合成データ生成の出力を評価するメンバーシップ推論攻撃フレームワーク

三浦 尭之[1,2,a)]　紀伊 真昇[1]　市川 敦謙[1]　岩花 一輝[1]
芝原 俊樹[1]　奥田 哲矢[1]　山本 充子[1]　矢内 直人[2]

概要：合成データ生成技術のプライバシー脅威として，特定個人の元データへの所属の有無を推論するメンバーシップ推論攻撃がある．本稿では，訓練済みモデルは使用せず，出力された合成データからどの程度のプライバシー漏洩が起きうるのかを評価するフレームワークを提案した．また，そのフレームワークの下，元データの統計的な情報に着目した具体的な攻撃手法も構成した．さらに，統計量，ベイジアンネットワーク，あるいはニューラルネットワークを用いた合成データ生成で得られた出力を用いて，元データに対するメンバーシップ推論攻撃を公開データセットを用いた実験で検証した．実験の結果，ランダムな攻撃成功率が 0.5 に対して，ターゲットサンプルを統計的な外れ値になるように選ぶと，いくつかの方式では正答率が 0.7〜0.96 に上昇しメンバーシップ推論攻撃の精度が向上することが確認できた．また，合成手法ごとにメンバーシップ推論がしやすい攻撃方法が異なることも明らかにした．

キーワード：合成データ生成, メンバーシップ推論攻撃, 統計的な外れ値, プライバシー脅威

A Membership Inference Attack Framework for Evaluating Output of Synthetic Data Generation

Takayuki Miura[1,2,a)]　Masanobu Kii[1]　Atsunori Ichikawa[1]　Kazuki Iwahana[1]
Toshiki Shibahara[1]　Tetsuya Okuda[1]　Juko Yamamoto[1]　Naoto Yanai[2]

Abstract: A membership inference attack, which infers whether or not a particular individual belongs to the original data, is known as a privacy threat to synthetic data generation. In this paper, we propose a framework to evaluate privacy leakage from output of synthetic data without using a trained model. We also demonstrate a concrete attack method that focuses on statistical information of the original data under the framework. Furthermore, we conduct experiments on public datasets to verify membership inference attacks on data by synthetic data generation using statistics, Bayesian networks, or neural networks on the original data. As a result, when a target sample is selected as a statistical outlier, in contrast to an attack success rate of 0.5 in a random way, the attack success rate increases from 0.7 to 0.96 for several methods. Thereby, we confirm that the accuracy of the membership inference attack could be improved. We also found that the results of membership inference attacks are different for each synthetic method.

Keywords: synthetic data generation, membership inference attack, statistical outlier, privacy risk

1. はじめに

個人にかかわるデータの利活用の際は，プライバシー保護への十分な配慮が重要である．しかし，多属性の表形式のデータなどの 1 レコード当たりの情報が高次元化すると，従来の k 匿名化 [5] などのプライバシー保護手法では保護されたデータの有用性が著しく落ちてしまうことが知られている．それに対して，高次元のデータに対しても

[1] NTT 社会情報研究所, NTT Social Informatics Laboratories
[2] 大阪大学, Osaka University
[a)] takayuki.miura.br@hco.ntt.co.jp

表 1: 生成モデルへのメンバーシップ推論攻撃の設定比較

	Model	Target Model	Dataset（推論時）	Dataset（target 作成時）
LOGAN [1]	○	GAN, VAE	×	-
GAN-Leaks(White, Grey, Partial Black Box) [2]	○	GAN（VAE）	×	-
GAN-Leaks(Full Black Box) [2]	×	GAN（VAE）	×	-
Stadler，Oprisanu らの手法 [3,4]	×	一般	○	×
本稿	×	一般	○	○

元のデータと類似した性質を持つデータを作り出せる合成データ生成技術に注目が集まっている [6–9]．一般的な合成データ生成技術は，保護すべきデータセットから生成パラメータを抽出し，生成パラメータを用いて元のデータセットと同様の特徴を持つレコードやデータセットを生成する技術である．具体的な手法としては，統計量に基づく手法 [6]，ベイジアンネットワークを用いた手法 [10]，深層ニューラルネットワークを用いた手法 [8] などがある．

こうした合成データ生成技術のプライバシー保護性では未解明なことが多く，攻撃可能性の方向からその安全性を検証する研究も多く提案されている [1,2,4,11,12]．既存の攻撃の中でも特に基本的なものが，特定の個人が訓練データに含まれていたかを当てるメンバーシップ推論攻撃である [13]．例えば，あるサービスを提供する会社が利用者のデータを用いて機械学習モデルや合成データを作成し，第三者が触れるようになっているとする．そうしたモデルやデータに対してメンバーシップ推論攻撃が可能であることは，特定の個人がそのサービスを利用していたか否かを第三者が推論できることを意味し，プライバシーの漏洩になりうる．こうした攻撃の多くの研究は対象モデルを敵対的生成ネットワーク (GAN) [14] としており，特に代表的な攻撃手法は GAN-Leaks [2] である。この中で論じられている攻撃は訓練済みモデルへのアクセスを部分的，もしくは完全に可能にした設定である．

しかし，表形式のデータの匿名化のような文脈で合成データを用いる場合，攻撃者が訓練済みモデルを部分的に得られるという設定は現実的ではない．そうした中間生成物となるモデルパラメータは開示されずに破棄されるからである．それを受けて本稿では，攻撃者が訓練済みモデルにはアクセスできないという設定でも，生成モデルの出力をヒントに行うメンバーシップ推論攻撃に焦点を当てる．この設定下で，合成データに対するメンバーシップ推論攻撃を定量評価するためのゲーム形式のフレームワークを新たに提案する．本稿のフレームワークは既存のもの [3] と比べて，攻撃者は評価に用いるサンプルの生成に，元のデータセットの知見を活用することが可能となる．このため，従来は表現できていないより強い攻撃者を表現することもでき，より包括的な脅威の検討が可能になっている．本稿と既存研究の違いを表 1 に示す。また，既存研究では攻撃者の推論を機械学習モデルを用いて Black Box 的に行っていたが，統計的外れ値に基づく解釈しやすい推論方法を提案した．

本稿では，こうしたフレームワークや提案した推論方法を用いて，公開データセットを用いた実験を行った．実験では，既存研究とは異なり，統計量ベースの外れ値を定義し，それに基づいて攻撃対象となるターゲットレコードを選択した．また，合成手法として，統計量に基づく手法 [6]，ベイジアンネットワークを用いた手法 [10]，深層ニューラルネットワークを用いた手法 [8] を実装し評価した．

実験の結果，ランダムにターゲットサンプルを選んだ場合の攻撃成功率が 0.5 程度なのに対して，ターゲットサンプルを攻撃者が有利なように選ぶと，いくつかの方式では正答率が 0.7〜0.96 に上昇しメンバーシップ推論攻撃の精度が向上することが確認できた．また，合成方法が統計量ベースの場合と機械学習・深層学習ベースの場合で，有効な攻撃手法が異なることも明らかにした．

さらに，発展的な研究として差分プライバシー [15] の適用やターゲットサンプルを増やした設定の対照実験や，Oprisanu らの先行研究 [4] との比較実験も行った．

本稿の貢献をまとめると下記の通りになる；

- 訓練済みモデルを用いずに，出力された合成データを用いて行うメンバーシップ推論攻撃において，従来より強い攻撃者を表現できるフレームワークを提案した．
- 具体的な攻撃手法として，統計的な外れ値に注目したサンプルの作成方式や推論方式も示した．従来の機械学習モデルによる推論より解釈性があり，また攻撃の成功率も高いものとなった．
- 実験を通して，ターゲットサンプルを統計的な外れ値にすることは攻撃の成功率を高めることが確かめられた．また，統計量に基づく合成方法には統計量の誤差に注目した推論方法が有効で，機械学習や深層学習に基づく合成方法は統計的外れ値に注目する推論方法が有効であることを明らかにした．

2. 関連研究

2.1 生成モデルに対するプライバシー攻撃

生成モデルに対して，その訓練データに関する情報を推定する攻撃は，多くが敵対的生成ネットワーク（GAN）[14] などのニューラルネットワークによる生成モデルを対象としていて，代表的な攻撃は大きく 4 種類に分けられる．

まず，訓練済み生成モデルから訓練データの一部のサンプルを部分的，もしくは完全に復元するモデル反転攻撃，データ復元攻撃があげられる [16]．次に，訓練済み生成モデルから，訓練データ全体の性質を推測する Property 推論攻撃という脅威がある [11]．3 つ目が，表形式のデータで学習した生成モデルの出力に対して，特定の個人を狙うのではなく元データにも含まれていたであろう個人を推論するメンバーシップ突合攻撃である [12]．そして 4 つ目が，訓練済みモデルなどの情報から訓練データに特定の個人が含まれていたか否かを推論するメンバーシップ推論攻撃である [2]．上 3 つの脅威は本稿でスコープを当てるメンバーシップ推論攻撃とは攻撃者の目的が異なる．

Chen らの GAN-Leaks [2] や Hayes らの Logan [1] では，画像形式のデータを生成する GAN や VAE などの深層学習ベースの生成モデルに対するメンバーシップ推論攻撃を提案して検証している．また，攻撃の手掛かりとして訓練済みモデルを使用しているため本稿の提案するフレームワークとは攻撃者の能力が異なる．

Stadler，Oprisanu [3,4] らは本稿と同様に，攻撃者が訓練済みモデルにはアクセスできないが，元データにはアクセスできる設定で合成データに対するメンバーシップ推論攻撃を行っている．しかし，これらは攻撃者の仮定が部分的に弱くなっていて，また推論の方式もそれがなぜ有効と考えられるのかが不明瞭である．それに対して，本稿ではより強い攻撃者も表現できるフレームワークを提案しており，また，推論の方式も統計的に意味が明瞭でより解釈しやすいものになっている．

また，こうした攻撃への対策として，通常のメンバーシップ推論攻撃と同様に差分プライバシー [15] に基づく対策があるが，それ以外にもメンバーシップ推論に耐性を持つよう学習フレームワークを設計した PrivGAN [17] など提案されている．本稿では攻撃に関する検討を行っているがこうした対策が有効であるか否かも今後の検討事項である．

2.2 メンバーシップ推論攻撃

メンバーシップ推論は，与えられたサンプルが学習データセットに含まれていたか推論する攻撃である [13,18,19]．基本的な攻撃対象は分類を行う機械学習モデルで，これらの攻撃では GAN は対象にしていない．

メンバーシップ推論攻撃は差分プライバシー [15] により軽減または防ぐことが可能である [20,21]．しかし，差分プライバシーの適用は作成される合成データの品質への影響が大きいことから，近年ではノイズの量を抑えてメンバーシップ推論の対策をする研究が注目されている [22,23]．差分プライバシーによる対策の検討は今後の課題である．

また，モデル抽出やポイズニングなどの別の攻撃を組み合わせることでメンバーシップ推論攻撃が効果的になることも示されている [24–27]．本稿の問題設定をこれらと組み合わせることで，より強力な攻撃の検討も期待できる．

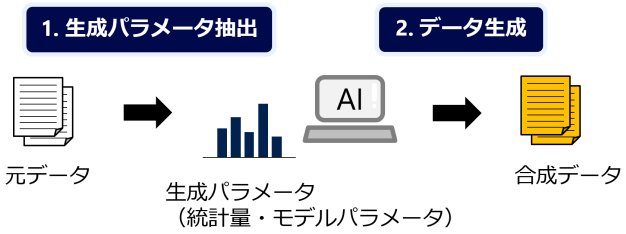


図 1: 合成データ生成の処理手順

3. 提案フレームワーク

本節では合成データに対するメンバーシップ推論攻撃を定量評価するフレームワークを提案する．1 節で述べたとおり，以下に示すフレームワークは既存研究 [3] と比べて，より強い攻撃者を想定したゲーム形式となっている．また，そのゲームを通じた本稿の主たる問いについても述べる．

3.1 合成データ生成

本稿でスコープを当てる合成データ生成は，図 1 のように二段階に分解することができる．一段目の生成パラメータ抽出は，元データ（データセットの取りうる候補の集合を $\mathcal{D}$ とおく）から統計量を計算する，あるいは学習によって機械学習パラメータを得る操作である．ここで，得られた統計量やパラメータを生成パラメータと呼ぶこととする（生成パラメータの取りうる値の集合を $Param$ とおく）．この操作は「生成パラメータ抽出 $\mathrm{EXT} : \mathcal{D} \to Param$」と表せる．合成データ生成技術を差分プライバシー [15] 化する際は，一般のこの生成パラメータ抽出関数をノイズ付加などによってランダム化する．

次に二段目として，その生成パラメータ $\theta \in Param$ を用いてランダムにデータを生成する「データ生成 $\mathrm{Gen}_\theta : \mathbb{R}^* \to \mathcal{D}$」がある．データ生成部では任意のレコード数のデータを生成することができるが，本稿では問題を単純化するため，入力データセットと同じ数のデータを出力することとする．

3.2 合成データ生成に対するメンバーシップ推論攻撃

本稿で提案するゲーム形式のメンバーシップ推論フレームワークの具体的な手法を紹介する．

3.2.1 ゲーム形式

本稿ではメンバーシップ推論の可否をゲーム形式の枠組みで評価する．そのゲームのやり方は次の (1), (2), (3) のステップからなり（図 2），登場人物は Challenger と Adversary の 2 者である．

(1) **Create Datasets**: Challenger がデータセット D を用意[*1]し，Adversary に渡す．Adversary は見分けやす

[*1] ここのデータセットも Adversary が作成する方がより Adversary

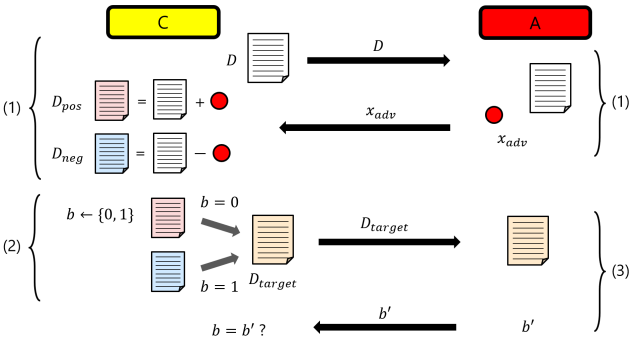


図 2: ゲーム概要：(1) Create Datasets，(2) Data Synthesis，(3) Inference

いサンプル x_{adv} を選択・作成し，Challenger に渡す．Challenger は $D_{pos} = D \cup \{x_{adv}\}$ と $D_{neg} = D \backslash \{x_{adv}\}$ を用意する．

(2) **Data Synthesis**: Challenger がコイントスをし，表が出たら D_{pos} を用いて合成データセット D_{target} を生成，裏が出たら D_{neg} を用いて合成データセット D_{target} を生成し，Adversary に渡す．

(3) **Inference**: 攻撃者は D_{target}，D_{pos}，D_{neg}，x_{adv} の情報から，D_{target} がどちらから来たか（コイントスの結果がどっちだったか）を推測し回答する．

このゲームは全く情報を使わずランダムに回答すれば 1/2 の確率で正解することができるため，Adversary の正答率が 1/2 よりどれくらい大きいかを Adversary の利益とすることで安全性を定量的に評価できる．

既存研究のゲーム形式によるフレームワーク [3] との違いについて述べる．まず，(1) Create Datasets のステップにおいて，既存研究では Adversary が先にサンプル x_{adv} を作成し，Challenger に渡してから，Challenger が合成データセット D_{target} を生成する．これは Adversary の観点からはデータセットの知識がない状態でサンプルを作成することになるため，利益を得にくい．一方，本稿のゲームでは Adversary はデータセット D を渡してからサンプルを作るため，データセットに適したサンプル生成が可能である．このため，本稿の攻撃者の方がより強い設定といえる．

3.2.2 具体的手法

上記フレームワークのメンバーシップ推論で攻撃者がステップ 1, 3 として取りうる具体的な手法を紹介する．

ステップ 1（Create Datasets）：このステップでは Adversary がより見分けやすいサンプル x_{adv} を作成するが，本稿では，より見分けやすいということを統計的な外れ値であることとし，その測り方としてマハラノビス距離 [28] を採用した．

定義 3.1 (マハラノビス距離 [28])．データセット $D = \{x_i \in \mathbb{R}^d\}_{i=1,\dots,n}$ に対して，その平均ベクトル $\mu_D = \frac{1}{n}\sum_{i=1}^{n} x_i$ と分散共分散行列 $\Sigma_D = \frac{1}{n}\sum_{i=1}^{n} x^t x - \mu^t \mu$ を用いて，各点 $x \in \mathbb{R}^d$ につき

$$M(D,x) := \sqrt{{}^t(x-\mu_D)\Sigma_D^{-1}(x-\mu_D)}$$

と定め，この距離を点 x の D に対するマハラノビス距離と呼ぶ．この距離の直観的な意味は「データセット D に対して，相関関係も加味した中で，平均ベクトル μ_D からの距離を測ったもの」である．つまり，$M(D,x)$ が大きい x ほど，データセット D の中では外れ値になっていると考えることができる．

これを用いて行える x_{adv} の作成の仕方は下記の 3 通りである；

- **selective 式：**与えられたデータセット D の中で最も $M(D,x_i)$ が大きいサンプル $x_i \in D$ を x_{adv} とする．
- **adaptive 式：**与えられたデータセット D に対して，データの定義域内で $M(D,x)$ をより大きくする点を探索し x_{adv} とする．
- **random 式：**与えられたデータセット D の中からランダムに 1 レコード選びそれを x_{adv} とする．

ステップ 3（Inference）の詳細：このステップでは攻撃者は D_{target} が D_{pos} から来たのか D_{neg} から来たのかを当てる．まず，平均と分散共分散行列を用いたデータセット間の距離を定義する．

定義 3.2 (mean-variance-loss)．データセット D_1, D_2 と実数 $\lambda > 0$ に対して，

$$MVL(D_1,D_2,\lambda) := (1-\lambda)||\mu_{D_1}-\mu_{D_2}||_2+\lambda||\Sigma_{D_1}-\Sigma_{D_2}||_F$$

と定める．本実験ではデフォルトは $\lambda = \frac{1}{2}$ とする．ここで，$||\cdot||_2$ はベクトルの $L2$ ノルム，$||\cdot||_F$ は行列のフロベニウスノルムである．

次に x_{adv} 近傍データ関数を下記のように定義する．

定義 3.3 (近傍データ関数)．データセット D とサンプル x，正の整数 m について，D の中で x にユークリッド距離で近いサンプル m 個を用意し，その距離の平均を $N(D,x,m)$ と表す．

ステップ 3 での攻撃方法は下記 2 種類（3 方式）である；

- MVL 法
  - orig 式：D_{target}，D_{pos}，D_{neg} に対し，$MVL(D_{target},D_{pos},\frac{1}{2})$ と $MVL(D_{target},D_{neg},\frac{1}{2})$ を計算し，小さい方のデータセットを回答する．
  - syn 式：D_{pos}，D_{neg} から合成データ $D_{pos-syn}$，$D_{neg-syn}$ を作成し，$MVL(D_{target},D_{pos-syn},\frac{1}{2})$ と $MVL(D_{target},D_{neg-syn},\frac{1}{2})$ の小さい方のデータセットを回答する．
- x_{adv} 近傍データ法：x_{adv} を含むデータセットから生成された合成データは x_{adv} に近いデータを含みうるという観察から x_{adv} に近い他のレコードの距離の平均を $t_{target} := N(D_{target},x_{adv},10)$ とする．また，

に有利な攻撃になるがこの場合の検証は今後の課題とする．

表 2: 実験で用いた合成データ生成方式

	生成パラメータ
STAT [6]	統計量
BN [30]	グラフ構造・確率テーブル
CTGAN [8]	深層学習モデルパラメータ

D_{pos}, D_{neg} から合成データ $D_{pos-syn}$, $D_{neg-syn}$ を作成し，$t_{pos} := N(D_{pos-syn}, x_{adv}, 10)$，$t_{neg} := N(D_{neg-syn}, x_{adv}, 10)$ とおく．$t_{target} \leq \frac{t_{pos}+t_{neg}}{2}$ なら D_{pos} と回答，そうでない場合は D_{neg} と回答する．

3.3 今回の問い

本稿の主たる問いは，前節で述べたフレームワークを通じて，合成データ生成がメンバーシップ推論攻撃に対してどのような安全性を満たすか明らかにすることにある．とくに，Create Datasets のステップにおいて， Adversary がデータセット D を見た後にサンプル x_{adv} を生成することで，どのように Adversary の利益が変わるか確認する．

実は既存のフレームワーク [3] では，学習データセットのなかの外れ値に対してのみ攻撃が成功していた．その理由は，Adversary がデータセット D を見る前にサンプル x_{adv} を作ることに起因する．直観的には，データセット D の分布に合わせたサンプルを用意できないためである．

これに対して，外れ値に寄らないより一般的なサンプルに対して攻撃が成功するか，また，どのようなサンプルの生成方法ならば Adversary の利益が向上するか確認する．直観的には，本稿ではデータセット D を見た後にサンプル x_{adv} を作成することで，Adversary はより適応的にサンプル x_{adv} を生成できるようになる．具体的には，データセット D と統計的に最も距離のあるサンプルを用意するなど，D の定義域が取りうる空間をより正確に探索することが可能となる．とくに一般的な統計距離，あるいは，決定木など解釈性の高い機械学習モデルを具体的手法として用いることで，攻撃の成功可否と効果的なサンプルの生成方法を明らかにする．

4. 実験

本実験では前節で提案したフレームワークの具体例を設計し，合成データ生成に対するメンバーシップ推論攻撃の可否を実験的に評価する．

4.1 実験設定

実験で用いたデータセットは Adult Dataset である [29]．これはカテゴリー値，連続値からなる 14 種類の属性を用いて，目的変数である「年収が 5 万ドル以上か否か」を推論するデータセットである．本実験では 3 通りの合成手法を用いて実験を行った（表 2）．実装は Python で行った．1 つ目が統計量をベースとした方式 [6]（STAT と呼ぶ），ベイジアンネットワークを用いた方式 [7]（BN と呼ぶ），深層学習をベースとした方式 [8] である（CTGAN と呼ぶ）．今回の実験では，生成した合成データのレコード数は入力データと同じものとした．

実験のパターンは「x_{adv} の作成方法（random, selective, adaptive）」「合成データ生成手法（STAT, BN, CTGAN）」「Inference の方法（MVL-orig, MVL-syn, x_{adv} 近傍データ法）」の 3 つの組み合わせで記述できる．組み合わせは $3 \times 3 \times 3$ の 27 通りあるが，それぞれのパターンで 500 回ずつ実験を行って，500 回の攻撃の中で正解した割合（正答率）を求めた．ただし，BN 方式と CTGAN 方式はモデルの学習に時間がかかるため，500 回の試行は 5 回の学習とそれぞれ 100 回のデータの生成を組み合わせて実現するものとする．また，偏りを減らすためにステップ 2 の Data Synthesis ではコイントスの結果は D_{pos}, D_{neg} が選ばれる回数がちょうど 250 回ずつになるようにした．

4.2 実験結果

本項では，4.1 項で説明した実験結果について報告する．ステップ 3 の Inference の方式が MVL 法 orig 式であるものが図 3a，MVL 法 syn 式が図 3b，x_{adv} 近傍データ法が表 図 3c である．

図 3a では，STAT 方式以外はほとんどランダムな攻撃と変わらない結果になった．STAT に関しては x_{adv} の作成方法に関して random 式，selective 式，adaptive 式と作成する x_{adv} が統計的な外れ値になればなるほど攻撃の成功率が上がっている．

図 3b については，攻撃の方法が図 3a の手法とほとんど同じにもかかわらず CTGAN 以外はほとんどランダム攻撃と変わらない結果になった．

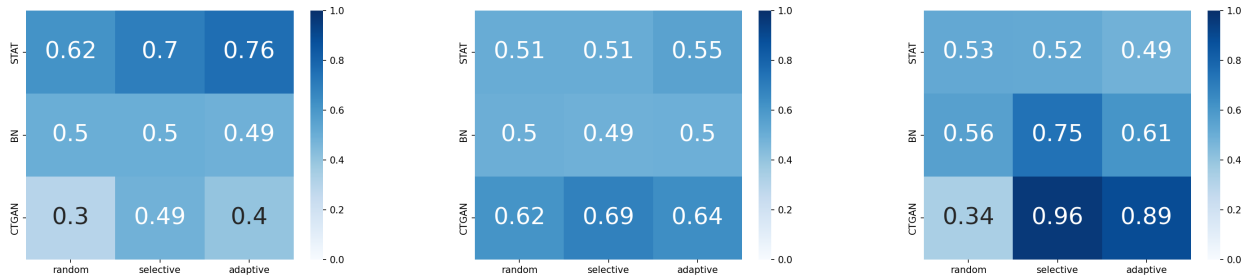
図 3c に関しては，STAT はランダムな攻撃と変わらない結果になった．BN や CTGAN に関しては図 3a, 図 3b のときより高い正答率となった．これは，BN や CTGAN による合成方法では訓練データの外れ値レコードに対して，類似した性質を持つレコードを生成する傾向があることを表す．CTGAN は特に selective 式で正答率が 0.962 という非常に高い結果となった．

実験結果をおおまかにまとめると，

- x_{adv} の作成方法に関しては random より，selective, adaptive の方が攻撃が成功する傾向がある．
- Inference の方式は x_{adv} 近傍データ法の方が成功する傾向がある．
- 合成方法に関しては STAT は MLV 法 orig 式，BN や CTGAN は x_{adv} 近傍データ法で成功する傾向がある．

5. 考察

上記の実験結果を受けて，いくつかの追加検討を行った．実行時間の関係で，本節では合成方法を Bayesian Network



(a) MVL 法 orig 式　(b) MVL 法 syn 式　(c) x_{adv} 近傍データ法

図 3: 各攻撃方法による正答率

表 3: $\varepsilon = 1$ の差分プライベートな BN に対する攻撃

attack 手法	x_{adv} の作成方法	正答率	最大正答率
MVL-orig	random	0.5	0.5
MVL-orig	selective	0.5	0.5
MVL-orig	adaptive	0.5	0.5
MVL-syn	random	0.232	0.232
MVL-syn	selective	0.566	0.686
MVL-syn	adaptive	0.51	0.51
x_{adv} 近傍データ法	random	0.566	-
x_{adv} 近傍データ法	selective	1	-
x_{adv} 近傍データ法	adaptive	0.412	-

による実装に絞って検討を行った．これらの実験も 4 節と同様にそれぞれのパターンにつき，5 回の学習と 100 回のデータの生成を組み合わせて 500 回の試行を行った．

5.1　差分プライバシー適用時

4 節では攻撃可能性の調査の観点から，理論的なプライバシー保護がない状態でメンバーシップ推論の検討を行ったが，一般には差分プライバシー [15] の観点から保護した合成データ生成を用いることが多い．本項では差分プライバシー化すると攻撃の精度がどの程度低下するのかを調査した．差分プライバシー化については $\varepsilon = 1$ とし，Zhang らの実装 [7] を用いた．また，MVL 法 orig 式，syn 式において今回の実験では $\lambda = \frac{1}{2}$ として正答率を測っていたが，λ を $[0, 0.01]$ の区間で動かした場合の正答率の最大値を最大正答率として記載した[*2]．

実験結果は表 3 である．もともとの攻撃性能があまりよくなったこともあり，ほとんどの場合ランダムな攻撃と同様の結果になった．また，差分プライベート化しない場合の実験で高い正答率を出した selective 式で Inference が x_{adv} 近傍データ法による場合が正答率 1 という結果になった．これは差分プライバシーの意味的にはノイズをかけることで識別しやすくなるということはないので，より詳細な追加の検証を行う予定である．

[*2] MVL における平均ベクトルの誤差と分散共分散の誤差の比が 1:10000 程度あったため，計算時間の観点から探索範囲を $[0, 0.01]$ とした

表 4: 機械学習モデル精度：分類精度

モデル	ε	STAT	BN	CTGAN
XGBoost	∞	0.79272	0.83664	0.81504
	1	-	0.79742	-
決定木	∞	0.77526	0.83394	0.80246
	1	-	0.80796	-
SVM	∞	0.78904	0.82484	0.81314
	1	-	0.79012	-

5.2　データの品質について

本稿で提案した攻撃は生成される合成データの品質とも密接に関係している．各種 Adalt Dataset の半数を元データとして合成手法で合成データを生成し，それを訓練データとして学習させた分類器に，残りの半分の Adult Dataset の推論を行わせた推論精度が表 4 である．目的変数は年収が「>50K」か「≤50K」かの 2 値である分類器は XGBoost，決定木，サポートベクターマシン（SVM）を用いた．

結果としては，BN，CTGAN，STAT の順に良い品質となった．ここでいう良い品質とは「それを用いた機械学習モデルの精度がよい」という意味であるが，図 3c からは BN や CTGAN は，STAT と比べると x_{adv} 近傍データ法に弱いことが見て取れる．外れ値の情報を保存することと，機械学習モデル精度という観点からの合成データの品質は，今後より詳細な調査をする意義があると考えられる．

5.3　発展的な攻撃

4 節で検証したメンバーシップ推論は Adult Dataset の 3 万件以上のサンプル[*3]から一人分の変化を見抜けるかという攻撃であった．これ自体は難しい攻撃であるので攻撃成功率が 0.7 程度でも有効な攻撃と考えられるが，この攻撃を少し緩和してより高い成功率を出せる状況を考察した．緩和として，本稿では D_{pos} と D_{neg} の差となる x_{adv} の数を増やす方法を検証した．またこの実験の x_{adv} の作成方

[*3] 欠損値があるものは取り除いて実験を行った．

表 5: サンプル数を増やした際の攻撃成功率：x_{adv} の作成方法は adaptive，合成手法は BN による．

	MVL-orig		x_{adv} 近傍データ法
x_{adv} の数	正答率	最大正答率	正答率
1	0.49	0.5	0.61
10	0.57	0.57	0.702
25	0.696	0.696	0.884
50	1	1	1
75	1	1	1
100	0.806	0.806	1
250	1	1	1
500	1	1	1
750	1	1	1
1000	1	1	1

法はすべて adaptive 方式で行った．ここで，複数の x_{adv} を作成する際，そのまま複数作成しようとするとすべて同じサンプルになってしまうのでまずカテゴリー値の組み合わせを探索し，マハラノビス距離を大きくする異なる組み合わせを必要なサンプル数分用意した．そののちに連続最適化の枠組みで数値属性を決めるという手法で行った．

結果は表 5 である．Inference は MVL-orig と x_{adv} 近傍データ法の 2 種類を用いたが，どちらもサンプル数が増えるに従って攻撃成功率が上昇していて，サンプル数が 50 になったときから攻撃成功率が 1 になるよう非常に良い結果になった．サンプル数 100 のときの正答率 0.806 については，回答の内訳が，5 回行った学習のうち 3 回は 100 回生成された合成データに 100%正答しており，残りの 2 回でほぼランダムな回答（すべての問題に D_{pos} と回答する）となっていたため，何らかの理由により学習がうまくいかなかったと考えられる．

5.4 Stadler，Oprisanu らの攻撃との比較

本項で，Oprisanu ら [3,4] との比較を行う．本稿の手法では，4 節のステップ 3 で述べたように，データセット間の統計的な距離から与えられた合成データの元データセットが D_{neg} か D_{pos} かを出力する．一方，Oprisanu らはそのデータセットの見分け方に機械学習を用いている。本実験を通じて，統計的距離と機械学習それぞれの方法での正答率の違いを確認する．なお，Oprisanu らの x_{adv} の選び方は再現性が取れないため，本項では x_{adv} の選び方を本手法における random, selective, adaptive を採用している．

実験設定：本実験では，合成データ生成アルゴリズムをベイジアンネットワーク，合成データ数を 10 とする．また，機械学習手法として Oprisanu らと同様にランダムフォレスト，ロジスティック回帰，k-近傍法を用いる．

以上の実験設定を踏まえて，実験結果を表 6 に示す．どの手法も分類結果が 0.5 付近，すなわちランダムな推論になっている．これは，上述した設定において，機械学習ベースの手法では合成データの識別が難しいことを示している．また，図 3c の BN の項と表 6 を比較すると，本稿の手法の x_{adv} 近傍データ法の方が精度良く推論できることが確認できる．これは機械学習によるデータセット間の特徴に着目するよりも，攻撃者が選択した x_{adv} の特徴に着目する方が高い精度で推論できることを示唆している．

表 6: Oprisanu ら [3,4] の実験結果

	ランダムフォレスト	ロジスティック回帰	k-近傍法
random	0.51	0.50	0.50
selective	0.43	0.39	0.45
adaptive	0.44	0.55	0.52

5.5 提案フレームワークの応用

前述したフレームワークはメンバーシップ推論攻撃をブラックボックスとして扱っている．つまり，このフレームワークを用いて実際に評価する際は，任意のメンバーシップ推論攻撃を導入することで，様々な応用が可能である．詳細な検討は今後の課題として，以下その応用例を述べる．

例えば代表的な手法として，合成データ生成に着目したメンバーシップ推論攻撃が挙げられる [1,2,17,31]．これらの攻撃では攻撃対象のモデルに特化したホワイトボックスあるいは攻撃者が学習データの母集団の知識を持つような状況を想定できる．直観的には，D が Adversary に与えられることが，母集団の知識に相当する．

また，攻撃者が攻撃対象のモデルを模倣するシャドウモデルを用意する攻撃 [13,32] も利用できる．この攻撃では，D_{target} を与えられた後に，Adversary は D を複数のサブセットに分け，それらをから合成データ D_{pos} と D_{neg} を作成し，自らのシャドウ生成モデルを複数用意し，それぞれ学習させる．各シャドウ生成モデルが生成したサンプルを集め，ラベル付けする．この処理をサンプル x_{adv} を変えながら繰り返し行う．最後に Adversary はメンバーシップ推論用モデルをラベル付けしたデータセットで学習することで，攻撃を行う．

6. 結論

本稿では，出力された合成データからどの程度のプライバシー漏洩が起きうるのかを評価するフレームワークを提案した．これは既存研究 [3,4] より強い攻撃者も表現できるフレームワークであり，また，具体的な統計的外れ値に基づく推論手法も提案し，実験を通してその有効性を確かめた．特にベイジアンネットワークや深層学習ベースの合成データ生成に対しては，統計的な外れ値を用いた提案攻撃の成功率が高くなることが確認できた．

参考文献

[1] Hayes, J., Melis, L., Danezis, G. and De Cristofaro, E.: Logan: Membership inference attacks against generative models, *arXiv preprint arXiv:1705.07663* (2017).

[2] Chen, D., Yu, N., Zhang, Y. and Fritz, M.: Gan-leaks: A taxonomy of membership inference attacks against generative models, *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*, pp. 343–362 (2020).

[3] Stadler, T., Oprisanu, B. and Troncoso, C.: Synthetic Data – Anonymisation Groundhog Day, *Proc. of USENIX Security 2022*, pp. 1451–1468 (2022).

[4] Oprisanu, B., Ganev, G. and De Cristofaro, E.: On Utility and Privacy in Synthetic Genomic Data, *Proceedings 2022 Network and Distributed System Security Symposium. NDSS*, Vol. 22, pp. 1–17 (2022).

[5] Sweeney, L.: k-anonymity: A model for protecting privacy, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol. 10, No. 05, pp. 557–570 (2002).

[6] 岡田莉奈，正木彰伍，長谷川聡，田中哲士：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 論文集，Vol. 2017, No. 2 (2017).

[7] Zhang, J., Cormode, G., Procopiuc, C. M., Srivastava, D. and Xiao, X.: PrivBayes: Private Data Release via Bayesian Networks, *ACM Trans. Database Syst.*, Vol. 42, No. 4 (online), DOI: 10.1145/3134428 (2017).

[8] Xu, L., Skoularidou, M., Cuesta-Infante, A. and Veeramachaneni, K.: Modeling Tabular data using Conditional GAN, *Advances in Neural Information Processing Systems*, Vol. 32, pp. 7335–7345 (2019).

[9] Takagi, S., Takahashi, T., Cao, Y. and Yoshikawa, M.: P3GM: Private high-dimensional data release via privacy preserving phased generative model, *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, IEEE, pp. 169–180 (2021).

[10] Holohan, N., Braghin, S., Mac Aonghusa, P. and Levacher, K.: Diffprivlib: the IBM differential privacy library, *ArXiv e-prints*, Vol. 1907.02444 [cs.CR] (2019).

[11] Zhou, J., Chen, Y., Shen, C. and Zhang, Y.: Property Inference Attacks Against GANs, *Proceedings 2022 Network and Distributed System Security Symposium. NDSS*, Vol. 22, pp. 1–17 (2022).

[12] Hu, A., Xie, R., Lu, Z., Hu, A. and Xue, M.: TableGAN-MCA: Evaluating Membership Collisions of GAN-Synthesized Tabular Data Releasing, *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pp. 2096–2112 (2021).

[13] Shokri, R., Stronati, M., Song, C. and Shmatikov, V.: Membership inference attacks against machine learning models, *2017 IEEE Symposium on Security and Privacy (SP)*, IEEE, pp. 3–18 (2017).

[14] 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*, Vol. 27 (2014).

[15] Dwork, C.: Differential privacy, *International Colloquium on Automata, Languages, and Programming*, Springer, pp. 1–12 (2006).

[16] Xia, W., Zhang, Y., Yang, Y., Xue, J.-H., Zhou, B. and Yang, M.-H.: Gan inversion: A survey, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

[17] Mukherjee, S., Xu, Y., Trivedi, A., Patowary, N. and Ferres, J. L.: privGAN: Protecting GANs from membership inference attacks at low cost to utility., *Proc. Priv. Enhancing Technol.*, Vol. 2021, No. 3, pp. 142–163 (2021).

[18] Salem, A., Zhang, Y., Humbert, M., Berrang, P., Fritz, M. and Backes, M.: ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models, *Proc. of NDSS 2019*, The Internet Society (2019).

[19] Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A. and Tramèr, F.: Membership Inference Attacks From First Principles, *Proc. of IEEE S&P 2022*, IEEE, pp. 1897–1914 (2022).

[20] Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K. and Zhang, L.: Deep learning with differential privacy, *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318 (2016).

[21] Yeom, S., Giacomelli, I., Fredrikson, M. and Jha, S.: Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting, *Proc. of CSF 2018*, IEEE, pp. 268–282 (2018).

[22] Jayaraman, B. and Evans, D.: Evaluating Differentially Private Machine Learning in Practice, *Proc. of USENIX Security 2019*, USENIX Association, pp. 1895–1912 (2019).

[23] Jagielski, M., Ullman, J. and Oprea, A.: Auditing Differentially Private Machine Learning: How Private is Private SGD?, *Advances in Neural Information Processing Systems*, Vol. 33, Curran Associates, Inc., pp. 22205–22216 (2020).

[24] Tramèr, F., Shokri, R., Joaquin, A. S., Le, H., Jagielski, M., Hong, S. and Carlini, N.: Truth Serum: Poisoning Machine Learning Models to Reveal Their Secrets, *arXiv preprint arXiv:2204.00032* (2022).

[25] Mahloujifar, S., Ghosh, E. and Chase, M.: Property Inference from Poisoning, *Proc. of IEEE S&P 2022*, IEEE, pp. 1569–1569 (2022).

[26] HIDANO, S., MURAKAMI, T., KATSUMATA, S., KIYOMOTO, S. and HANAOKA, G.: Model Inversion Attacks for Online Prediction Systems: Without Knowledge of Non-Sensitive Attributes, *IEICE Transactions on Information and Systems*, Vol. E101.D, No. 11, pp. 2665–2676 (2018).

[27] Lyu, L., He, X., Wu, F. and Sun, L.: Killing Two Birds with One Stone: Stealing Model and Inferring Attribute from BERT-based APIs, *CoRR*, Vol. abs/2105.10909 (online), available from ⟨https://arxiv.org/abs/2105.10909⟩ (2021).

[28] Chandra, M. P. et al.: On the generalised distance in statistics, *Proceedings of the National Institute of Sciences of India*, Vol. 2, No. 1, pp. 49–55 (1936).

[29] Dua, D. and Graff, C.: UCI Machine Learning Repository (2017).

[30] Pearl, J.: *Probabilistic reasoning in intelligent systems: networks of plausible inference*, Morgan kaufmann (1988).

[31] Hilprecht, B., Härterich, M. and Bernau, D.: Monte Carlo and Reconstruction Membership Inference Attacks against Generative Models, *Proceedings on Privacy Enhancing Technologies*, Vol. 2019, No. 4, pp. 232–249 (2019).

[32] Pyrgelis, A., Troncoso, C. and Cristofaro, E. D.: Knock Knock, Who's There? Membership Inference on Aggregate Location Data, *Proc. of NDSS 2018* (2018).