Computer Security Symposium 2019
21 - 24 October 2019

集合の分割に基づく仮名化手法の実装

福嶋 雄也[1,a)]　古坂 浩輝[1]　満保 雅浩[2]

概要：パーソナルデータの利活用が進む一方，データ提供者のプライバシが侵害される懸念が強まってきている．パーソナルデータの匿名化を行うにあたり，データの匿名性を定量的に評価する指標の検討は重要な課題である．筆者らはこれまで，時系列データにおける仮名の更新頻度の安全性に関して考察をおこなってきたが，当該手法は非常に限定的な状況下でしか利用できなかった．そこで本論文では，より一般的な多重仮名化の安全性評価手法について，集合の分割操作を用いて検討する．

キーワード：プライバシ保護, Web アクセス履歴, 匿名性評価

Consideration of Set-Partition Based Pseudonymisation Method

Kazuya FUKUSHIMA[1,a)]　Hiroki FURUSAKA[1]　Masahiro MAMBO[2]

Abstract: Importance of privacy preservation to personal data have pointed out recently. In this paper, we consider a safety evaluation method for multi-pseudonymisation based on set-partition, which is different from time dependent pseudonymisation.

Keywords: Privacy preserving, Web-access history, Anonymity evaluation

1. はじめに

IoT デバイスの普及等によりパーソナルデータ・ビッグデータの利活用が進んでいる一方，データ所有者のプライバシに関する懸念が強まっている．2017 年には我が国で個人情報保護法が改正されたほか，2018 年には EU で GDPR(一般データ保護規則) が施行され，更に米カリフォルニア州では CCPA2018(2018 年カリフォルニア州消費者プライバシ法) が州議会にて可決されるなど，パーソナルデータの保護法制が世界中で導入され始めている．

このような情勢下で，パーソナルデータの匿名化を行う技術の重要性がより増しており，その中においてデータの匿名性を定量的に評価する指標・手法を確立することは非常に重要である．筆者らはこれまで，事前に定めた一定の頻度ごとに仮名を更新する手法について検討してきた（[1], [2], [3] 等）が，当該手法においては，更新頻度を設定する以外の加工を想定していなかった．そこで本論文では，集合の分割に基づくより一般的な仮名化の安全性について評価を行う手法を提案する．

[1] 金沢大学大学院 自然科学研究科
Graduate School of Natural Science and Technology, Kanazawa University
[2] 金沢大学 理工研究域
Institute of Science and Engineering, Kanazawa University
[a)] k-fukushima@stu.kanazawa-u.ac.jp

2. 事前知識

2.1 仮名化

データセット上のあるユーザの識別子を，容易に復元することが出来ない別の文字列に置き換える手法を仮名化といい，置換した文字列を仮名という．また，特に 1 つの識別子に対して複数の仮名が存在するような仮名化を多重仮名化と呼ぶ．

また [8] において，仮名化は，ユーザの再識別リスクを考慮し以下の性質を満たすべきだとされている．

不可逆性

仮 ID から，仮名化前の識別子が特定されないよ

うにする必要がある．具体的には，ハッシュ化やUUID(Universally Unique Identifier) を用いる手法等が考えられる．

不連続性

履歴データや継続的にデータが提供される場合，同一の識別子に対して割り振られる仮 ID はある一定の期間で更新する必要がある．

非同一性

異なるデータ間において，同一の識別子に対して同一の仮 ID が割り振られないようにする必要がある．

本提案手法では，集合の分割操作を用いて仮名化を実現し，対象とするデータは集合値データ (set-valued data) の形式で表されるものとする．

2.2 データの形式と集合の分割

本論文では，[1] における想定と異なり，単純な属性値データを対象にすることとする．属性値データの例を表 1 に示す．

表 1 本論文で用いるデータの形式

User	Attirubute Value
Tony	A, B, C, E, F, G, I
Gordon	B, C, F, G, J, K, M
David	A, C, D, E, H, K, L, M, N
Theresa	D, E, I, J, L, N, O
Boris	C, D, G, H, J, K, M, O

表 1 の例として Web アクセス履歴データを考えると，Attribute Value の値は各ユーザがアクセスしたドメインを表す．さらに，表 1 で示す形式のデータセット D を以下のように定義する:

$$D = \{(u, A_u)\} \in User \times 2^A$$

ここで $User, A$ はそれぞれユーザ，属性値の全体集合を表す．

また，本論文では集合の分割を次のように定義する．

定義 1 (集合族と添字づけられた集合族).
集合の集合を集合族という．
また，添字集合 I に対し，添字 $i \in I$ が集合族 $\{A_i\}_{i\in I}$ の各要素 A_i に対応しているとき，集合族 $\{A_i\}_{i\in I}$ を，I を添字集合とする添字づけられた集合族という．

定義 2 (部分集合族).
全体集合 Ω に対し，その冪集合 2^Ω の任意の元からなる集合を，Ω 上の部分集合族という．

定義 3 (集合の分割).
全体集合 Ω と Ω 上の添字づけられた部分集合族 $\{A_i\}_{i\in I}$ の各要素 A_i に対して

$$\Omega = \bigcup_{i\in I} A_i, \bigcap_{i\in I} A_i = \emptyset$$

が成り立つとき，$\{A_i\}_{i\in I}$ を Ω の分割であるという．

2.3 既存研究

集合値データに対する匿名化手法として [4], [5], [6] などが提案されているが，いずれも一般化 (Generalization) や抑制 (Suppression) を用いて実現されており，本論文で考察するような分割操作による仮名化に適用するのは難しい．

筆者らの既存研究 [1] においては，時系列データにおいて仮名の更新頻度のみを設定できるような状況を想定しており，本提案手法はこれをより一般化した手法と捉えることができる．

3. 分割による仮名化の定義と安全性

3.1 多重仮名化

2.2 節で定義したデータセット D に対する多重仮名化を次のように定義する．

定義 4 (多重仮名化).
あるユーザ $u \in User$ のデータ A_u と u に対応する添字集合 P_u に対してその分割を生成する操作

$$\mathrm{Pse}_u : A_u \to \{a_u(p)\}_{p\in P_u}$$

を u の多重仮名化という．さらに添字集合 P_u を u の仮名集合といい，P_u の任意の要素 p を u の仮名という．

表 1 を用いて多重仮名化の例を示す．表 1 においてユーザ Gordon に対応する属性値は B, C, F, G, J, K, M であるから

$$A_{Gordon} = \{B, C, F, G, J, K, M\}$$

と書ける．この A_{Gordon} を例えば 3 つの集合に分割するとし，それぞれ

$$a_{Gordon}(G1) = \{B, F\},$$
$$a_{Gordon}(G2) = \{C, K, M\},$$
$$a_{Gordon}(G3) = \{G, J\}$$

とする．これら 3 つの集合の集合

$$\{a_{Gordon}(G1), a_{Gordon}(G2), a_{Gordon}(G3)\}$$

は A_{Gordon} の分割であるから，これを Gordon の多重仮名化データという．さらに $G1, G2, G3$ はそれぞれ Gordon の

仮名であり，集合 $P_{Gordon} = \{G1, G2, G3\}$ が Gordon の仮名集合となる．

また，D の各行の u, A_u を全て対応する仮名 p と仮名化データ $a_u(p)$ で置き換えたデータ

$$D^{\mathrm{P}} = \{(p, a_u(p))\}$$

を D の多重仮名化データという．

3.2 安全性の定義: k-包含

3.1 節では集合の分割に基づいた多重仮名化の形式的な定義について述べたが，本節ではこの多重仮名化の安全性について，多重仮名化データセット D^{P} が満たすべき性質を定義する．本論文では，以下で定義する k-包含という性質を安全性の評価指標として用いる．

定義 5 (k-包含)**.**
あるユーザ $u \in User$ の任意の仮名 $p \in P_u$ と自然数 k に対し

$$|\{x \in User \mid a_u(p) \subset A_x\}| \geq k$$

が成り立つとき，u の多重仮名化データ $\{a_u(p)\}_{p \in P_u}$ は k-包含であるという．また，任意の u に対して k-包含が成り立つとき，多重仮名化データ D^{P} は k-包含であるという．

ここで定義した k-包含という性質は，多重仮名化を行った際に，どの仮名に対応するデータも，少なくとも k 人のユーザのデータの部分集合となることを保証する．
以下，k-包含を，表 1 及び 3.1 節で示した例を用いて説明する．
表 1 における Gordon のデータが $a_{Gordon}(G1) = \{B, F\}, a_{Gordon}(G2) = \{C, K, M\}, a_{Gordon}(G3) = \{G, J\}$ の 3 つの集合に分割されている．このとき，多重仮名化の定義から $a_{Gordon}(G1), a_{Gordon}(G2), a_{Gordon}(G3)$ はいずれも A_{Gordon} の部分集合である．さらに，この 3 つの集合を表 1 の Gordon 以外のユーザのデータと比較すると，それぞれ

$$a_{Gordon}(G1) = \{B, F\} \subset A_{Tony}$$
$$a_{Gordon}(G2) = \{C, K, M\} \subset A_{David}, A_{Boris}$$
$$a_{Gordon}(G3) = \{G, J\} \subset A_{Boris}$$

となることが分かる．よって k-包含の定義から，表 1 で示したデータ D において，ユーザ Gordon のデータを上記のように分割する場合その分割は 2-包含であることになる．

ある多重仮名化データが k-包含 (ただし $k \geq 2$ の場合) を満たす場合，攻撃者が当該データにおけるある仮名に対応するデータ (例えば $a_{Gordon}(G1)$) を入手し，かつオリジナルデータ D にアクセスできた場合でも，攻撃者は Gordon の他の属性値 (すなわち G2, G3 に対応する値) を一意に求めることはできなくなる．

3.3 k-包含の性質

まず，ある多重仮名化データが k-包含であるための必要条件を考える．定義から容易に導ける必要条件の一つとして以下があげられる．

定理 6 (k-包含の必要条件)**.**
多重仮名化データ D^{P} が k-包含となるための必要条件は

$$\forall a \in A,\ |\{x \in User \mid a \in A_x\}| \geq k$$

Proof. D^{P} が k-包含であるとき，すべての $a_u(p)$ に対して

$$|\{x \in User \mid a_u(p) \subset A_x\}| \geq k$$

が成り立つ．このときある $a_u(p)$ について

$$\forall a \in a_u(p),\ |\{x \in User \mid a \in A_x\}| \geq k$$

となるため元の命題は成り立つ． □

上記の必要条件より，ある k の値に対して多重仮名化データが k-包含となる場合，データ上のどの属性値も必ず k 人以上のユーザのデータに属していなければならない．よってより大きな k の値に対して k-包含データを生成しようとすると，出現回数の少ない属性値を削除する必要があるため，そのぶんデータの有用性は低下すると考えられる．逆に，そのような属性値の削除を防ぐためには，k の値は極力小さく取る必要がある．

次に，k-包含データの存在性を示す．まず，無名化を次のように定義する．

定義 7 (無名化)**.**
多重仮名化 $\mathrm{Pse}_u : A_u \to \{a_u(p)\}_{p \in P_u}$ に対して

$$\forall p,\ |a_u(p)| = 1$$

が成り立つ場合を特に無名化という．

無名化されたデータでは，仮名と属性値が一対一で対応する．このとき以下が成り立つ．

定理 8 (無名化データの k-包含性)**.**
定理 6 の必要条件を満たす無名化データは k-包含である．

Proof. 無名化データは仮名と属性値が一対一に対応することから，定理 6 の条件式における a を対応する $a_u(p)$ に置き換えて

$$|\{x \in User \mid a_u(p) \in A_x\}| \geq k$$

とできる．任意の仮名に対してこれが成り立つことから，

無名化データは k-包含である． □

無名化とは，ある仮名に対応する属性値が 1 つしかないような場合を指している．定理 8 より任意の k に対して定理 6 の必要条件を満たすように出現回数の少ない属性値を削除することで無名化データは必ず k-包含となる．すなわち，どのようなデータ D に対しても，定理 6 で示した必要条件を満たす場合は少なくとも一つ k-包含を満たすようなデータを生成する分割が存在する．

4. まとめ

本論文では，集合の分割に基づくより一般的な仮名化の安全性について評価を行う手法を提案した．

謝辞 本研究成果の一部は，情報通信研究機構 (NICT) 委託研究『Web 媒介型攻撃対策技術の実用化に向けた研究開発』の一環として得られたものです．

参考文献

[1] 福嶋 雄也, 北島 祥伍, 満保 雅浩 : 仮名の再同定率に基づく仮名更新頻度の定量的評価, 情報処理学会論文誌, Vol.60, No.9 (印刷中).

[2] 福嶋 雄也, 北島 祥伍, 名古谷 悠樹, 満保 雅浩 : Web 閲覧履歴における仮名更新頻度のプライバシリスク評価, 2019 年暗号と情報セキュリティシンポジウム (SCIS2019), 3C2-2 (2019).

[3] Yuma Takeuchi, Shogo Kitajima, Kazuya Fukushima, Masahiro Mambo: Privacy Risk Evaluation of Re-identification of Pseudonyms, The 14th Asia Joint Conference on Information Security (AsiaJCIS 2019).

[4] He, Yeye, and Jeffrey F. Naughton: Anonymization of set-valued data via top-down, local generalization, Proceedings of the VLDB Endowment 2.1 (2009): 934-945.

[5] Xue, Mingqiang, et al.: Anonymizing set-valued data by nonreciprocal recoding, Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM (2012): 1050-1058.

[6] Chen, Liuhua, et al. : A Sensitivity-Adaptive ρ-Uncertainty Model for Set-Valued Data, International Conference on Financial Cryptography and Data Security, Springer, Berlin, Heidelberg (2016): 460-473.

[7] Gunawan, Dedi, and Masahiro Mambo: Hiding Personal Tendency in Set-valued Database Publication, Proceedings of the 2019 2nd International Conference on Information Science and Systems, ACM (2019): 284-289.

[8] 経済産業省: 事業者が匿名加工情報の具体的な作成方法を検討するにあたっての参考資料（「匿名加工情報作成マニュアル」）Ver1.0, 2016 年 8 月. (http://www.meti.go.jp/policy/it_policy/privacy/downloadfiles/tokumeikakou.pdf, 2019/08/10 閲覧)