# PWSCUP2020コンテスト：AMIC（“Anonymity against Membership Inference” Contest）

千田 浩司[1,a)]　荒井 ひろみ[2]　井口 誠[3]　小栗 秀暢[4]　菊池 浩明[5]　黒政 敦史[6]　中川 裕志[2]
中村 優一[7]　西山 賢志郎[8]　野島 良[9]　長谷川 聡[1]　波多野 卓磨[10]　濱田 浩気[1]　古川 諒[11]
山田 明[12]　渡辺知恵美[13]

**概要**：本稿では，2020 年 8 月 27 日から 10 月 20 日にかけて開催される，匿名化とその攻撃の技術を競うコンテスト PWSCUP2020（通称：AMIC）の設計について述べる．今年で 6 回目となる PWSCUP では，擬似データの安全性指標として機械学習分野等で近年注目されている，「メンバシップ推定」をテーマとした．メンバシップ推定は，匿名化されたデータから，誰のデータが含まれているか推定する攻撃である．これまでの PWSCUP では擬似データの安全性を評価することが困難だったが，メンバシップ推定の導入により可能となる．例えば新型コロナウイルス感染者のデータ分析等，学習データに含まれる対象者の存在自体が機微な場合の安全性指標としてもメンバシップ推定は有用と考えられる．

**キーワード**：PWSCUP，メンバシップ推定，擬似データ，匿名化

# PWSCUP2020 Contest : AMIC (“Anonymity against Membership Inference” Contest)

Koji Chida[1,a)]　Hiromi Arai[2]　Makoto Iguchi[3]　Hidenobu Oguri[4]　Hiroaki Kikuchi[5]
Atsushi Kuromasa[6]　Hiroshi Nakagawa[2]　Yuichi Nakamura[7]　Kenshiro Nishiyama[8]
Ryo Nojima[9]　Satoshi Hasegawa[1]　Takuma Hatano[10]　Koki Hamada[1]　Ryo Furukawa[11]
Akira Yamada[12]　Chiemi Watanabe[13]

***Abstract:*** We introduce the design of PWSCUP2020 (a.k.a. AMIC) contest, which will be held from August 27th to October 20th, 2020, to compete in technologies for *de-identification* and attacks. In the 6th PWSCUP this year, we focus on *membership inference*, which has recently attracted attention in research fields such as machine learning for a privacy measure of synthetic data. Membership inference is an attack to identify whose data is sampled from de-identified data. Unlike the conventional PWSCUP rules, membership inference can evaluate a privacy level of synthetic data. In addition, a privacy measure based on membership inference is also useful when the membership of the target person to the training data set is sensitive, such as analyses using the data of patients infected with a new coronavirus.

***Keywords:*** PWSCUP, Membership Inference, Synthetic Data, De-Identification

[1] NTT セキュアプラットフォーム研究所
NTT Secure Platform Laboratories
[2] 国立研究開発法人 理化学研究所
RIKEN
[3] Kii 株式会社
Kii Corporation
[4] 株式会社富士通研究所
Fujitsu Laboratories Ltd.
[5] 明治大学
Meiji University
[6] 富士通クラウドテクノロジーズ株式会社
FUJITSU CLOUD TECHNOLOGIES LIMITED
[7] 早稲田大学
Waseda University
[8] 株式会社ビズリーチ
BizReach, Inc.

## 1. はじめに

EU による GDPR の施行や我が国の個人情報保護法の改正[*1]等を受け，パーソナルデータの保護は個人，事業者ともに一層関心が高まっている．特に国内外においてパーソナルデータの活用ニーズが急速に高まる中，パーソナルデータの活用と保護の両立が強く求められていると言えよう．このような背景の下，情報処理学会 コンピュータセキュリティ研究会（CSEC）では，2015 年に PWS(Privacy Workshop) 組織委員会を立ち上げ，パーソナルデータの活用と保護を効果的に両立できる技術や規準の発展に資する各種活動を行っている [1]．例えば，パーソナルデータセットの効果的な匿名化方法の探求のため，匿名化とその攻撃の技術を競うコンテスト PWSCUP を毎年開催している．

本稿では，2020 年開催の PWSCUP2020（通称：AMIC (“Anonymity against Membership Inference” Contest)）の設計について述べる．AMIC は，擬似データ[*2]の安全性指標として機械学習分野等で近年注目されている，「メンバシップ推定」をテーマとした．メンバシップ推定は，匿名化されたデータから，誰のデータが含まれているか推定する攻撃である．これまでの PWSCUP では主にレコードリンケージを攻撃の指標，言い換えれば安全性指標としていたが，擬似データ生成等にはそのままでは適用できない．メンバシップ推定を新たに安全性指標とすることで，擬似データ生成等の安全性を PWSCUP で評価できるようになる．メンバシップ推定は，例えば新型コロナウイルス感染者のデータ分析等，学習データに含まれる対象者の存在自体が機微な場合の安全性指標としても有用であると考えられる．

なお偶然にも，機械学習分野のトップ会議 NeurIPS 2020 の competition track として実施されている hide-and-seek privacy challenge[*3]においても，擬似データに対するメンバシップ推定を安全性指標としている [2]．hide-and-seek privacy challenge は，メンバシップ推定に耐性がある有用な擬似データ生成方法を競うコンテストであり，高次元のパーソナルデータセットに対する有効な匿名化方法として擬似データ生成に着目しているようである．

本稿の以降の構成は次のとおりである．2 節で主要な用語や AMIC で用いる既存技術を説明する．3 節，4 節でそれぞれ AMIC の概要および詳細について説明する．5 節で考察を行い，6 節で関連研究を紹介する．最後に 7 節で本稿をまとめる．

## 2. 準備

### 2.1 用語

パーソナルデータセットとは**図 1** に記載のとおり，各行（レコード）に個人のデータが記載された表形式のデータとする．個人と ID は 1 対 1 に対応し，本稿では簡単のため複数行に同一 ID は無いものとする．

攻撃とは，図 1 のレコードリンケージやメンバシップ推定を指す．正解率が高いほど強力な攻撃となる．

匿名化とは，攻撃を防ぐためのパーソナルデータセットの加工を指し，加工後のデータを匿名化データと呼ぶ．図 1 における①から③あるいは⑤までの加工は匿名化である．②や④もパーソナルデータセットであるため，②から③，および④から⑤の加工も匿名化となる．

安全性指標は，攻撃に対する耐性を定量的に評価したものとする．図 1 のレコードリンケージやメンバシップ推定の正解率が安全性指標の例であり，この場合正解率が低いほど安全性が高いと評価できる．

有用性指標は，パーソナルデータセットとその匿名化データからそれぞれ得られる統計量や機械学習の予測・分類結果等の類似性を定量的に評価したものとする．類似しているほど有用性が高いと評価できる．

擬似データ生成とは，パーソナルデータセットの分布に類似した別のデータセットを生成する方法を指し，それによって生成されたデータを擬似データと呼ぶ．擬似データ生成を図 1 の④から⑤への匿名化として用いた場合，メンバシップ推定によって擬似データ生成の安全性を評価することができる．

### 2.2 擬似データ生成

AMIC で利用した，岡田らによる擬似データ生成アルゴリズム（OMTH17 方式と呼ぶ）[3] を説明する．OMTH17 方式は，線形回帰と相関分析の有用性が高い擬似データの生成を目的としている．特に各数値属性の平均および分散共分散行列を指定できるため，パーソナルデータセットの各数値属性の平均および分散共分散行列と完全一致した擬似データを生成できる（ただし実際には離散化や最大小値補正により多少異なる）．カテゴリ属性はダミー変数化することで数値属性として扱えばよい．擬似データのレコード数は任意に設定できる．OMTH17 方式のアルゴリズム概要を以下に記す．

[9] 国立研究開発法人 情報通信研究機構
NICT
[10] 日鉄ソリューションズ株式会社
NS Solutions Corporation
[11] NEC
NEC Corporation
[12] 株式会社 KDDI 総合研究所
KDDI Research, Inc.
[13] 筑波技術大学
Tsukuba University of Technology
[a)] koji.chida.eb@hco.ntt.co.jp

[*1] 2015 年に成立・公布．2017 年 5 月 30 日に施行．その後更に改正法案が 2020 年 3 月 10 日に閣議決定され，同年 6 月 5 日の国会において可決，成立．同年 6 月 12 日に公布．

[*2] 本稿では Synthetic Data（合成データ，人工データ，模造データ）の意訳として用いる．なお「擬似」ではなく「疑似」と表記した文献もあるが，本稿では区別しない．

[*3] https://www.vanderschaar-lab.com/privacy-challenge/

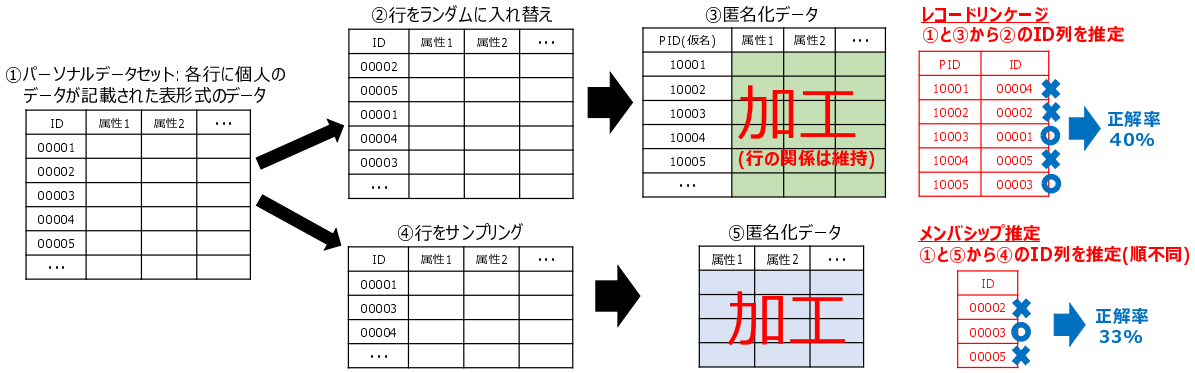


**図 1** レコードリンケージとメンバシップ推定の差異
**Fig. 1** Differences of record linkage and membership inference.

Input: 各属性の平均のベクトル $\boldsymbol{\mu}_D$，分散共分散行列 $\Sigma_D$，各属性のヒストグラム，および擬似データのレコード数
Output: 擬似データ $D'$

(1) 各属性のヒストグラムとレコード数が入力と一致するランダムなデータセット $Y$ を生成する．
(2) $Y$ を白色化し，各属性の平均が 0，分散共分散行列が単位行列となるデータセット $Y'$ を生成する．
(3) $\Sigma_D = U_D \Lambda_D U_D^T$ となる回転行列 $U_D$ および拡大縮小行列 $\Lambda_D^{1/2}$（$\Lambda_D$ の各要素に対して平方根を取った行列）を求める．
(4) $Y'(U_D \Lambda_D^{1/2})^T$ を計算し，各行に $\boldsymbol{\mu}_D$ を足したデータセット $D^*$ を計算する．
(5) $D^*$ に離散化や最大小値補正を行ったものを擬似データ $D'$ として出力する．

上記アルゴリズムにおいて，$D^*$ の各属性の平均，および分散共分散行列はそれぞれ $\boldsymbol{\mu}_D, \Sigma_D$ と完全一致することが証明されている．

## 3. コンテストの概要

AMIC では従来の PWSCUP 同様，各参加チームは「匿名化フェーズ」と「攻撃フェーズ」の両方に参加して得点を競う．各参加チーム（**加工者**と呼ぶ）がそれぞれ匿名化データを作成・提出し，その匿名化データに対して他の全ての参加チーム（**攻撃者**と呼ぶ）が攻撃を行う．全体像を**図 2** に，各処理フローを**図 3** にそれぞれ示す．また以降の説明で用いる記号の一覧を**表 1** に示す．

### 3.1 準備

出題者は，匿名化フェーズで加工者 $i$ が受け取るサンプリングデータ $C_i$ を生成する．パーソナルデータセット $A$ は公開データの Census Income Data Set[4] とし，OMTH17 方式により 10 万レコードの擬似データ $B$ を生成した．なおルールの簡略化のため，$B$ は重複レコードを含まないようにした．

**表 1** 記号
**Table 1** Symbols.

| | |
|---|---|
| $A$ | パーソナルデータセット |
| $B$ | $A$ の擬似データ |
| $C_i$ | 加工者 $i$ のサンプリングデータ |
| $D_i$ | 加工者 $i$ の匿名化データ |
| $E_{ij}$ | 攻撃者 $j$ の加工者 $i$ に対するメンバシップ推定データ |
| $\mathcal{G}$ | 擬似データ生成関数 （$B = \mathcal{G}(\mathsf{seed}, A)$） |
| $\mathcal{H}$ | サンプリング関数 （$C_i = \mathcal{H}(i, B)$） |
| $\mathcal{Q}_i$ | 加工者 $i$ の匿名化関数 （$D_i = \mathcal{Q}_i(C_i)$） |
| $\mathcal{R}_j$ | 攻撃者 $j$ のメンバシップ推定関数 （$E_{ij} = \mathcal{R}_j(B, D_i)$） |
| $S_{ij}$ | 加工者 $i$ の攻撃者 $j$ に対する安全性指標の評価値 |
| $T_k$ | 有用性指標 $k$ の閾値 |
| $U_{ik}$ | 加工者 $i$ の有用性指標 $k$ の評価値 |
| $\mathcal{V}_k$ | 有用性指標 $k$ の評価関数 （$U_{ik} = \mathcal{V}_k(C_i, D_i)$） |
| $\mathcal{W}$ | 安全性指標の評価関数 （$S_{ij} = \mathcal{W}(C_i, E_{ij})$） |

ここで注意すべき点として，AMIC で用いるデータセットは，パーソナルデータセット $A$ そのものではなく，その擬似データ $B$ としている．その理由は，$A$ を用いてサンプリングした場合，加工者は $A$ の非サンプリングデータを用いて匿名化を行い，不正に攻撃者を欺ける可能性を否定できないためである．対策として，匿名化フェーズでは加工者に対して擬似データ $B$ を秘匿する．

サンプリングは，擬似データ $B$ から加工者毎に異なる 1 万レコードをランダムに抽出し，サンプリングデータ $C_i$ として提供する．出題者は $C_i$ を記録しておく（実際には $B$ からサンプリングしたレコードの行番号のみでよい）．

### 3.2 匿名化フェーズ

各加工者 $i$ は，サンプリングデータ $C_i$ を出題者から受け取り，どのレコードがサンプリングされたか識別困難となるよう加工し，匿名化データ $D_i$ として出題者に提出する．なお加工に関しては，

- 所定の有用性指標を満たす，

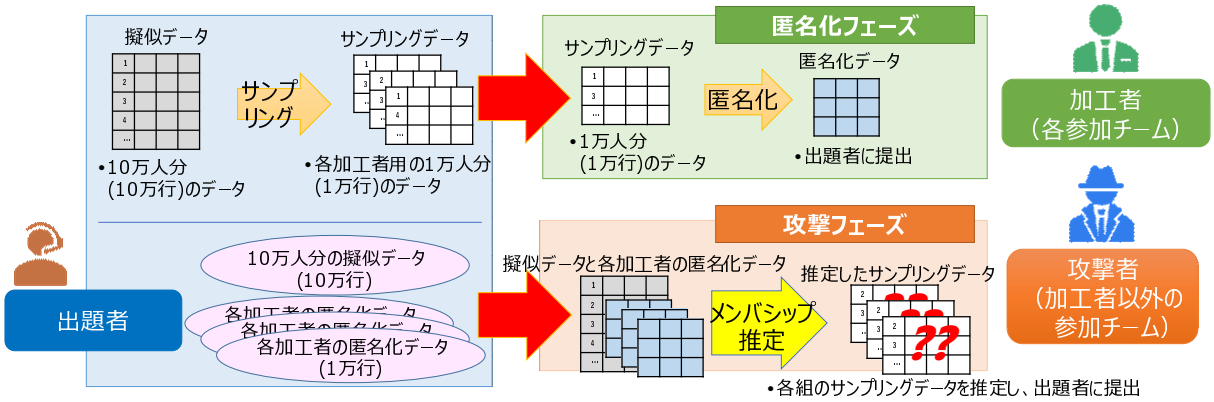


**図 2**　AMIC の全体像
**Fig. 2**　An overall image of AMIC.

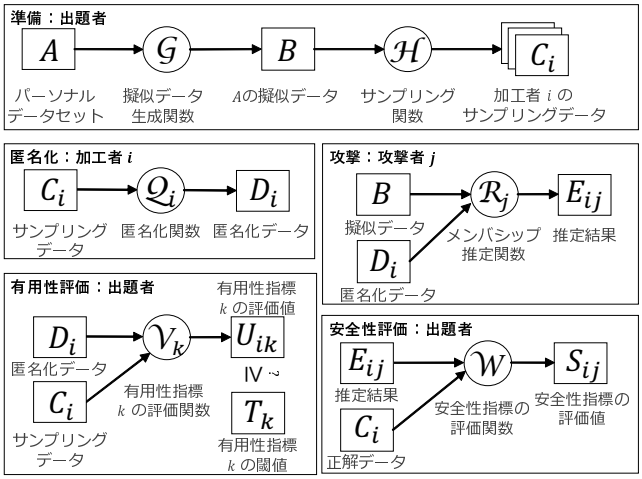


**図 3**　処理フロー
**Fig. 3**　Processing flows.

- 有用性指標の都合上，属性値の値域を変更しない[*4]，

ことが制約条件となる．匿名化データ $D_i$ のレコード数はサンプリングデータ $C_i$ と同一でなくてもよいが，増減するほど一般に有用性が不利になるよう有用性指標を設計している．詳細は 4.2 節で説明する．

### 3.3　攻撃フェーズ

各攻撃者 $j$ は，擬似データ $B$ と，各加工者 $i$ が提出した匿名化データ $D_i$ を出題者から受け取り，各加工者のサンプリングデータ $C_i$ を推定したデータ $E_{ij}$ を提出する．$E_{ij}$ は $C_i$ のレコード全てを推定したものではなく，確信度の高い 100 レコードのみとする．これにより加工者が任意のデータについて推定されないよう匿名化する効果を狙っている．

### 3.4　採点

AMIC では，匿名化部門と攻撃部門のそれぞれで参加チームの勝敗を決めるルールとした．匿名化部門では，攻撃者が推定したサンプリングデータの正解率が低いほど高得点となる．具体的には，全攻撃者の中の最高正解率を 1 から引いた値を得点とする．攻撃部門では，匿名化部門の優勝チームの匿名化データに対して，正解率が高いほど高得点となる．コンテストは従来同様，予備戦（2020/08/27(木)〜2020/09/18(金)）と本戦（2020/09/24(木)〜2020/10/20(火)）からなり，それぞれの得点を 1:9 の割合で合計する．

## 4.　コンテストの詳細（補足）

### 4.1　準備

パーソナルデータセット $A$ に用いた Census Income Data Set は，機械学習への適用を想定した 15 属性からなる訓練用データ 32,561 レコード，テスト用レコード 16,281 レコードのデータセットである．AMIC では各属性の値域や重複度等を考慮し，**表 2** に示す 9 属性を用いた．訓練用データのうち欠損値の無い 30,162 レコードを入力として，OMTH17 方式により重複の無い 10 万レコードの擬似データを生成する．擬似データ生成関数 $\mathcal{G}$ を実装したコード synthetic.py およびサンプリング関数 $\mathcal{H}$ を実装したコード randomsampling.py をホームページ[*5]に公開しており，参加チームは Census Income Data Set からテスト用の擬似データおよびサンプリングデータを作成することができる．

### 4.2　匿名化フェーズ

匿名化関数 $\mathcal{Q}_i$ は，例えば個人情報保護委員会事務局のレポート [5] の「図表 4-3 代表的な加工手法」に記載の手法[*6]のいくつかを効果的に組み合わせた処理が挙げられる．

---

[*4] 例えば Census Income Data Set の属性 age の値域は 17 歳から 90 歳までの年齢だが，これを 20 代，30 代といった年代等に値域を変更（再符号化／一般化）することはできない．

[*5] https://www.iwsec.org/pws/2020/cup20.html

[*6] 項目削除，レコード削除，セル削除，一般化，トップ（ボトム）コーディング，レコード一部抽出（サンプリングを含む），項目一

**表 2** AMIC で用いる Census Income Data Set の属性
**Table 2** Attributes of Census Income Data Set using AMIC.

| 属性名 | 種別 | 値域 |
|---|---|---|
| $X_1$: age | 整数値 | 17 – 90 |
| $X_2$: workclass | カテゴリ | Private 等 8 種類 |
| $X_3$: education | カテゴリ | Bachelors 等 16 種類 |
| $X_4$: marital-status | カテゴリ | Married-civ-spouse 等 7 種類 |
| $X_5$: occupation | カテゴリ | Tech-support 等 14 種類 |
| $X_6$: relationship | カテゴリ | Wife 等 6 種類 |
| $X_7$: sex | カテゴリ | Female, Male |
| $X_8$: hours-per-week | 整数値 | 1 – 99 |
| $X_9$: income | カテゴリ | >50K, <=50K |

加工者 $i$ は，安全性指標の評価値 $S_{ij}$ が高くなるよう匿名化関数 $\mathcal{Q}_i$ を設計し，匿名化データ $D_i$ を生成する．匿名化関数のサンプルコードとして，擬似データ生成関数を実装したコード synthetic.py が利用可能である．その他いくつかの匿名化関数のサンプルコードも用意する予定である．

匿名化データ $D_i$ は，所定の有用性指標の評価値 $U_{ik}$ が閾値 $T_k$ 以上でなければいけない（満たさない場合は失格となる）．有用性指標として，ヒストグラム，分散共分散行列，決定木分析を用いた．ヒストグラム，分散共分散行列，決定木分析の閾値 $T_{\mathsf{histogram}}, T_{\mathsf{COV}}, T_{\mathsf{DTA}}$ はコンテスト開始までに決定，公開する．

ヒストグラムは，匿名化データ $D_i$ とサンプリングデータ $C_i$ の各属性の値域毎の度数を求める．そして $D_i$ と $C_i$ の各度数の差の割合を 1 から引いた値を評価値とする．すなわち，属性 $X_l$ の属性値 $X_{lm}$ について，匿名化データ $D_i$ とサンプリングデータ $C_i$ の度数をそれぞれ $X^D_{lm}, X^C_{lm}$ とすると，評価値は以下のように与えられる．

$$U_{i\,\mathsf{histogram}} = 1 - \frac{\sum_{l,m} \left| X^D_{lm} - X^C_{lm} \right|}{2C_i^{\mathsf{Rec}} C_i^{\mathsf{Att}}} \tag{1}$$

ただし $C_i^{\mathsf{Rec}}$ は $C_i$ のレコード数，$C_i^{\mathsf{Att}}$ は $C_i$ の属性数とする．

分散共分散行列は，匿名化データ $D_i$ とサンプリングデータ $C_i$ の分散共分散行列を求め，各要素の差の合計の逆数を評価値とする．すなわち，簡単のため属性 $X_l$ はカテゴリ属性の場合ダミー変数化されているものと仮定し，属性 $X_l$ の分散 $\sigma_{ll}$，属性 $X_l, X_{l'}$ の共分散 $\sigma_{ll'}$ について，匿名化データ $D_i$ とサンプリングデータ $C_i$ の分散/共分散をそれぞれ $\sigma^D_{ll'}, \sigma^C_{ll'}$ とすると，評価値は以下のように与えられる．

$$U_{i\,\mathsf{VCM}} = \left( \sum_{l,l'} \left| \sigma^D_{ll'} - \sigma^C_{ll'} \right| \right)^{-1} \tag{2}$$

ただし $\sum_{l,l'} \left| \sigma^D_{ll'} - \sigma^C_{ll'} \right| = 0$ のときは $\infty$ を返す．

決定木分析は，匿名化データ $D_i$ とサンプリングデータ $C_i$ からそれぞれ決定木関数 $\mathcal{Y}^D_{X_l}, \mathcal{Y}^C_{X_l}$（$X_l$ は目的変数）を作成し，Census Income Data Set のテスト用データ 16,281 レコードの分類結果の一致数を評価値とする．目的変数は $X_6$: relationship と $X_9$: income の 2 パターンとし，二値分類とするため $X_6$ は Husband とそれ以外に分類する．テスト用データの各レコードを $Z_m$ とすると，評価値は以下のように与えられる．

$$U_{i\,\mathsf{DTA}} = \sum_m F_{\mathsf{EQ}}(\mathcal{Y}^D_{X_l}(Z_m), \mathcal{Y}^C_{X_l}(Z_m)) \tag{3}$$

ただし $F_{\mathsf{EQ}}$ は二入力の値が等しければ 1 を返し，そうでなければ 0 を返す関数とする．

前記のヒストグラム，分散共分散行列，決定木分析の有用性指標は，有用性評価コード utilityfunc.py により加工者自身で評価可能である．ただし，匿名化データとサンプリングデータの属性数や属性値の値域が異なる場合に対応していないため，本コンテストでは属性の項目削除や属性値の値域が変わるような匿名化[*7]を禁止する．匿名化データとサンプリングデータのレコード数が一致しない場合，その差分がヒストグラムの度数の差として積み上がるので，一般にレコード数の増減が大きいほど評価値が下がることに注意されたい．

### 4.3 攻撃フェーズ

攻撃者は，各加工者 $i$ の匿名化データ $D_i$ に対する安全性指標の評価値 $S_{ij}$ が低くなるようメンバシップ推定関数 $\mathcal{R}_j$ を設計し，推定結果を作成する．メンバシップ推定関数は加工者毎に異なる関数としてもよい．

メンバシップ推定関数 $\mathcal{R}_j$ のサンプルコードとして，擬似データ $B$ と匿名化データ $D_i$ の各レコードの距離に基づきサンプリングデータ $C_i$ を推定する関数を実装したコード attack.py が利用可能である．attack.py は，レコード間のユークリッド距離（整数値の属性であれば属性値の差，カテゴリ属性であれば属性値が一致すれば 0，そうでなければ 1 とし，それらの合計）を計算し，匿名化データの各レコードと最も距離が小さい擬似データ $B$ のレコードの行番号を返す．

安全性指標の評価関数 $\mathcal{W}$ は，単純にメンバシップ推定データ $E_{ij}$（100 レコード）に対する $C_i$ との不一致数を返す．すなわち最高値は 100，最低値は 0 となる．実際には擬似データ $B$ のサンプリングされた行番号とメンバシップ推定した行番号のみを入力とする．

### 4.4 アンケート

匿名化フェーズの終了後に，各加工者にどのような匿名化手法を用いたのかアンケートを行う．どのような匿名化

部抽出，ミクロアグリゲーション，丸め（ラウンディング），データ交換（スワッピング），ノイズ（誤差）付加，疑似データ生成．

[*7] 具体的には，セル削除，一般化，丸め等．ただし丸めた値が元の値域に含まれる場合や，セルの値を一般化や削除する代わりに値域内のランダムな値に置き換える加工は問題ない．

**表 3** アンケートの項目
**Table 3** Items of questionnaire.

| | |
|---|---|
| 加工方法 | |
| レコード削除 | |
| ランダム化 | |
| トップ（ボトム）コーディング | |
| レコード一部抽出 | |
| ミクロアグリゲーション | |
| 丸め（ラウンディング） | |
| データ交換（スワッピング） | |
| ノイズ（誤差）付加 | |
| PRAM | |
| 擬似データ生成 | |
| その他： | |
| 安全性指標 | |
| 差分プライバシー | $\epsilon$, $\delta$ の値 |
| $k$-匿名性 | $k$ の値 |
| $Pk$-匿名性 | $k$ の値 |
| $\delta$-存在性 | $\delta$ の値 |
| その他： | |

手法を用いたかは，匿名化データの利用者にとっても重要な情報と考えられるためである．逆に言えば匿名化手法が分かっていても安全性の高い匿名化データの生成が求められる．アンケートは選択式（複数選択可）とし，**表 3** の項目を予定している．利用した加工方法はそれぞれどの属性に対して適用したのかについても回答してもらう予定である．

## 5. 考察

### 5.1 安全性と有用性のバランス

匿名化データは理想的に言えば，元のパーソナルデータセットと遜色の無い有用性の高いデータであり，かつ十分な安全性を有することが望ましい．しかし基本的に有用性と安全性はトレードオフの関係があり，適切なバランスの匿名化データを作成する必要がある．そこで AMIC では，許容レベルと考えられる有用性指標の閾値を設定し，閾値以上の有用性指標を満たす範囲で，どこまで安全性を高められるかを競うルール設計とした．

4 節で紹介した各種実装コードを用いて有用性と安全性の評価を試行したところ，許容レベルと考えられる有用性指標の閾値を満たす匿名化データが生成可能であることを確認できた．具体的な試行結果や閾値については，コンテスト開始までにホームページに公開する．

### 5.2 攻撃

メンバシップ推定の具体的な攻撃について 4.3 節で例示した．ここではその他の攻撃の可能性について簡単に述べる．

まず考えられるのは，有用性の評価値を利用した攻撃だろう．加工者は与えられたサンプリングデータに対し，ヒストグラム，分散共分散行列，決定木分析の結果が類似する匿名化データを生成する．逆に言えば，匿名化データとヒストグラム，分散共分散行列，決定木分析の結果が類似するような擬似データのサブセットがサンプリングデータの候補となり得る．AMIC ではヒストグラム，分散共分散行列，決定木分析の有用性の閾値を公開するため，少なくとも閾値を満たさないような擬似データのサブセットはサンプリングデータではないことが分かる．しかし擬似データのサブセットは ${}_{100000}C_{10000}$ 通りの組み合わせがあるため，単純な検証は容易ではないと考えられる．

前記の攻撃は AMIC の設計に特有のものでなく，一般にメンバシップ推定における匿名化データの安全性と有用性のトレードオフの問題と捉えることができる．すなわち各種統計量や機械学習の予測・分類結果等が元のパーソナルデータセットとほぼ一致するような匿名化データのリスクを示唆している．

### 5.3 レコードリンケージとの関係性

メンバシップ推定は，レコードリンケージでは評価できない匿名化手法の安全性を評価できることを示した．ここではメンバシップ推定とレコードリンケージの安全性と有用性の差異について簡単に考察する．

パーソナルデータセット $A$ について，レコードリンケージおよびメンバシップ推定で安全性を評価可能な匿名化データをそれぞれ $D^L$, $D^M$ とする．そして図 1 にあるように，$A$ の行をランダムに入れ替えたデータセットを $A^L$，同様に $A$ の行をランダムにサンプリングしたデータセットを $A^M$ とする．このとき，$A$ と $A^L$ の情報量は一般に変わらないが，$A^M$ は $A$ に対して情報量が下がるため，有用性も一般に低下する．すなわち有用性に関しては，扱える匿名化手法が豊富なメンバシップ推定において，$A^M$ から $D^M$ を生成する加工がどれだけ有用性を維持できるかがポイントになる．

一方安全性について，レコードリンケージによって再識別されたレコードは，個人とレコードの対応関係が特定される．しかしメンバシップ推定されたレコードは，そのレコードが匿名化データに用いられたことが分かるだけであり，直感的にはメンバシップ推定の方がより高い安全性指標と言えるだろう．実際，$D^M$ の安全性がレコードリンケージでも評価可能であれば，その結果を用いてメンバシップ推定を行う方法が考えられる．すなわち $D^M$ の各レコードについて，レコードリンケージを用いて対応する $A$ のレコードを再識別し，再識別できたレコードをメンバシップ推定したレコードとする．その場合，メンバシップ推定はレコードリンケージの安全性に帰着できる．

### 5.4 差分プライバシーとの関係性

差分プライバシーは，識別不能性に基づく安全性指標の一種であり，あるプライバシー保護方式 $Q' : \mathcal{D} \to \mathcal{R}$ とパラメータ $\epsilon$ について

$$\Pr[Q'(D_0) \in S] \leq e^{\epsilon} \cdot \Pr[Q'(D_1) \in S] \quad (4)$$

となるとき，$Q'$ は $\epsilon$-差分プライバシー（$\epsilon$-DP）を満たすと定義される [6]．ただし $D_0, D_1 \in \mathcal{D}$ は任意の隣接したデータベース，$S$ は $\mathcal{R}$ の任意の部分空間とする．例えば $D_1$ は $D_0$ のレコード $R$ のみ置き換えたものとすれば，$Q'$ が $\epsilon$-DP を満たし $\epsilon$ が十分小さければ，$Q'(D_b)$ $(b \in \{0, 1\})$ から $R$ のメンバシップ推定は困難となる．実際，6 節で紹介するように，ランダムサンプリングを用いて $\epsilon$-DP を満たす $Q'$ の構成法も提案されている．

一方で，差分プライバシーのパラメータとメンバシップ推定の困難性の関係は自明でない．これは理論的な側面と実験的な側面があり，後者については AMIC で，差分プライバシーのメカニズムを適用して匿名化データを生成し，安全性を評価することで知見が得られると考えられる．

## 6. 関連研究

パーソナルデータを保護する技術を競うコンテストとして，NIST 主催による Differential Privacy Synthetic Data Challenge [7] や 1 節で紹介した hide-and-seek privacy challenge [2] が知られる．Differential Privacy Synthetic Data Challenge では，参加チームは $\epsilon$-DP を満たす擬似データ生成アルゴリズム（コード）を開発し，有用性を競う．一方 hide-and-seek privacy challenge では，擬似データ生成アルゴリズム（コード）を開発し，所定の有用性を満たしつつメンバシップ推定を安全性指標とした安全性を競い，AMIC と類似している点が多い．hide-and-seek privacy challenge と AMIC の比較を**表 4** にまとめる．

Nergiz, Clifton は，$\delta$-存在性（$\delta$-Presence）と呼ばれるメンバシップ推定の安全性指標および匿名化アルゴリズムを提案した．$\delta$-存在性は $\delta := \{\delta_{min}, \delta_{max}\}$ とし，匿名化データが与えられたときに任意のレコードの条件付き所属確率が $\delta_{min}$ 以上 $\delta_{max}$ 以下であるとき，その匿名化アルゴリズムは $\delta$-存在性を満たすと定義する．ランダムサンプリングと $k$-匿名化を用いて $\delta$-存在性を満たす匿名化アルゴリズムが提案されている [8]．

Li, Qardaji, Su は，ランダムサンプリングおよびある条件を満たす $k$-匿名化により，差分プライバシーを満たすことを示した [9]．興味深い点は，サンプリング率 $\beta$，差分プライバシーのパラメータ $\epsilon$（および $\delta$），そして $k$-匿名化のパラメータ $k$ の関係を明らかにしたことである．例えば $\beta = 0.1, k = 5, \delta = 10^{-6}$ としたとき，$\epsilon$ はおよそ 0.8 となり，AMIC のサンプリング率を 0.1 とした判断材料となっている．

Wang, Balle, Kasiviswanathan は，ランダムサンプリングおよび Gaussian Mechanism により差分プライバシーを満たすアルゴリズムを提案した [10]．本アルゴリズムは TensorFlow Privacy で実装されている．

菊池は，サンプリングを用いた場合にどのような攻撃者に対してどの程度レコードリンケージに対する効果があるか評価を行っている [11]．$Pk$-匿名性と呼ばれる安全性指標を基に評価し，PRAM による匿名化手法と組み合わせることで任意の背景知識を持つ攻撃者によるレコードリンケージに対して有効であることを示した．

一方，擬似データ生成とメンバシップ推定に関する研究も盛んに行われている．擬似データ生成は学習モデルを生成するための訓練データを効果的に増やす手段として注目されており，GAN(Generative Adversarial Networks) が有名である [12]．しかし学習結果から擬似データを生成する元となるパーソナルデータセットのメンバシップ推定のリスクが指摘されており，様々な対策が提案されている [13]．

## 7. おわりに

本稿では，匿名化とその攻撃の技術を競うコンテスト PWSCUP2020（通称：AMIC）の設計について述べた．AMIC で新たに匿名化の安全性指標として導入したメンバシップ推定は，より多くの匿名化手法が評価可能であるため，匿名化技術の更なる進展に寄与することが期待できる．

今後の課題としては大きく二点挙げられる．一つは，パーソナルデータセット（またはその擬似データ）に対する匿名化データの有用性評価である．AMIC では擬似データではなくそのサンプリングデータに対して匿名化データの有用性を評価している．これはサンプリングデータが加工者毎に異なるため，擬似データに対する匿名化データの有用性を公平に評価することが難しいことによる．類似のコンテスト hide-and-seek privacy challenge では，加工者と攻撃者はデータでなくコードを提出することで，出題者が多数のデータを生成して評価し公平性を向上させている．一方でコードの提出は敷居が高いとの声も聞かれる．

もう一つは，新型コロナウイルス感染者のデータ分析等，対象者の存在自体が機微な場合の安全性指標の定式化である．AMIC ではサンプリングを匿名化の一部と位置付けており，ランダムサンプリングを用いている．しかし特定個人のデータをサンプリングする場合は一般にランダムサンプリングとはならない．そのため，ある条件に基づいてサンプリングされたデータに対するメンバシップ推定の定式化が求められる．

最後に AMIC の日程を**表 5** に記す[*8]．パーソナルデータの活用と保護を効果的に両立できる技術や規準の発展に資する，画期的なアイデアが産まれることを期待したい．

[*8] 変更される可能性があるため，最新版はホームページ `https://www.iwsec.org/pws/2020/cup20.html` を参照されたい．

**表 4** コンテストの比較
**Table 4** Comparison of Two Contests.

| | hide-and-seek privacy challenge [2] | AMIC (ours) |
|---|---|---|
| 元データ | Amsterdam UMCdb（非公開） | Census Income Data Set[4] |
| 入力（匿名化） | サンプリングデータ | 擬似データのサンプリングデータ |
| 出力（匿名化） | 擬似データ生成アルゴリズム（コード） | 匿名化データ |
| 入力（攻撃） | 元データ | 擬似データ，匿名化データ |
| 出力（攻撃） | メンバシップ推定アルゴリズム（コード） | サンプリングデータの推定データ |
| 安全性評価 | メンバシップ推定 | メンバシップ推定 |
| 有用性評価 | 元データとの類似度（閾値を満たせばよい） | サンプリングデータとの類似度（閾値を満たせばよい） |

**表 5** AMIC の日程
**Table 5** Schedule of AMIC.

| | |
|---|---|
| 2020/08/07(金) – 2020/08/26(水) | エントリー受付 |
| 2020/08/27(木) – 2020/09/18(金) | 予備戦 |
| 2020/09/22(火) | 予備戦結果発表 |
| 2020/09/24(木) – 2020/10/20(火) | 本戦 |
| 2020/10/27(火) | CSS2020 にて最終結果発表 |

## 参考文献


[1] PWS 組織委員会：プライバシーワークショップ (PWS)，PWS 組織委員会（オンライン），入手先 ⟨https://www.iwsec.org/pws/⟩（参照 2020-08-14）.

[2] Jordon, J., Jarrett, D., Yoon, J., Barnes, T., Elbers, P., Thoral, P., Ercole, A., Zhang, C., Belgrave, D. and van der Schaar, M.: Hide-and-Seek Privacy Challenge, *arXiv preprint arXiv: 2007.12087* (2020).

[3] 岡田莉奈，正木彰伍，田中哲士，長谷川聡：統計値を用いたプライバシ保護擬似データ生成手法，コンピュータセキュリティシンポジウム 2017(CSS2017) 論文集 3F3-4，情報処理学会 (2017).

[4] Kohavi, R. and Becker, B.: UCI Machine Learning Repository, Census Income Data Set, UC Irvine (online), available from ⟨https://archive.ics.uci.edu/ml/datasets/census+income⟩ (accessed 2020-08-16).

[5] 個人情報保護委員会事務局：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，個人情報保護委員会（オンライン），入手先 ⟨https://www.ppc.go.jp/files/pdf/report_office.pdf⟩（参照 2020-08-13）.

[6] 寺田雅之：差分プライバシーとは何か，システム／制御／情報，Vol. 63, No. 2, pp. 58–63 (2019).

[7] NIST: 2018 Differential Privacy Synthetic Data Challenge, NIST (online), available from ⟨https://www.nist.gov/ctl/pscr/open-innovation-prize-challenges/past-prize-challenges/2018-differential-privacy-synthetic⟩ (accessed 2020-08-13).

[8] Nergiz, M. E. and Clifton, C.: $\delta$-Presence without Complete World Knowledge, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 22, pp. 868–883 (2010).

[9] Li, N., Qardaji, W. H. and Su, D.: On Sampling, Anonymization, and Differential Privacy or, $k$-Anonymization Meets Differential Privacy, *7th ACM Symposium on Information, Compuer and Communications Security (ASIACCS'12)*, ACM, pp. 32–33 (2012).

[10] Wang, Y.-X., Balle, B. and Kasiviswanathan, S. P.: Subsampled Renyi Differential Privacy and Analytical Moments Accountant, *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019*, pp. 1226–1335 (2019).

[11] 菊池亮：サンプリングを用いた際の個人識別リスクの評価，コンピュータセキュリティシンポジウム 2016(CSS2016) 論文集 1A4-3，情報処理学会 (2016).

[12] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y.: Generative Adversarial Nets, *Advances in Neural Information Processing Systems 27 (NIPS'14)*, Curran Associates, Inc., pp. 2672–2680 (2014).

[13] Chen, D., Yu, N., Zhang, Y. and Fritz, M.: GAN-Leaks: A Taxonomy of Membership Inference Attacks against GANs, *arXiv preprint arXiv: 1909.03935* (2019).