# 安全な社会をデジタルでストレスなく守るセキュリティ

## Digital Security Systems Protecting Society from Potential Threats

● 山岡　裕司

**あらまし**

近年，データ活用がもたらす価値が注目されている。世界のビッグデータ市場は2桁成長が続き，2020年までに約20兆円の規模に達すると言われている。それに合わせて，適正なデータの流通・活用を推進するための法整備が世界的に進み，その法律に対応するための技術も実用化され始めている。しかし，個人がリスクの高さに気付かずにデータの流通に同意してしまったり，事業者が匿名性の低いデータを流通させた場合にプライバシー問題を起こし，損害賠償などの大きな損失を被ってしまったりするなど，パーソナルデータの流通におけるリスクが判断できないことによる不安の声もある。これに対して，富士通研究所ではパーソナルデータが暴露するプライバシーのリスクを定量化・金額化する技術を開発した。また，従来不足していた匿名加工後のデータの特定性(匿名性の低さ)を算定するモデルを開発し，実際のデータに適用できることを確認した。更に，特定性を高速に算定する技術を開発し，一般的な性能のパソコンを使用した場合，100万人規模のデータセットを約1時間で算定できる十分な実用性を確認した。

本稿では，パーソナルデータのリスク評価を可能とする技術と，それによってデータの価値をより引き出せる社会を実現する構想について述べる。

**Abstract**

As the value brought about by data utilization is attracting attention, the global big data market has been making double-digit growth and is estimated to reach a scale of about 20 trillion yen by 2020. In line with this trend, development of laws for promoting proper data distribution and utilization has progressed on a global basis, and technologies to meet the needs of those laws are beginning to come into practical use. However, there are also anxieties expressed arising from the inability to make decisions regarding risks in personal data distribution; individuals may agree to data distribution without realizing how high the risk is, or business owners may cause privacy issues by distributing data with low anonymity, possibly resulting in compensation for damages and other major losses. To deal with these issues, Fujitsu Laboratories has developed a technology to quantify privacy risks from personal data disclosures in terms of monetary value. We have also developed a model for calculating the specificity (how low anonymity is) of data after anonymity processing, which was insufficient in the past, and confirmed that these are applicable to real data. Furthermore, we have developed a high-speed specificity calculation technology that allows for the calculation of data sets on a scale of 1 million people in about an hour with an ordinary PC, confirming adequate practicability. This paper describes the technology that allows for risk evaluations regarding personal data and the concept of realizing a society that can better extract the value of data by utilizing this technology.

## ま え が き

近年，クラウド，AI（人工知能），IoTなどの技術の進展に伴い，データ活用がもたらす価値が注目されている。

世界のビッグデータ市場は年率11.9％の成長が続き，2020年までに約20兆円の規模に達すると言われている。[1] 特に，パーソナルデータは2011年に世界経済フォーラム[2]で「新しい石油，通貨」と例えられ，その価値が注目されている。そういった動向に鑑み，データ流通・活用が適正に行われるようにするための法整備が世界的に進んでいる。

日本では，超少子高齢社会における諸課題の解決にデータ活用が有効であるとされ，データ流通の拡大を狙った法整備が行われている。2015年には個人情報保護法が改正（2017年に全面施行）され，匿名加工情報（定められた規則を守って特定の個人が識別できないように加工したパーソナルデータ）を自由に活用可能とする制度が制定された。2016年には官民データ活用推進基本法が施行され，国民の権利権益を保護しながらデータの流通・活用が官民で推進されている。

欧州では，個人データの保護に対する権利を基本的人権と位置付け，それを保護する法律であるGDPR（General Data Protection Regulation）が2018年5月25日より適用された。この法律のもとでは，個人データの収集および利用目的について，原則的に本人の同意を明確に取得しなければならない。

そうした中，富士通はお客様の持つデータから価値を最大限に引き出す「つながるサービス」[3]を推進している。価値を最大限に引き出すためには，しばしば他組織のデータと組み合わせてデータを分析する必要がある。例えば，ある場所への出店効果を予測したいという要求に対して，その場所周辺の消費傾向などの情報を用いてお客様のデータを分析できれば，予測精度を上げられる。

消費者の年代，性別，時間帯，消費額，人数などは，その場所周辺の他店から入手できると望ましい。そのような情報授受を円滑にするために，データの売買などによって組織間でデータを流通させる場が必要となり，富士通はその実現にも取り組んでいる。

組織間のデータ流通において，パーソナルデータを取り扱う際には，法規制への対応が必要になる。日本を含む多くの国で合法となるパーソナルデータの流通方式には，本人から取得した同意の範囲でのみ流通させる方式（同意取得方式）と，個人を特定できないデータに匿名加工して流通させる方式（匿名化方式）がある。これらの方式は，データ活用の目的に応じて使い分けられる。

富士通は，これまで両方式の技術開発に取り組み，データ流通の拡大を先導している。同意取得方式では『FUJITSU Cloud Service for OSS「Personiumサービス」』[4]を，匿名化方式では「FUJITSU ビジネスアプリケーション NESTGate 匿名化」[5]や「FUJITSU Software Symfoware Analytics Server」[6]を提供している。

しかし，一部の個人や事業者からはプライバシーに対する不安の声が挙がっている。例えば，個人がリスクの高さに気付かずにデータ流通に同意してしまう不安や，事業者が匿名性の低いデータを流通させた場合にプライバシー問題を起こし，損害賠償などの大きな損失を被ってしまう不安である。このように，リスクの程度が判断できないことが，漠然とした不安を生じさせている。

この不安を解消し，データの流通を更に拡大するために，富士通研究所ではパーソナルデータがもたらすプライバシーリスクを定量化・金額化するリスク評価技術を開発した。匿名加工の有無に関わらずリスクを評価できるため，同意取得方式でも匿名化方式でも，漠然とした不安を具体的な数値として見える化できる。これにより，事業者はデータを安心して流通できるようになる。また，別の事業者はそのデータと組み合わせることで，自組織のデータが持つ価値を更に引き出せるようになる。

本稿では，リスク評価技術について詳述する。

## 従来の評価技術：匿名加工に非対応

パーソナルデータのリスクを定量化する従来技術に，JOモデル（JNSA Damage Operation Model for Individual Information Leak）[7]やk-匿名性[8]がある。前者は匿名加工されたデータには対応しておらず，後者は定量化する対象の決め方自体を評価できないという問題点がある。

JOモデルは複数の個人情報漏えいの判例を基に，保険や情報セキュリティなどの専門家が作成した，個人情報漏えい時の予想損害賠償額の算定式である。一人あたりのリスクを金額化できることが特長であり，その主要部分は以下の式（1）で表すことができる。

漏えい個人情報価値＝

機微情報度×本人特定容易度×500円　　（1）

機微情報度は，当該個人情報が漏えいし本人が特定された際に，本人に与える影響を定量化したものである。本人特定容易度は，当該個人情報が漏えいした際に，本人が特定される可能性を定量化したものである。

しかし，このモデルは匿名加工処理を想定して設計されていない。このため，匿名加工しても本人特定容易度が変わらない場合が多く，匿名加工によるリスクの低減は評価できないという問題があった。例えば，匿名加工では**図-1**に示すように，項目「メールアドレス」などの識別子の削除や，項目「年齢」の階級化，項目「職業」の一般化などが行われる。これによって個人の特定は困難になるが，JOモデルの算定金額は加工前後で変化しない。

なお，機微情報度は代表的な項目の種類ごとに値が決められているほか，項目を組み合わせた場合の算定式もあり，おおむね各項目の値の最大値に近くなる。例えば，項目「職業」なら2と，項目「本籍」なら101となり，これらの組み合わせでは101となる。

k-匿名性は，世界的に使われているパーソナルデータの匿名性の指標である。ある人Xについて，Xと同じデータを持つほかの人が多いほど，Xは特定されにくいとする考え方による指標である。指標利用者が評価対象として設定した複数の項目の組（準識別子と呼ばれる）について，同一データを持つ人の数を用いて，匿名性を定量化する。

同一データを持つ人の数は，対象のデータセット内だけで数えることが多い。これは，その人数が確実にいると言えるためである。また，匿名加工情報の要件で強調されている，データセット内での差異の勘案（個人情報の保護に関する法律施行規則 第19条第5号）に対応する考え方になっている。

k-匿名性はリスクを評価する指標ではないが，これをJOモデルの本人特定容易度として使用することで，匿名加工後データにも対応したリスク算定モデルを作ることができる。しかし，単純に組み合わせただけでは適切なモデルにならない。

例えば，どの情報を準識別子とするかを決めるのは一般的には困難である。また，このモデルは準識別子の設定が不適切であるリスクを捉えられないという問題がある。全項目を準識別子とすればこの問題は避けられるが，匿名加工情報では一般にそこまで強い匿名性は求められない。このため，年齢の階級化などの項目単位での加工の効果も評価する必要があり，全項目を準識別子としてしまうと実用的な評価は行えない。

例えば図-1において，全項目を準識別子としてそのデータセット内でk-匿名性を評価する場合，加工前だけでなく加工後でも他人と同じデータとなる人はいないため，算定金額は加工前後で変化しない。しかし，図-1の加工前は匿名加工情報の規則に明確に違反している一方で，加工後はそうではないため，加工前後でリスクが同一となるのは

| メールアドレス | 年齢 | 職業 | 本籍 |
|---|---|---|---|
| alice@... | 12 | ピアニスト | 本町1 |
| bob@... | 18 | 中小企業社員 | 山奥村 |
| charlie@... | 19 | 中小企業社員 | 本町1 |
| dave@... | 22 | 画家 | 山奥村 |
| ellen@... | 24 | 大企業社員 | 本町1 |

（a）加工前データセット

匿名加工

| 年齢 | 職業 | 本籍 |
|---|---|---|
| 10代 | 芸術家 | 本町1 |
| 10代 | 会社員 | 山奥村 |
| 10代 | 会社員 | 本町1 |
| 20代 | 芸術家 | 山奥村 |
| 20代 | 会社員 | 本町1 |

（b）加工後データセット

図-1　匿名加工の例

不適当である。この問題を解決するためには，利用者による準識別子の設定が不要なモデルが必要である。

## リスク評価技術

富士通研究所は，これまで培ってきた高速匿名化技術を応用し，匿名性を精緻かつ高速に定量化することで，匿名加工前後のパーソナルデータのリスクを評価できる技術を開発した。本技術は，k-匿名性の考え方をより一般的な匿名加工に対応できるように拡張して特定性（匿名性の低さ）を算定する。更に，JOモデルの本人特定容易度をこの特定性で置き換えることによって，リスクを金額化する。

本技術の二つの特長を以下に述べる。

### ● 特定性の算定モデル

一つ目の特長は，匿名加工前後のパーソナルデータのリスクをより適切に定量化できるモデルを使用することである。

本モデルは，各人のデータについて，データセット内で最も特定しやすい項目の組み合わせで，特定性を定量化する。JOモデルの本人特定容易度をこれに置き換えることで金額化する。

特定しやすさは，以下の二つの性質をモデル化している。

(1) 性質1：情報の入手しやすさ

特定のために必要な情報が入手しやすいほど，特定しやすい。例えば，図-1（b）において一人目のデータは年齢と職業で特定でき，二人目のデータは年齢と本籍で特定できる。多くの場合，本籍より職業の方が情報を入手しやすいため，ほかの条件が同一であれば前者の方が特定しやすい。

(2) 性質2：項目の少なさ

特定のために必要な項目が少ないほど，特定しやすい。例えば，図-1（b）において3人目のデータは年齢と職業と本籍の3項目を組み合わせることで，初めて特定できる。これに対して，ほかの4人は年齢と職業，あるいは年齢と本籍など，3項目のうちの2項目の組み合わせで特定できるため，より特定しやすい。

また一般的に，機微な情報ほど入手しにくいため，JOモデルにおける機微情報度が高いほど入手しにくいとした。例えば，本籍は職業より機微情報度が高いため，本籍の方が入手しにくい情報となる。

本モデルで特定性（値域は0から1）を算定した結果を**図-2**に示す。なお，10代芸術家，20代芸術家，20代会社員の特定性はいずれも0.9であるが，例えば一般市民に20代会社員が多いとすると，この評価に違和感があるかもしれない。しかし，このデータセット内にデータが登録され得る個人には20代会社員が少ないという可能性を考える必要があり，妥当な結果であると考える。

### ● 高速探索

二つ目の特長は，本モデルの特定性を高速に算定できることである。

特定性の算定において，各人のデータについて，データセット内で最も特定しやすい項目の組み合わせを効率的に探索する。上述した二つの性質に基づいて，より特定性が高くなる組み合わせの中から各データを特定できるかどうかを判定していくことで不必要な判定を省略でき，高速に算定できる。例えば，職業で特定できるデータは，性質1から本籍で特定できるかどうかを省略できるため，職業による特定を本籍より先に探索する。また，

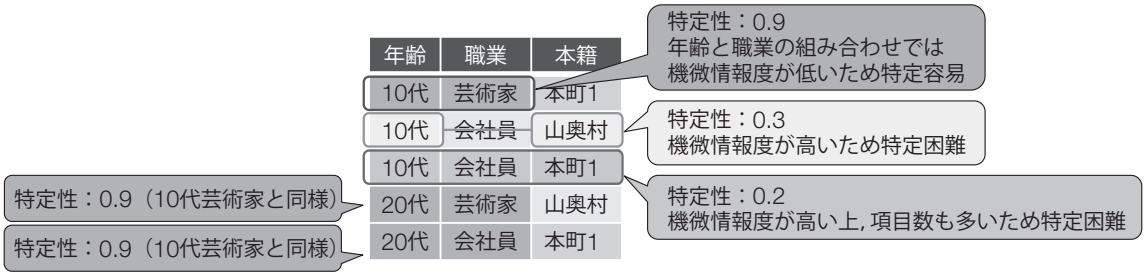



図-2　特定性算定の例

年齢と職業だけで特定できるデータは，性質2から年齢と職業と本籍による探索は省略できる。

## 性　能　評　価

本技術が従来技術と異なり，匿名加工前後のリスクを定量化できることおよび実用的な処理時間であることを確認するために，実際のデータで実験を行った。

### ●　算定モデルの確認

リスク定量化の確認は，匿名化のベンチマークとして使用されているAdultデータセット[9]に適用して行った。全48,842人に対して，先行研究でよく用いられている9項目（age，workclass，education，marital-status，occupation，race，sex，native-country，INCOME）のデータを使用した。データセットに対して，以下に示す3種類の匿名加工を施したデータセットを作成し，加工前と併せて4種類のデータセットに対してリスク定量化を行った。

（1）項目ageを10歳間隔で階級化
（2）項目occupationと項目INCOME以外を準識別子としてk-匿名化（k-匿名性を達成するように加工）（k＝3）
（3）項目INCOME以外を準識別子としてk-匿名化（k＝3）

このうち，（2）と（3）は先行研究でよく見られる加工であり，理論的には準識別子がより少ない（2）の方が（3）よりもリスクが高いはずである。

また，各データセットの情報量を算出した。一般的に，情報量が少なくなればリスクも減少する関係にあるため，その分布を見ることでモデルの妥当性をある程度確認できる。情報量は，統計コミュニティでよく使用される情報利得（Kullback-Leiblerダイバージェンス）に基づいて算出した。

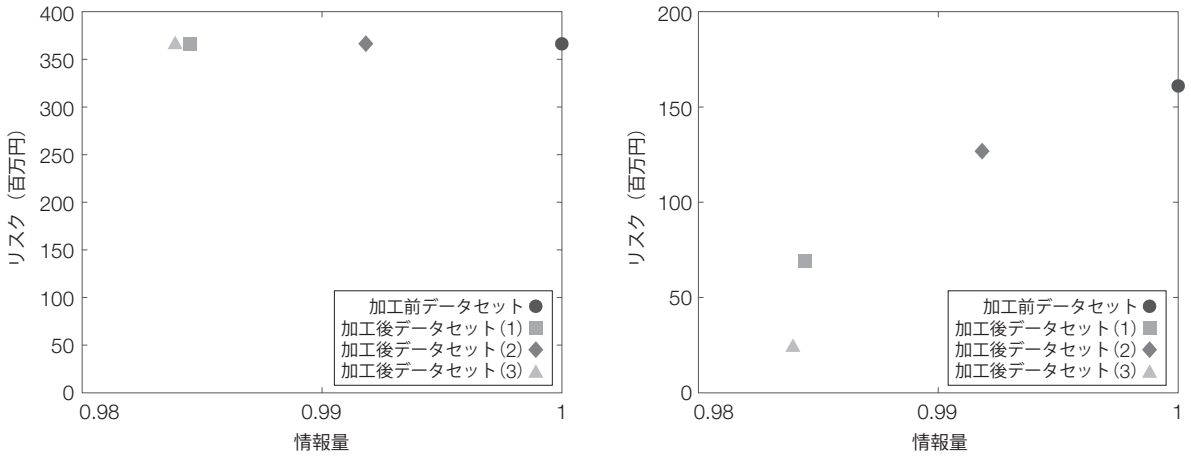
（1）～（3）の適用結果を**図-3**に示す。リスクは，式（1）で示したJOモデルにおける漏えい個人情報価値を全員分合算したものである。情報量は多い方が良く，リスクは低い方が良いため，各グラフで右下にプロットされるほど理想的な匿名加工といえる。

図-3（a）に示したとおり，JOモデルでは加工前後でリスクが変化していない。一方，図-3（b）では，本技術のモデルではリスク（の小ささ）と情報量（の大きさ）のトレードオフが見られ，理論どおり加工ターゲット（2）は加工ターゲット（3）よりもリスクが高い。このように，本技術は従来技術と異なり，加工前後のリスクを定量化できることが確認できた。

### ●　処理時間の確認

処理時間の確認は，事業者が実際に取り扱っている複数のデータセットに適用して行った。

一般的な性能のパソコンで処理した結果を**表-1**に示す。100万人規模のデータセットを約1時間で処理できており，全日本国民のデータセットであっても数日で処理できるレベルである。一般的に，ビッグデータ分析は試行錯誤が必要となるため，




（a）JOモデル　　（b）本技術のモデル

図-3　各匿名加工後データセットの情報量と，各リスク算定モデルでのリスクとの関係

表-1　実際のデータセットでの処理時間

| データセット | レコード数 | 項目数 | サイズ | 処理時間 |
|---|---|---|---|---|
| A | 0.6万 | 49 | 1 Mバイト | 2秒 |
| B | 50万 | 27 | 107 Mバイト | 30分 |
| C | 94万 | 29 | 241 Mバイト | 40分 |
| D | 202万 | 22 | 376 Mバイト | 1時間 |
| E | 290万 | 49 | 865 Mバイト | 1時間 |

数週間以上かかるとされている。一方で，本技術のリスク定量化は数時間から数日で完了するため，データ流通から活用まで全体の作業時間に占める割合は少なく，実用性が高いと考えられる。

## むすび

本稿では，富士通研究所が開発したパーソナルデータを安心して流通できるようにするリスク評価技術について述べた。

本技術によって，これまでプライバシーに対して不安があった事業者も，より安心してパーソナルデータの流通による新しいビジネスを始められるようになる。例えば，ある事業者は消費者にリスクを提示しながらデータ流通の同意を取得することで安心感を与え，より同意を得やすくなる。また，ある事業者は許容できる上限までリスクを取ることで，これまで以上に他事業者からの需要が高い匿名加工データを作り，ビジネスを推進できる。

今後は，本技術を製品化することでパーソナルデータの安心な流通を拡大させるとともに，事業者の持つデータから価値を更に引き出すために，必要となるデータが活発に流通する社会の実現を目指す。これによって，「つながるサービス」の提供価値の更なる向上につなげていきたい。

## 参考文献


(1) IDC：Big Data and Business Analytics Revenues Forecast to Reach $150.8 Billion This Year, Led by Banking and Manufacturing Investments, According to IDC.
https://www.idc.com/getdoc.jsp?containerId=prUS42371417

(2) World Economic Forum：Personal Data: The Emergence of a New Asset Class.
https://www.weforum.org/reports/personal-data-emergence-new-asset-class

(3) 富士通：Fujitsu Technology and Service Vision.
http://www.fujitsu.com/jp/vision/

(4) 富士通：Personiumサービス.
http://jp.fujitsu.com/solutions/cloud/k5/function/paas/personium/

(5) 富士通：FUJITSU ビジネスアプリケーション NESTGate 匿名化.
http://www.fujitsu.com/jp/solutions/business-technology/intelligent-data-services/bigdata/ba-solutions/nestgate/anony/

(6) 富士通：情報活用を支えるDWH専用データベース FUJITSU Software Symfoware Analytics Server.
http://www.fujitsu.com/jp/products/software/middleware/database/symfoware/products/analyticsserver/

(7) NPO日本ネットワークセキュリティ協会：2016年情報セキュリティインシデントに関する調査報告書～個人情報漏えい編～.
http://www.jnsa.org/result/incident/2016.html

(8) L. Sweeney：k-anonymity: a model for protecting privacy. Int. J. Uncertainty Fuzziness Knowl.-Based Syst., Vol.10, p.557-570, October 2002.

(9) C. Blake et al.：UCI Machine Learning Repository.
http://archive.ics.uci.edu/ml/


## 著者紹介


**山岡　裕司**（やまおか　ゆうじ）
（株）富士通研究所
セキュリティ研究所
データ・プライバシー保護技術の研究開発に従事。