作成できます。これは、実パーソナルデータに基づく疑似パーソナルデータとも言えるでしょう。Pk-匿名化はk-匿名性と同様の性質をもちながら、ビッグデータの匿名化に有効だと私たちは考えています。

佐藤　Pk-匿名化は、どんなデータにも活用できるのですか？

高橋　とくに、多数の項目をもつ「長い」パーソナルデータの匿名化に有効です。長いデータの匿名化では、データの必要な項目を選択して匿名化を行う「オーダーメイド匿名化」が繰り返されることが多いと思いますが、複数の匿名化データを突き合わせることでk-匿名性が損なわれる可能性が指摘されていました。対して、Pk-匿名化はそうしたk-匿名性の喪失にも耐性をもつため、繰り返しオーダーメイド匿名化ができ、多項目のパーソナルデータに関しても、プライバシーを保護しつつ利用価値の高い分析用データに加工できるのが特長です。

技術と制度の両面でプライバシーを保護

佐藤　今後、個人情報保護法の改正等に伴い、さまざまな場面でk-匿名性が大きく関わってくると考えています。データの第三者提供の許諾等に際して、k-匿名性がその判断の指標の1つとなるでしょうか？　その場合、kの値がいくつだったら、安全になるということは言えるのでしょうか？

高橋　それは難しい問題ですね。実際にkの値だけで安全性を判断できるのかという面もありますし、k=1でもリスクがほとんどないケースもあるでしょう。kの数値の妥当性は、今後、法制度の整備を含めた社会的な運用によって定まっていくでしょう。

佐藤　これからの匿名化データの適正な利活用にあたっては、技術だけでなく制度も含めた両輪で回していくことが必要というわけですね。

高橋　個人情報については、法制度に基づき取り扱えばいい。また、明らかに特定の個人と切り離されたデータは、第三者も含め自由に使うことができる。これからの課題は、この2つ以外の「匿名化データ」の扱いです。その提供や利用に際して、社会的なルールを策定していくことで、プライバシー保護が可能になると考えています。プライバシー上のリスクがあるデータは、信頼できる相手にしか提供できません。すなわち、定められたルールを守っている相手であれば許可するけれども、ルールを守れなかった場合には、許可を取り下げたり罰したりするといった法制度を策定、運用していくことで、匿名化データをとりまく課題も徐々に解決していくのではないでしょうか。

（取材・文＝伊藤秀樹）

※1 ランダム化　データ分析に影響を与えない程度に元データにノイズデータを付加する匿名化手法。
※2 ベイズ推定　観測された事象から、それの原因となる事象の確率を推定するための確率論的方法。