Vol.2018-CSEC-82 No.18
Vol.2018-SPT-29 No.18
2018/7/25

複数回にわたる匿名加工情報の提供に対する再識別リスク評価方法：PWSCUP2017における安全性指標の考察と分析から

チャン クワン カイ[1,a)]　坂本 一仁[1]　松永 昌浩[1]

概要：PWSCUP2017 では，長期間にわたる購買履歴データの匿名加工情報の提供を想定し，月毎に異なる仮名が付与されたデータの匿名加工手法と再識別リスクに関する技術的検討が行われた．PWSCUP2017 における安全性指標のルール設計では，12 ヶ月分の仮名を全て正しく推定できたユーザを，再識別されたユーザと定義していた．しかしながら，毎月その月の購買履歴データの匿名加工情報を複数回にわたって提供するといったユースケースでは，PWSCUP2017 の安全性指標をそのまま導入した場合，再識別リスクを正確に評価できない可能性がある．本稿では，PWSCUP2017 の安全性指標について再考するとともに，複数回にわたる匿名加工情報の提供を想定した場合においても再識別リスクを正確に評価可能な新たな安全性指標を提案する．そして，PWSCUP2017 本戦のデータセットに対し，提案した安全性指標を利用した再識別リスク評価実験を実施した．本稿で提案する安全性指標は，データ提供毎に再識別リスクの増減を評価できるものであり，実験や議論から匿名加工情報の提供は，ユースケースや提供先の利用方法を考慮し，再識別リスクの評価方法を適切に選択することが重要であると結論付けた．

An Evaluation Method for Re-identification Risks of Anonymously Processed Information Provided Multiple Times: From a Consideration and an Analysis of Privacy Metrics in PWSCUP 2017

Tran Quang Khai[1,a)]　Takahito Sakamoto[1]　Masahiro Matsunaga[1]

1. はじめに

PWSCUP は，コンピュータセキュリティシンポジウム (CSS) のプライバシーワークショップ (PWS) において開催される匿名加工と再識別の技術を競うコンテストである．PWSCUP では、主に個人情報保護法の改正 [1]（以降，改正法と記載する．）によって新設された「匿名加工情報」における加工手法の技術的検討を，コンペティション形式で行うことが目的とされている．第 3 回大会の PWSCUP2017 では，改正法における匿名加工情報の加工規則である委員会規則 [2] 第 19 条の各号に対応するようにルールが設計された．特徴は，長期間にわたる購買履歴データの匿名加工情報の提供を想定し，月毎に異なる仮名が付与されたデータの匿名加工手法と再識別リスクに関する技術的検討を行った点である．

本稿では，PWSCUP2017 において設定された匿名加工情報提供のユースケースと再識別リスク評価方法を再考することを目的とする．PWSCUP2017 の論文 [3] では，「1 年間の履歴を月ごとの 12 個の期間に分けて提供する．」，「加工は短期間（1 か月）で定期的に行われて，第三者に提供される．」と記載があるように，毎月その月の履歴データの匿名加工情報を複数回にわたって提供することを想定している．そして，PWSCUP2017 における最終的な再識別

[1] セコム株式会社 IS 研究所
Intelligent Systems Laboratory, SECOM CO., LTD.
[a)] ku-chan@secom.co.jp

リスクの評価方法 [4] は，匿名加工データから元データの「12 ヶ月分の仮名が全て正しく推定できたユーザの割合」であった．しかしながら，複数回にわたって匿名加工情報を提供するユースケースや，匿名加工情報の提供先における利用方法を再考すると，PWSCUP2017 の安全性指標や評価方法では，再識別リスクを正確に評価できない可能性が発見された．

そこで本稿では，複数回にわたる匿名加工情報の提供を想定した場合においても再識別リスクを正確に評価可能な新たな安全性指標を提案する．そして，PWSCUP2017 本戦のデータセットに対し，提案した安全性指標を利用した再識別リスク評価実験を実施した．本稿の実験や議論を通じ，1 つの再識別リスクの評価のみでなく，ユースケースや提供先の利用方法を考慮し，再識別リスクの評価方法を適切に選択することが重要であることがわかった．

本稿の構成は以下の通りである．2 節では，本稿で想定する匿名加工情報提供のユースケースを説明する．3 節では，PWSCUP2017 の安全性指標を説明し，想定したユースケースに適応した場合の問題点を示す．4 節では，新たな再識別リスク評価方法および安全性指標を示し，PWSCUP2017 のデータセットに対して 5 節において実験方法を説明する．6 節において実験結果と考察を示し，7 節において議論を展開する．8 節でまとめとする．

2. 想定するユースケース

PWSCUP2017 では長期間の履歴データの再識別リスクの評価が目的とされ，12 ヶ月分の購買履歴データのデータセット [5] を使い，競技が開催された．PWSCUP2017 において，加工された購買履歴データの詳細な提供方法は説明されていないが，論文 [3] の「1 年間の履歴を月ごとの 12 個の期間に分けて提供する.」,「加工は短期間（1 か月）で定期的に行われて，第三者に提供される.」，学会誌 [6] の「毎月の購買データを第三者に提供することを考え,」という記述から，複数回にわたって匿名加工情報を提供するユースケースが想定されていたと考える．本節では図 1 のように，PWSCUP2017 の 12 ヶ月のデータを反映し，毎月その月分のデータを加工して複数回にわたって匿名加工情報を提供するユースケースを詳細に説明する．

2.1 複数回にわたる匿名加工情報の提供

図 1 では，毎月その月のデータを加工して匿名加工情報を提供する様子を示す．提供期間は PWSCUP2017 のデータセットの期間を踏襲している．図 1(a) の 2010 年 12 月から匿名加工情報の提供を始める場合，まず提供元は 2010 年 12 月のデータを加工して提供先に提供する．

次の月には図 1(b) のように 2011 年 1 月データを，2010 年 12 月の仮名や加工状態を鑑みてデータを加工し，提供先に提供する．このとき，個人情報保護委員会の事務局レポート [7] や PWSCUP のルール設計で推奨されているように，同じユーザであっても前月の仮名と異なる仮名を付与し，再識別リスクを軽減することができる．そして提供先は，2010 年 12 月と 2011 年 1 月のデータが同様のデータ構造かつ期間に重複が無ければ，提供元は単純に連結して 2 ヶ月分のデータを扱えることになる [*1]．

同様に，図 1(c) や (d) において，毎月その月のデータを加工し，匿名加工情報を提供先に提供していく．最終的に，図 1(f) では，2011 年 11 月のデータ提供を行い，提供元は 12 ヶ月分の匿名加工情報を扱える状態になる．

3. PWSCUP2017 の安全性指標

PWCUP2017 の再識別リスクの評価は，最大知識攻撃者モデル [8] を基準として安全性指標を設計し，実施している．PWSCUP2017 では，最大知識攻撃者モデルを拡張した部分知識攻撃者モデルを導入しているが，本稿では簡単化のため最大知識攻撃者モデルのみを参照して議論を進める．

PWSCUP2017 の再識別リスクの評価方法は，競技の進行とともにいくつかの変更が加えられた．まず論文 [3] では，下記のような数式を安全性指標として示し，匿名加工データの再識別リスクを評価していた．数式中の変数等の詳細は付録 A.1 を参照されたい．

$$reid(F,\hat{F}) = \frac{|\ \{(i,l)\mid l\in\{1,\ldots,12\}, f^{(l)}(c_{i,1}) = \hat{f}^{(l)}(c_{i,1})\}\ |}{12n} \quad (1)$$

式 (1) では，元データにおいてあるユーザのある月は商品を購入していないが，商品を購入していると誤って仮名を推定してしまった場合，再識別率が下がってしまう．元データのユーザが存在していないにも関わらず，再識別率が下がってしまうため，PWSCUP2017 の予備戦の競技ルール version 1.0[9] では安全性指標が下記のように修正された．

$$\begin{aligned} match =& |\ \{(i,l)\mid l\in\{1,\ldots,12\}, f^{(l)}(c_{i,1}) = \hat{f}^{(l)}(c_{i,1})\}\ | \\ del =& |\ \{(i,l)\mid l\in\{1,\ldots,12\}, f^{(l)}(c_{i,1}) = \mathrm{DEL}\}\ | \\ reid(F,\hat{F}) =& \max\left(\frac{match-del}{12n-del}, 0\right) = MM \end{aligned} \quad (2)$$

式 (2) では，ユーザが存在している月の仮名を 1 つでも識別した場合に，再識別成功と定義している．本稿では，式 (2) を *Month Matching (MM)* 方式と呼ぶ．しかしながら，予備戦前半では上記の式で再識別リスクの評価を行っていたが，予備戦後半から本戦にかけては，下記の安全性指標に変更され，競技が進行した [4].

[*1] 再識別目的での匿名加工情報同士の連結は禁止されているが，統計処理目的での連結は許可されている [7].

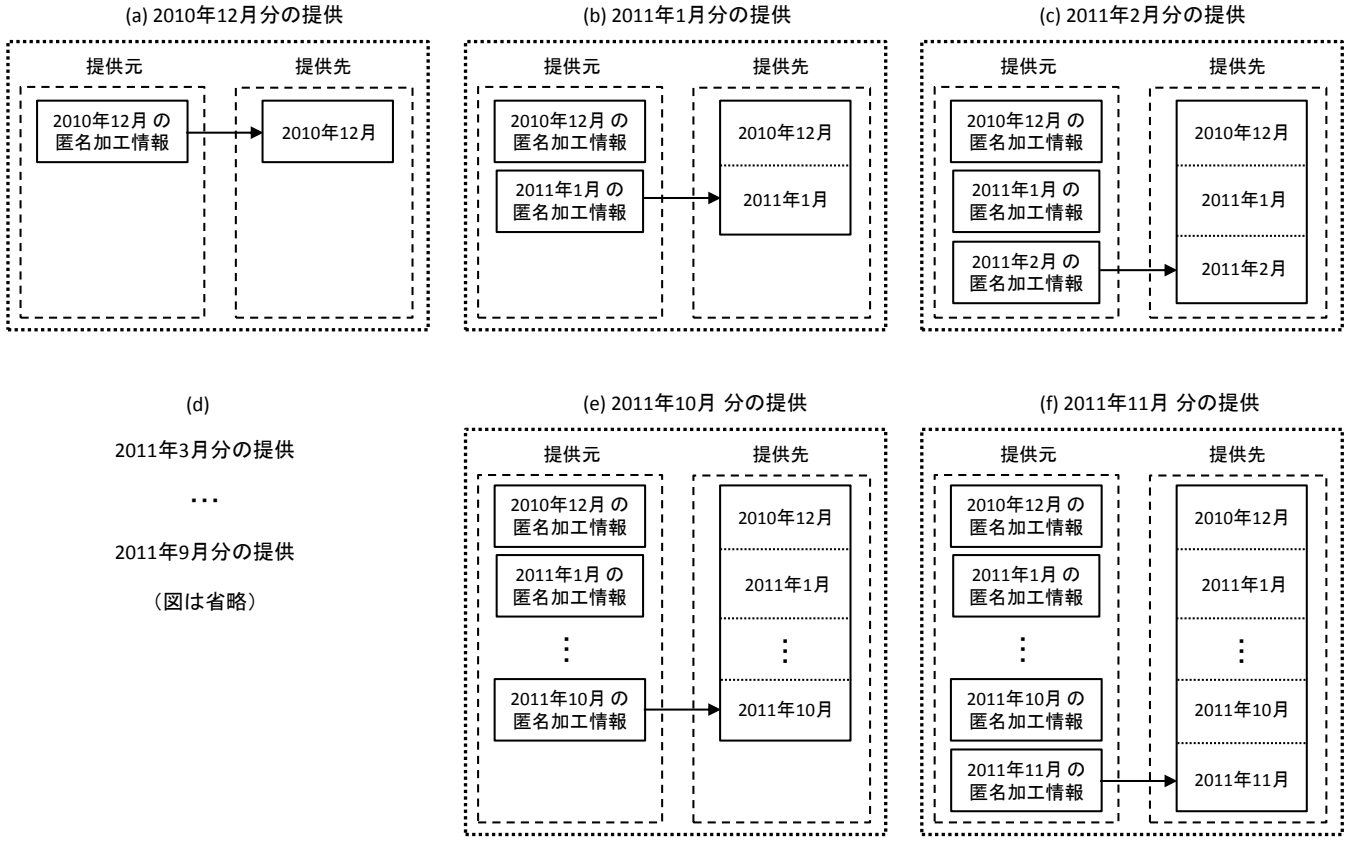

図 1 複数回にわたる匿名加工情報の提供

$$reid(F, \hat{F}) = \frac{|\{i \mid \forall l \in \{1, \ldots, 12\}, f^{(l)}(c_{i,1}) = \hat{f}^{(l)}(c_{i,1})\}|}{n} = UM \quad (3)$$

式 (3) では，「DEL を含めた 12 ヶ月分の仮名がすべて正しく推定できたユーザの割合」が再識別率として評価される．本稿では，式 (3) を *User Matching (UM)* 方式と呼ぶ．

黒政ら [10] は，UM 方式（[10] では And 方式），MM 方式（[10] ではセル数方式）を比較し，UM 方式が最終的な安全性指標として採用された経緯や議論を紹介し，UM 方式では 1 つの月の仮名がわからなければ再識別とみなされないというルール上の死角について言及している．実際に PWSCUP2017 の上位チームは，ある 1 つの月の仮名が低い確率でしか推定できないデータを作成し，競技ルールにおいて高い成績を獲得していた [11]．すなわち，黒政らも指摘しているように高い確率で仮名を推定できる月が多く存在している．

UM 方式と MM 方式のどちらが正確に再識別リスクを評価できるかについて議論は収束していないが，2 節で示したユースケースを想定すると，UM 方式ではいくつかの問題点が指摘できる．下記に 2 つの例を示す．

3.1 例 1：提供先における利用範囲

PWSCUP2017 の再識別リスクの評価は，図 1(f) の時点における評価であり，暗黙的に提供先が 12 ヶ月すべてのデータを常に利用することが想定されている．しかしながら，提供先は分析内容によって利用するデータの範囲を選択することが考えられる．

例えば，提供元が (a) の 2010 年 12 月のデータに対して，強い匿名加工を施し，(b) 以降は有用性を保つためほとんど加工しなかったとする．式 (3) の UM 方式では，2010 年 12 月に含まれるユーザの再識別リスクは低いと判断されるかもしれない．しかし，もし提供元が (f) の時点で 2010 年 12 月のデータは古いため，2011 年 1 月からのデータ（11 ヶ月分）のみを利用する判断をした場合，その 11 ヶ月はほとんど加工されていないため，再識別リスクは高くなると予想される．

また，提供先がある連続した 2 ヶ月分のみのデータを抽出し，仮名が同一のユーザの月ごとの差異を分析したいとする．その場合，抽出した 2 ヶ月に属するユーザが十分に加工されたデータとなっていない場合，高い再識別リスクとなる場合がある．

以上のように，式 (3) の UM 方式は，提供先が 12 ヶ月すべてのデータを常に利用する場合に限り，仮名分割によってすべての仮名からユーザを一意に識別することが困難になっていることを評価できるが，提供先が 12 ヶ月すべて

のデータを利用しない場合は，再識別リスクを正確に評価できない．そのため，提供先の匿名加工情報の利用をより柔軟に想定して再識別リスクを評価するならば，式 (2) のように MM 方式による評価が良いと言える．

3.2 例 2：複数回にわたる提供

また，PWSCUP2017 では図 1(f) のデータ提供時点において再識別リスクを評価しており，(e) の時点の再識別リスク，(c) や (b) の時点の再識別リスクは評価されていない．2 節で示したように，毎月その月のデータを提供することを考えると，各データ提供時点で提供先が保持しているデータを勘案して再識別リスクの評価を都度実施することが望ましい．

4. 安全性指標の提案

本稿では，3 節で示した点を考慮し，新たな安全性指標を下記のように設計する．まず，$l(l \leq d)$ 回目のデータ提供に対する再識別率を下記の数式のように定義する．

$$MM^{(l)} = \max\left(\frac{match^{(l)} - del^{(l)}}{l \times n^{(l)} - del^{(l)}}, 0\right) \tag{4}$$

ここで $match^{(l)}$ と $del^{(l)}$ は，

$$match^{(l)} = \mid \{(i,l) \mid l \in \{1,\ldots,d\}, f^{(l)}(c_{i,1}) = \hat{f}^{(l)}(c_{i,1})\} \mid$$
$$del^{(l)} = \mid \{(i,l) \mid l \in \{1,\ldots,d\}, f^{(l)}(c_{i,1}) = \mathrm{DEL}\} \mid$$

である．式 (4) は MM 方式と同様である．そして，全てのデータ提供 $MM^{(l)}$ から安全性指標を下記の数式のように定義する．

$$EMM = \max_{l \in \{1,\ldots,d\}} MM^{(l)} \tag{5}$$

式 (5) は，MM 方式による再識別率を採用し，さらにデータ提供毎に評価を行い，最も再識別率が高いデータ提供 $MM^{(l)}$ を再識別のリスクとする安全性指標である．本稿では，式 (5) を *Extended Month Matching (EMM)* 方式と呼ぶ．

5. 実験方法

UM 方式，MM 方式と EMM 方式の各安全性指標を，PWSCUP2017 のデータセットと競技内容を利用して評価し，比較する．

5.1 データセット

データセットは，PWSCUP2017 の本戦で使用されたトランザクションデータ T および，PWSCUP2017 の終了後に公開された 10 チームの匿名加工データ $\mathcal{S}$，顧客 ID と仮名 ID の対応関係を表す正解の仮名表 F を利用する [12]．チーム名に関しては，No1 から No10 までの仮名をランダムに付与している．

5.2 実験内容

10 チームの匿名加工データ $\mathcal{S}$ に対し，PWSCUP2017 の再識別フェーズを再現し，UM，MM，EMM の各方式における再識別リスクの評価を比較する実験を行う．PWSCUP2017 では再識別リスクの評価として 6 種類の再識別アルゴリズム（S_1〜S_6）が用意されていた．S_1〜S_6 は主に商品 ID や単価等を示す $t_{.,3} \sim t_{.,7}$ の値の組み合わせをキーとしてマッチングさせ，仮名 ID と顧客 ID の対応関係を推定している．S_1〜S_6 の詳細は競技ルール [4] や PWSCUP2017 の Web サイトを参照されたい．さらに，PWSCUP2017 では他チームからの再識別アルゴリズムによる攻撃も，最終的な再識別リスクの評価に反映された．そのため，本稿では他チームからの攻撃を考慮し，再識別アルゴリズムとして S_7, S_8 を加えて評価を行う．

S_7 は単価平均，S_8 はレコード数の特徴に基づく再識別アルゴリズムであり，月単位で仮名 ID と顧客 ID のマッチングを行うものである．PWSCUP2017 では，E1〜E6 の有用性を保ちつつ再識別リスクを低く抑える必要があった．E5 はレコード同士の単価 $(t_{.,6})$ の比率の平均であり，E6 は削除されたレコード数の割合である．E5, E6 ともに単純な評価方法であるため，単価への少しのノイズ付与やレコード削除で評価値が大きく悪化する傾向があった．その特性を利用し，再識別アルゴリズムに単価やレコード数の特徴を利用するチームが多く見られたため，本稿においても単価とレコード数に対する単純な攻撃を，S_7, S_8 として追加している．

S_7, S_8 の再識別アルゴリズムは下記の式で与えられる．

$$S_7, S_8 = \hat{f}^{(l)}(c_{i,1}) = \underset{j \in Dom(\mathcal{S}_1^{(l)})}{\mathrm{argmax}} \frac{\min(a_i, b_j)}{\max(a_i, b_j)} \tag{6}$$

ここで，S_7 の単価平均の特徴に基づく再識別アルゴリズムの場合，

$$a_i = \mathrm{avg}(\{t_{k,6}^{(l)} \mid t_{k,1}^{(l)} = c_{i,1}\})$$
$$b_j = \mathrm{avg}(\{s_{k,6}^{(l)} \mid s_{k,1}^{(l)} = j, j \in Dom(\mathcal{S}_1^{(l)})\})$$

であり，月単位で仮名 ID が持つ単価平均と顧客 ID が持つ単価平均の比率から類似度を求める．

また，S_8 のレコード数の特徴に基づく再識別アルゴリズムの場合，

$$a_i = \mathrm{count}(\{t_{k,6}^{(l)} \mid t_{k,1}^{(l)} = c_{i,1}\})$$
$$b_j = \mathrm{count}(\{s_{k,6}^{(l)} \mid s_{k,1}^{(l)} = j, j \in Dom(\mathcal{S}_1^{(l)})\})$$

であり，月単位で仮名 ID が持つレコード数と顧客 ID が持つレコード数の比率から類似度を求める．そして，類似度が最も大きい仮名 ID と顧客 ID の組み合わせによって，再識別を実施する．

PWSCUP2017 では，他チームによる再識別も含めた複

表 1　チーム毎の各安全性指標に対して
$S_1 \sim S_8$ の中で最大となる再識別率

チーム	UM 方式	MM 方式	EMM 方式	EMM/MM
No1	0.832	0.942	1.000	1.062
No2	0.022	0.442	0.468	1.059
No3	0.316	0.466	0.474	1.017
No4	0.002	0.620	0.633	1.021
No5	0.020	0.307	0.433	1.410
No6	0.008	0.047	0.053	1.127
No7	0.982	0.994	1.000	1.006
No8	0.226	0.442	0.516	1.167
No9	0.090	0.319	0.324	1.016
No10	0.082	0.192	0.247	1.286

数の再識別アルゴリズムによって再識別率を算出し，最も再識別率が高いアルゴリズムによる評価を採用している [4]．本実験においても UM, MM, EMM の各指標それぞれで S_1〜S_8 の再識別率を算出し，それぞれの最大値を評価値として採用する．

6. 結果と考察

6.1 各安全性指標の比較

UM，MM，EMM の各指標における再識別リスクの評価値を表 1 に示す．UM 方式は，式 (3) から，DEL を含めた 12 ヶ月分の仮名を全て正しく推定できたユーザの割合を再識別リスクとしている．MM 方式は，式 (2) から，DEL 以外の月で仮名を正しく推定できた月の割合を再識別リスクとしている．そして，EMM 方式は，MM 方式によるデータ提供毎（今回のデータセットでは 12 回）に行い，最大値を再識別リスクとしている．

UM 方式と MM 方式は評価方法が異なるが，表 1 に示すように全てのチームにおいて UM と MM の評価値が大きく異なっている．今回のデータセットは UM 方式による評価をよくするために加工されたデータであり，「あるユーザのひと月のみ仮名を推定できない」ようにチューニングされている．そのため，MM 方式のように仮名を推定できる月の割合で評価すると，再識別リスクの評価値が大きくなる．

EMM 方式は MM 方式を内包している．すなわち，今回のデータセットでは MM 方式は 2011 年 11 月のデータ提供において，12 ヶ月分のデータのみを評価しているが，EMM 方式では 2010 年 10 月のデータ提供から 2011 年 11 月のデータ提供まで，データ提供の度に評価を行い，再識別率が最も高いデータ提供時点の値を採用している．EMM 方式と MM 方式には $EMM \geq MM$ の関係がある．

表 1 から全てのチームにおいて $EMM \geq MM$ の関係が確認できる．これは，2 節で示したユースケースを考慮すると，2011 年 11 月に 12 ヶ月分を MM 方式で評価した時よりも，再識別リスクが高かったデータ提供が存在していることを意味する．EMM/MM を見ると，チームによっては 1.41 倍の差が存在している．

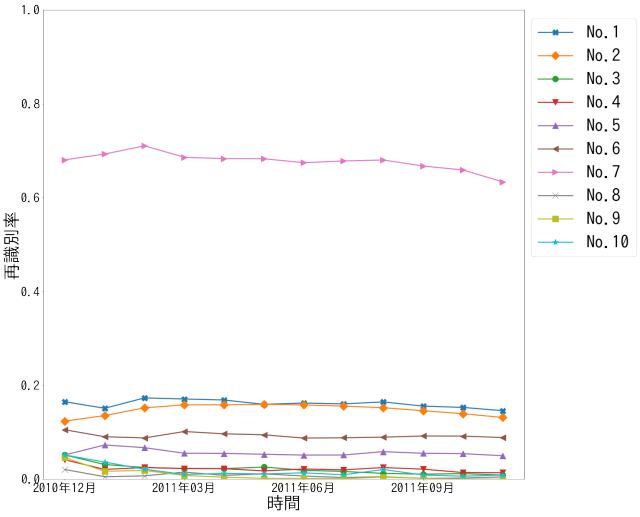

図 2　$S_1 \sim S_6$ の中で最大となる再識別率

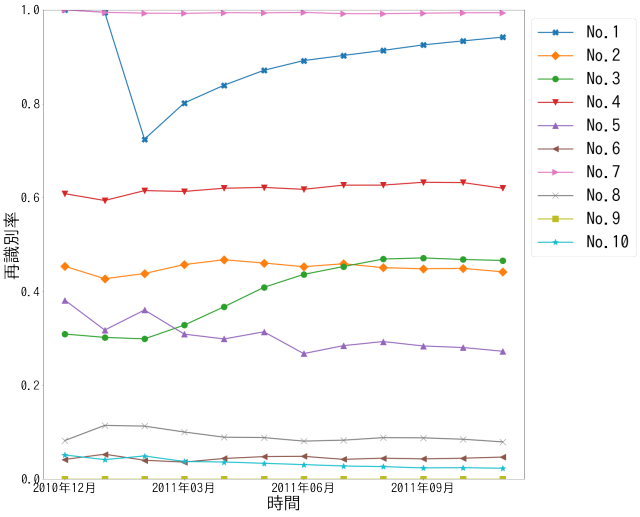

図 3　S_7 の再識別率

6.2 EMM 方式におけるデータ提供時の再識別率

図 2〜4 に，EMM 方式における各データ提供時点（横軸）の再識別率（再識別）を示す．本稿の実験では S_1〜S_8 の再識別アルゴリズムを採用している．ここでは，データ提供時点における再識別率の違いと，再識別アルゴリズムにおける再識別の違いの両方を考察する．

図 2 は，競技ルールとして用意されている S_1〜S_6 の再識別アルゴリズムを EMM 方式で評価した結果である．図

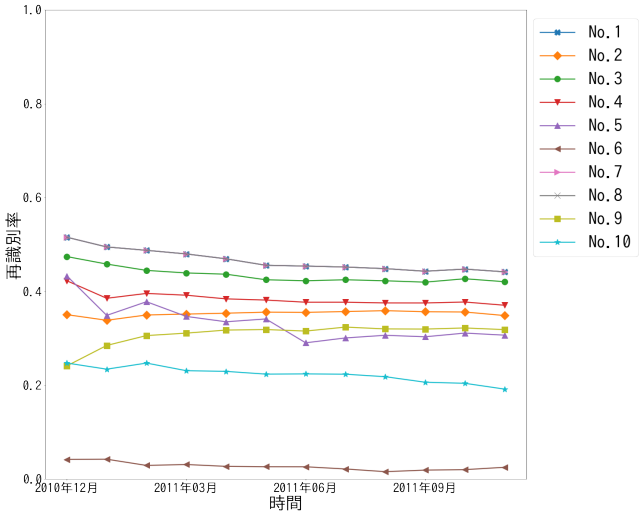

図 4　S_8 の再識別率

では，各データ提供時点における S_1〜S_6 で算出した再識別率の最大値を示している．図 3 は，平均単価の類似度を利用した再識別アルゴリズム S_7 を EMM 方式で評価した結果である．図 4 は，レコード数の類似度を利用した再識別アルゴリズム S_8 を EMM 方式で評価した結果である．以降，それぞれの図について結果と考察を述べる．

6.2.1　競技ルールにおける再識別アルゴリズム

図 2 に，競技ルールとして用意されていた再識別アルゴリズム S_1〜S_6 による，各データ提供時点での再識別率を示す．1 チームを除いては，ほとんどのチームが競技ルールでの再識別アルゴリズムへ対応していたことがわかる．なお，S_1〜S_6 はデータの各項目への少ないノイズ付与で対応できるものが多く，参加チームの対応は比較的容易であったと考えられる．

6.2.2　平均単価に対する再識別アルゴリズム

図 3 に，著者らが用意したユーザの単価平均を利用した再識別アルゴリズムによる，各データ提供時点での再識別率を示す．単純な再識別アルゴリズムであるが，全体として S_1〜S_6 よりも再識別率が高く，再識別率が 1.0 のものもいくつか見られた．しかしながら，S_1〜S_6 よりも低い再識別率のチームも存在していた（チーム No.9）また，データ提供時点によって単価に対する加工の程度が大きく異なるデータも存在していた（チーム No.1）．

6.2.3　レコード数に対する再識別アルゴリズム

図 4 に，著者らが用意したユーザの単価平均を利用した再識別アルゴリズムによる，各データ提供時点での再識別率を示す．レコード数の比較という単純なアルゴリズムではあるが，ほとんどのチームのデータが S_1〜S_6 における再識別率よりも，高い再識別率を示している．しかしながら，S_7 のようにデータ提供時点での大きな違いは確認できない．また，レコード数による再識別が常に低いデータも存在していた（チーム No.6）．

7.　議論

本節では，UM, MM, EMM の各方式がどのような再識別リスクの評価が可能であるかを議論する．

7.1　UM 方式における再識別リスク評価

UM 方式では，12 ヶ月全ての仮名が正しく識別されたユーザの割合で，再識別リスクを評価する．提供元がある特定の期間（例えば 1 ヵ月）のみ強い匿名加工を施した場合，UM 方式ではユーザの再識別リスクが低いと判断されるかもしれない．しかしながら，提供先が加工されている 1 ヵ月を使用しないようなデータの利用を行った場合，そのユーザの再識別リスクは評価されていないこととなる．UM 方式は再識別されるユーザの割合を評価できるが，提供先がデータの全期間を利用することを想定した場合の再識別リスク評価となっている．なお，仮名が提供された全データにおいて統一されている場合，提供先のデータ利用を想定せずに，再識別リスクを評価できる可能性がある．

7.2　MM 方式における再識別リスク評価

MM 方式では，仮名が存在する月における仮名が正しく識別された月の割合で，再識別リスクを評価する．提供元は，データの全期間にわたって十分な匿名加工を行っていなければ，MM 方式で低い再識別リスクの評価を得ることは難しい．そのため，提供先がデータのどのような期間を選択的に利用したとしても，再識別リスクの評価が実施されているといえる．しかし，2 節のようなユースケースを想定すると，MM 方式での評価は複数回にわたるデータ提供の最終時点での評価であり，データ提供毎には評価が行われていない．

7.3　EMM 方式における再識別リスク評価

著者らが提案した EMM 方式では，データ提供毎に MM 方式による再識別率を算出し，全データ提供で最も再識別率が高いものを再識別リスクとして評価する．EMM 方式は，データ提供毎に再識別リスクの増減を評価できる．

複数回にわたるデータ提供では，データ提供の最終時点の判断が不明確な場合が多い．そのため，データ提供の度に再識別リスクを評価し，再識別リスクが増加しないようなデータ加工を施しつつ，データ提供を行うことが望まれる．

8.　まとめ

本稿では，複数回にわたる購買履歴データの匿名加工情報を提供するユースケースを想定し，PWSCUP2017 のデータ

セットを利用して，安全性指標を再考した．PWSCUP2017では，仮名が全て正しく推定されたユーザの割合（UM 方式）で再識別リスクが評価されたが，UM 方式は限られた範囲でのリスク評価であることを指摘した．本稿では，仮名が正しく推定できた月の割合（MM 方式）での再識別リスクの評価方法を拡張し，データ提供毎に再識別リスクを評価する方式（EMM 方式）を提案した．評価実験から，EMM 方式は MM 方式よりも再識別リスクが高いデータ提供を発見することができた．匿名加工情報に対する再識別リスクの評価は，提供のユースケースや提供先の利用方法を考慮して，適切なリスク評価を行うことが重要である．今後においても時系列データの再識別リスクの評価についてさらに発展した議論が期待される．

参考文献

[1] 個人情報の保護に関する法律（平成 29 年 5 月 30 日時点）：https://www.ppc.go.jp/files/pdf/290530_personal_law.pdf. (accessed 2018-06-15).
[2] 個人情報の保護に関する法律施行規則（平成 28 年 10 月 5 日個人情報保護委員会規則第 3 号）：https://www.ppc.go.jp/files/pdf/290530_personal_commissionrules.pdf. (accessed 2018-06-15).
[3] 菊池浩明，小栗秀暢，中川裕志， 野島良，波多野卓磨，濱田浩気，村上隆夫ほか：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， Vol. 2017 (2017).
[4] PWSCUP2017 匿名加工・再識別コンテスト 競技ルール Ver.1.3：https://pwscup.personal-data.biz/web/pws2017/data/PWSCUP2017_ContestRules.pdf. (accessed 2018-06-15).
[5] Chen, D., Sain, S. L. and Guo, K.: Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining, *Journal of Database Marketing & Customer Strategy Management*, Vol. 19, No. 3, pp. 197–208 (2012).
[6] 小栗秀暢：匿名加工とプライバシ保護：4．匿名加工・再識別コンテスト-世界唯一の対戦型データ匿名加工コンテスト PWS Cup-，情報処理， Vol. 59, No. 5, pp. 452–456 (2018).
[7] 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. (accessed 2018-06-15).
[8] 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*, IEEE, pp. 28–35 (2015).
[9] PWSCUP2017 匿名加工・再識別コンテスト 競技ルール Ver.1.0：. (accessed 2018-06-15).
[10] 黒政敦史，小栗秀暢，門田将徳：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〜，技術報告 9，富士通クラウドテクノロジーズ株式会社, 富士通クラウドテクノロジーズ株式会社, 東京大学大学院学際情報学府 (2017).
[11] 濱田浩気：優勝チーム解説と Challenge，https://pwscup.personal-data.biz/web/pws2017/data/PWSMeetup_2_hamada.pdf. (accessed 2018-06-15).
[12] PWSCUP2017 本戦用再識別データ：https://pwscup.personal-data.biz/web/pwscup2017/data/Final.zip. (accessed 2018-06-15).

Vol.2018-CSEC-82 No.18
Vol.2018-SPT-29 No.18
2018/7/25

付　　録

A.1 PWSCUP2017 の設計概要

本付録では，PWSCUP2017 の論文 [3]，および競技ルール [4] から，本稿で利用する変数等を引用し説明する．

マスターデータ $\mathcal{M}$ は，n 人の顧客の情報を格納した n 行 4 列の行列

$$\mathcal{M} = \begin{pmatrix} c_{1,1} & \cdots & c_{1,4} \\ \vdots & \ddots & \vdots \\ c_{n,1} & \cdots & c_{n,4} \end{pmatrix} \tag{A.1}$$

である．ここで，式 (A.1) の列は，表 A·1 に示される顧客 ID，性別，誕生日，国籍を表す．

マスターデータ $\mathcal{M}$ の例を表 A·1 にそれぞれ示す．

表 A·1　マスターデータ $\mathcal{M}$ の例

$c_{\cdot,1}$ 顧客 ID	$c_{\cdot,2}$ 性別	$c_{\cdot,3}$ 誕生日	$c_{\cdot,4}$ 国籍
12360	f	1950/1/1	Others
12361	m	1960/1/1	Germany
12362	m	1950/1/1	France
12363	f	1970/1/1	United Kingdom

トランザクションデータ $\mathcal{T}$ は，期間 $l = 1, \cdots, 12$ についての履歴 $\mathcal{T}^{(1)}, \cdots, \mathcal{T}^{(12)}$ から構成される．期間 l の履歴 $\mathcal{T}^{(l)}$ は m 行 7 列の行列

$$\mathcal{T}^{(l)} = \begin{pmatrix} t^{(l)}_{1,1} & \cdots & t^{(l)}_{1,7} \\ \vdots & \ddots & \vdots \\ t^{(l)}_{m,1} & \cdots & t^{(l)}_{m,7} \end{pmatrix} \tag{A.2}$$

である．ここで，行は，m 個のレコード（履歴），式 (A.2) の列は，表 A·2 の順番に対応する 7 つの属性 (顧客 ID，伝票 ID，購入日，購買時，商品 ID，単価，数量）を表す．

トランザクションデータ $\mathcal{T}$ の例を表 A·2 にそれぞれ示す．

表 A·2　トランザクションデータ $\mathcal{T}$ の例

$t_{\cdot,1}$ 顧客 ID	$t_{\cdot,2}$ 伝票 ID	$t_{\cdot,3}$ 購入日	$t_{\cdot,4}$ 購入時	$t_{\cdot,5}$ 商品 ID	$t_{\cdot,6}$ 単価	$t_{\cdot,7}$ 数量
12362	0	2011/2/17	10:30	21913	3.75	4
12362	0	2011/2/17	10:30	22431	1.95	6
12361	0	2011/2/25	13:51	22630	1.95	12
12361	0	2011/2/25	13:51	22555	1.65	12
12362	0	2011/4/28	9:12	21866	1.25	12
12362	0	2011/4/28	9:12	20750	7.95	2
12360	0	2011/5/23	9:43	21094	0.85	12
12360	0	2011/5/23	9:43	23007	14.95	6

匿名加工データ $\mathcal{S}$ は，期間 $l = 1, \cdots, 12$ についての履歴 $\mathcal{T}^{(1)}, \cdots, \mathcal{T}^{(12)}$ を加工した匿名加工データ $\mathcal{S}^{(1)}, \cdots, \mathcal{S}^{(12)}$ から構成される．期間 l の履歴 $\mathcal{S}^{(l)}$ は m 行 7 列の行列

$$\mathcal{S}^{(l)} = \begin{pmatrix} s^{(l)}_{1,1} & \cdots & s^{(l)}_{1,7} \\ \vdots & \ddots & \vdots \\ s^{(l)}_{m,1} & \cdots & s^{(l)}_{m,7} \end{pmatrix} \tag{A.3}$$

である．$\mathcal{S}$ の形式が $\mathcal{T}$ と同様である．

仮名表 F は，データを加工する際に n 人の顧客 ID の仮名 ID を格納した n 行 13 列の行列

$$F = \begin{pmatrix} c_{1,1} & f^{(1)}(c_{1,1}) & \cdots & f^{(12)}(c_{1,1}) \\ \vdots & \vdots & \ddots & \vdots \\ c_{n,1} & f^{(1)}(c_{n,1}) & \cdots & f^{(12)}(c_{n,1}) \end{pmatrix} \tag{A.4}$$

である．ここで，$f^{(l)}(c_{i,1})$ は，期間 l に，i 番目の顧客 ID$c_{i,1}$ に割り当てた仮名であり，列は，表 A·3 に対応する顧客 ID とその 12 期間の仮名である．ただし，仮名が存在しない場合は，DEL と記載する．

推定仮名表 $\hat{F}$ は，匿名加工データ $\mathcal{S}$ を攻撃する際に n 人の顧客 ID にマッチングした仮名を格納した n 行 13 列の行列

$$\hat{F} = \begin{pmatrix} c_{1,1} & \hat{f}^{(1)}(c_{1,1}) & \cdots & \hat{f}^{(12)}(c_{1,1}) \\ \vdots & \vdots & \ddots & \vdots \\ c_{n,1} & \hat{f}^{(1)}(c_{n,1}) & \cdots & \hat{f}^{(12)}(c_{n,1}) \end{pmatrix} \tag{A.5}$$

である．ここで，$\hat{f}^{(l)}(c_{i,1})$ は，期間 l に，i 番目の顧客 ID$c_{i,1}$ を推定できた仮名を格納する．ただし，仮名が存在しない推定に対して，DEL と記載する．推定仮名表 $\hat{F}$ が仮名表 F と同様な形式を持つ．

表 A·3　仮名表 F の例

$c_{\cdot,1}$ 顧客 ID	$f^{(1)}(c_{\cdot,1})$ 期間 1 仮名 ID	$f^{(2)}(c_{\cdot,1})$ 期間 2 仮名 ID	··· ··· ···	$f^{(11)}(c_{\cdot,1})$ 期間 11 仮名 ID	$f^{(12)}(c_{\cdot,1})$ 期間 12 仮名 ID
12360	61	61	···	61	63
12361	62	62	···	DEL	DEL
12362	31	DEL	···	DEL	31
12363	10	20	···	DEL	40