British Embassy Kuwait

المركز الوطني للأمن السيبراني
National Cyber Security Center

cerc 2024

THE 6TH CYBERSECURITY EDUCATION & RESEARCH CONFERENCE

Privacy-Preserving Machine Learning

Aydin Abadi

Newcastle University

KUWAIT COLLEGE OF SCIENCE & TECHNOLOGY
كلية الكويت للعلوم والتكنولوجيا

جامعة الكويت
KUWAIT UNIVERSITY

UK Science & Innovation Network

125 عاماً من الشراكة الكويتية البريطانية
125 YEARS OF KUWAITI-BRITISH PARTNERSHIP

Aydin Abadi

My Research in Security & Privacy

cerc 2024
THE 6TH CYBERSECURITY EDUCATION & RESEARCH CONFERENCE

Privacy-Enhancing Technologies
including **financial** cryptography

Primitives (theoretical)

- Time-Lock Encryption — FC' 21
- Threshold Cryptography — AsiaCrypt'24
- Timed Zero-knowledge Proofs — ACNS'20
- Private Set Intersection (PSI) — FC' 22, TDSC'19, FC'16

Systems (applied)

- COVID-19 Contact Tracing — SGX + Functional Encryption
- Privacy-preserving Private Banking — OT + Secret Sharing
- Two-Factor Authentication — To securely accessing online banking account
- Economics of Private Databases — Under review by AAAI'25
- Security of Calendar — CCS'24—poster
- Proofs of Data Retrievability

Cryptanalysis

- Attacks on PSIs — ESORICS'21
- Attacks on Payment System
- Attacks on "Fair Exchange" Systems — EURO S&P'23

Blockchain & Cryptocurrency

- Transparent Online Banking — ASIA CCS'23
- Decentralised Fair Payment Platforms — EURO S&P'23
- Insurance for Cryptocurrency Users — Work in progress
- Fairly Rewarding Parties in Secure Multi-party Computation
- Decentralised Marketplace — ICSA-C'23

Federated Learning (FL)

- Privacy preserving FL to Deal with Financial Fraud
- Enhancing FL through Secure Deduplication — Under review by NDSS'25
- Delegated FL — Work in progress

Post-Quantum Security

- Post-quantum secure Oblivious Transfer (OT) Protocols

UK Science & Innovation Network

125 عاما من الشراكة الكويتية البريطانية
125 YEARS OF KUWAITI-BRITISH PARTNERSHIP

Machine Learning and Privacy Concerns

- Machine Learning (ML):
    - ML is a subset of artificial intelligence (AI) that enables systems to learn and make decisions from data without being explicitly programmed.
        - There are various types of ML, including:
            - Supervised ML: Learning from labeled data
            - Unsupervised ML: Finding patterns in unlabeled data
- Challenges with data **privacy** in ML:
    - As ML relies heavily on data, there is an inherent risk of privacy breaches associated with this process
    - ML requires accessing to raw data
    - If there are many sources of data, then different sources must share their data (often in plaintext), which can violate their privacy.

(Privacy-Preserving) Federated Learning

- Federated Learning (FL):
  - FL enables training across decentralized devices without sharing raw data
  - FL aims to preserve data privacy of different data providers while enabling machine learning

Source: DALL.E

(Privacy-Preserving) Federated Learning

- Federated Learning (FL):
  - In FL, devices (or any data contributors) compute local models based on their data and then share the local model updates with a central server
    - This server aggregates the updates to derive a global model that encapsulates the features of all the local data held by the individual devices

Federated Learning's General Procedure

1: **Server:**
2: Initialize global model θ
3: **for** each round $k = 1, 2, 3, \ldots, K$ **do**
4: Broadcast θ to all participating devices
5: **Clients:**
6: **for** each client i (where $1 \leq i \leq n$) in parallel **do**
7: Receive global model θ
8: Compute local update g_i using local data
9: Send g_i to the server
10: **Server:**
11: Aggregate local updates: $G_k = \sum_{i=1}^{n} g_i$
12: Update global model: $\theta_{k+1} = \text{UpdateModel}(\theta_k, G_k)$

Source: https://arxiv.org/pdf/2401.10765

Source: DALL.E

Problem of Data Duplication in Federated Learning

- In general, the quality of the training data significantly influences the accuracy of an ML model
- To ensure meaningful learning, the collected data must undergo a thorough data cleaning process
- Duplicated sequences are prevalent in text datasets
- Duplicated sequences can adversely affect the training process of Language Models [1]

[1] K. Lee, D. Ippolito, A. Nystrom, C. Zhang, D. Eck, C. Callison-Burch, and N. Carlini, “Deduplicating training data makes language models better,” 2022.

Affects of data duplicates on machine learning

- Language models need clean and duplicate-free data for training
- Duplicate data can reduce model accuracy and cause issues such as:
  - Negative impact on model accuracy
  - Increased learning costs

Large language models memorize duplicate data, which reduces the quality and accuracy of learning

More processing time is required compared to situations without duplicate data

Challenges of removing data duplicates in federated learning

- Removing duplicate data in federated learning is **complicated** because direct data sharing is not possible
- Removing duplicate data is **easy** when data privacy is not a concern:
    - all devices send their data to the server, and the server finds and removes duplicate data
- In federated learning, it is not expected that every device will send its data to the server or another device without protection

Our Solution

Efficient Privacy-Preserving Multi-Party Deduplication (EP-MPD)

- We have developed a new protocol (called EP-MPD) that removes duplicate data in federated learning environments without revealing sensitive information
- Our proposed solution (EP-MPD):
    - improves learning accuracy by up to 19%
    - reduces learning time by up to 27%

Our Solution

G-PSI a building block of EP-MPD

- Our proposed solution EP-MPD is based on “Private Set Intersection” (PSI) protocols
- We introduced the new concept of **Group** Private Set Intersection (G-PSI)
- PSI is a cryptographic protocol that allows two or more users to privately share their data sets without revealing anything about the elements of the sets beyond the result

Background on PSI

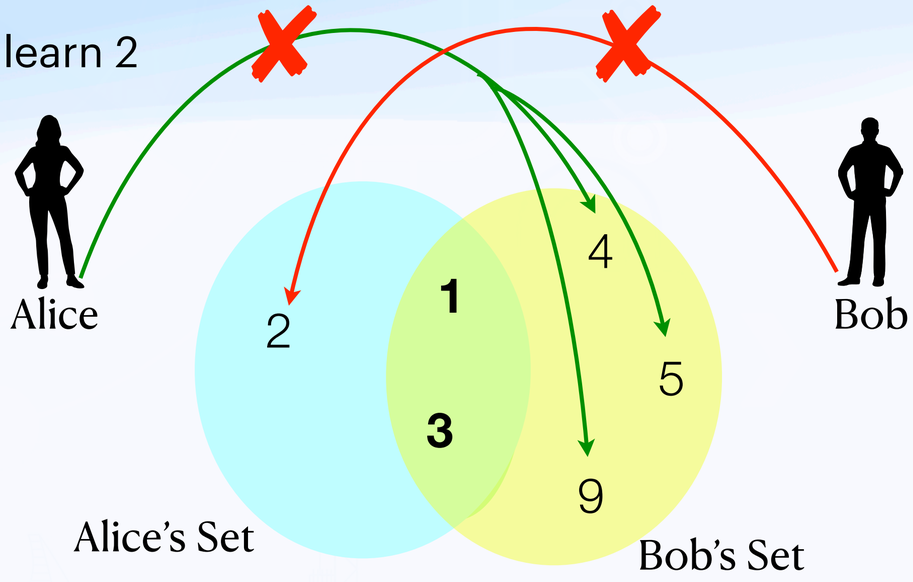
- PSI is a cryptographic protocol that allows two or more users to privately share their data sets without revealing anything about the elements of the sets beyond the result

- A PSI lets mutually distrustful parties compute the intersection of their private sets such that nothing about the sets' elements, beyond the result, is revealed

- According to the PSI's definition, in this example:
  - during and after the computation, Bob must not learn 2 and Alice must not learn 4, 5, and 9.

Intersection of sets A and B: 1, 3

$A \cap B$



Our Solution

G-PSI a building block of EP-MPD

- In general, G-PSI allows each user (client) in a group to efficiently find the intersection of their set with the set of every user from another group, without learning anything beyond that

A set belonging to a client in $\mathcal{G}_0$

A set belonging to a client in $\mathcal{G}_1$

$$\vec{v}_{j,i} = \Big[[S_{j,i} \cap S_{1-j,1}], \ldots, [S_{j,i} \cap S_{1-j,m}]\Big]$$

$$f_{\text{G-PSI}} = \Big((S_{0,1}, \ldots, S_{0,m}, \emptyset), (S_{1,1}, \ldots, S_{1,m}, \emptyset)\Big) \rightarrow \Big((\vec{v}_{0,1}, \ldots, \vec{v}_{0,m}), (\vec{v}_{1,1}, \ldots, \vec{v}_{1,m}), \emptyset\Big)$$

The main functionality

Our Solution

G-PSI a building block of EP-MPD

- We presented two separate protocols that securely meet the requirements of G-PSI:
  - EG-PSI $^{(I)}$: only uses private key encryption and requires a Trusted Execution Environment (TEE) to find shared encrypted data, thus requiring very little processing time
  - EG-PSI $^{(II)}$: uses public key encryption (Oblivious Pseudorandom Function) and requires a TEE to encrypt data. Although it requires more processing time, the TEE plays a smaller role during the protocol

Our Solution

G-PSI a building block of EP-MPD

- *Parties*. Trusted execution environment $\mathcal{TEE}$, clients in group $\mathcal{G}_0$: $\{\mathcal{C}_{0,1}, \ldots, \mathcal{C}_{0,m}\}$, and clients in group $\mathcal{G}_1$: $\{\mathcal{C}_{1,1}, \ldots, \mathcal{C}_{1,m}\}$.
- *Inputs*. Sets $S_{0,1}, \ldots, S_{0,m}, S_{1,1}, \ldots, S_{1,m}$, where each $S_{i,j}$ belongs to client $\mathcal{C}_{i,j}$, $0 \leq j \leq 1$ and $1 \leq i \leq m$.
- *Outputs*. $\vec{v}_{j,i}$ to $\mathcal{C}_{j,i}$, where $\vec{v}_{j,i} = \Big[[S_{j,i} \cap S_{1-j,1}], \ldots, [S_{j,i} \cap S_{1-j,m}]\Big]$.

---

1) *Setup*.
   a) each client $\mathcal{C}_{0,i}$ in $\mathcal{G}_0$ agrees with every client $\mathcal{C}_{1,l}$ in $\mathcal{G}_1$ on a secret key $k_{i,l}$, by picking a random key $k_{i,l}$ and sending it to $\mathcal{C}_{1,l}$. Client $\mathcal{C}_{0,i}$ stores this key as $k_{i,l}$ while $\mathcal{C}_{1,l}$ stores this key as $k_{l,i}$.
   b) each $\mathcal{C}_{j,i}$ takes the following steps:
      i) encrypts its set elements under keys $k_{i,l}$ $(\forall l, 1 \leq l \leq m)$ as follows, $\forall e \in S_{j,i}$: $\mathrm{PRP}(k_{i,l}, e) \rightarrow e'_{i,l}$. Let set $S'_{j,i}$ contain the encrypted set elements of $\mathcal{C}_{j,i}$ and let set $T_{j,i}$ contains all triples of the form $(e'_{i,l}, k_{i,l}, l)$.
      ii) sends $S'_{j,i}$ to $\mathcal{TEE}$ and locally keeps $T_{j,i}$.

In Phase 1:

a) Each user agrees on a private key with each other user in the other group

b) Each user encrypts their data (set elements) using the private keys agreed upon with other users.

Each user sends all their encrypted data to the TEE

Our Solution

G-PSI a building block of EP-MPD

2) *Finding Encrypted Intersection.* $\mathcal{TEE}$ takes the following steps for each $\mathcal{C}_{j,i}$.

a) appends to an empty set, $R_{j,i}$, every ciphertext that satisfy the following conditions hold:

- it appears more than once in the set $S = \sum_{j=0}^{1} \sum_{i=1}^{m} S'_{j,i}$.
- it appears in set $S'_{j,i}$.

b) sends $R_{j,i}$ to $\mathcal{C}_{j,i}$.

In Phase 2:

a) The TEE finds duplicate encrypted data

b) The TEE sends the found duplicate data to the respective users

Our Solution

G-PSI a building block of EP-MPD

3) *Extracting Plaintext Intersection.* Each $\mathcal{C}_{j,i}$ takes the following steps.

a) constructs a vector $\vec{v}_{j,i} = [\vec{v}_{j,i,1}, \ldots, \vec{v}_{j,i,m}]$, where each vector in $\vec{v}_{j,i}$ is initially empty.

b) decrypts each element of $R_{j,i}$ as follows. $\forall e' \in R_{j,i}$:

i) retrieves decryption key $k_{i,l}$ and index l from $T_{j,i}$ using e'.

ii) calls $\mathtt{PRP}^{-1}(k_{i,l}, e') \rightarrow e$ and appends e to l-th vector in $\vec{v}_{j,i}$.

c) considers $\vec{v}_{j,i}$ as the result.

In Phase 3:

a) Each user decrypts their duplicate data

b) All share data (the intersection) is considered as duplicated data

Our Solution

Efficient Privacy-Preserving Multi-Party Deduplication (EP-MPD)

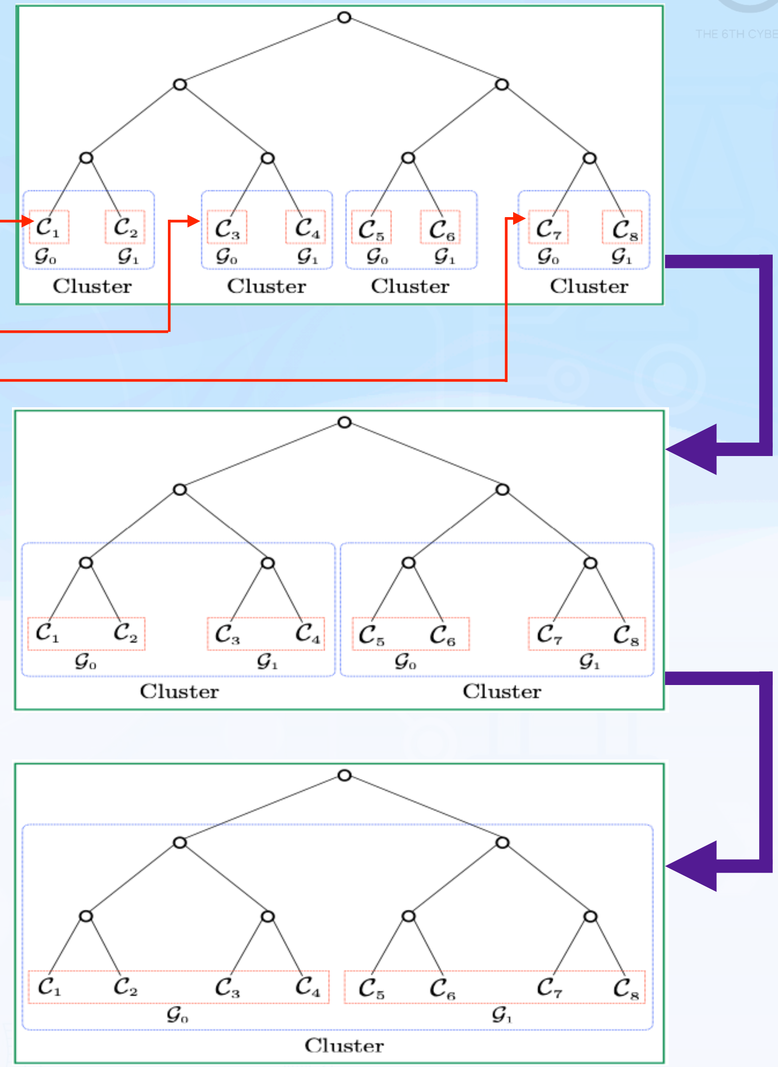
Our Solution

EP-MPD

The idea behind EP-MPD involves constructing a binary tree where the leaf nodes contain user identifiers

At each level, clusters are formed with two different groups of users, named G0 and G1

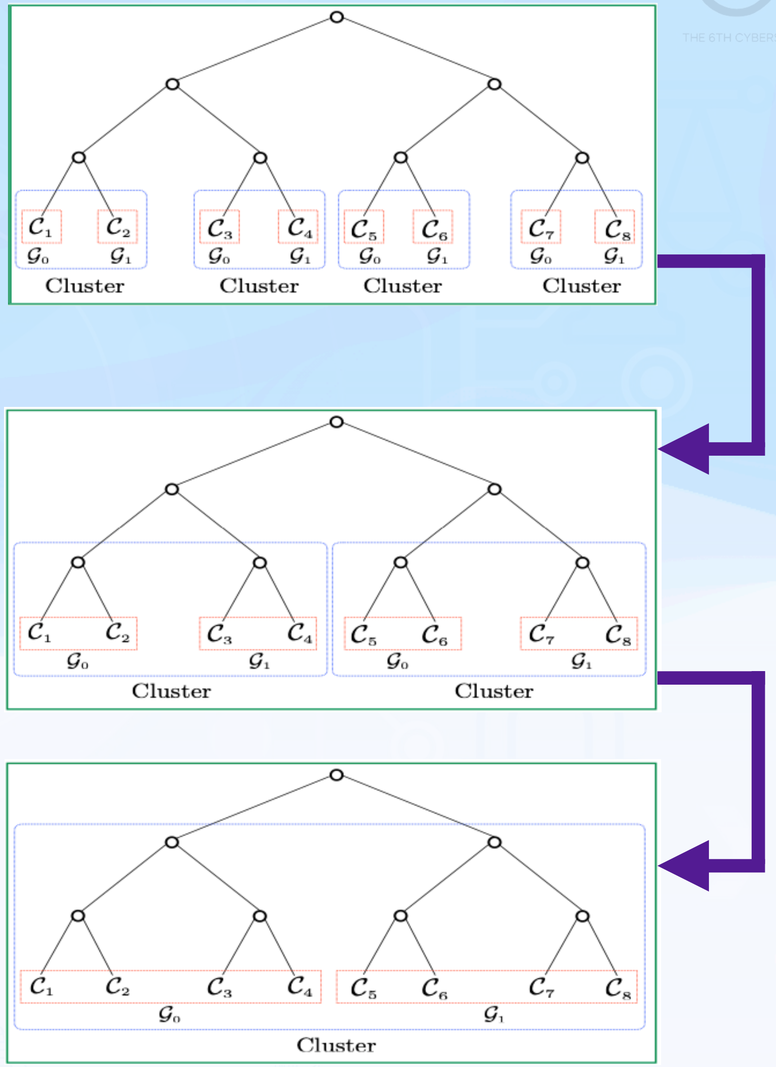
EG-PSI is applied to the sets of users sharing a cluster until we reach the root of the tree



Our Solution
EP-MPD

After each call to EG-PSI, users in group G0 update their sets by removing intersections returned by EG-PSI

These updated sets are then used as input for the next call to EG-PSI



Our Solution

Federated learning equipped with deduplication

1. Each user locally removes duplicate data

2. All users collaborate to remove duplicate data

3. All users join the federated learning protocol

- *Parties*. A set of clients $\{\mathcal{C}_1, \ldots, \mathcal{C}_m\}$.
- *Server*. Holds the initial model $\boldsymbol{\theta}$.
- *Inputs*. Sets $S_1, \ldots, S_m$, where each S_i belongs to client $\mathcal{C}_i$, $1 \leq i \leq m$, and m is a power of two.
- *Outputs*. A global updated model Θ.

---

1) *Local Deduplication.* Each client runs a deduplication algorithm on their local dataset. At the end, client $\mathcal{C}_i$ receives an updated dataset S'_i.
2) *Global Deduplication.*
   a) All the clients participate in the EP-MPD as described in Figure 4.
   b) Each client $\mathcal{C}_i$ gets updated set S''_i, such that

$$\sum_{i=1}^{m} S''_i = \bigcup_{i=1}^{m} S'_i.$$

3) *Federated learning.*
   a) The server and clients agree upon an FL protocol for training.
   b) The server initiates the learning by sharing the initial model $\boldsymbol{\theta}$ with each client.
   c) Each client $\mathcal{C}_i$ trains on their local dataset S''_i and updates $\boldsymbol{\theta}$ to $\boldsymbol{\theta}_i$.
   d) The clients and server aggregate the local models $\boldsymbol{\theta}_i$ trained by the clients.
   e) The server outputs the global updated model Θ for the next training round.

Our Solution
EP-MPD

We implemented the EP-MPD protocol in Python

Our experiments showed that the maximum improvement in learning quality from deduplication is about 19%

Test set perplexity (PP) and improvement rate (IR) of perplexity after deduplication.

Model	Dataset	Duplication Percentage						Deduplicated
		30%		20%		10%		
		PP	IR (%)	PP	IR (%)	PP	IR (%)	PP
GPT-2 Medium	Haiku	3.73	5.36	3.69	4.3	3.6	1.94	3.53
	Rotten Tomatoes	2.4	3.75	2.36	2.18	2.35	1.7	2.31
	Short Jokes	3.95	5.31	3.89	3.85	3.83	2.34	3.74
	Poetry	5.46	8.42	5.59	10.55	5.33	6.19	5.00
	IMDB	12.81	4.57	12.71	3.75	12.5	2.0	12.25
	Sonnets	15.83	13.64	15.63	12.5	14.22	4.02	13.67
	Plays	34.31	18.27	34.88	19.61	28.12	–	28.04
GPT-2 Large	Haiku	3.27	8.86	3.26	8.58	3.23	7.73	2.98
	Rotten Tomatoes	2.65	16.98	2.6	15.38	2.53	13.00	2.2
	Short Jokes	4.11	7.78	4.02	5.72	3.95	4.05	3.79
	Sonnets	8.51	5.53	8.4	4.28	8.02	–	8.04

Our Solution
EP-MPD

Our experiments also showed that deduplication reduces learning time by up to 27%

Total GPU training time (minutes) of all clients and improvement rate (IR) of time after deduplication.

Model	Dataset	Duplication Percentage						Deduplicated
		30%		20%		10%		
		Time	IR (%)	Time	IR (%)	Time	IR (%)	Time
GPT-2 Medium	Haiku	111.92	22.96	105.03	17.91	95.62	9.83	86.22
	Rotten Tomatoes	162.79	21.7	151.54	15.89	138.76	8.14	127.46
	Short Jokes	396.62	27.85	338.69	15.51	313.35	8.68	286.15
	Poetry	114.28	22.65	105.48	16.2	96.85	8.74	88.39
	IMDB	2133.36	22.56	2006.94	17.68	1788.11	7.61	1652.04
	Sonnets	33.13	27.95	28.53	16.33	26.14	8.68	23.87
	Plays	31.48	22.9	29.38	17.39	26.95	9.94	24.27
GPT-2 Large	Haiku	20.89	22.98	19.28	16.55	17.7	9.1	16.09
	Rotten Tomatoes	70.74	23.08	65.26	16.63	59.91	9.18	54.41
	Short Jokes	340.75	22.93	313.65	16.27	288.86	9.08	262.63
	Sonnets	13.91	20.92	12.89	14.66	11.87	7.33	11.0

cerc 2024
THE 6TH CYBERSECURITY EDUCATION & RESEARCH CONFERENCE

Privacy-Preserving Data Deduplication for Enhancing Federated Learning of Language Models

Aydin Abadi

Newcastle University

المركز الوطني للأمن السيبراني
National Cyber Security Center
British Embassy Kuwait
KUWAIT COLLEGE OF SCIENCE & TECHNOLOGY
كلية الكويت للعلوم والتكنولوجيا
جامعة الكويت
KUWAIT UNIVERSITY
UK Science & Innovation Network
125 عاماً من الشراكة الكويتية البريطانية
125 YEARS OF KUWAITI-BRITISH PARTNERSHIP