連合学習のための知識抽出法に対するデータ再構築攻撃

水門 巧実[1,a)]　小泉 佑揮[1,b)]　武政 淳二[1,c)]　長谷川 亨[2,d)]

概要：分散データに対するプライバシーを保護した知識抽出手法として，連合特異値分解が提案されている．このフレームワークは，分散データに対する特異値分解を可能にし，元データを秘匿しながら分解結果のみを参加者に公開する．この分解結果は，分散データに対する主成分分析を実現するだけでなく，連合学習と従来の集中型学習の間に存在するギャップを埋めるための補助データとして利用することも可能である．しかし，既存研究は連合特異値分解を行うための秘密計算プロトコルの設計にのみ焦点が当てられており，最終的に共有される分解結果によって損なわれるプライバシーについては議論されていない．本稿では，連合特異値分解のプロトコルに従うセミオネストな敵対的参加者が，アクセス可能な分解結果から他の参加者のデータを高精度に再構築するデータ再構築攻撃を提案する．主なアイデアは，特異値分解に固有の数学的性質を悪用することであり，敵対者はデータ再構築のための最適化問題について限定された探索空間で解析することができる．最後に，複数の現実的なデータセットで攻撃手法の評価を行う．

キーワード：連合学習，知識抽出，特異値分解，プライバシー

A Data Reconstruction Attack against a Knowledge Extraction Method for Federated Learning

Takumi Suimon[1,a)]　Yuki Koizumi[1,b)]　Junji Takemasa[1,c)]　Toru Hasegawa[2,d)]

Abstract: Federated Singular Value Decomposition (Federated SVD) has been proposed as a privacy-preserving knowledge extraction method for distributed data. This framework enables SVD on distributed data, revealing only decomposition results to participants while concealing raw data. These results can be used for Principal Component Analysis on distributed data and as auxiliary data to bridge the gap between federated and centralized learning. However, existing research focuses solely on designing secure computation protocols for Federated SVD, neglecting privacy risks associated with shared decomposition results. This paper proposes a data reconstruction attack by semi-honest adversarial participants following the Federated SVD protocol, who can accurately reconstruct other participants' data from accessible decomposition results. The main idea is to exploit mathematical properties inherent to SVD, allowing adversaries to analyze the optimization problem for data reconstruction within a constrained search space. Finally, the attack method is evaluated using multiple real-world datasets.

Keywords: Federated learning, Knowledge extraction, Singular value decomposition, Privacy

[1] 大阪大学 大学院情報科学研究科
Graduate School of Information Science and Technology, Osaka University, Japan
[2] 島根大学 材料エネルギー学部
Faculty of Materials for Energy, Shimane University, Japan
[a)] t-suimon@ist.osaka-u.ac.jp
[b)] ykoizumi@ist.osaka-u.ac.jp
[c)] j-takemasa@ist.osaka-u.ac.jp
[d)] t_hasegawa@mat.shimane-u.ac.jp

1. はじめに

大規模かつ高品質な学習データは高品質な機械学習モデルの構築に不可欠である．しかし，実社会では，このようなデータは一般的に複数のパーティに分散している．従来の集中型学習手法は，学習を司るサーバによる全ての学習データへのアクセスを前提とするため，医療・金融等の高

度なデータ機密性が要求される領域への適用が難しい．連合学習 [1] はこの問題を緩和し，各パーティの学習データをプライベートに保ちながらも協調したモデル訓練を実現するための手法として提案された．その一方で，連合学習ではサーバが学習データ全体にアクセスできないため，データ前処理やデータポイズニング攻撃の検出といった学習データに対する分析を必要とする操作が困難となり，集中型学習と比較して実用性が制限される．

複数パーティに分散したデータの活用を促進するための技術として，連合特異値分解 [2], [3], [4], [5] が注目されている．連合特異値分解は，各パーティの学習データを秘匿しつつ，学習データ全体に対して行列分解手法である特異値分解を適用し，その分解結果をパーティ間で共有する手法である．この分解結果は，分散した学習データ全体についての知識として機能し，上述した連合学習特有の課題の解決をサポートするデータとして活用できる．連合特異値分解を組み込んだ連合学習システムは実用性の面で確かな効果を示しており，今後も更なる発展と普及が期待される．

その有用性の一方で，既存研究では共有される分解結果から漏洩する学習データのプライバシーについては十分に検証されておらず，分解結果の一部のみ共有することだけをプライバシー保護の根拠としている．学習データの機密性を根本的な動機とする連合学習において，このアプローチの正当性は慎重に議論する必要がある．

本稿では，敵対者が，知識として共有される分解結果を用いて他のパーティの学習データを再構築するデータ再構築攻撃が可能であることを示す．提案攻撃は，特異値分解に固有の数学的特性である線形性と正規直交性に基づく．具体的には，元データと分解結果の間の線形性により，敵対者は少数のパラメータを推定するだけで他のパーティが提供した学習データを再構築することができる．加えて，分解結果の正規直交性により，敵対者が推定するパラメータについての探索空間が大幅に制限することができ，データ再構築のための最適化問題を効率的に解析することができる．

本稿の貢献は次の二点である．

- 我々の一連の研究は，連合特異値分解のプライバシーの脆弱性を指摘する最初の研究である．さらに，我々のこれまでの研究 [6] では，ニューラルネットワークベースの攻撃手法を採用していたのに対して，本稿では，脆弱性の本質が特異値分解に固有の線形性と正規直交性が脆弱性をもたらしていることを特定し，より効率的な攻撃手法を提案する．
- 複数の現実的なデータセットを使用して敵対者の知識別に提案攻撃の性能を評価し，連合特異値分解を組み込んだ連合学習システムに対して提案攻撃が脅威となることを示す．

本稿の構成は以下の通りである．2 章では，現在提案されている連合特異値分解プロトコルとそのアプリケーションとしての連合学習を紹介する．3 章では，必要な予備知識として連合特異値分解の定式化を行い，4 章で本稿で想定するシステムモデル，敵対者モデルの定義を行う．5 章で提案する提案攻撃手法を詳しく説明し，6 章でその性能を評価する．最後に，7 章で本稿をまとめる．

2. 関連研究

本章では，連合特異値分解プロトコルを提案する既存研究と，連合学習における連合特異値分解のアプリケーションについて紹介する．ここで紹介するプロトコルやアプリケーションは，本稿で提案するデータ再構築攻撃の標的になり得る．

2.1 連合特異値分解プロトコル

連合特異値分解は，データのプライバシーを保護しつつ分散したデータに対して特異値分解を実行し，その結果をパーティに配布するためのフレームワークである．

差分プライバシーを用いた手法 [7] では，各パーティは非可逆的なランダムノイズを共有データに付加することでそのプライバシーを保護する．この手法は簡潔であり利用が容易だが，ノイズによるデータの歪みによって特異値分解の精度が低下し，これは分解結果を活用するタスクの性能を著しく損なう．一方，準同型暗号を利用した手法 [8], [9], [10] は，データを計算可能な暗号文に変換することで精度を損なわずにデータプライバシーを保護する．しかし，この手法は計算と通信の双方に多大なオーバヘッドが伴い，現実的な運用には課題が多い．

Chai ら [5] は，これらの課題を同時に解決するフレームワークとして FedSVD を提案した．FedSVD は特異値分解専用の可逆ノイズを使用することで，高精度の特異値分解結果を導出および配布することが可能となる．また，計算・通信コストの面で効率的であるため，この手法は大規模な分散データに対しても軽量に特異値分解を行うことが可能である．さらに，Chai ら [2] は FedSVD の機能をサーバレスで実現する Excalibur を提案しており，サーバに起因するプライバシーの脅威を解決している．

連合特異値分解プロトコルは様々な秘密計算手法を活用して提案されているが，これらの研究の全ては特異値分解を秘密に行うためのプロトコルの設計に焦点を当てており，共有される分解結果から漏洩するプライバシーの検証は行われていない．本稿では，特異値分解の特性に注目し，これまで看過されてきた連合特異値分解におけるプライバシーの脆弱性を明らかにする．

2.2 連合特異値分解を組み込んだ連合学習

連合学習は，各パーティが学習データの代わりにモデルパラメータのみをサーバに送信することで，学習データを

秘匿しながら共同して学習を進める．連合学習はサーバが学習データを収集する従来の集中型学習の代替手法とみなされるが，サーバやパーティを含む誰もが学習データ全体にはアクセスできないというこの設計は，連合学習に固有の課題を引き起こしている．そこで，連合特異値分解を連合学習に組み込み，共有される分解結果を学習データ全体についての知識として活用することで，この課題を緩和することが可能となる．

連合特異値分解の代表的な活用法として，学習データの前処理が挙げられる．その目標は，学習データの次元削減 [5], [11] および冗長な特徴量の除去 [7], [10], [12] を通じて，学習モデルのオーバーフィッティングを回避し，汎化性能を向上させることにある．連合特異値分解で共有される分解結果を活用することで，各パーティは自身の所有する学習データに効果的な前処理を施すことが可能となる．

学習データの前処理の他にも，連合特異値分解を組み込むことで連合学習の実用性の向上を達成するシステムが複数提案されている．Feng ら [13] は，勾配反転攻撃が懸念によりサーバが集約されたモデルパラメータにしかアクセスできない場合に，集約されたモデルパラメータに対して連合特異値分解を施すことで，悪意のあるモデルパラメータを特定してデータポイズニング攻撃を検出する DPFLA を提案している．Huang ら [14] および Yin ら [15] は，連合特異値分解とオートエンコーダを組み合わせることで，連合学習における学習データ内の欠落値を補間し，訓練と推論に利用可能なデータ量を効果的に拡張させる手法を提案している．

3. 連合特異値分解

本章では，提案攻撃を理解するための準備として，連合特異値分解を説明する．本稿で主に使用する表記を表 1 にまとめる．

本稿では，データ行列 $\mathbf{X} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{F}|}$ を考える．この行列は，$|\mathcal{S}|$ 個のレコードと $|\mathcal{F}|$ 種類の特徴量で構成される表形式データであり，各行が特定のレコードに，各列が特定の特徴量に対応する．一般的に，表形式データはレコード数が特徴量数を大幅に上回る．そのため，本稿においても $|\mathcal{F}| < |\mathcal{S}|$ と仮定する．

特異値分解は，データ行列 $\mathbf{X}$ を 3 つの行列の積に分解する．

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{1}$$

ここで，行列 $\mathbf{U} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ と $\mathbf{V} \in \mathbb{R}^{|\mathcal{F}|\times|\mathcal{F}|}$ は共に直交行列であり，行列 $\mathbf{\Sigma} \in \mathbb{R}^{|\mathcal{F}|\times|\mathcal{F}|}$ は $\mathbf{X}$ の特異値を降順に含む対角行列である．

連合特異値分解は，データ行列 $\mathbf{X}$ の内容が複数のパーティに分散している場合に，それぞれのデータを他のパーティに公開することなく $\mathbf{X}$ に対して特異値分解を行うための手法である．実際のアプリケーションでは， $\mathbf{X}$ の主なデータ分散シナリオとして垂直分散と水平分散が考えられる．ここで，垂直分散は $\mathbf{X}$ を転置することで水平分散へと容易に変換できるため，本稿では一般性を失わずに垂直分散のみ考える．この場合，N $(N \geq 2)$ 人のパーティのそれぞれは，1 つ以上の異なる特徴量の $|\mathcal{S}|$ 個全てのレコードについてのデータを持つ．

表 1 本稿で使用されている表記

表記	説明
N	連合特異値分解 (連合学習) のネットワークに参加するパーティ数
$\mathbf{X}_\alpha, \mathbf{X}_\beta$	敵対者と被害者が所有するデータ行列
$\mathcal{F}_\alpha, \mathcal{F}_\beta$	$\mathbf{X}_\alpha$ と $\mathbf{X}_\beta$ を構成する特徴量の集合
$\mathbf{X}$	連合特異値分解で扱われる統合データ行列 ($\mathbf{X} = [\mathbf{X}_\alpha \vert \mathbf{X}_\beta]$)
$\mathcal{S}$	$\mathbf{X}$ を構成するサンプルの集合
$\mathcal{F}$	$\mathbf{X}$ を構成する特徴量の集合 ($\mathcal{F} = \mathcal{F}_\alpha \cup \mathcal{F}_\beta$)
$\mathbf{U}, \mathbf{V}^T$	$\mathbf{X}$ の特異値分解結果として得られる直交行列
$\mathbf{\Sigma}$	$\mathbf{X}$ の特異値分解結果として得られる対角行列
$\mathbf{V}_\alpha^T, \mathbf{V}_\beta^T$	$\mathcal{F}_\alpha$ と $\mathcal{F}_\beta$ それぞれに対応する $\mathbf{V}^T$ の部分行列
$\mathbf{N}$	$\mathbf{V}_\alpha^T$ の直交補空間に属するある行列
$\mathbf{R}$	回転行列

N 人のパーティに垂直分散したデータのそれぞれを $\mathbf{X}_1, \ldots, \mathbf{X}_N$ で表すとき， $\mathbf{X}$ はそれらの連結行列 $\mathbf{X} = [\mathbf{X}_1 | \cdots | \mathbf{X}_N]$ で表される．このとき，連合特異値分解は，

$$[\mathbf{X}_1 | \cdots | \mathbf{X}_N] = \mathbf{U}\mathbf{\Sigma}[\mathbf{V}_1^T | \cdots | \mathbf{V}_N^T] \tag{2}$$

で表される．実際のアプリケーションでは，式 (2) の右辺で表される分解結果 $\mathbf{U}, \mathbf{\Sigma}, \mathbf{V}^T$ のうち一部分が，分散データ全体についての知識としてパーティに公開される．具体的に公開される分解結果については，4.1 章のシステムモデルで説明する．

4. システムモデルと敵対者モデル

本章では，本稿では想定するシステムモデルと敵対者モデルの定義を行う．

4.1 システムモデル

N $(N \geq 2)$ 人のパーティから構成される連合特異値分解を考える．既存研究の一部は，第三者機関のサーバを含む中央集権的な連合特異値分解プロトコルを提案しているが，これらの機関は我々が提案する攻撃手法には関連しないため，本稿ではこれらの要素に関する説明は省略する．

i $(1 \leq i \leq N)$ 番目のパーティは部分的なデータ行列 $\mathbf{X}_i$ を所有しているとする．各パーティは自身が所有するデータ行列を連合特異値分解に提供し，最終的にその分解結果のいくらかを分散データ全体についての知識として取得する．ここで，各パーティが取得する分解結果は，連合学習において連合特異値分解が何を目的として使用されている

かに応じて，主に以下の 2 パターンに分けられる．

4.1.1 (ケース 1) $\mathbf{U}, \mathbf{\Sigma}, \mathbf{V}_i^T$ を取得できる場合

このケースでは，i 番目のパーティは，分散データについての知識として $\mathbf{U}, \mathbf{\Sigma}$ と，このパーティが提供したデータ行列 $\mathbf{X}_i$ に対応する部分行列 $\mathbf{V}_i^T$ を取得できる．対応する具体的なアプリケーションとしては，データ前処理としての次元削減 [5], [11] や，データポイズニング攻撃の検出 [13] が挙げられる．

4.1.2 (ケース 2) $\mathbf{U}$ のみを取得できる場合

このケースでは，全てのパーティは，分散データについての知識として $\mathbf{U}$ のみを取得できる．対応する具体的なアプリケーションとしては，データ前処理としての冗長な特徴量の除去 [7], [10], [12] や，オートエンコーダを用いた欠落データの補間 [14], [15] が挙げられる．

4.2 敵対者モデル

一般性を失わずに，N 人のパーティのうち 1 番目のパーティを敵対者とし，残りの $N-1$ 人のパーティを被害者とする．本稿では簡単のために単一の敵対者を想定するが，提案攻撃は複数の敵対者が存在するシナリオでも攻撃手法を変更せずに拡張できる．この場合，データ行列 $\mathbf{X} = [\mathbf{X}_1|\cdots|\mathbf{X}_N]$ は，敵対者が所有する部分的なデータ行列 $\mathbf{X}_\alpha = \mathbf{X}_1$ $(\mathbf{X}_\alpha \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{F}_\alpha|})$ と被害者が所有する残りのデータ行列 $\mathbf{X}_\beta = [\mathbf{X}_2|\cdots|\mathbf{X}_N](\mathbf{X}_\beta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{F}_\beta|})$ に分けられる．ここで，$\mathcal{F}_\alpha$ と $\mathcal{F}_\beta$ のそれぞれは敵対者と被害者がデータを所有する特徴量の集合であり，$\mathcal{F} = \mathcal{F}_\alpha \cup \mathcal{F}_\beta$ である．

敵対者の目標は，連合特異値分解プロトコル内で，被害者が所有するデータ行列 $\mathbf{X}_\beta$ を正確に再構築することである．また，敵対者はセミオネストであると想定する．すなわち，敵対者は連合特異値分解や連合学習で規定されたプロトコルを忠実に遵守しながらも，同時にプライバシー攻撃の実行を試みる．この仮定は提案攻撃の隠密性を強調するものであり，被害者は進行中のプライバシー攻撃を検出することが難しいことを意味する．また，提案攻撃は敵対者が独立して行うものであり，他のパーティとの結託はないものとする．

敵対者が攻撃に利用可能なデータは，敵対者が所有しているデータ行列 $\mathbf{X}_\alpha$ と再構築対象となるデータ行列 $\mathbf{X}_\beta$ についての補助データに加えて，連合特異値分解のプロトコルを通して取得可能な分解結果で構成される (4.1 章)．すなわち，ケース 1 の場合は $\mathbf{U}, \mathbf{\Sigma}, \mathbf{V}_\alpha^T$ $(\mathbf{V}_\alpha^T = \mathbf{V}_1^T)$ を，ケース 2 の場合は $\mathbf{U}$ のみをプロトコルから取得する．

5. 提案攻撃手法

5.1 概要

図 1 で図示されているように，本稿で提案するデータ再構築攻撃は，5.2 章で説明する特異値分解に固有の数学的特性である線形性と正規直交性を悪用することで効率的に攻撃を実現する．具体的には，線形性から，敵対者は被害者のデータ行列 $\mathbf{X}_\beta$ の逆算のために，$\mathbf{X}_\beta$ と比較して小さい行列 $\mathbf{\Sigma}, \mathbf{V}_\beta^T$ だけを推測する．加えて，正規直交性から，敵対者は $\mathbf{V}_\alpha^T$ の直交補空間に含まれる行列だけを $\mathbf{V}_\beta^T$ の候補として探索すればよい．以上から敵対者は，少量かつ探索空間の狭いパラメータについての探索によって $\mathbf{X}_\beta$ 全体を再構築することができる．

以降では，線形性と正規直交性について詳しく説明し，提案するデータ再構築攻撃の具体的な手法について説明する．

5.2 データ再構築攻撃の理論的基礎

5.2.1 線形性

本稿において線形性は，式 (2) で示されるように，データ行列 $\mathbf{X}$ と共有される知識データ $\mathbf{U}$ の間の線形関係を指す．したがって，被害者のデータ行列 $\mathbf{X}_\beta$ も $\mathbf{U}$ との間に線形関係を維持している．

$$\mathbf{X}_\beta = \mathbf{U}\mathbf{\Sigma}\mathbf{V}_\beta^T \tag{3}$$

敵対者はこの線形性に注目し，右辺の $\mathbf{V}_\beta^T$ や $\mathbf{\Sigma}$ を推測して $\mathbf{U}$ に乗算することで，$\mathbf{X}_\beta$ を逆算することができる．$|\mathcal{F}| < |\mathcal{S}|$ の条件下では，$\mathbf{V}_\beta^T \in \mathbb{R}^{|\mathcal{F}|\times|\mathcal{F}_\beta|}$ や $\mathbf{\Sigma} \in \mathbb{R}^{|\mathcal{F}|\times|\mathcal{F}|}$ は再構築を試みるデータ行列 $\mathbf{X}_\beta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{F}_\beta|}$ と比べてサイズの小さい行列であるため，敵対者は線形性を悪用することで再構築のために推測しなければならないパラメータ数を大幅に削減することができる．

5.2.2 正規直交性

本稿において正規直交性は，$\mathbf{V}^T$ が以下の性質を満たすことを指す．

- 各列 (行) ベクトルが他の全ての列 (行) ベクトルと直交する
- 各列 (行) ベクトルのノルムが 1 である

$\mathbf{V}^T$ が正規直交性を満たす場合，その部分行列 $\mathbf{V}_\beta^T$ の各列ベクトルは他の全ての列ベクトルの直交し，かつその列ベクトルはもう一方の部分行列 $\mathbf{V}_\alpha^T$ の全ての列ベクトルと直交する．このような部分行列 $\mathbf{V}_\beta^T$ は，$\mathbf{V}_\alpha^T$ の直交補空間に含まれる．敵対者はこの性質に注目することで，被害者のデータ行列 $\mathbf{X}_\beta$ を逆算するために推測しなければならない $\mathbf{V}_\beta^T$ の探索空間を大幅に狭めることができる．

5.3 データ再構築攻撃

初めに，敵対者が知識データとして $\mathbf{U}, \mathbf{\Sigma}, \mathbf{V}_\beta^T$ を取得できるケース 1 の場合における攻撃を詳細に説明する．次に．ケース 1 における攻撃の拡張として，敵対者が $\mathbf{U}$ だけを取得できるケース 2 の場合における攻撃を説明する．

5.3.1 直交補空間 $\mathbf{N}$ の計算

5.2.1 章で説明した線形性に基づいて，敵対者は $\mathbf{X}_\beta$ の

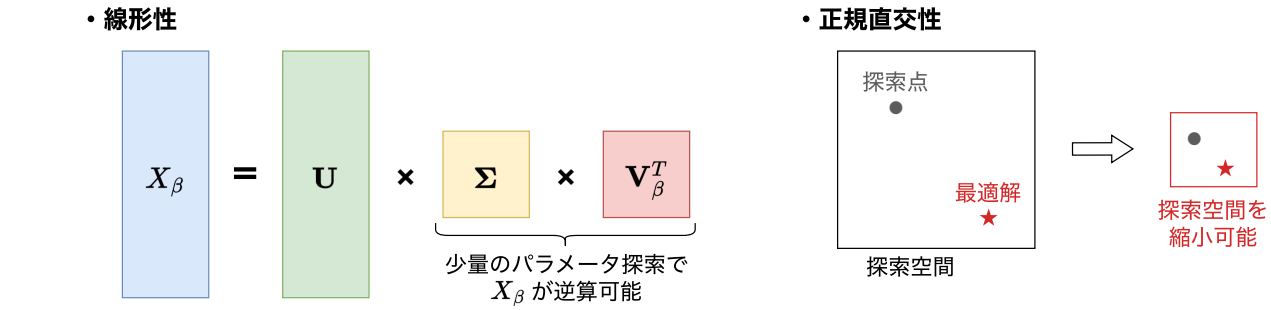

図 1 線形性と正規直交性によるデータ再構築の効率化

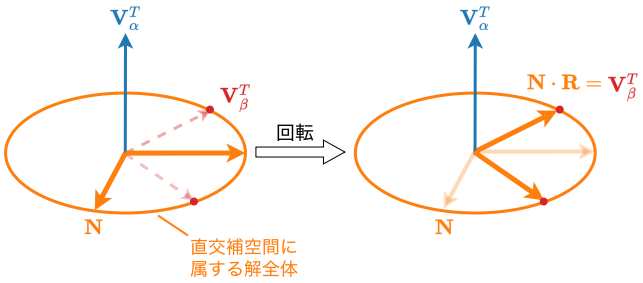

図 2 $|\mathcal{F}|=3$ の場合における $\mathbf{N}$ の自由度

再構築のために $\mathbf{V}_\beta^T$ の推測だけを行う．ここで，5.2.2 章で説明したように，$\mathbf{V}_\beta^T$ は $\mathbf{V}_\alpha^T$ と直交し，かつそれぞれの列ベクトルが互いに直交する直交補空間に属した行列である．したがって，最初に敵対者は $\mathbf{V}_\alpha^T$ の直交補空間に属する一つの行列 $\mathbf{N}$ を計算する．これは，以下の方程式を $\mathbf{N}$ について解くことで容易に得られる．

$$\mathbf{V}_\alpha \cdot \mathbf{N} = O \tag{4}$$

ここで，$O \in \mathbb{R}^{|\mathcal{F}_\alpha|\times|\mathcal{F}_\beta|}$ はその要素が全て 0 の行列 (零行列) である．

式 (4) を解いて得られる $\mathbf{N}$ は，$\mathbf{V}_\beta^T$ と同様に $\mathbf{V}_\alpha^T$ の直交補空間に含まれる行列だが，$\mathbf{N}=\mathbf{V}_\alpha^T$ とは限らない．これは，(4) の解として求まる $\mathbf{N}$ は非一意性を持ち，$\mathbf{V}_\beta^T$ は無数に存在する解のうちの一つに過ぎないためである．したがって敵対者は，$\mathbf{X}_\beta$ を再構築する一意の行列 $\mathbf{V}_\alpha^T$ を探索する必要がある．

5.3.2 回転行列 R の探索

$\mathbf{V}_\alpha^T$ の直交補空間から $\mathbf{X}_\beta$ を再構築するために必要な一意の行列を探索するために，提案攻撃において敵対者は，$\mathbf{N}$ を回転させる $|\mathcal{F}_\beta|$ 次元の回転行列に焦点を当てて探索を行う．これを説明するために，式 (4) における解 $\mathbf{N}$ の非一意性の原因となる自由度について説明する．

単純な例として，$|\mathcal{F}|=3$，$|\mathcal{F}_\alpha|=1$，$|\mathcal{F}_\beta|=2$ の場合を考える．このとき，式 (4) の解の一つである $\mathbf{N}\in\mathbb{R}^{3\times2}$ は，$\mathbf{V}_\alpha^T \in \mathbb{R}^{3\times1}$ に直交する平面を張る，ある 2 つの列ベクトルを持つ．図 2 が示すように，この平面を張る列ベクトルは無数に存在し，これらは $\mathbf{V}_\alpha^T$ を軸とした回転の自由度を持つ．すなわち，この軸で回転させた列ベクトルの全てが，式 (4) の解 $\mathbf{N}$ となり得る．一方，$\mathbf{X}_\beta$ の再構築に必要な $\mathbf{V}_\beta^T \in \mathbb{R}^{3\times2}$ は，その無数の列ベクトルのうちの一意の列ベクトルに対応する．

以上の議論を一般化する．$\mathbf{N} \in \mathbb{R}^{|\mathcal{F}|\times|\mathcal{F}_\beta|}$，$\mathbf{V}_\alpha^T \in \mathbb{R}^{|\mathcal{F}|\times|\mathcal{F}_\alpha|}$ の場合，$\mathbf{N}$ は $\mathbf{V}_\alpha^T$ に垂直な部分空間の基底を列ベクトルとして持つ．$\mathbf{N}$ は $\mathbf{V}_\alpha^T$ を軸とする回転の自由度を持っており，その自由度の数は $\frac{|\mathcal{F}_\beta|\cdot(|\mathcal{F}_\beta|-1)}{2}$ である．この回転の自由度が，式 (4) の解 $\mathbf{N}$ の非一意性をもたらす．

この観察に基づいて，提案攻撃では，$\mathbf{N}$ の持つ回転の自由度だけに着目して適切な行列を探索する．具体的には，$|\mathcal{F}_\beta|$ 次元の回転行列 $\mathbf{R}\in\mathbb{R}^{|\mathcal{F}_\beta|\times|\mathcal{F}_\beta|}$ を探索し，この $\mathbf{R}$ を用いて $\mathbf{N}$ に対して回転操作 $\mathbf{N}\cdot\mathbf{R}$ を適用することで，$\mathbf{N}$ を自由度の軸で回転させる．$\mathbf{N}\cdot\mathbf{R}$ は依然として $\mathbf{V}_\alpha^T$ の直交補空間に属する行列であり，式 (4) の解 $\mathbf{N}$ に対して適切な $\mathbf{R}$ を乗算することで，$\mathbf{N}\cdot\mathbf{R}$ は $\mathbf{X}_\beta$ の再構築に必要な $\mathbf{V}_\beta^T$ と一致する．また，$\mathbf{N}$ についての回転行列 $\mathbf{R}$ のみを探索することで，$\mathbf{V}_\alpha^T$ の直交補空間に属する行列だけを候補として探索することが可能となり，$\mathbf{N}$ 自体を探索する場合と比較して，探索空間を大幅に狭めることが可能となる．

5.3.3 回転行列 R の評価

被害者のデータ行列 $\mathbf{X}_\beta$ を再構築するためには，敵対者は，適切な回転行列 $\mathbf{R}$ を見つけ出し，

$$\mathbf{X}_\beta = \mathbf{U}\mathbf{\Sigma}(\mathbf{N}\cdot\mathbf{R}) \tag{5}$$

から $\mathbf{X}_\beta$ を逆算しなければならない．探索アルゴリズムから導出した $\mathbf{R}$ の評価として，4.2 章で述べたように，敵対者は $\mathbf{X}_\beta$ についての補助データを使用する．本稿では，$\mathbf{X}_\beta$ に含まれる特徴量 $\mathcal{F}_\beta$ のそれぞれについての平均値，分散値，最大値を補助データとする．敵対者はこの補助データと $\mathbf{U}\mathbf{\Sigma}(\mathbf{N}\cdot\mathbf{R})$ から求められる各列の平均値，分散値，最大値を比較することで $\mathbf{R}$ の評価を行う．

これらの補助データはデータ全体の性質を表す基本統計量であり，特定のレコードについてのデータ値と比較して敵対者は容易にアクセスが可能である．データ再構築攻撃のために必要な補助データが基本統計量で十分である理由は次の通りである．式 (5) から，$\mathbf{X}_\beta$ の基本統計量は，

$\mathbf{X}_\beta$ よりも大幅にサイズの小さい行列 $\mathbf{R}$ の値によって大きく変動する．したがって，敵対者が $\mathbf{X}_\beta$ の基本統計量にアクセスできる場合，敵対者はその値を $\mathbf{R}$ の探索に役立てることが可能となる．

5.3.4 探索アルゴリズム

本研究では，回転行列 $\mathbf{R}$ の探索アルゴリズムとして差分進化を採用する．差分進化は古典的なメタヒューリスティックスであり，連続空間における大域最適化のための効果的な進化的計算手法である．差分進化は網羅的な探索アプローチのため，勾配法などと比較して探索には時間がかかる傾向にあるが，本研究では探索時間の短縮よりも高品質な解を得ることを優先するため，この探索アルゴリズムを使用する．注目すべきは，差分進化で探索する $\mathbf{R}$ はサイズが $|\mathcal{F}_\beta| \times |\mathcal{F}_\beta|$ の行列であるが，実際に探索が必要なパラメータ数はその回転行列の自由度の数 $\frac{|\mathcal{F}_\beta|(|\mathcal{F}_\beta|-1)}{2}$ と一致することである．これにより，差分進化で探索するパラメータ数を抑えることができる．

5.3.5 ケース 2 の場合における攻撃

4.1.2 章で述べたように，ケース 2 において敵対者が知識データとして取得できるのは $\mathbf{U}$ のみであり， $\mathbf{\Sigma}$ にアクセスすることはできない．したがって，ケース 2 では回転行列 $\mathbf{R}$ と同時に $\mathbf{\Sigma}$ の探索も行う必要がある．

初めに，$\mathbf{\Sigma}$ の候補 $\hat{\mathbf{\Sigma}}$ と敵対者が保有するデータ行列 $\mathbf{X}_\alpha$ を用いて， $\mathbf{V}_\alpha^T$ を導出する．

$$\mathbf{V}_\alpha^T = (U\Sigma)^{-1}\mathbf{X}_\alpha = \Sigma^{-1}U^T\mathbf{X}_\alpha \tag{6}$$

以降はケース 1 と同様である．

$\mathbf{R}$ に加えて $\mathbf{\Sigma}$ の探索も伴うため，ケース 2 はケース 1 よりも探索するパラメータ数が増加し，攻撃はより複雑になる．しかし， 3 章 で述べたように， $\mathbf{\Sigma}$ は $\mathbf{X}$ についての $|\mathcal{F}|$ 個の特異値を降順に含む対角行列であることから，ケース 2 で増加する探索パラメータ数は $|\mathcal{F}|$ 個に収まる．

6. 評価実験

本章では, 5 章で議論した攻撃手法の性能を評価する．まず，提案攻撃の評価のために採用したデータセットと評価指標について説明した後に，いくつかの条件下で提案攻撃の評価結果を示し，その有効性を詳細に実証する．最後に，データ行列に差分プライバシー戦略が適用されている場合の攻撃性能を示す．なお，本稿で使用したコードは [16] から入手することができる．

6.1 実験構成

6.1.1 データセット

評価には MIMIC-III (Medical Information Mart for Intensive Care) [17] と Bank Marketing [18] を使用する．MIMIC-III は，2001 年から 2012 年までに収集された匿名化された医療データであり，27 種類の特徴量を持った 58,976 レコードで構成されている．実験では，9 種類の連続値特徴量を選択し，無作為抽出した 4,000 レコードからデータ行列 $\mathbf{X} \in \mathbb{R}^{4000\times 9}$ を構成する．Bank Marketing は，銀行の顧客の個人情報から最適なマーケティング戦略を決定するための金融データセットであり，17 種類の特徴量を持った 45,211 レコードで構成されている．実験では，7 種類の連続値特徴量を選択し，無作為抽出した 4,000 レコードからデータ行列 $\mathbf{X} \in \mathbb{R}^{4000\times 7}$ を構成する．

MIMIC-III を使用した実験では，使用する $|\mathcal{F}| = 9$ 種類の特徴量のうち $9 - |\mathcal{F}_\beta|$ 種類の特徴量を敵対者に，残りの $|\mathcal{F}_\beta|$ 種類の特徴量を被害者に割り当て， $|\mathcal{F}_\beta|$ の値を 1 から 8 まで変化させて攻撃性能を評価する．同様に，Bank Markrting を使用した実験では，使用する $|\mathcal{F}| = 7$ 種類の特徴量のうち $7 - |\mathcal{F}_\beta|$ 種類の特徴量を敵対者に，残りの $|\mathcal{F}_\beta|$ 種類の特徴量を被害者に割り当て， $|\mathcal{F}_\beta|$ の値を 1 から 6 まで変化させる．両方のデータセットの各特徴量は，Yao-Jonson 変換によって標準正規分布に近似するようスケーリングする．これはデータ分析や連合学習の場面で一般的に採用されるデータ前処理手法である．

6.1.2 評価指標

提案攻撃の性能を評価するための指標として，本稿では RMSE (Root Mean Squared Error) と Accuracy Metric を使用する．

RMSE は MSE (Mean Square Error) の平方根であり，データ再構築攻撃を評価する最も一般的な指標である．$|\mathcal{S}|$ 個のレコードを含む $|\mathcal{F}_\beta|$ 種類の特徴量について，再構築したデータ $\hat{x}$ とその元データ x との間の RMSE は以下のように計算される．

$$\mathrm{RMSE} = \frac{1}{|\mathcal{F}_\beta|}\sum_{i=1}^{|\mathcal{F}_\beta|}\sqrt{\frac{1}{|\mathcal{S}|}\sum_{j=1}^{|\mathcal{S}|}(x_{ij}-\hat{x}_{ij})^2} \tag{7}$$

RMSE は $\hat{x}$ と x の距離を元データと同じスケールで測定し，スコアが 0 に近いほどデータ再構築の攻撃性能が高いことを示す．

実際の攻撃シナリオでは，元データの値を正確に再構築すること以外にも，元データに近い値が再構築データとして導出されることもデータプライバシーの侵害としてみなされる．これを考慮して，Vero ら [19] が提案した Accuracy Metric も同時に評価指標として使用する．Accuracy Metric (ACC) は次のように測定される．

$$\mathrm{ACC} = \frac{1}{|\mathcal{F}_\beta|}\sum_{i=1}^{|\mathcal{F}_\beta|}\left(\frac{1}{|\mathcal{S}|}\sum_{j=1}^{|\mathcal{S}|}\mathbb{I}\{\hat{x}_{ij}\in[x_{ij}-\epsilon_i, x_{ij}+\epsilon_i]\}\right) \tag{8}$$

ここで， $\mathbb{I}$ は指示関数であり， $\hat{x}_{ij}$ が x_{ij} の許容誤差 ϵ_i 以内にあるかどうかを判定し，条件を満たす場合には 1，それ以外の場合は 0 を返す．本稿では，元データの周囲で 25%の誤差確率範囲に入る限りデータ再構築をプライ

表 2　MIMIC-III において再構築する特徴数 $|\mathcal{F}_\beta|$ を変化させた場合の RMSE と ACC

特徴数	ケース 1		ケース 2		ランダム	
$\|\mathcal{F}_\beta\|$	RMSE	ACC	RMSE	ACC	RMSE	ACC
1	0.000	1.000	0.000	1.000	3.444	0.079
2	0.000	1.000	0.096	0.997	2.624	0.119
3	0.000	1.000	0.119	0.982	2.334	0.128
4	0.000	1.000	0.140	0.978	2.348	0.120
5	0.074	1.000	0.292	0.727	2.441	0.116
6	0.227	0.838	0.785	0.337	2.311	0.121
7	0.363	0.725	0.677	0.391	2.273	0.118
8	0.423	0.651	0.722	0.356	2.186	0.122

表 3　Bank Marketing において再構築する特徴数 $|\mathcal{F}_\beta|$ を変化させた場合の RMSE と ACC

特徴数	ケース 1		ケース 2		ランダム	
$\|\mathcal{F}_\beta\|$	RMSE	ACC	RMSE	ACC	RMSE	ACC
1	0.000	1.000	0.000	1.000	1.628	0.121
2	0.087	0.974	0.135	0.991	1.640	0.115
3	0.100	0.999	0.123	0.992	1.590	0.125
4	0.105	0.995	0.156	0.959	1.630	0.134
5	0.195	0.892	0.311	0.715	1.637	0.126
6	0.237	0.821	0.289	0.734	1.621	0.134

バシー漏洩として取り扱うこととし，$\epsilon = 0.319$ と定める．これは，6.1.1 章で述べたように，各特徴量は標準正規分布に近似されており，標準正規分布に従う確率変数 x は $\Pr[-0.319 < x < 0.319] \approx 0.25$ であるためである [19]．

6.2　評価結果

6.2.1　ケース 1 とケース 2 の比較

敵対者の知識は，連合特異値分解のアプリケーションによって，プロトコルから $\mathbf{U}, \mathbf{\Sigma}, \mathbf{V}_\alpha^T$ を取得するケース 1 と，$\mathbf{U}$ のみを取得するケース 2 に分けられる．表 2 と表 3 はそれぞれ，MIMIC-III と Bank Marketing をデータセットとして使用した場合のケース 1 とケース 2 両方の攻撃性能を示している．この結果からわかるように，ケース 1 とケース 2 の両方において，再構築する特徴量の数 $|\mathcal{F}_\beta|$ が少量である場合，敵対者はその特徴量についてのデータを高精度に再構築することができる．これにより，連合特異値分解のネットワークに参加するパーティ内で被害者の規模が小さい状況 (すなわち，被害者の数が少ないか，あるいは被害者の所有する特徴量の種類が少ない場合) において，提案攻撃は特に脅威であることが示される．

ケース 1 では，$|\mathcal{F}_\beta| = 5$ 以上から攻撃性能は徐々に低下傾向にある．この要因として以下の 2 つが考えられる．

- 探索アルゴリズムで探索するパラメータ数 (5.3.4 章) が増大し，大域最適解の探索が困難になる
- 基本統計量を補助データとするだけでは最適な回転行列 $\mathbf{R}$ を一意に決定できない

しかし，ランダムデータと比較して明らかなように，$|\mathcal{F}_\beta|$ が大きい場合であっても，提案攻撃は被害者のデータ行列 $\mathbf{X}_\beta$ に近いデータを生成する．

ケース 2 では，$|\mathcal{F}_\beta|$ の増大よる攻撃性能の低下がより顕著に確認される．これは，ケース 2 がケース 1 と比較して探索パラメータ数が多く，大域最適解の探索がより難しいためであると考えられる．結論として，提案攻撃はケース 1 とケース 2 の両方で被害者のプライバシーを侵害する可能性があるが，ケース 1 ではそのリスクはより深刻になるといえる．

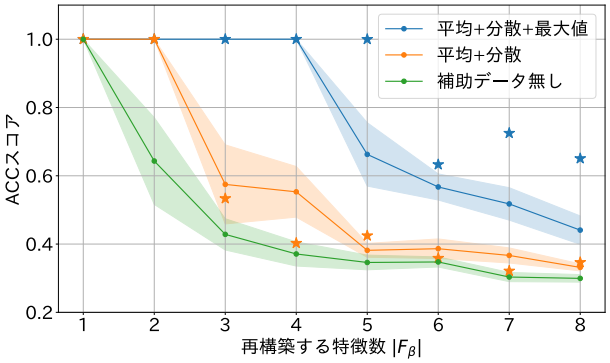

図 3　MIMIC-III において $|\mathcal{F}_\beta|$ と補助データを変化させた時のケース 1 での ACC と 95% 信頼区間 (試行回数: 20 回, ⋆: 損失最小時の ACC)

表 4　MIMIC-III においてガウシアンノイズ付加時の攻撃性能とその際の特異値分解の誤差 (ケース 1，$|\mathcal{F}_\beta| = 4$)

ノイズスケール	攻撃性能 (RMSE)	攻撃性能 (ACC)	特異値分解の誤差
0.0	0.000	1.000	1.03×10^{-15}
10^{-2}	0.048	1.000	7.97×10^{-3}
10^{-1}	0.067	1.000	7.96×10^{-2}

6.2.2　補助データ別の性能

図 6.2.1 は，攻撃に使用する補助データを変えた場合の攻撃性能を，平均の ACC スコアとその 95%信頼区間で示している．補助データを使用しない場合と比較して，基本統計量を補助データとして使用することで提案攻撃の攻撃性能が向上していることがわかる．特に，平均値と分散値に加えて最大値を補助データとして使用する場合において，提案攻撃は $|\mathcal{F}_\beta|$ の増大に対しても攻撃性能を維持することできる．

6.2.3　ノイズ付加時の性能

データに対するノイズ付加は，差分プライバシー戦略として行われる一般的なプライバシー保護アプローチである．以下では，この差分プライバシー戦略が提案攻撃にどのように影響するかを評価する．

特異値分解の誤差は，既存研究 [5] を参考にして計算す

る．具体的には，ノイズのないデータ行列 $\mathbf{X}$ と，ノイズ付加後 $\tilde{\mathbf{X}}$ のデータ行列に対する特異値分解結果から再構成した行列 $\tilde{\mathbf{U}}\tilde{\mathbf{\Sigma}}\tilde{\mathbf{V}}^T$ との間の距離 $\|\mathbf{X} - \tilde{\mathbf{U}}\tilde{\mathbf{\Sigma}}\tilde{\mathbf{V}}^T\|$ として導出する．

表 4 は，提案攻撃がガウシアンノイズの影響をほとんど受けず，高い攻撃性能を維持することを示している．これは，提案攻撃においてガウシアンノイズの影響は限定的だからである．具体的には，再構築データから求めた基本統計量と，補助データとして使用している真値の基本統計量との差としてのみ現れ，攻撃性能に大きく影響しない．その一方で，データ行列へのガウシアンノイズ付加は，特異値分解結果に大きな誤差をもたらす．既存研究 [5] が指摘するように，この誤差はデータの有用性を本質的に損ない，連合特異値分解の応用を妨げる．したがって，差分プライバシー戦略は提案攻撃に対して効果的な対策とはならず，むしろ連合特異値分解アプリケーションの性能を大きく低下させる可能性がある．

7. おわりに

本稿では，異なるパーティに分散したデータを活用するために有効な技術である連合特異値分解にプライバシーの脆弱性があることを示した．提案した攻撃手法は，特異値分解の線形性と正規直交性を悪用し，敵対者が制限された少量のパラメータ探索から被害者の大規模なデータを再構築するものである．現実的なデータセットと評価指標を用いた評価実験で，被害者の所有する特徴量の数 $|\mathcal{F}_\beta|$ が少なく，敵対者が分解結果のうち $\mathbf{U}, \mathbf{\Sigma}, \mathbf{V}_\alpha^T$ にアクセスできるケース 1 の場合に，提案攻撃は特に高い性能を発揮することを示した．また，それ以外の状況でも，提案攻撃は被害者のデータに近いデータを再構築可能であることが明らかとなった．

謝辞　本研究は，JSPS 科研費挑戦的萌芽研究 24K22295 によるものである．

参考文献

[1] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proc. Artif. Intell. Statist. Conf. (AISTATS)*, pp. 1273–1282, April 2017.

[2] Di Chai, Junxue Zhang, Liu Yang, Yilun Jin, Leye Wang, Kai Chen, and Qiang Yang. Efficient decentralized federated singular vector decomposition. In *Proc. USENIX Annu. Tech. Conf.*, pp. 1029–1047, July 2024.

[3] Xiang Li, Shusen Wang, Kun Chen, and Zhihua Zhang. Communication-efficient distributed SVD via local power iterations. In *Proc. Int. Conf. Mach. Learn.*, Vol. 139, pp. 6504–6514, July 2021.

[4] Shuguo Han, Wee Keong Ng, and Philip S Yu. Privacy-preserving singular value decomposition. In *Proc. IEEE Int. Conf. Data Eng.*, pp. 1267–1270, March 2009.

[5] Di Chai, Leye Wang, Junxue Zhang, Liu Yang, Shuowei Cai, Kai Chen, and Qiang Yang. Practical lossless federated singular vector decomposition over Billion-Scale data. In *Proc. ACM SIGKDD Conf. Knowl. Discov. Data Mining*, pp. 46–55, August 2022.

[6] Takumi Suimon, Yuki Koizumi, Junji Takemasa, and Toru Hasegawa. A data reconstruction attack against vertical federated learning based on knowledge transfer. In *Proc. IEEE INFOCOM Workshops*, May 2024.

[7] Andreas Grammenos, Rodrigo Mendoza Smith, Jon Crowcroft, and Cecilia Mascolo. Federated principal component analysis. In *Proc. Adv. Neural Inf. Process. Syst.*, pp. 6453–6464, December 2020.

[8] Mohammad Al-Rubaie, Pei-Yuan Wu, J Morris Chang, and Sun-Yuan Kung. Privacy-preserving PCA on horizontally-partitioned data. In *Proc. IEEE Conf. Dependable Secure Comput.*, pp. 280–287, August 2017.

[9] Bowen Liu and Qiang Tang. Privacy-preserving decentralised singular value decomposition. In *Proc. Inf. Commun. Secur.*, pp. 703–721, December 2019.

[10] David Froelicher, Hyunghoon Cho, Manaswitha Edupalli, Joao Sa Sousa, Jean-Philippe Bossuat, Apostolos Pyrgelis, Juan R Troncoso-Pastoriza, Bonnie Berger, and Jean-Pierre Hubaux. Scalable and Privacy-Preserving federated principal component analysis. In *Proc. IEEE Symp. Secur. Privacy*, pp. 1908–1925, May 2023.

[11] Wenquan Cui, Yue Zhao, Jianjun Xu, and Haoyang Cheng. Federated sufficient dimension reduction through high-dimensional sparse sliced inverse regression. *Commun. Math. Statist.*, May 2023.

[12] William Briguglio, Waleed A Yousef, Issa Traore, and Mohammad Mamun. Federated supervised principal component analysis. *IEEE Trans. Inf. Forensics. Secur.*, Vol. 19, pp. 646–660, October 2024.

[13] Xia Feng, Wenhao Cheng, Chunjie Cao, Liangmin Wang, and Victor S Sheng. DPFLA: Defending private federated learning against poisoning attacks. *IEEE Trans. Serv. Comput.*, Vol. 17, No. 4, pp. 1480–1491, March 2024.

[14] Chung-Ju Huang, Leye Wang, and Xiao Han. Vertical federated knowledge transfer via representation distillation for healthcare collaboration networks. In *Proc. ACM Web Conf.*, pp. 4188–4199, April 2023.

[15] Zilong Yin, Haoyu Wang, Bin Chen, Xin Zhang, Xiaogang Lin, Hangling Sun, Anji Li, and Chenyu Zhou. Federated semi-supervised representation augmentation with cross-institutional knowledge transfer for healthcare collaboration. *Knowledge-Based Sys.*, Vol. 300, pp. 1–21, September 2024. Art. no. 112208.

[16] Data Leakage Attack against SVD. `https://github.com/jargonargon/Data-Leakage-Attack-against-SVD`, 2024. Last accessed: Aug. 22, 2024.

[17] Alistair E W Johnson, Tom J Pollard, Lu Shen, Li-Wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Sci. Data*, Vol. 3, , May 2016. Art. no. 160035.

[18] Sérgio Moro, Paulo Cortez, and Paulo Rita. Bank marketing. UCI Machine Learning Repository, 2012.

[19] Mark Vero, Mislav Balunovic, Dimitar Iliev Dimitrov, and Martin Vechev. TabLeak: Tabular data leakage in federated learning. In *Proc. Int. Conf. Mach. Learn.*, pp. 35051–35083, July 2023.