Vol.2017-DPS-172 No.9
Vol.2017-SPT-26 No.9
Vol.2017-EIP-78 No.9
2017/11/29

匿名加工情報の加工方法と有用性・安全性指標の考察 ～匿名加工・再識別コンテスト 2017 から～

黒政 敦史[†1]　小栗 秀暢[†1]　門田 将徳[†2]

概要：2017 年 5 月 30 日に施行された改正個人情報保護法により，「匿名加工情報」という新たな情報の類型が導入された．一方，2015 年よりコンピュータセキュリティシンポジウム（CSS）において匿名加工・再識別コンテスト（PWSCUP）を行い，匿名化データの加工手法および有用性と安全性の定量指標による評価方法の取り組みがなされている．本稿では今年開催された PWSCUP の概要と結果を参照しながら，匿名加工情報の加工方法と有用性指標・安全性指標の社会実装について検討する．

キーワード：匿名加工情報，匿名化，プライバシー，個人情報，個人情報保護法，統計情報，***k*-匿名性**，PWSCUP

1. はじめに

2017 年 5 月 30 日に施行された改正個人情報保護法「個人情報の保護に関する法律及び行政手続における特定の個人を識別するための番号の利用等に関する法律の一部を改正する法律」[1]（以後，「改正法」という）により，匿名加工情報という新たな情報の類型が導入された．

匿名加工情報は，一定の条件の下で，本人の同意がなくても第三者提供や目的外利用が可能となる．それによって同意取得が困難なセンサーデータをはじめとするＩｏＴ分野や利用目的の特定が難しい AI や機械学習の教師データ等において新しいデータ利活用が期待できる．

個人情報保護委員会は，匿名加工情報の加工基準については「個人情報の保護に関する法律施行規則[2]」（以後，「規則」という）に定め，個人情報保護法ガイドライン匿名加工情報編[3]（以後，「ガイドライン」という），「個人データの漏えい等の事案が発生した場合等の対応について」に関するＱ＆Ａ[4]（以後，「Ｑ＆Ａ」という），個人情報保護委員会事務局レポート[5]（以後，「事務局レポート」という）を公開した．個人情報保護委員会は，事業者が具体的にどのような加工を行うかについては，取り扱う個人情報の性質，取扱い実態等に応じて定めることが望ましいことから，認定個人情報保護団体（以後，「認定団体」という）が作成する個人情報保護指針（以後，「指針」という）等の自主的なルールに委ねることとしており，呼応して認定団体から指針が順次公開されている．

一方，個人情報を含むパーソナルデータに関して個人の識別可能性などを減少させる技術的な措置「匿名化処理」と，その評価指標に関する研究の伸展のための取り組みとして，2015 年からコンピュータセキュリティシンポジウム（CSS）プライバシーワークショップ(PWS)にて匿名加工・再識別コンテスト（Privacy Work Shop CUP，以後「PWSCUP」という）が行われている．

PWSCUP は，用意されたデータセットに対して匿名化処理されたデータ「匿名化データ」を生成し，その有用性と安全性を競うコンテストである．PWSCUP の成果は匿名加工情報の加工・評価技術の発展に寄与し，データ利活用が促進することを期待している．

本稿では，2017 年 10 月 23 日に開催された PWSCUP2017（以降 「CUP'17」という）の概要と結果をまとめ，匿名加工情報を作成，運用する上で必要となる加工基準，安全性指標について考察する．なお，CUP'17 では，本稿著者のうち，小栗は PWS2017 実行委員会 PWSCUP 担当委員（以後，「実行委員」という）および運営システム開発/運用，門田は運営スタッフおよび CUP'17 参加者チーム"M-OND-A"の構成員として関わった．

2. 匿名加工情報の作成に係る制度等

まず，匿名加工情報の作成において，事業者が参照するべき法令，規則等について述べる．

2.1 法令・規則

・個人情報の保護に関する法律

（定義）
匿名加工情報とは「*特定の個人を識別することができないように個人情報を加工して得られる個人に関する情報であって，当該個人情報を復元することができないようにしたものをいう*」（改正法第 2 条 9 項[1]）．

（匿名加工情報の作成等）
「個人情報取扱事業者は，匿名加工情報（匿名加工情報データベース等を構成するものに限る．以下同じ．）を作成

†1 富士通クラウドテクノロジーズ株式会社 FUJITSU CLOUD TECHNOLOGIES LIMITED，〒169-8333 東京都新宿区北新宿 2-21-1 Shinjuku，Tokyo 169-8333 Japan.

†2 東京大学大学院 学際情報学府 The University of Tokyo Graduate School of Interdisciplinary Information Studies，〒113-0033 東京都文京区本郷 7-3-1 Bunkyo，Tokyo 113-0033 Japan.

Vol.2017-DPS-172 No.9
Vol.2017-SPT-26 No.9
Vol.2017-EIP-78 No.9
2017/11/29

するときは，特定の個人を識別すること及びその作成に用いる個人情報を復元することができないようにするために必要なものとして個人情報保護委員会規則で定める基準に従い，当該個人情報を加工しなければならない.」（改正法第 36 条**[1]**）

・個人情報保護委員会規則

保護委員会は匿名加工情報の作成について以下の基準を定めている.

（匿名加工情報の作成の方法に関する基準）

法第 36 条第 1 項の個人情報保護委員会規則で定める基準は、次のとおりとする。

(1) 個人情報に含まれる特定の個人を識別することができる記述等の全部又は一部を削除すること（当該全部又は一部の記述等を復元することのできる規則性を有しない方法により他の記述等に置き換えることを含む.）.

(2) 個人情報に含まれる個人識別符号の全部を削除すること（当該全部又は一部の記述等を復元することのできる規則性を有しない方法により他の記述等に置き換えることを含む.）.

(3) 個人情報と当該個人情報に措置を講じて得られる情報とを連結する符号（現に個人情報取扱事業者において取り扱う情報を相互に連結する符号に限る.）を削除すること（当該全部又は一部の記述等を復元することのできる規則性を有しない方法により他の記述等に置き換えることを含む.）.

(4) 特異な記述等を削除すること（当該全部又は一部の記述等を復元することのできる規則性を有しない方法により他の記述等に置き換えることを含む.）.

(5) 前各号に掲げる措置のほか，個人情報に含まれる記述等と当該個人情報を含む個人情報データベース等を構成する他の個人情報に含まれる記述等との差異その他の当該個人情報データベース等の性質を勘案し，その結果を踏まえて適切な措置を講ずること.（規則第 19 条[2]）

規則第 19 条の加工基準を解釈するためにはガイドライン[3]およびＱ＆Ａ[4]を参照する必要がある.

2.2 個人情報保護委員会ガイドライン、Ｑ＆Ａ

個人情報保護委員会は，匿名加工情報の適正な取扱いの確保に関して行う活動を支援すること，及び当該支援により事業者が講ずる措置が適切かつ有効に実施されることを目的として，法が定める事業者の義務のうち，匿名加工情報の取扱いに関する部分に特化して分かりやすく一体的に示す観点から，個人情報の保護に関する法律についてのガイドライン（匿名加工情報編）[3]を公開した.

解釈が難しい規則第 19 条 4 号 5 号に関連する「特異な記述等」に関しては以下の記述がある.

ここでいう「特異な記述等」とは、特異であるがために特定の個人を識別できる記述等に至り得るものを指すものであり、他の個人と異なるものであっても特定の個人の識別にはつながり得ないものは該当しない。実際にどのような記述等が特異であるかどうかは、情報の性質等を勘案して、個別の事例ごとに客観的に判断する必要がある。

～（中略）～

なお、規則第 19 条第 4 号の対象には、一般的なあらゆる場面において特異であると社会通念上認められる記述等が該当する。他方、加工対象となる個人情報に含まれる記述等と当該個人情報を含む個人情報データベース等を構成する他の個人情報に含まれる記述等とで著しい差異がある場合など個人情報データベース等の性質によるものは同第 5 号において必要な措置が求められることとなる。（ガイドライン, 3-2-4 特異な記述等の削除[4]）

また,「個人情報の保護に関する法律についてのガイドライン」及び「個人データの漏えい等の事案が発生した場合等の対応について」に関するＱ＆Ａ[4]を公開し, 規則第 19 条５号の解釈について補足している.

ここでの「当該個人情報を含む個人情報データベース等」とは、当該個人情報取扱事業者が匿名加工情報を作成する際に加工対象とする個人情報データベース等を想定しています。すなわち、匿名加工情報を作成する個人情報取扱事業者が保有する加工とは無関係の個人情報を含む全ての個人情報データベース等の性質を勘案することを求めるものではありません。（Ｑ＆Ａ，A11－９[4]）

2.3 個人情報保護委員会事務局レポート

個人情報保護委員会は，認定団体による保護指針の作成又は事業者団体が自主ルール等の策定，個別の事業者や関係団体等が匿名加工情報を作成しようとする場合の参考資料として，事務局レポート[5]を公開した.

これには，5 つの匿名加工情報のユースケースと加工例を紹介しており，ID-POS データを用いた購買履歴のユースケースが，CUP'17 と類似している.

いずれのユースケースも定性的な加工手法を紹介するに止まり，具体的なデータセットを用いた加工手法やリスク評価手法は記載されていない.

3. PWSCUP 匿名加工・再識別コンテスト

PWS（プライバシーワークショップ）はコンピュータセ

Vol.2017-DPS-172 No.9
Vol.2017-SPT-26 No.9
Vol.2017-EIP-78 No.9
2017/11/29

キュリティシンポジウムにおける 1 トラックとして発足したが，年間を通じてプライバシーと法制度，技術的な措置についての研究を促進する活動を行っている．

特にその加工技術の向上を促すために実施されているのが PWSCUP 匿名加工・再識別コンテストである．

PWSCUP は 2015 年[6]，2016 年[7]，2017 年[8]と続けて実施されており，それぞれにコンテストの目的が定義されている．

3.1 CUP'16 までの課題

CUP '17 の企画にあたり，前年度までの課題としてルール論文[8]にて次の点が挙げられた．

- 1 年間もの長期間に，顧客に単一の仮名を割り当てるのは識別リスクが高い．識別される前に，別の新たな仮名を割り当てることが出来ない．
- 独自の有用性，安全性の基準を用いているため，CUP'16 の後に公開された規則第 19 条の加工基準に照らしてみると，不適合とみなされる可能性がある．

その他，技術的課題がいくつか挙げられたが，本稿では割愛する．詳細な課題，及びルールについては CUP'17 のルール論文[8]を参照されたい．

3.2 CUP'17 のルールの特徴

本節では，3.1 節で述べた CUP'16 までで得られた課題を元に，CUP'17 で新たに加えられたルールの特徴について紹介する．

3.2.1 長期間履歴の分割.

ルール論文[8]において次のように記載されている．「*元データは，1 年間の履歴を月ごとの 12 個の期間に分けて提供するものと考える．仮名の割り当ては期間ごとに行い，別の期間で変更することを認める．単一の顧客に全ての期間で同じ仮名を割り当てれば有用性が高いが，識別されるリスクが高まるため，参加者のこの有用性と再識別リスクのトレードオフの中で，仮名制御を最適化する．*」

長期間履歴データの加工については，事務局レポート[5]においても記述があり，CUP'17 では定期的な仮 ID 変更が再識別攻撃に与える影響について観察することを目的の 1 つとした．

3.2.2 規則第 19 条への対応.

参加者には，2 章で述べた個人情報保護委員会規則第 19 条の条文を提示し，その中で求められている加工を参加者の解釈によって実践することを求めた．

実行委員による事前の議論では実行委員による規則第 19 条の解釈に基づいて，各号の加工方法についてコンテストルールで縛り，一定以上の加工を行わないデータにデメリットを与える方法も検討されたが，定量的な基準に落とし込むには，まだ国内における議論が足りないと判断し，各参加者チームに解釈を委ねることとした．

それぞれの解釈は，有用性・安全性の総合点で上位 10 位に入ったチームによる最終プレゼンテーションの中で報告する機会を設けた．さらに，規則第 19 条を十分に検討し，最も適切な措置を行ったチームに「匿名加工基準賞」を与えることとし，参加者，聴講者よる投票を行い最も得票数の多いチームを対象とした．

これは，現時点における匿名加工基準の技術的措置が確定されていない状況においては，解釈と実践の情報交換，情報共有の場として有益であったと考える．

3.3 コンテストのユースケース

CUP'17 におけるユースケースはルール論文[8]やその他 HP などでの明確な記述はない．しかし，元データの提供元や，匿名加工および再識別フェイズのモデルから考えられるプレイヤー像について筆者らが検討したものを述べる．

3.3.1 匿名加工フェイズ

匿名加工フェイズでは，次の 2 者が想定される．

○データ提供者＝オンラインショップ事業者（DP）

オンラインショップ事業者は，CUP'17 で利用された，元データを所有する事業者である．その事業者が経営するオンラインショップから，顧客のマスターデータとトランザクションデータ（購買データ）が，毎月蓄積され，これを他社に提供する．

○データ利活用者＝レコメンドエンジンの利用者（DU）

データ利活用者は，データ提供者から得た匿名化データを用いてレコメンドエンジンを作成し，販売された商品と商品の関係性を計測して正確な結果を得るために改良する．そのレコメンドエンジンの結果が，元となるトランザクションデータを用いた結果と比較して，正確であることが重要視される．

DP は図 1 に示すように，オンラインショップから収集されたデータ($T^{(n)}$と示す)を月ごとに匿名加工を施し，さらにその提供先を想定した有用性を評価した上で，匿名加工情報 S として DU に提供する．

DU として想定しているのは，流通などの様々な事業者や特定クラスタの消費者向けなど複数種類あり，その作成における結果の精度を元にした有用性指標を匿名加工情報の作成条件として DP に伝えているものとする．また，その他のデータ利用者についても想定し，合計 6 種類の有用性指標を設定した．具体的な有用性の評価方法については，4.1 節, 4.2 節で述べる．

Vol.2017-DPS-172 No.9
Vol.2017-SPT-26 No.9
Vol.2017-EIP-78 No.9
2017/11/29

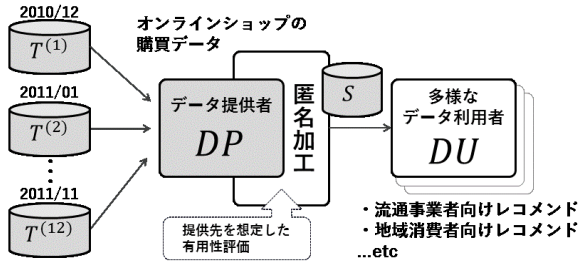

図 1 有用性指標の考え方

3.3.2 再識別フェイズの背景ストーリー

再識別フェイズでは主に以下の 2 者が想定される．

○オンラインショップ事業者：DP

このオンラインショップ事業者は，匿名加工フェイズで想定された者（データ提供者：DP）と同じ者を指す．

○再識別攻撃者：ReAD

再識別攻撃者は，オンラインショップ事業者によって公開された，匿名化データに加えて，オンラインショップ事業者が誤って流出させてしまった部分データを所有している

図 2 に再識別攻撃者の考え方を示す．再識別フェイズでは，元のデータを所有する DP が，ミスやハッキングなどの理由により，オリジナルデータの一部（以下，部分知識と呼ぶ）を流出させてしまったという状況を想定する．

すなわち，再識別攻撃者は，匿名化データと部分知識を組み合わせることによって，匿名化データにおけるある仮名を元の識別子と接続することが目的である．元データに接続することが出来れば，具体的な購入品目などを知ることも可能である．

CUP'17 の再識別攻撃者 *ReAD* は，匿名化フェイズで設定された，月ごとに変化する仮名を含めて，元データとの再識別攻撃に成功することが求められる．また，その際に利用する知識は，元データ全てを知っているという最大知識攻撃者[10]を想定せず，行が削減され，商品コードが削られている部分知識を用いる攻撃者を想定することが大きな特徴である．

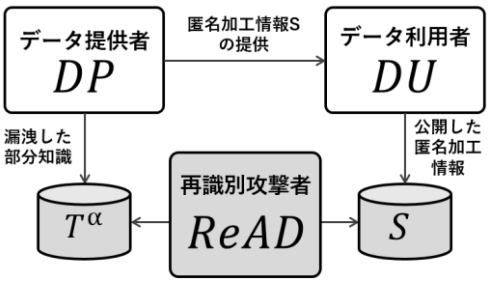

図 2 再識別攻撃者の考え方

CUP'17 における有用性は，データ提出時に計測した有用性指標 $E^i=(E^1,...,E^6)$の最大値とした．安全性は実行委員が作成した再識別アルゴリズム $S^k = (S^1,...,S^6)$に加えて，他の参加者による再識別試行の結果の最大値 S^{user} を計測し，安全性指標 $S = \max(S^k,S^{user})$と設定した．総合順位は U＋S で定めた．

また，最も総合順位の高い匿名加工データに対して，最も再識別したレコード数が多いチームに対して「再識別賞」を与える．

4. PWSCUP 2017 の評価指標

4.1 有用性指標

CUP'17 では表 1 の有用性指標が採用された[9].

表 1　CUP'17 で採用された有用性指標

有用性指標	概要	作成者（所属）
E1-ItemCF-s	対問屋用のレコメンドエンジン作成のための有用性を評価	村上 隆夫 (産業技術総合研究所) 門田 将徳 (東京大学)
E2-ItemCF-r	対小売店用のレコメンドエンジン作成のための有用性を評価	
E3-topk	購入した顧客数が最も多い上位 k 個の商品集合の差分を評価	
E4-diff-date	匿名加工前後の購入日の差の絶対値を評価	野島 良 (情報通信研究機構)
E5-diff-price	匿名加工前後の単価の比率の和を評価	
E6-nrow	消去されたレコード(行)の割合を評価	小栗 秀暢 (富士通クラウドテクノロジーズ)

今年度の有用性指標は，レコメンドエンジンの作成を目的としたデータ利用者を想定した指標(E1-E3)と，時系列分析や価格変動の分析を目的としたデータ利用者を想定した指標(E4,E5)に大別される．また，過度なレコード消去を抑制するための指標として E6 も採用された．E1-E3 については以下で詳細に説明を行う．

複数の有用性指標を設定した理由は，作成された匿名加工情報が複数の目的で利活用されることを想定したためである．さらにこれは，昨年度と同じデータセットを利用しているが，昨年度とはユースケース，有用性が異なることも留意されたい．

4.2 レコメンドエンジン作成を想定した有用性指標

上述した有用性指標のうち，レコメンドエンジンを作成することを目的としたデータ利用者のための有用性指標(E1-E3)について詳細に説明する．CUP'17 では，ある商品を買う人が追加して買う商品を推薦するためのレコメンドエンジンをユースケースの 1 つとして想定した．オリジナルのトランザクションデータ T と，匿名加工トランザクションデータ A(T)からそれぞれ商品同士の類似度行列を作成し，それらの距離によってレコメンドエンジンにおける有用性の評価を行う．

Vol.2017-DPS-172 No.9
Vol.2017-SPT-26 No.9
Vol.2017-EIP-78 No.9
2017/11/29

今回用いられたデータセットを調査すると，各商品を 12 個，24 個，…と「ダース買い」している顧客が多いことが判明した．図 3 は，合計購入点数が 12 個，24 個，36 個，48 個の時に，それぞれ顧客×商品の組み合わせ数が飛び抜けて多いことを表している．

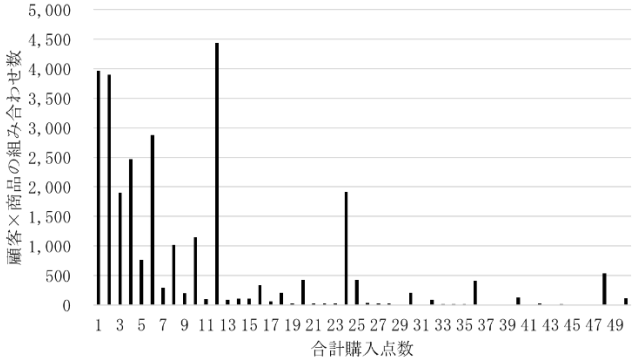

図 3　合計購入数の傾向

そのため，データ利用者として「問屋」と「小売店」の 2 種類の流通事業者を想定し，異なる指標を用意した．E1-ItemCF-s は，匿名加工前後で顧客と商品の組み合わせごとの合計購入点数が，12 個以上の組み合わせを対象として類似度行列を作成し，それらの距離を評価するもので，E2-ItemCF-r は合計購入点数が 12 個以下のものを対象として評価を行うものである．

E3-topk は，購入した顧客数が最も多い上位 k 個の商品集合の差分をパーセンテージで評価し，さらに，T における上位 k 商品について匿名加工前後で類似度行列を作成し，その距離を評価する．

4.3 安全性指標としての再識別率の定義

CUP'17 の安全性指標の特徴は，過去における再識別の定義を変更したことにある．CUP'15，CUP'16 における再識別の定義は，元データ M と匿名加工した M' において，元の識別子と，匿名加工された仮名とを紐付ける行為を再識別，と定義していた．

この方法は，仮名を 1 対 1 で使用する場合は有効であるが，本年度の目的にあるような，ある仮名を特定期間利用した後に他に変更するような，1 対多の紐付けには不向きである．

この紐付けの概要について図 4 にて示す．上は元データにおける識別子(ID)と匿名化データにおける仮名(CID)が 1 対 1 で対応しているため，Alice=A という紐付けが可能である．しかし，下は Alice に対応する仮名を A1 と A2 と設定しており，対応が一意でない．そのため，A2 と B，A1 と C のデータは，攻撃者にとって同じ値になることから，再識別攻撃者はどちらが正しい Alice であるかを判別できない．この処理を仮名の分割処理とする．

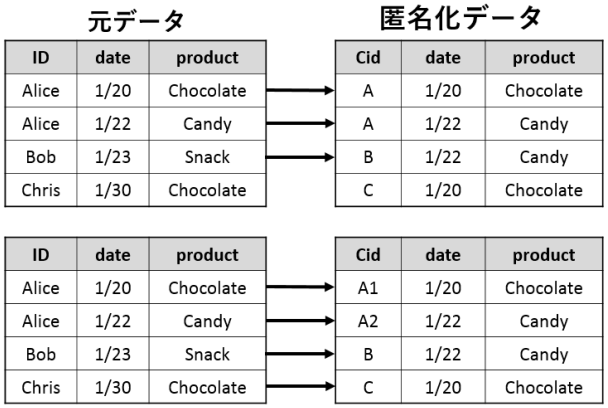

図 4　CUP'16(上)と CUP'17(下)の違い

この仮名の分割処理によって，仮名の全数は最大でトランザクションデータ T の全量まで増加させることが可能となり，その効果の検証が困難となるため，大会ルールとして不適切であるという議論となった．

そこで，CUP'17 では，今年度のユースケースとあわせ，仮名に対して 1 ヶ月単位でのみ仮名を変更可能とした．それら月ごとの仮名の変更を示すため，表 2 に示すような仮名表 F を生成した．表 2 の場合，Alice の仮名は，1 月には A1，2 月には A2 という仮名に分割したが，また 3 月には A1 に戻している．また，空白になっている値はその月に決済したデータが無いことを示す．ただし，実際の大会ルールでは，その場合 DEL という値を入れている．

表 2　仮名表 F のイメージ

ID	12月	1月	2月	3月	・・・	10月	11月
Alice		A1	A2	A1	・・・		A1
Bob	B1			B1	・・・	B1	
Chris		C1			・・・		
David			D1		・・・		D11
・・・	・・・	・・・	・・・	・・・	・・・	・・・	・・・

再識別アルゴリズム，及び再識別攻撃者は，この仮名表 F を推定して提出し，その一致パターンをもって安全性指標＝再識別率と設定した．

最終的には，ある顧客について全ての月の仮名を再識別された場合，1 人のデータが再識別された，と定義した．

この一致率の定義と課題については 6 章にて述べる．

5. 結果データの検討

図 5 に，本戦における再識別フェイズ終了後の参加者チームの順位,安全性と有用性の指標値分布を示す．最も有用性(UTL)と安全性(SEC)の和が小さいものが勝者である．今年度の優勝チームは，有用性:0.0476，安全性:0.0195 で

Vol.2017-DPS-172 No.9
Vol.2017-SPT-26 No.9
Vol.2017-EIP-78 No.9
2017/11/29

あった．

今年度の安全性は，全ての部分知識における平均値を示している．そのため最大再識別リスクを別途確認すると，0.038 であり，これは人数にして 500 人中 19 人が再識別された計算となる．一方，50%以上再識別されたデータも 5 チームあり，上位と下位における安全性には大きな差が見られた．

これは，同じ基準，同じデータセットで生成された匿名化データ間において，安全性と有用性に大きな差があり，それぞれの加工ロジックが異なっていることを示している．

5.1 節にて上位の参加者チームにおける特徴的な処理方法について述べる．

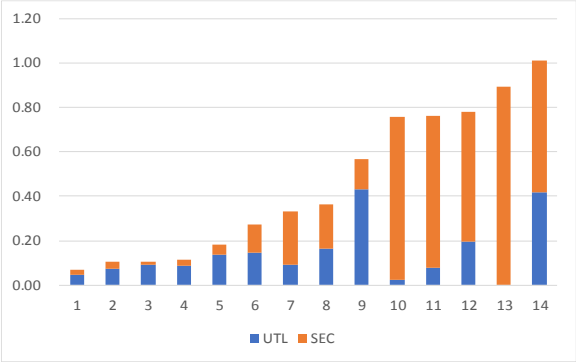

図 5　CUP'17 本戦における安全性と有用性の分布

5.1 CUP´17 上位チームに特徴的な匿名化手法

4.3 節で記述した通り，CUP'17 は再識別の定義が「年間を通して全ての仮名を特定された場合再識別されたとする」であったため，各顧客のトランザクションのうち，ある 1 つの月だけでも仮名を特定されなければ再識別には該当しない．

本戦参加 14 チームのうち上位 9 チームは，トランザクションデータの属性のうち，[購入日，単価]と[商品 ID，数量]を切り離して加工を行った．その理由は，前者はトランザクションデータの行番号に紐づけられたまま有用性が評価され(E4，E5)，後者は年間を通した統計値で評価される(E1，E2，E3)ためである．したがって，Item-CF によって評価される［商品 ID，数量］は，顧客と商品の組み合わせごとの合計購入点数を保持していれば，月や顧客 ID を超えた加工を行っても E1,E2,E3 への評価には影響を与えない．

その点を考慮し，顧客ごとに E1,E2,E3 の有用性に与える影響が最も小さい月の全てのレコードに，それ以外の月とは別の仮名を付与することで，その月を分割した．

その上で，分割し別の仮名を付与したレコードが，本来の顧客の他のレコードと同一人物のものだと特定されないように，購入日や単価，さらにはレコード数，購入商品集合などの個人を特定されうる情報を撹乱した．

以上の手法によって，定義された有用性を高く保ちながら，再識別を防止するデータに加工することができた．

ここで，表 3 と表 4 を用いて有効な匿名加工手法について紹介する．表 3 と表 4 を比較すると，6 行目のレコードに他の 5 つのレコードと異なる仮名が付与されている．

顧客 ID「12345」による商品 ID「B500C」の購買履歴が行番号 2 のレコードに集計され（数量を合計値に変換），購入日，商品 ID，単価も異なる値に変更されている．

このように加工することで，顧客 ID「12345」に関する 1，2 月のレコードは特定されても，3 月のレコードは特定されず，今回の PWSCUP の再識別条件の定義のもとでは，顧客 ID「12345」は再識別されないことになる．

表 3 トランザクションデータの例

	顧客 ID	購入日	商品 ID	単価	数量
1	12345	1/5	A400X	10.5	2
2	12345	1/23	B500C	2.5	14
3	12345	2/12	A400X	10.5	15
4	12345	2/14	A400X	10.5	3
5	12345	2/27	P230A	40.0	1
6	12345	3/9	B500C	2.5	5

表 4 匿名加工したデータの例

	顧客 ID	購入日	商品 ID	単価	数量
1	**Ab4xd**	1/5	A400X	10.5	2
2	**Ab4xd**	1/23	B500C	2.5	**19**
3	**Ab4xd**	2/12	A400X	10.5	15
4	**Ab4xd**	2/14	A400X	10.5	3
5	**Ab4xd**	2/27	P230A	40.0	1
6	**BB89Q**	**3/20**	**P230A**	**40.0**	**20**

PWSCUP においては，データセット，ユースケース，再識別条件の 3 つが全て実行委員会によって定義されており，その条件のもとでの最適な匿名加工処理が競われた．そのため，本結果をそのまま実社会に適用することはできない．

しかし，CUP'17 においては，再識別率の定義について大きくルールを変更し，その定義に沿った加工が行われたことも大きく影響している．本課題について 6 章にて述べる．

5.2 個人情報保護委員会規則第 19 条の解釈と措置

CUP'17 で，各参加者チームによる規則第 19 条 4 号 5 号の解釈とそれに対して講じた措置について，「匿名加工基準賞」を得た，東京大学のチーム"M-OND-A"による発表内容を以下に紹介する．

チーム"M-OND-A"による規則の解釈と措置

- 各号に対する解釈

  4 号は，レコードに含まれる 1 つ 1 つの値のうち，特異な記述と解釈した．今回のデータセットでは商

Vol.2017-DPS-172 No.9
Vol.2017-SPT-26 No.9
Vol.2017-EIP-78 No.9
2017/11/29

品は全て商品 ID によって管理されており，商品によって社会通念上の特異性を判断することはできず，単価や数量にも極端な外れ値はなく，特異性はないと判断した．また，4 号における「特異な記述」の有無は，母集団からのサンプリングの方法にも依存すると考えられる．

5 号は，各顧客に関するレコードの集計値に現れる特異性だと解釈した．今回のデータセットであれば，購入した月のパターンや，各月の合計購入商品数，レコード数なども考慮の対象とした．

・ **規則第 19 条と PWSCUP の比較**

PWSCUP ではデータセットとユースケースのみならず，再識別条件も定義されており，それらのルールのもとで参加者が技術を競い合うものである．それぞれの匿名化処理技術や再識別技術を客観的に評価するものとして，非常に意義のあるものであるが，その一方で，同じ方法で匿名化処理された実データを匿名加工情報として，流通させることは難しい．なぜなら，再識別条件やユースケースの定義は同じデータセットに対してもいくつか考えられるものがあり，規則第 19 条に対する措置はそれぞれ異なると考えられるためである．

これは，あくまでも加工に関する解釈の一例であり，他のチームにおいてはそれぞれの視点で措置を行ったことが報告された．その他のチームの解釈・措置については，別途 HP にて公表する予定である．今後の匿名加工基準の検討の参考となることを望む．

全チームの発表内容は事前のアンケート等でも確認され，その多くが個人情報保護委員会の担当者に届けられている．その個別のコメントについては公表できないが，それぞれの基準解釈については，明確に誤ったものでは無いことが，実行委員,各参加者に伝えられている．

6. 再識別率による安全性計測方式の検討

今年度に最も大きくルールが変更された部分として，再識別率に関する定義の問題がある．CUP'16 までの再識別率と異なり，仮名表を用いた再識別率をはじめて導入した CUP'17 では，その定義について多くの議論がなされた．本稿では議論の一部を紹介する．

まず，再識別の定義を行うにあたり，顧客数(M)，トランザクション数(T)，仮名表(F)の再識別率が存在することに留意されたい．それぞれの違いについて表 5 で述べる．

過去の再識別率の定義においては，顧客数(M)を用いて算定を行う．これは事業者等においてパーソナルデータを漏洩した場合，そのデータの機微性と漏洩した人数によって対応範囲が定まるためである．

表 5 検討された再識別定義の種類

種類	定義
顧客数(M)	元となるマスターデータに含まれる顧客数から，何人が識別されたかを示す．
トランザクション数(T)	元となるトランザクションデータから，何行のデータが再識別されたかを示す．その結果は人数に変換する場合としない場合がある．
仮名表(F)	顧客数 M に対して，仮名を複数付与した場合，その仮名表から考えて何個の仮名が元の顧客と再識別されたかを示す．

しかし，顧客の観点から考えるとデータの行数やそのデータの持つ意味のほうが重要である．1 行/1 ドル程度のデータが漏洩したリスクと，1000 行/1000 万ドルのデータが漏洩した場合では，個人が受け取るインパクトが異なる．それを表現する場合，基準はトランザクション数(T)とその値の機微性を計測する．

それに対して，今年度の再識別率の定義は，1 つの識別子に対して，複数の仮名が作成されることから仮名表(F)による再識別率が提案された．しかし，その場合でも定義は複数存在する．今年度検討された代表的な方式は，セル数方式，Or 型方式，And 型方式である．

図 6 セル数方式の再識別率の例

図 6 にセル数方式の再識別率の例を示す．この場合，推定された仮名表と比較して何セルが再識別されたかを知ることができる．これは再識別攻撃の成功数を正確に計測する手段ではあるが，個人の影響度や再識別された人数を示す指標ではない．

CUP'17 開始当初において，この指標が安全性指標として用いられていたが，実行委員内の議論によって，やはり人数によって再識別数は示すべき，との考えから，And 方式と Or 方式が考案された．

And 方式は，12 ヶ月全ての仮名が再識別された場合に，1 人再識別されたものとする方式であり，Or 方式は 12 ヶ月のうち n ヶ月(n>0)以上再識別された場合に 1 人再識別されたものとする方式である．

Vol.2017-DPS-172 No.9
Vol.2017-SPT-26 No.9
Vol.2017-EIP-78 No.9
2017/11/29

図 7　And 方式の再識別率の例

図 7 に And 方式の例を示す．12 ヶ月全てが再識別されないと再識別されない，という基準となったことから，5 章で示した結果データ全般について，全月のデータを守るのではなく，12 ヶ月の中の 1 ヶ月だけを守る，という戦略が効果的に利用できてしまった．

その反面，1 ヶ月でも再識別されるというのは，トランザクション数と同様に一般人にも解り易い指標である．しかし，そのしきい値を設定するのは困難であり，1 ヶ月以上，とする場合，全ての値を守る必要があることから作業が困難となり，今年度の目的のひとつでもある長期間の仮名分割による安全性検証が達成できない．

5 章で紹介した優勝チームのデータについて，再識別率をそれぞれ検証したところ，表 6 の結果を得た．And 方式で 19 人再識別されたデータは，セル数に換算すると約 53%の再識別されたことになる．また，その際に元となる識別子は 1087 の仮名まで分割された．

これらの数値をベースに，仮名分割された場合の再識別のあり方については技術的な議論の進展を期待する．

表 6　優勝データにおける再識別定義の違い

再識別定義	数値
And 方式	0.038 (19 人/500 人)
セル数方式	0.529 (773 セル/1460 セル)
仮名分割数	500 人 → 1087 人

7. まとめ

今年度の CUP'17 は，CUP'16 と同じデータセットを用いて実施されたが，有用性の設定変更と仮名の分割の概念を取り入れたことで，CUP'16 と大きくルールが異なるものとなった．

その結果として，比較的単純な指標として考えられていた再識別率という指標が，仮名の分割によって定義が拡張され，複数の定義で検討が必要になることが判明した点は今年度の大きな成果である．これにより規則第 19 条に対する解釈と措置はデータセットだけでなく，コンテストルールに相当するデータ提供者の運用ポリシーからも影響されることが浮き彫りになった．

つまり，事業者は提供するデータセットと有用性と安全性に関するデータ運用ポリシーを勘案しながら，規則第 19 条の加工措置を行う必要がある．加工を行うにあたっては，PWSCUP で採用された各指標は，既存の何らかのサービスに直接適用できるものではないが，多くの示唆に富んでおり参考になると考える，

今年度は海外のチームも参加したことから，これらの安全性と有用性の指標に関する議論は，海外にも広がることが期待できる．また，今後，国際的な匿名加工・再識別コンテストとして PWSCUP の取り組みは継続することが決定されている．

日本のみならず，国際的な視点での安全性検証のため，今後も，多くのデータ形式やユースケースに対する考察を深めながら，データ活用の議論が進展することを期待する．

謝辞

本稿執筆にあたり，東京大学 中川裕志先生，及び，明治大学 菊池浩明先生，及び，PWS 実行委員の皆様より，多くのご支援を頂きました．厚く御礼申し上げます．

参考文献

[1] 個人情報の保護に関する法律及び行政手続における特定の個人を識別するための番号の利用等に関する法律の一部を改正する法律(平成 27 年法律第 65 号)
[2] 個人情報保護委員会, "個人情報の保護に関する法律施行規則", [https://www.ppc.go.jp/files/pdf/290530_personal_commissionrules.pdf], (2017).
[3] 個人情報保護委員会, "個人情報の保護に関する法律についてのガイドライン(匿名加工情報編)", [http://www.ppc.go.jp/files/pdf/guidelines04.pdf], (2016).
[4] 個人情報保護委員会, "「個人情報の保護に関する法律についてのガイドライン」及び「個人データの漏えい等の事案が発生した場合等の対応について」に関するＱ＆Ａ", [https://www.ppc.go.jp/files/pdf/kojouhouQA.pdf], (2016).
[5] 個人情報保護委員会, "個人情報保護委員会事務局レポート：匿名加工情報「パーソナルデータの利活用促進と消費者の信頼性確保の両立に向けて」", [https://www.ppc.go.jp/files/pdf/report_office.pdf], (2017).
[6] 菊池 浩明, 山口 高康, 濱田 浩気, 山岡 裕司, 小栗 秀暢, 佐久間 淳, "匿名加工・再識別コンテスト Ice & Fire の設計", コンピュータセキュリティシンポジウム 2015 論文集, 2015(3), pp.363-370, (2015).
[7] 菊池 浩明, 小栗 秀暢, 野島 良, 濱田 浩気, 村上 隆夫, 山岡 裕司, 山口 高康, 渡辺 知恵美, "PWSCUP: 履歴データを安全に匿名加工せよ", コンピュータセキュリティシンポジウム 2016 論文集, 2016(2), pp.271-278, (2016).
[8] 菊池 浩明, 小栗 秀暢, 中川 裕志, 野島 良, 波多野 卓磨, 濱田 浩気, 村上 隆夫, 門田 将徳, 山岡 裕司, 山田 明, 渡辺 知恵美, "PWSCUP2017:長期間の履歴データの再識別リスクを競う", コンピュータセキュリティシンポジウム 2017 論文集, (2017).
[9] PWSCUP 実行委員会, "PWSCUP 2017 匿名加工・再識別コンテスト 競技ルール, Ver1.3", [https://pwscup.personal-data.biz/web/pws2017/index.php], (2017).
[10] Domingo-Ferrer, J., Ricci, S. and Soria-Comas, J., "Disclosure risk assessment via record linkage by a maximum-knowledge attacker", Privacy, Security and Trust (PST), 2015 13th Annual Conference on, pp.28-35, (2015).