Versión final de los componentes básicos de MLEDGE

MLEDGE - Aprendizaje automático en la nube y en el borde

(Cloud and Edge Machine Learning)

Diciembre de 2024

Información sobre el entregable

Nombre del documento:

Versión final de los componentes básicos de MLEDGE

Versión actual: 1.0

Proyecto: MLEDGE - Aprendizaje automático en la nube y en el borde (Cloud and Edge Machine Learning)

Paquete de trabajo: P2 - Implementación de componentes básicos de MLEDGE

Tareas: El entregable es resultado del trabajo en los diversos componentes técnicos:

- A2.1: Uso eficiente de Federated Learning (FL) en nubes híbridas y protección contra ataques, incluidos ataques de envenenamiento e inferencia (FedSecure)
- A2.2: Protección de datos sensibles o confidenciales que sean intercambiados entre diferentes dominios administrativos en la nube, incluido su borde que es potencialmente menos seguro (FedWM)
- A2.3: Equidad en términos de distribución de costos y ganancias cuando la computación en el borde se usa para entrenar de forma colaborativa modelos de aprendizaje automático (FLaaS Manager)
- A2.4: Gestión de los desafíos de portabilidad de datos en el borde de la nube (DataEdge)
- A2.5: DevOps y desarrollo continuo para servicios de aprendizaje automático que se ejecutan en borde de la nube (FLaaS)

Entregable: E2.1 - M12 - Requisitos y diseño de la arquitectura y casos de uso (preliminar)

Autores: Javad Dogani (IMDEA), Devriş Işler (IMDEA), Tianyue Chu (IMDEA), Alexander Goultiaev (IMDEA), Naicheng Li (IMDEA)

Revisores: Nikolaos Laoutaris (IMDEA)

Historial de Versiones

Versión	Fecha	Resumen de modificaciones
Version 1.0	31-12-2024	Versión inicial del documento

Índice

Información sobre el entregable 3
Historial de Versiones 3
Índice 4
1. Introducción 6
2. Arquitectura de MLEDGE 8
3. FLaaS 10
4. FedSecure 11
4.1. Protección de la Clasificación de Temas Sensibles en Aprendizaje Federado contra Ataques de Envenenamiento 11
4.2. Cuantificación y Preservación de la Privacidad en Aprendizaje Federado con Poda 12
4.3. FedQV: Aprovechando la Votación Cuadrática en Aprendizaje Federado 13
5. FedWM 14
6. FLaaS Manager 15
6.1. Entendiendo el Precio de los Datos en Mercados Comerciales de Datos 15
6.2. Pruébalo Antes de Comprarlo 15
6.3. De RAGs a Riquezas: Recuperación Descentralizada y Medición de la Influencia de Documentos en Sistemas de Generación con Recuperación Mejorada (RAG) 16
7. Conclusión 17
Anexo 1: FLTorrent progress report 19
Summary 19
Objectives 19
Methodologies 19
Decentralized Federated Learning Model 19
Privacy-Preserving Techniques 19
Bandwidth-Constrained Performance Optimization Heuristics 20
Evaluation 20
Conclusion 20
Anexo 2: The report on optimizing and real-world implementation of FLTorrent 21
Optimized Chunk Transmission 21
Future Directions 22
Anexo 3: Securing Federated Sensitive Topic Classification against Poisoning Attacks 24
Anexo 4: PriPrune: Quantifying and Preserving Privacy in Pruned Federated Learning 43
Anexo 5: FedQV: Leveraging Quadratic Voting in Federated Learning 74

Anexo 6: FreqyWM - Frequency Watermarking for the New Data Economy 105

Anexo 7: Understanding the price of data in commercial Data Marketplaces 116

Anexo 8: Try Before You Buy 128

Anexo 9: RAGs to Riches: Decentralized Retrieval and Measuring Document Influence in Retrieval-Augmented Generation (RAG) Systems 135

1. Introducción

En diciembre de 2022 fue adjudicado a IMDEA Networks el proyecto “MLEDGE - Aprendizaje automático en la nube y en el borde (Cloud and Edge Machine Learning)” (REGAGE22e00052829516, en adelante el ‘Proyecto’ o MLEDGE) por parte del Ministerio de Asuntos Económicos y Transformación Digital del Gobierno de España, con fondos de la Unión Europea dentro del Plan de Recuperación, Transformación y Resiliencia (European Union - NextGenerationEU/PRTR). El proyecto tiene como objetivo habilitar un ecosistema próspero de servicios FL en el borde seguros y eficientes capaces de facilitar el uso de datos personales y B2B confidenciales para entrenar modelos de ML para consumidores mientras se protege la privacidad de los datos y de sus propietarios.

Los **objetivos generales del proyecto** se pueden resumir en los siguientes:

1. Hacer del aprendizaje federado una funcionalidad accesible y de fácil uso en el borde mediante el desarrollo de una capa de software intermedio y componentes que escondan la complejidad del procesamiento y el intercambio de datos que supone.
2. Resolver problemas técnicos asociados al aprendizaje federado en el borde de la nube.
3. Demostrar esta funcionalidad en casos de uso que reflejen problemas reales de la industria que pueden ser resueltos con estas tecnologías.
4. Explotar los resultados del proyecto involucrando a agentes externos y comunicar los hallazgos al público potencial en general.

Para alcanzar estos objetivos, se han catalogado una serie de **objetivos técnicos específicos** del proyecto que se resumen en los siguientes:

1. Diseñar un marco de desarrollo de servicios de aprendizaje federado (FLaaS) en el borde de la nube y componentes que ayuden a popularizar este tipo de servicios
2. Diseñar y desarrollar soluciones de seguridad (FedSecure) contra ataques de envenenamiento o inferencia lanzados desde servidores de borde rebeldes y/o nodos de agregación “honestos pero curiosos”.
3. Gestionar los desafíos de la portabilidad de datos en el borde de la red (DataEdge)
4. Crear un esquema de marca de agua (FedWM) para proteger contra la redistribución de los datos o metadatos que se intercambien entre servidores en el borde en el marco del FLaaS.
5. Crear una capa de lógica económica y de negocio (FLaaS Manager) que implemente una distribución justa de costes e ingresos entre las partes cuando colaboren en el entrenamiento de modelos de ML.
6. Soporte a DevOps y al desarrollo continuo de servicios de aprendizaje automático en la nube, optimizando los costes mediante su monitoreo, predicción y asignación inteligente y energéticamente eficiente de los trabajos de computación.

Finalmente, uno de los propósitos fundamentales de este proyecto es diseñar, desarrollar y divulgar demostradores públicos que manejen datos personales sensibles y que nutran modelos de aprendizaje automático aplicables a distintos sectores industriales. Para ello, en la primera fase del proyecto, se seleccionaron diversas empresas para colaborar en el desarrollo de la plataforma FLaaS (Federated Learning as a Service) y en la supervisión de los costes computacionales, así como para el diseño e implementación de casos prácticos de negocios que se beneficien del aprendizaje distribuido en la periferia de la nube.

Este documento corresponde al entregable E2.1, titulado “Versión final de los componentes básicos de MLEDGE”, y tiene como fin presentar los progresos alcanzados en los componentes fundamentales de MLEDGE durante el primer año del proyecto. La estructura del documento es la siguiente: La sección 2 resume la arquitectura propuesta para el proyecto y cómo esta contribuye al logro de los objetivos planteados. De la sección 3 a la 7, se detallan los avances logrados en los diferentes componentes. Dado que el equipo trabaja predominantemente en inglés, estas secciones incluyen material en dicho idioma, como publicaciones y documentos técnicos. Finalmente, la sección 8 ofrece las conclusiones y los próximos pasos a seguir en el proyecto.

2. Arquitectura de MLEDGE

En esta sección, se presenta la arquitectura del proyecto MLEDGE junto con una descripción de los diferentes componentes de la misma. El proyecto busca:

i) Avanzar en el estado del arte de los componentes de acuerdo con los objetivos científicos

ii) Demostrar estos avances sobre plataformas y casos de uso comerciales de forma que se facilite la explotación posterior de los resultados del proyecto en la economía real.

El diagrama mostrado en la figura siguiente resume la arquitectura de MLEDGE y los bloques del proyecto que se presentó como propuesta del proyecto.

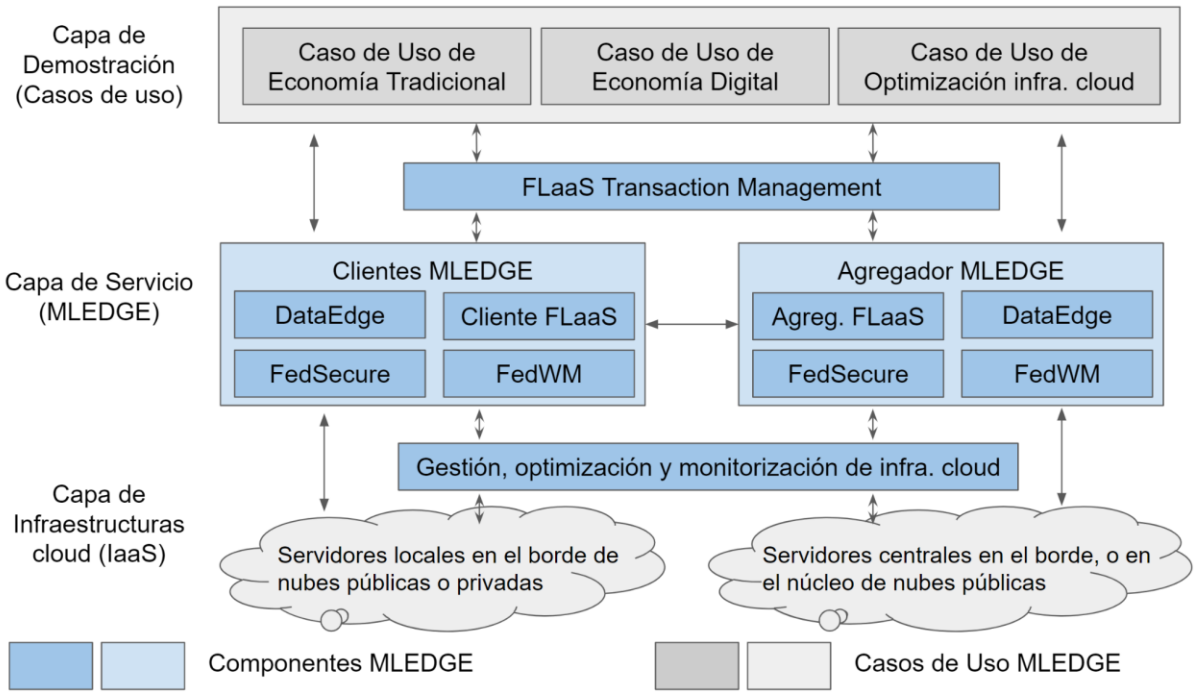


Figura 1. Diagrama de bloques de MLEDGE

La arquitectura de MLEDGE se articula en torno a tres capas:

- La **capa de infraestructuras** dispone los recursos de computación y comunicación necesarios para la ejecución del proyecto. Dicha capa de infraestructuras puede incluir recursos a diferentes niveles de la red, incluyendo clouds públicas o privadas "centralizadas" en el núcleo de la nube, nodos de computación en el borde de la nube, e incluso infraestructuras en casa o terminales de los usuarios.
- La **capa de servicios MLEDGE** busca integrar una serie de componentes que habiliten servicios FLaaS en el borde de la nube, y que puedan integrar componentes innovadores que provengan del desarrollo de los objetivos científicos del proyecto.
- La **capa de demostración** incluye casos de uso reales de empresas de la economía tradicional y digital, así como un caso de uso específico de optimización de infraestructuras cloud. Los casos de uso buscarán demostrar el uso del aprendizaje federado en el borde de la nube y de los componentes de MLEDGE en escenarios

reales y con datos reales. Para ello, durante los primeros seis meses de proyecto se ha realizado un screening de empresa y se han elaborado pliegos para realizar licitaciones de estos tres casos de uso a empresas españolas externas a IMDEA.

A continuación, se describen los componentes de la capa de servicios en los que se está trabajando durante el proyecto, su relación con los objetivos científicos, y los casos de uso que se buscará incorporar de la industria mediante estas licitaciones.

La capa de servicios de MLEDGE trabajará sobre la base de alguna de las soluciones comerciales de FLaaS que se describieron en la sección 4.2. Adicionalmente, se incluye una serie de componentes perfectamente alineados con los objetivos científicos del proyecto que buscan extender el estado del arte de algunos componentes del aprendizaje federado en el borde de la nube. Además, su integración con casos de uso específicos de la industria permitirá probar estos componentes directamente sobre casos reales y acelerar el tiempo hasta la explotación de los mismos por parte de la industria.

La siguiente tabla relaciona los componentes de la capa de servicios de MLEDGE con los objetivos científicos del proyecto:

Tabla 1. Relación entre objetivos científicos y componentes de la capa de servicios de MLEDGE

Objetivo científico	**Componente MLEDGE**
1. DevOps y desarrollo continuo para servicios de aprendizaje automático (FLaaS) que se ejecutan en borde de la nube	FLaaS
2. Uso eficiente de FL en nubes híbridas y protección contra ataques	FedSecure
3. Protección de datos sensibles o confidenciales que sean intercambiados entre dominios administrativos en la nube y el borde de la nube	FedWM
4. Equidad en términos de distribución de costos y ganancias cuando la computación en el borde se usa para entrenar de forma colaborativa modelos de ML	FLaaS Manager
5. Gestión de los desafíos de portabilidad de datos en el borde	DataEdge

En las siguientes secciones se presentan los avances de cada uno de los componentes MLEDGE, except DataEdge. Para este componente, tras el estudio del estado del arte, se decidió que se usarían formatos estándar de representación de datos e interfaces como XML, RDF, REST API, adaptados a las recomendaciones de DCAT, IDSA y GAIA-X para asegurar mecanismos estándar y controlados de intercambio de datos entre las partes que intervienen en el aprendizaje federado.

3. FLaaS

La investigación del componente de innovación de FLaaS ha abordado la creación de un sistema avanzado de aprendizaje federado distribuido, utilizando como base el protocolo BitTorrent. Este enfoque busca superar los desafíos asociados con los sistemas centralizados tradicionales, evitando que una única entidad controle o acceda a la información completa de todos los modelos entrenados por los diferentes clientes. Este sistema permite también el entrenamiento simultáneo de múltiples modelos que son propuestos y personalizados según las necesidades específicas de cada cliente. Hemos denominado a esta iniciativa como FLTorrent, un nombre que refleja su inspiración en la eficiencia y descentralización característica del protocolo BitTorrent.

Hasta la fecha de diciembre de 2024, hemos logrado progresos significativos en este componente innovador, los cuales están detalladamente descritos en el Anexo 1, redactado en inglés. Además, los esfuerzos futuros para FLTorrent incluyen la optimización de la transmisión de datos dentro de la aplicación FLTorrent, asegurando una implementación efectiva en situaciones prácticas y en escenarios del mundo real.

El siguiente paso se centra en perfeccionar y expandir las capacidades de FLTorrent, como se describe en el Anexo 2. Este componente crucial será sometido a pruebas rigurosas en la plataforma FLaaS, que ha sido integrada al proyecto gracias a la colaboración de una de las empresas que ganó los lotes de trabajo de MLEDGE. Este componente no solo se empleará en los estudios de caso diseñados para demostrar la viabilidad y eficacia del proyecto, sino que también se presentará al final del proyecto como una de las principales innovaciones desarrolladas, mostrando su aplicabilidad y los beneficios en entornos empresariales y operativos reales.

4. FedSecure

La integración del aprendizaje federado (FL) con entornos de nube híbrida ofrece beneficios significativos en términos de privacidad de datos y entrenamiento descentralizado de modelos. Sin embargo, esta configuración es inherentemente vulnerable a una variedad de amenazas cibernéticas, lo que hace necesario implementar medidas de protección robustas. Nuestra investigación aborda los desafíos duales de optimizar la eficiencia del aprendizaje federado en nubes híbridas y mejorar la seguridad contra posibles ataques. Detallamos nuestros hallazgos y metodologías en dos documentos subsiguientes, cada uno centrado en estrategias innovadoras para agilizar las operaciones de FL y fortalecerlas contra violaciones de seguridad en entornos tan complejos.

4.1. Protección de la Clasificación de Temas Sensibles en Aprendizaje Federado contra Ataques de Envenenamiento

Uno de los problemas de que adolece el aprendizaje federado es que es vulnerable a una serie de ataques realizados por parte de los clientes, el servidor o entidades externas. El módulo FedSecure tiene como objetivo trabajar para mejorar la seguridad del aprendizaje federado en el borde de la nube. Este trabajo se realizará en dos direcciones: 1) el desarrollo de modelos de reputación de los clientes para detectar ataques de envenenamiento, 2) la mejora de los criterios de agregación del modelo global.

El componente FedSecure ha comenzando trabajando por este segundo componente. Como resultado se ha publicado el siguiente artículo en NDSS, conferencia tier-1 en seguridad de sistemas distribuidos:

> T Chu, A Garcia-Recuero, C Iordanou, G Smaragdakis, N Laoutaris. Securing Federated Sensitive Topic Classification against Poisoning Attacks. 30th Annual Network and Distributed System Security Symposium, {NDSS} 2023
>
> Enlace a la presentación del paper en NDSS

El paper presenta una solución basada en aprendizaje federado para construir un clasificador distribuido capaz de detectar URLs contenido sensible, es decir, contenido relacionado con categorías como como la salud, las creencias políticas, la orientación sexual, etc. Aunque las limitaciones de los anteriores clasificadores offline/centralizados, sigue siendo centralizados, sigue siendo vulnerable a los ataques de envenenamiento de maliciosos que pueden intentar reducir la precisión de los usuarios benignos difundiendo actualizaciones defectuosas del modelo. Para evitarlo desarrollamos un esquema de agregación robusto basado en la lógica subjetiva y la detección de ataques basados en residuos. Empleando una combinación de de análisis teórico, simulación basada en trazas y validación experimental validación experimental con un prototipo y usuarios reales, demostramos que nuestro clasificador puede detectar contenidos sensibles con gran precisión, aprender nuevas etiquetas con rapidez, y seguir siendo robusto ante ataques de envenenamiento envenenamiento por parte de usuarios maliciosos, así como de entradas imperfectas de usuarios no maliciosos. no maliciosos.

El Anexo 2 del entregable incluye el paper (en inglés).

4.2. Cuantificación y Preservación de la Privacidad en Aprendizaje Federado con Poda

De Modelos La poda de modelos ha sido propuesta como una técnica para reducir el tamaño y la complejidad de los modelos de aprendizaje federado (FL). Al hacer los modelos locales más simples mediante la poda, se espera intuitivamente mejorar la protección contra ataques a la privacidad. Sin embargo, el nivel de protección de la privacidad esperado no había sido previamente caracterizado ni optimizado conjuntamente con la utilidad. En este documento, primero caracterizamos la privacidad ofrecida por la poda. Establecemos límites superiores teóricos de la información sobre la fuga de información desde FL podado y validamos experimentalmente estos límites bajo ataques a la privacidad de última generación en diferentes esquemas de poda de FL. En segundo lugar, introducimos PriPrune: un algoritmo consciente de la privacidad para la poda en FL. PriPrune utiliza máscaras de poda de defensa, que se pueden aplicar localmente después de cualquier algoritmo de poda, y adapta la tasa de poda de defensa para optimizar conjuntamente la privacidad y la precisión. Otra idea clave en el diseño de PriPrune es la Pseudo-Poda: realiza la poda de defensa dentro del modelo local y solo envía el modelo podado al servidor; mientras que los pesos eliminados por la máscara de defensa se retienen localmente para entrenamientos locales futuros en lugar de ser eliminados. Demostramos que PriPrune mejora significativamente el equilibrio entre privacidad y precisión en comparación con los esquemas de FL podados de última generación. Por ejemplo, en el conjunto de datos FEMNIST, PriPrune mejora la privacidad de PruneFL en un 45.5% sin reducir la precisión. Como resultado, se ha publicado el siguiente artículo en la revista ACM Transactions on Modeling and Performance Evaluation of Computing Systems, una conferencia de primer nivel sobre seguridad de sistemas distribuidos.

Chu, T., Yang, M., Laoutaris, N., & Markopoulou, A. (2023). PriPrune: Quantifying and Preserving Privacy in Pruned Federated Learning. *ACM Transactions on Modeling and Performance Evaluation of Computing Systems*.

Link to the paper

Este estudio cuantifica teórica y empíricamente las fugas de privacidad en la poda de modelos. Derivamos límites superiores teóricos sobre la cantidad de información revelada acerca del conjunto de datos de un usuario individual a través de las actualizaciones del modelo, en cualquier esquema de FL podado. Esta es la primera caracterización teórica de las fugas de privacidad en modelos de FL podados. El estudio también realiza una evaluación empírica exhaustiva que cuantifica la cantidad de fuga de privacidad de seis esquemas de poda, considerando varios ataques a la privacidad de última generación. Además, diseña un nuevo ataque a la privacidad (denominado Inversión de Gradientes Escasos, o SGI), específicamente diseñado para explotar vulnerabilidades inherentes a la poda de modelos en FL. Esta evaluación, combinada con nuestro análisis teórico, proporciona percepciones valiosas sobre las elecciones y parámetros que afectan la privacidad proporcionada por la poda.

Basándonos en estos hallazgos, hacemos nuestra segunda contribución: diseñamos PriPrune, un mecanismo de poda que preserva la privacidad. PriPrune defiende contra ataques de inversión de gradientes en FL podado, realizando una poda local adicional (de defensa), después de cualquier esquema de poda (base) en FL. PriPrune aplica una máscara de defensa personalizada y adapta la tasa de poda de defensa, de modo que optimiza conjuntamente la precisión del modelo y la privacidad, utilizando retropropagación aumentada con Muestreo de Softmax Gumbel. Otra idea clave en el diseño de PriPrune es lo que denominamos Pseudo-Poda: implica podar con la máscara de defensa dentro del modelo local y transmitir el modelo

podado al servidor, mejorando así la privacidad. Sin embargo, los pesos eliminados por la máscara de defensa en el modelo local no se descartan; en cambio, se retienen localmente para rondas subsiguientes de entrenamiento local, preservando así la precisión del modelo.

4.3. FedQV: Aprovechando la Votación Cuadrática en Aprendizaje Federado

El Aprendizaje Federado (FL) permite que diferentes partes colaboren en el entrenamiento de un modelo global sin revelar sus etiquetas locales respectivas. Un paso crucial del FL, el de agregar modelos locales para producir uno global, comparte muchas similitudes con la toma de decisiones públicas, y en particular con las elecciones. En ese contexto, una debilidad mayor del FL, específicamente su vulnerabilidad a los ataques de envenenamiento, puede interpretarse como una consecuencia del principio de un persona un voto (en adelante 1p1v), que sustenta la mayoría de las reglas de agregación contemporáneas.

En este documento, presentamos FedQV, un nuevo algoritmo de agregación basado en la votación cuadrática, recientemente propuesta como una alternativa mejor a las elecciones basadas en 1p1v. Nuestro análisis teórico establece que FedQV es un mecanismo veraz en el que ofertar de acuerdo con la valoración real de uno es una estrategia dominante que logra una tasa de convergencia equiparable a la de los métodos más avanzados. Como resultado, el siguiente artículo ha sido publicado en la revista ACM Transactions on Modeling and Performance Evaluation of Computing Systems, una conferencia de primer nivel sobre seguridad de sistemas distribuidos.

Chu, T., & Laoutaris, N. (2024). FedQV: Leveraging Quadratic Voting in Federated Learning. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, *8*(2), 1-36.

Link to the paper

Además, nuestro análisis empírico utilizando múltiples conjuntos de datos del mundo real valida el rendimiento superior de FedQV frente a ataques de envenenamiento. También muestra que combinar FedQV con "presupuestos" de votación desiguales según una puntuación de reputación aumenta aún más sus beneficios de rendimiento. Finalmente, demostramos que FedQV puede combinarse fácilmente con mecanismos de preservación de privacidad robustos a ataques bizantinos para mejorar su robustez tanto contra ataques de envenenamiento como de privacidad.

El Anexo 5 del entregable incluye el artículo (en inglés).

5. FedWM

El objetivo el componente es crear un esquema de marca de agua o similar para proteger de la propiedad de los datos cuando se requiera la redistribución de los datos o metadatos que se necesiten intercambiar entre servidores en el borde en el marco del FLaaS. Como primer avance, se una nueva técnica para modular la frecuencia de aparición de unos pocos tokens en un conjunto de datos para codificar una marca de agua invisible que puede utilizarse para proteger derechos de propiedad sobre los datos. Desarrollamos algoritmos heurísticos algoritmos heurísticos óptimos y rápidos para crear y verificar esas marcas de agua.

La técnica se presentará en una conferencia internacional Tier-1 en materia de ingeniería de datos:

> D. Isler, E. Cabana, A. Garcia-Recuero, G. Koutrika, N. Laoutaris, “FreqyWM: Frequency Watermarking for the New Data Economy,” Aceptado para publicación en IEEE International Conference of Data Engineering 2024.

En el artículo también demostramos la robustez de nuestra técnica contra contra varios ataques y derivamos límites analíticos para la probabilidad de "detectar" erróneamente una marca de agua en un conjunto de datos que no la contiene. Nuestra técnica es aplicable tanto a datos unidimensionales a conjuntos de datos unidimensionales y multidimensionales. tipo de token, permite un control fino de la distorsión introducida y se puede utilizar en una variedad de casos de uso que implican la compra y en los mercados de datos actuales.

El anexo 3 del entregable incluye el paper (en inglés)

6. FLaaS Manager

De alguna manera, el aprendizaje federado asume que los diferentes nodos que participan en el entrenamiento del modelo de aprendizaje distribuido tienen incentivo suficiente en mejorar el modelo para participar activamente en el proceso. Esto ha sido así en los primeros modelos de aprendizaje federado, como los empleados por el teclado de Google. Sin embargo, según se desarrolle la tecnología y aumenten los casos de uso, puede haber ocasiones en que la entidad que dispone los datos para entrenar los modelos no necesariamente es la que está interesada en el desarrollo de estos modelos.

Por el contrario, los mercados de datos buscan diseñar entidades que sean capaces de mediar entre los proveedores y los consumidores de datos sin que medie ningún interés de los primeros en el uso que los segundos hacen de los mismos. El diseño de estos mercados de datos distribuidos tiene una serie de desafíos técnicos, algunos de los cuales son objeto de investigación en el proyecto. En este sentido, se está trabajando en dos frentes:

6.1. Entendiendo el Precio de los Datos en Mercados Comerciales de Datos

Facilitar el establecimiento de precios de los datos, para dinamizar y reducir la incertidumbre en las transacciones, facilitando la toma de decisiones de precios por parte de la parte vendedora, y proporcionando información para la toma de decisiones de compra. En esta dirección se ha conseguido una primera publicación del primer modelo de predicción de precios basado en información de mercado y que se presentó en 2023 en una conferencia internacional Tier-1 en materia de ingeniería de datos en Anaheim, California::

> Santiago Andrés Azcoitia, Costas Iordanou, and Nikolaos Laoutaris. Understanding the Price of Data in Commercial Data Marketplaces. April 2023. 39th IEEE International Conference on Data Engineering (ICDE 2023)
>
> Enlace a la presentación (Talk)
>
> Enlace a las transparencias (Slides)
>
> Enlace al conjunto de datos compartido con la comunidad (Dataset)

En la actualidad, se trabaja en la federación de este modelo de precios, de forma que la información de oferta y transacciones resida en los proveedores y mercados de datos, pero puedan compartir modelos en la nube para establecer los precios en base a su conocimiento conjunto.

6.2. Pruébalo Antes de Comprarlo

Se han propuesto nuevos mecanismos de mercado de datos (DM) para entrenar colaborativamente modelos compartidos por compradores potenciales a través de arquitecturas distribuidas utilizando datos bajo el control de la plataforma. En la mayoría de los casos, los DM requieren que los compradores compartan información y modelos sensibles, lo que compromete su escalabilidad y la privacidad y propiedad intelectual de los compradores, además de sobrecargar a la plataforma con el costo de procesamiento no despreciable para

seleccionar o evaluar los datos. Además, estos mecanismos imponen a la plataforma un costo de procesamiento significativo para seleccionar o evaluar los datos, el cual demostramos que está lejos de ser insignificante, basándonos en los costos reales de las nubes públicas y los DM comerciales, poniendo en riesgo su viabilidad.

Definimos una nueva arquitectura de mercado de datos que permite a los compradores probar datos contra su modelo para seleccionar y comprar activos que mejor se ajusten a sus necesidades, y propone cobrar a los compradores por los costos de procesamiento asociados con las transacciones de datos. A diferencia de diseños anteriores, nuestra propuesta obliga a los compradores a preocuparse por la cantidad de procesamiento solicitado de la plataforma. Demostramos técnicas para que los compradores optimicen el costo de los procesos de selección y compra de datos, a saber, estrategias de compra inteligentes novedosas para reducir el número de solicitudes de evaluación, y el uso de "modelos de valoración títeres" (puppet valuation models o PVM) y "funciones de valoración" (valuation functions o VFs) para reducir su complejidad. Los VFs y PVM también sortean el problema de compartir propiedad intelectual sensible por parte de los compradores. Creemos que nuestro mercado de datos está listo para ser implementado utilizando la funcionalidad de sandbox existente de entidades comerciales, y estamos trabajando en demostrar su viabilidad utilizando casos de uso de clasificación de imágenes y predicción basados en datos de movilidad.

En el Anexo 5 presentamos el progreso en este componente, que llamamos preliminarmente Entre la Suerte y el Abuso de Devoluciones Gratuitas.

6.3. De RAGs a Riquezas: Recuperación Descentralizada y Medición de la Influencia de Documentos en Sistemas de Generación con Recuperación Mejorada (RAG)

Los sistemas de Generación con Recuperación Mejorada (RAG) representan una mejora innovadora para los modelos de lenguaje grandes (LLMs) tradicionales al incorporar la capacidad de recuperar datos externos relevantes durante el proceso de generación. Esta técnica aumenta significativamente la precisión y relevancia de los resultados, especialmente en campos especializados como finanzas, derecho o medicina. Sin embargo, la integración de la recuperación de datos externos introduce desafíos complejos que incluyen preocupaciones de privacidad, problemas de escalabilidad y la necesidad de incentivar de manera efectiva a los proveedores de datos de terceros.

Estamos avanzando en un estudio que propone un marco novedoso que aborda estos desafíos introduciendo la recuperación descentralizada (DR) y la medición de la influencia de documentos. Nuestro enfoque busca equilibrar los intereses de los propietarios de LLM y los proveedores de datos de terceros. Los propietarios de LLM pueden compensar solo por los datos valiosos que realmente se utilizan, sin comprometer la privacidad de las consultas de los usuarios. Simultáneamente, los proveedores de datos reciben una compensación justa sin necesidad de revelar prematuramente su información propietaria. Esta estrategia dual no solo mejora la privacidad y la escalabilidad, sino que también fomenta un modelo de negocio sostenible para los sistemas RAG, promoviendo un mercado de datos más equitativo.

En el Anexo 7 presentamos el progreso en este componente, que llamamos preliminarmente De RAGs a Riquezas.

7. Conclusión

Este documento detalla el progreso actual y una versión preliminar de los componentes dentro del proyecto MLEDGE. Un objetivo principal de este proyecto, y en particular de la iniciativa PRTR, es aprovechar los avances científicos producidos por la industria. Para este fin, el documento elabora sobre tres casos de uso específicos. Estos casos de uso se han iniciado a través de licitaciones separadas como parte del marco del proyecto para involucrar a agentes industriales. Estos agentes tienen la tarea de demostrar la practicidad del aprendizaje federado en el borde de la nube. Además, estos casos de uso actúan como pruebas de concepto vitales para los diversos componentes científicos desarrollados durante la duración del proyecto.

El documento también describe los próximos pasos en el proyecto MLEDGE, que están estructurados en torno a varios objetivos principales:

Integración de Socios Industriales: El proyecto tiene como objetivo incorporar a empresas responsables de desarrollar e implementar los casos de uso descritos. Esta integración es crucial para impulsar las aplicaciones prácticas de los avances científicos del proyecto.

Investigación y Desarrollo Continuos: Se dirigirá un esfuerzo continuo hacia la investigación y el desarrollo adicional de los componentes del proyecto. Esto involucrará refinar tecnologías existentes y posiblemente introducir nuevas innovaciones para mejorar la producción y eficiencia del proyecto.

Integración Técnica y Exploración de Aplicaciones: Se ha planificado un estudio para explorar cómo estos componentes técnicos recién desarrollados de MLEDGE pueden integrarse con la plataforma existente de Federated Learning as a Service (FLaaS). Además, este paso implica evaluar la utilidad de estos componentes dentro de los casos de uso del proyecto y más allá, con el objetivo de maximizar su aplicabilidad e impacto en varios escenarios.

Evaluación de la Eficacia del Componente: Evaluar el rendimiento y el impacto de los componentes recién desarrollados dentro de entornos operativos. Esto involucra procesos de prueba y validación rigurosos para asegurar que cumplan con los estándares y requisitos necesarios.

Escalado y Despliegue: Estrategias para escalar las soluciones desarrolladas dentro de MLEDGE para un despliegue más amplio, incluyendo la identificación de nuevos mercados y aplicaciones potenciales para las tecnologías.

Medidas de Cumplimiento y Seguridad: Desarrollar y hacer cumplir estrictos protocolos de cumplimiento y seguridad para proteger los datos y operaciones asociados con el aprendizaje federado en el borde, asegurando que todas las actividades se alineen con los estándares legales y éticos.

Al abordar estos pasos, el proyecto MLEDGE busca consolidar sus contribuciones a las tecnologías de aprendizaje federado en el borde de la nube, asegurando que el proyecto no solo avance el conocimiento científico sino que también brinde beneficios tangibles liderados por la industria. A medida que el proyecto MLEDGE avanza y se determinan los oferentes exitosos

para los casos de uso, una mayor visibilidad de sus planes influirá significativamente en el contenido de este entregable. El propósito de alinear este entregable con las estrategias de los oferentes exitosos es enriquecer los detalles que rodean los casos de uso, incluyendo sus requisitos específicos y el diseño de la plataforma que apoyará su desarrollo. Este proceso de alineación asegurará que los documentos de análisis de requisitos y diseño de casos de uso reflejen con precisión las necesidades matizadas de cada caso de uso. Además, aclarará cómo estos requisitos se integran con los objetivos más amplios del proyecto, facilitando un proceso de desarrollo más cohesivo.

Además, estos entregables proporcionarán información detallada sobre qué módulos específicos serán probados en cada caso de uso y los demostradores prácticos utilizados para evaluar su funcionalidad. Este enfoque tiene como objetivo delinear un marco de pruebas claro para cada módulo, resaltando las demostraciones prácticas que verificarán la efectividad y robustez de las soluciones desarrolladas. Al hacerlo, el proyecto puede asegurarse de que cada componente no solo cumpla con sus requisitos funcionales previstos, sino que también se alinee con las realidades prácticas de los entornos industriales para los que están diseñados. Esta metodología estructurada e informada respalda una evaluación sistemática, asegurando que todos los componentes desarrollados entreguen sus beneficios previstos de manera efectiva dentro de aplicaciones del mundo real.

Anexo 1: FLTorrent progress report

Summary

The FLTorrent project has made substantial progress in the realm of decentralized federated learning, emphasizing a delicate balance between security, performance, and bandwidth optimization. Decentralized Federated Learning (FL) is a variant of FL, where the central server is eliminated and clients can send local gradients to others by peer-to-peer communication. Despite their pioneering contributions to decentralized FL, most of these works focus on improving the convergence speed of the model. Decentralized FL still has privacy issues that cannot be ignored since the clients can investigate their neighbors' gradients directly. FLTorrent, with its foundation in BitTorrent, introduces a novel dimension to the distributed learning architecture. By leveraging the inherent efficiency and scalability of BitTorrent's peer-to-peer communication model, FLTorrent enhances its ability to distribute and synchronize model updates across a decentralized network of peers. This not only fosters collaboration but also facilitates the seamless exchange of information, contributing to the democratization of machine learning processes.

This report encapsulates the detailed efforts to design, implement, and evaluate the FLTorrent system with a particular focus on optimizing loss while adhering to constraints on upload and download links' bandwidth. Notably, our efforts also include incorporating insights from FLtorrent, a groundbreaking approach that leverages BitTorrent for the efficient communication of model chunks.

Objectives

The overarching objectives of the FLTorrent project have been intricately crafted to address the multifaceted challenges associated with decentralized federated learning:

- **Decentralized FL Model Development**: Develop a collaborative and secure decentralized federated learning model.
- **Privacy-Preserving Techniques**: Implement advanced techniques to ensure privacy, focusing on anonymizing data chunks during the learning process.
- **Bandwidth-Constrained Performance Optimization**: Optimize model loss while considering constraints on upload and download links' bandwidth.

Methodologies

Decentralized Federated Learning Model

The FLTorrent decentralized FL model utilizes a collaborative learning approach, distributing the training process across a network of peers. This ensures a robust model while promoting privacy, security, and efficient bandwidth utilization.

Privacy-Preserving Techniques

Our commitment to privacy is underscored by the implementation of multiple advanced techniques, including the anonymization of data chunks. By employing a combination of privacy-preserving methods, the FLTorrent system fortifies itself against potential security

threats and breaches. This approach not only safeguards sensitive information during data exchange but also ensures the integrity of the decentralized federated learning process.

Bandwidth-Constrained Performance Optimization Heuristics

In the pursuit of optimal performance, FLTorrent incorporates performance optimization heuristics that explicitly consider constraints on upload and download links' bandwidth. This strategic approach minimizes loss while ensuring efficient bandwidth utilization. The system employs adaptive chunk distribution heuristics, considering constraints on upload and download links' bandwidth. This approach optimizes the learning process by adapting to the unique bandwidth capabilities of individual peers, minimizing loss, and enhancing overall performance. Importantly, we draw inspiration from FLTorrent, a paradigm that utilizes BitTorrent for the efficient communication of model chunks across distributed peers.

Evaluation

The current phase of the FLTorrent project revolves around a meticulous evaluation process, focusing on:

1. **Security Evaluation**: Ensuring the effectiveness of privacy-preserving techniques against potential privacy breaches and attacks.
2. **Bandwidth-Constrained Performance Evaluation**: Assessing the impact of performance optimization heuristics with a focus on minimizing loss while adhering to constraints on upload and download links' bandwidth.

Conclusion

The FLTorrent project stands at the forefront of decentralized federated learning, emphasizing not only security and privacy but also performance optimization within the constraints of upload and download links' bandwidth. It extends beyond conventional methodologies, offering a solution that is not only secure and privacy-conscious but also highly performant in bandwidth utilization. The integration of privacy-preserving techniques, which encompass a sophisticated blend of multiple advanced methods, ensures that sensitive information remains confidential during the decentralized learning process. Moreover, FLTorrent 's commitment to performance optimization is underscored by the deployment of bandwidth-constrained heuristics. These heuristics, inspired by the dynamic distribution principles of FLTorrent, adaptively allocate model chunks based on the bandwidth constraints of individual peers. This approach not only minimizes loss in the learning process but also optimizes the use of available bandwidth resources, contributing to the overall efficiency of the federated learning model.

Drawing inspiration from Bit Torrent network and incorporating its principles, it emerges not just as a solution but as a transformative force in the realm of decentralized federated learning. This amalgamation of cutting-edge technologies, privacy-preserving methodologies, and innovative bandwidth utilization strategies positions FLTorrent as a beacon of innovation and a pioneering solution in achieving the delicate equilibrium between security, performance, and efficient bandwidth utilization in the ever-evolving landscape of machine learning. The successful integration of privacy-preserving techniques and bandwidth-constrained performance optimization heuristics positions FLTorrent as a pioneering solution in achieving a delicate equilibrium between security, performance, and efficient bandwidth utilization.

Anexo 2: The report on optimizing and real-world implementation of FLTorrent

To overcome the shortcomings inherent in purely simulation-based federated learning (FL) research, I have developed and implemented a practical FL framework that enables testing on real-world servers. This framework leverages the peer-to-peer (P2P) capabilities of BitTorrent to run federated learning processes, enhancing the realism and relevance of our research. By conducting tests on actual servers, we gain vital insights into the operational scalability of our FL solutions and their adaptability to real network conditions. This approach not only helps in understanding how FL systems perform outside controlled environments but also in identifying potential bottlenecks and challenges that may not be apparent in simulations.

Expanding on this foundation, the use of a P2P platform like BitTorrent for running federated learning allows us to simulate a more decentralized and distributed network environment, closely mirroring the actual deployment scenarios where FL is most beneficial. This method provides a robust test bed for assessing the resilience and efficiency of FL strategies under varied network reliabilities and data distribution conditions. It also facilitates a deeper understanding of the security implications and data privacy measures necessary in a P2P federated learning setup. By integrating real-world testing early in the research phase, we can iteratively refine our FL models and protocols, ensuring they are not only theoretically sound but also practically viable and ready for broader adoption.

In the pursuit of refining federated learning (FL) systems, especially those operating on peer-to-peer (P2P) platforms like BitTorrent, it's crucial to address the efficiency and adaptability of data transmission. This involves not only the technical execution of transferring model updates (chunks) but also ensuring these systems can dynamically adjust to meet user-specific requirements. Our research has led to the development of a robust framework that enhances both the strategic dissemination of data and the system's responsiveness to user preferences and constraints.

Optimized Chunk Transmission

1. Chunk Importance Evaluation: Recognizing the varied impact of different chunks on the overall model performance, I introduced a method that utilizes similarity metrics and group testing strategies to rank the importance of each chunk. This evaluation process ensures that chunks with the most significant influence on model accuracy are identified and prioritized during transmission. This approach is particularly valuable in networks where bandwidth is limited, ensuring that the most critical updates are disseminated first.
2. Linear Programming for Scheduling: To effectively manage the complexities of chunk transmission across a distributed network, a scheduling algorithm based on linear programming was developed. This algorithm optimizes the distribution of chunks by considering both the bandwidth limitations of the network and the potential accuracy improvements each chunk offers. By balancing these factors, the algorithm facilitates

efficient use of network resources while maximizing the performance gains from each transmitted chunk.

User-Centric Adaptability

The framework is designed to be highly adaptive, catering to varying user demands that range from minimizing training time to maximizing model accuracy:

- Adaptive Transmission Strategies: Depending on the user's priorities, the framework can adjust its transmission strategy. For users focused on achieving the highest accuracy, the system prioritizes the transfer of chunks that offer the most substantial improvements. Conversely, for users where training time is a constraint, the system adapts to transmit smaller or less complex chunks to speed up the overall process without substantial losses in performance.
- Incremental Updates: The framework supports incremental updates, a method that allows for meaningful accuracy improvements even when only limited chunk transfers are feasible. This capability is essential for scenarios where network conditions are suboptimal or where users cannot afford the bandwidth required for full model updates. Incremental updates ensure that each transmitted chunk still contributes positively towards the model's learning, thus optimizing both resource usage and learning outcomes.

Future Directions

As federated learning (FL) continues to evolve, particularly in the context of peer-to-peer (P2P) platforms, there are several avenues for further research and development. These future directions aim to enhance privacy, optimize system performance, and expand testing capabilities to ensure the robustness and applicability of FL systems in diverse environments.

1. Privacy Protection Research:
    - Advanced Attack Strategies: In light of increasing cyber threats, there is a critical need to design and test advanced attack strategies specifically targeting FL systems. This includes exploring potential vulnerabilities through membership inference and reconstruction attacks, which can expose sensitive information about participants in FL systems.
    - Enhanced Privacy Mechanisms: Given the distributed nature of chunk exchanges in FL, developing enhanced privacy-preserving mechanisms tailored to these scenarios is paramount. This involves creating sophisticated methods that protect data during transmission and processing, ensuring that participant privacy is maintained without compromising the integrity and usability of the shared model.
2. System Optimization:
    - Dynamic Node Behaviors and Heterogeneous Environments: FL systems must address the challenges posed by dynamic node behaviors and heterogeneous environments, which can affect the consistency and efficiency of data exchanges. Optimizing the scheduling model to accommodate these variables will improve system resilience and performance.

- Expanding Privacy Guarantees: By integrating differential privacy techniques with existing chunk scheduling methods, FL systems can offer stronger privacy guarantees. This combination can help obscure individual data contributions while still allowing for the aggregation of useful, generalized insights from the distributed data.

3. Broader Testing:
    - Scaling to Diverse Environments: To ensure the practicality and effectiveness of FL systems, it is essential to scale the framework to support diverse real-world settings. This includes deployment on edge and IoT devices, which often operate under constraints such as limited computational power and intermittent connectivity.
    - Extensive Model Testing: Broader testing also involves validating the framework across various models and configurations to ascertain its versatility and adaptability. This broad spectrum testing will help identify potential issues in real-world applications and guide further improvements.

By focusing on these areas, future research can significantly advance the capabilities of federated learning systems, making them more secure, efficient, and applicable across a wider range of environments and scenarios. These efforts will not only strengthen the theoretical foundations of FL but also enhance its practical implementations, driving forward the next generation of distributed machine learning technologies.

Anexo 3: Securing Federated Sensitive Topic Classification against Poisoning Attacks

Securing Federated Sensitive Topic Classification against Poisoning Attacks

Tianyue Chu
IMDEA Networks Institute
Universidad Carlos III de Madrid

Alvaro Garcia-Recuero
IMDEA Networks Institute

Costas Iordanou
Cyprus University of Technology

Georgios Smaragdakis
TU Delft

Nikolaos Laoutaris
IMDEA Networks Institute

***Abstract*—We present a Federated Learning (FL) based solution for building a distributed classifier capable of detecting URLs containing sensitive content, i.e., content related to categories such as health, political beliefs, sexual orientation, etc. Although such a classifier addresses the limitations of previous offline/centralised classifiers, it is still vulnerable to poisoning attacks from malicious users that may attempt to reduce the accuracy for benign users by disseminating faulty model updates. To guard against this, we develop a robust aggregation scheme based on subjective logic and residual-based attack detection. Employing a combination of theoretical analysis, trace-driven simulation, as well as experimental validation with a prototype and real users, we show that our classifier can detect sensitive content with high accuracy, learn new labels fast, and remain robust in view of poisoning attacks from malicious users, as well as imperfect input from non-malicious ones.**

I. Introduction

Most people are not aware that tracking services are present even on sensitive web domains. Being tracked on a cancer discussion forum, a dating site, or a news site with non-mainstream political affinity can be considered an "elephant in the room" when it comes to the anxieties that many people have about their online privacy. The General Data Protection Regulation (GDPR) [33] puts specific restrictions on the collection and processing of sensitive personal data *"revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, also genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural persons sex life or sexual orientation"*. So do other public bodies around the world, *e.g.* in California (California Consumer Privacy Act (CCPA) [34]), Canada [35], Israel [36], Japan [37], and Australia [38].

In a recent paper, Matic et al. [4] showed how to train a classifier for detecting whether the content of a URL relates to any of the above-mentioned sensitive categories. The classifier was trained using 156 thousand sensitive URLs obtained from the Curlie [32] crowdsourced web taxonomy project. Despite the demonstrated high accuracy, this method has limitations that stem from being *centralised* and *tied to a fixed training set*. The first limitation means that the method cannot be used "as is" to drive a privacy-preserving distributed classification system. The second limitation implies that it is not straightforward to cover new labels related to yet unseen sensitive content. For example, in their work the Health category could be classified with accuracy greater than 90%. However, the training labels obtained from Curlie in 2020 did not include any labels related to the COVID-19 pandemic. Therefore, as will be shown later, this classifier classifies COVID-19 related sites with only 53.13% accuracy.

Federated Learning (FL) [5], [13] offers a natural solution to the above two mentioned limitations, namely, centralized training and training for a fixed training set. FL allows different clients to train their classification models locally without revealing new or existing sensitive URLs that they label, while collaborating by sharing model updates that can be combined to build a superior global classification model. FL has proved its value in a slew of real-world applications, ranging from mobile computing [46]–[48] to health and medical applications [49]–[51]. However, due to its very nature, FL is vulnerable to so-called *poisoning attacks* [12], [26] mounted by malicious clients that may intentionally train their local models with faulty labels or backdoor patterns, and then disseminate the resulting updates with the intention of reducing the classification accuracy for other benign clients. State-of-the-art approaches for defending against such attacks depend on robust aggregation [8], [15], [16], [20], [27], [60] which, as we will demonstrate later, are slow to converge, thereby making them impractical for the sensitive-content classification problem that we tackle in this paper.

Our Contributions: In this paper, we employ FL for sensitive content classification. We show how to develop a robust FL method for classifying arbitrary URLs that may contain GDPR sensitive content. Such a FL-based solution allows building a distributed classifier that can be offered to end-users in the form of a web browser extension in order to: (i) warn them before and while they navigate into such websites, especially when they are populated with trackers, and (ii) allow them to contribute new labels, e.g., health-related websites about COVID-19, and thus keeping the classifier always up-to-date. To the best of our knowledge this method represents the first use of FL for such task.

Network and Distributed System Security (NDSS) Symposium 2023
27 February - 3 March 2023, San Diego, CA, USA
ISBN 1-891562-83-5
https://dx.doi.org/10.14722/ndss.2023.23112
www.ndss-symposium.org

Our second major contribution is the development of a reputation score for protecting our FL-based solution from poisoning attacks [12], [26]. Our approach is based on a novel combination of subjective logic [3] with residual-based attack detection. Our third contribution is the development of an extensive theoretical and experimental performance evaluation framework for studying the accuracy, convergence, and resilience to attacks of our proposed mechanism. Our final contribution is the implementation of our methods in a prototype system called *EITR* (standing for "Elephant In the Room" of privacy) and our preliminary experimental validation with real users tasked to provide fresh labels for the accurate classification of COVID-19 related URLs.

Our findings: Using a combination of theoretical analysis, simulation, and experimentation with real users, we:

- Demonstrate experimentally that our FL-based classifier achieves comparable accuracy with the centralised one presented in [4].
- Prove analytically that under data poisoning attacks, our reputation-based robust aggregation built around subjective logic, converges to a near-optimal solution of the corresponding Byzantine fault tolerance problem under standard assumptions. The resulting performance gap is determined by the percentage of malicious users.
- Evaluate experimentally our solution against state-of-the-art algorithms such as Federated Averaging [5], Coordinate-wise median [20], Trimmed-mean [20], FoolsGold [8], [15], Residual-based re-weighting [16] and FLTrust [60], and show that our algorithm is robust under Byzantine attacks by using different real-world datasets. We demonstrate that our solution outperforms these popular solutions in terms of convergence speed by a factor ranging from $1.6\times$ to $2.4\times$ while achieving the same or better accuracy. Furthermore, our method yields the most consistent and lowest Attack Success Rate (ASR), with at least 72.3% average improvement against all other methods.
- Validate using our *EITR* browser extension that our FL-based solution can quickly learn to classify health-related sites about COVID-19, even in view of noisy/inconsistent input provided by real users.

The remainder of the article is structured as follows: Section II introduces the background for our topic. Section III presents our reputations scheme for FL-based sensitive content classification, as well as its theoretical analysis and guarantees. Section IV covers our extensive performance evaluation against the state-of-the-art and Section V some preliminary results from our *EITR* browser extension. Section VI concludes the paper and points to on-going and future work including the generalization of our method to other topics.

II. Background

A. A Centralised Offline Classifier for Sensitive Content

Matic et al. [4] have shown how to develop a text classifier able to detect URLs that contain sensitive content. This classifier is centralised and was developed in order to conduct a one-off offline study aimed at estimating the percentage of the web that includes such content. Despite achieving an accuracy of at least 88%, utilising a high-quality training set meticulously collected by filtering the Curlie web-taxonomy project [32], this classifier cannot be used "as is" to protect real users visiting sensitive URLs populated by tracking services.

B. Challenges in Developing a Practical Classifier for Users

From offline to online: The classifier in [4] was trained using a dataset of 156 thousand sensitive URLs. Despite being the largest dataset of its type in recent literature, this dataset is static and thus represents sensitive topics up to the time of its collection. This does not mean, of course, that a new classifier trained with this data would never be able to accurately classify new URLs pertaining to those sensitive categories. This owes to the fact that categories such as Health, involve content and terms that do not change radically with time. Of course, new types of sensitive content may appear that, for whatever reason, may not be so accurately classified using features extracted from past content of the same sensitive category. Content pertaining to the recent COVID-19 pandemic is such an example. Although the Health category had 74,764 URLs in the training set of [4] which lead to a classification accuracy of 88% for Health, as we will see later in Figure 14 middle of Section V-C, the classifier of [4] classifies accurately as Health only 53.13% of the COVID-19 URLs with which we tested it. This should not come as a surprise since the dataset of [4] corresponds to content generated before the first months of 2020, during which COVID-19 was not yet a popular topic. Therefore, we need to find a way to update an existing classifier so that it remains accurate as new sensitive content appears.

From centralised to distributed: A natural way to keep a classifier up-to-date is to ask end-users to label new sensitive URLs as they encounter them. End-users can report back to a centralised server such URLs which can then be used to retrain the classification model. This, however, entails obvious privacy challenges of "Catch-22" nature, since to protect users by warning them about the presence of trackers on sensitive URLs, they would first be required to report to a potentially untrusted centralised server that they visit such URLs. Even by employing some methods for data scarcity, e.g., semi-supervised learning, the manual labelling from users remains sensitive and may be harmed by the untrusted server. Federated Learning, as already mentioned, is a promising solution for avoiding the above Catch22 by conducting a distributed, albeit, privacy-preserving, model training. In an FL approach to our problem, users would label new URLs locally, e.g., a COVID-19 URL as Health, retrain the classifier model locally, and then send model updates, not labelled data, to a centralised server that collects such updates from all users, compiles and redistributes the new version of the model back to them. In Section III we show how to develop a distributed version of the sensitive topic classifier of [4] using FL. The trade-off of using FL, is that the distributed learning group becomes vulnerable to attacks, such as "label-flipping" poisoning attacks discussed in Section IV. This paper develops a reputation scheme for mitigating such attacks. Other types of attacks and measures for preserving the privacy of users that participate in a FL-based classification system for sensitive content are discussed in Section VI.

C. Related Work

Privacy preserving crowdsourcing: Similar challenges to the ones discussed in the previous paragraph have been faced in

services like the *Price $heriff* [54] and *eyeWnder* [55] that have used crowdsourcing to detect online price discrimination and targeted advertising, respectively. Secure Multi-Party Computation (SMPC) techniques such as private k-means [56] are used to allow end-users to send data in a centralised server in a privacy-preserving manner. The centralised computation performed by *Price $heriff* and *eyeWnder* is not of ML nature, thus leaving data anonymisation as the main challenge, for which SMPC is a good fit. Classifying content as sensitive or not is a more complex ML-based algorithm for which FL is a more natural solution than SMPC.

General works on FL: FL [5], [13] is a compelling technique for training large-scale distributed machine learning models while maintaining security and privacy. The motivation for FL is that local training data is always kept by the clients and the server has no access to the data. Due to this benefit that alleviates privacy concerns, several corporations have utilised FL in real world services. In mobile devices, FL is used to predict keyboard input [46], human mobility [47] and behaviour for the Internet of Things [48]. FL is also applied in healthcare to predict diseases [49], [50], detect patient similarity [51] while overcoming any privacy constrains. For the classification, FL is not only implemented for image classification [52] but also text classification [53].

Resilience to poisoning attacks: Owing to its nature [12], [26], FL is vulnerable to poisoning attacks, such as label flipping [16] and backdoor attacks [12]. Therefore, several defence methods have been developed [8], [15], [16], [20]. While these state-of-the-art approaches perform excellently in some scenarios, they are not without limitations. First, they are unsuitable for our sensitive content classification, which necessitates that a classifier responds very fast to "fresh" sensitive information appearing on the Internet. In existing methods, the primary objective is to achieve a high classification accuracy. This is achieved via statistical analysis of client-supplied model updates and discarding of questionable outliers before the aggregation stage. However, since the server distrusts everyone by default, even if an honest client discovers some fresh sensitive labels, its corresponding updates may be discarded or assigned low weights, up until more clients start discovering these labels. This leads to a slower learning rate for new labels.

Second, recent studies [12], [26] have shown that existing Byzantine-robust FL methods are still vulnerable to local model poisoning since they are forgetful by not tracking information from previous aggregation rounds. Thus, an attacker can efficiently mount an attack by spreading it across time [31]. For example, [22] recently showed that even after infinite training epochs, any aggregation which is neglectful of the past cannot converge to an efficient solution.

The preceding studies demonstrate the importance of incorporating clients' previous long-term performance in evaluating their trustworthiness. Few recent studies have considered this approach [22], [60]. In [22], the authors propose leveraging historical information for optimisation, but not for assessing trustworthiness. In [60], a trust score is assigned to each client model update according to the cosine similarity between the client's and server's model updates, which is trained on the server's root dataset (details in Section IV-A3). However, it is impractical for a server to obtain additional data, such as a root dataset, in order to train a server-side model. In addition, because the server collects root data only once and does not update it throughout the training process, when new types of content emerge over time, the root data may become stale thereby harming the classifier's performance. Other recent studies employ spectral analysis [63], differential privacy [65], and deep model inspection [66] to guard against poisoning attacks, but, again, they do not use historical information to assess the reliability of clients. To measure client trustworthiness without collecting additional data at the server, in the next sections we show how to design a robust aggregation method to generate reputation automatically based on the historical behaviours of clients, which is a more realistic approach for a real FL-based decentralised system implemented as a browser extension for clients.

TABLE I: Notation

Abbreviation	Description
M	the total number of clients
N	the number of parameters of global model
Q	the number of samples of each client
T	the total number of iterations
$w_{i,n}^t$	the n-th parameter from client i in t iteration
$x_{i,n}^t$	the ranking of $w_{i,n}^t$ in w_n^t
A_n, B_n	the slope and intercept of repeated median linear regression
$e_{i,n}^t$	the normalised residual of the n-th parameter from client i in t iteration

III. A Robust FL Method for Classifying Sensitive Content on the Web

In this section, we first show how to build an FL-based classifier for sensitive content. Then we design a reputation score for protecting against poisoning attacks. We analyse theoretically the combined FL/reputation-based solution and establish convergence and accuracy guarantees under common operating assumptions.

A. FL Framework for Classifying Sensitive Content

Table I presents the notation that we use in the remainder of the paper. In FL, clients provide the server updated parameters from their local model, which the server aggregates to build the global model M.

Suppose we have M clients participating in our classification training task and the dataset $\mathcal{D} = \bigcup_{i=1}^{M} \mathcal{D}_i$, where $\mathcal{D}_i \sim \mathcal{X}_i(\mu_i, \sigma_i^2)$ denotes the local data of client i from non-independent and non-identically (Non-IID) distribution $\mathcal{X}_i$ with the mean μ_i and standard deviation σ_i. In our task, the clients' data is the textual content of URLs stripped of HTML tags. The objective function of FL, $\mathcal{L} : \mathbb{R}^d \rightarrow \mathbb{R}$ which is the negative log likelihood loss in our task, can be described as

$$\mathcal{L}(w) = \mathbb{E}_{D\sim\mathcal{X}}\left[l(w;\mathcal{D})\right]$$

where $l(w;\mathcal{D})$ is the cost function of parameter $w \in \mathcal{W} \subseteq \mathbb{R}^d$. Here we assume $\mathcal{W}$ is a compact convex domain with diameter d. Therefore, the task becomes

$$w^* = \arg\min_{w\in\mathcal{W}} \mathcal{L}(w)$$

To find the optimal w^*, we employ Stochastic Gradient Descent (SGD) to optimise the objective function.

During the broadcast phase, the server broadcasts the classification task and training instructions to clients. Then, the clients apply the following standard pre-processing steps

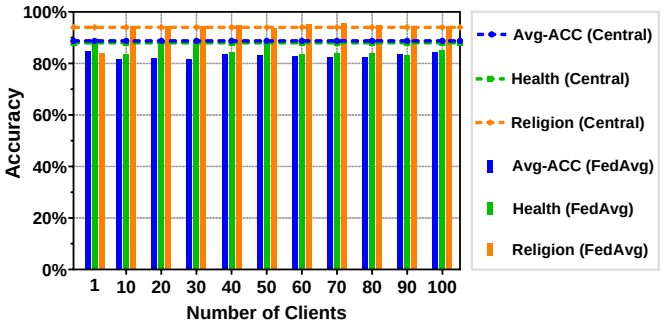


Fig. 1: Accuracy of FL classifiers and centralised classifiers in Health, Religion and all category.

on the webpage content, that is, transformation of all letters in lowercase and the removal of stop words. Next, the clients extract the top one thousand features utilising the Term Frequency-Inverse Document Frequency (TF-IDF) [58] as in [4]. At iteration t, the client i receives the current global model M_{global} and then following the training instructions from server, trains the local model on its training data $\mathcal{D}_i$ and optimises $w_i^t = \arg\min_w \mathcal{L}_i(w_i^t)$ by using SGD:

$$w_i^t \leftarrow w_i^{t-1} - r\frac{\partial \mathcal{L}_i(w_i^{t-1})}{\partial w}$$

where $\mathcal{L}_i(w) := \mathbb{E}_{\mathcal{D}_i \sim \mathcal{X}}[l(w;\mathcal{D}_i)] = \frac{1}{Q_i}\sum_{j=1}^{Q_i} l(w;\mathcal{D}_i^j)$, $\mathcal{D}_i^j$ and Q_i means the j-th sample and the number of samples of the client i respectively, and r is the learning rate.

In every iteration, after finishing the training process the clients send back their local updates to the server. Then, the server computes a new global model update by combining the local model updates via an aggregation method AGG as follows:

$$w^t = \texttt{AGG}\left(\{w_i^t\}_{i=1}^M\right)$$

Here we utilise the basic aggregation method (FedAvg) [5], which uses the fraction of each client's training sample size in total training samples as the average weights:

$$w^t = \sum_{i=1}^{M} \frac{Q_i}{Q} w_i^t$$

We introduce other robust aggregation methods in the next subsection. Subsequently, the server uses the global model update to renew the global model M_{global}.

Using the above FL-based framework we first evaluate how the number of users in the system affects the average accuracy of the classifier. The results for the sensitive categories, Health and Religion, as well as the overall average accuracy (Avg-ACC) are depicted in Figure 1. A first observation is that when a fixed size dataset is divided into multiple segments and distributed to more clients, the model's accuracy decreases since each client has less data for training. Compared to the centralised classifier, using the same data, the accuracy of the FL classifier is slightly lower, which is expected when the training is distributed to a larger number of clients. Overall, we observe that the average accuracy difference between the FL and the centralised classifier is 5.76%, and this remains steady as the number of clients grows. In addition, Looking at the different sensitive categories (Health and Religion), we see the FL-based classifier achieves an accuracy very close to

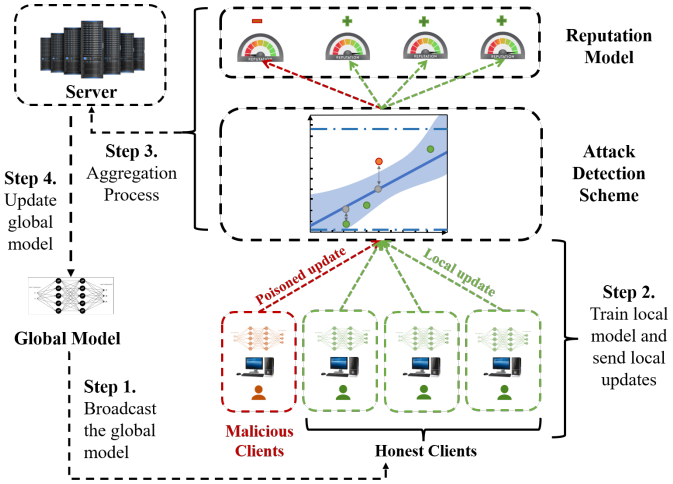


Fig. 2: Overview of reputation-based aggregation algorithm.

the corresponding one of the centralised classifier for these categories (on average 0.8533 vs. 0.88 and 0.9366 vs. 0.94, respectively).

B. A Reputation score for Thwarting Poisoning Attacks

Figure 2 shows an overview of our reputation-based aggregation algorithm consisting of three components: the *attack detection scheme*, the *reputation model*, and the *aggregation module*. The attack detection scheme re-scales and rectifies damaging updates received from clients. Then, the reputation model calculates each client's reputation based on their past detection results. Finally, the aggregation module computes the global model by averaging the updates of the clients using their reputation scores as weights. We detail each component in the following subsections.

1) Attack Detection Scheme: Our attack detection scheme aims to reduce the impact of suspicious updates by identifying them and applying a rescaling algorithm. At every iteration, when model updates from clients arrive at the server, we apply Algorithm 1 there to rescale the range of values for those parameters in the updates.

This restriction on the value range aims not only to minimise the impact of abnormal updates from attackers but also to limit the slope for the repeat median regression. Considering the n-th parameter in round t from all the participants, we calculate the standard deviation $\sigma(w_{i,n}^t)$ of this series. Then we sort them in ascending order and determine the range by subtracting the lowest value from the highest one. If the result is above the threshold ϖ, we rescale the highest and lowest value by deducting and adding its standard deviation respectively to further bound their range.

Then, a robust regression [14] is carried out to identify outliers among the updates in the current round. Outlier detection is a well-established topic in statistics. Robust regression methods often handle outliers by using the median estimators. Median-based aggregation methods have a rich and longstanding history in the area of robust statistics [21]. However, the methods developed by the traditional robust statistics can only withstand a small fraction of Byzantine clients, resulting in a low "breakdown point" [59]. Different from many other variations of the univariate median, the repeated median [14]

Algorithm 1: Rescale(w)

Input : $\{w_{i,n}^t\} \leftarrow$ Local Model parameters in round t

Output: $\{w_{i,n}^t\}$ with range of value less than ϖ

for $n \leftarrow 1$ **to** N **do**
  // Determine the maximum range
  $Rm = \max w_{i,n}^t - \min w_{i,n}^t = w_{i,n}^{t,(Max)} - w_{i,n}^{t,(Min)}$
  while $Rm > \varpi$ **do**
    // Rescale range based on standard deviation.
    $w_{i,n}^{t,(Max)} := w_{i,n}^{t,(Max)} - \sigma(w_{i,n}^t)$;
    $w_{i,n}^{t,(Min)} := w_{i,n}^{t,(Min)} + \sigma(w_{i,n}^t)$;
    // Updated Rm.
    $Rm = \max w_{i,n}^t - \min w_{i,n}^t = w_{i,n}^{t,(Max)} - w_{i,n}^{t,(Min)}$
  end while
end for

is impervious to atypical points even when their percentage is nearly 50%. The repeated median is defined as a modified U-statistic and the concept behind it is to utilise a succession of partial medians for computing approximation $\hat{\tau}$ of the parameter τ: For $k \in \mathbb{N}$, the value of parameter $\tau(z_1, \cdots, z_k)$ is determines by subset of k data points $z_1, \cdots, z_k$.

$$\hat{\tau} = \underset{z_1}{\text{median}} \left\{ \underset{z_2 \notin \{z_1\}}{\text{median}} \left\{ \underset{z_k \notin \{z_1, \cdots, z_{k-1}\}}{\text{median}} \tau(z_1, \cdots, z_k) \right\} \right\} \quad (1)$$

In our case, the intercept $\hat{A}$ and slop $\hat{B}$ are estimated by repeated median as bellow:

$$\hat{B}_n = \underset{i}{\text{median}} \left\{ \underset{i \neq j}{\text{median}} \{B_n(i,j)\} \right\} \quad (2)$$

$$\hat{A}_n = \underset{i}{\text{median}} \left\{ w_{i,n} - \hat{B}_n x_{i,n} \right\} \quad (3)$$

where $B_n(i,j) = \frac{w_{j,n} - w_{i,n}}{x_{j,n} - x_{i,n}}$, $x_{i,n}$ represents the index of $w_{i,n}$ in w_n which is sorted in ascending order.

Next, we employ the IRLS scheme [10] to generate each parameter's confidence score $s_{i,n}^t$ based on the normalised residual from repeated median regression, which is also utilised in a residual-based aggregation method [16]:

$$s_{i,n}^t = \frac{\sqrt{1 - diag(H_n^t)}}{e_{i,n}^t} \Psi \left(\frac{e_{i,n}^t}{\sqrt{1 - diag(H_n^t)}} \right) \quad (4)$$

where confidence interval $\Psi(x)$:

$$\Psi(x) = \max\{-\lambda\sqrt{2/M}, \min(\lambda\sqrt{2/M}, x)\}$$

and the hat matrix H_n^t:

$$H_n^t = x_n^t ({x_n^t}^T x_n^t)^{-1} {x_n^t}^T$$

with $e_{i,n}^t = \frac{25(M-1)\left(w_{i,n}^t - \hat{B}_n x_{i,n}^t - \hat{A}_n\right)}{37(M+4)\underset{i}{\text{median}}\left(|w_{i,n}^t - \hat{B}_n x_{i,n}^t - \hat{A}_n|\right)}$.

The distance between the point and the robust line is described by the confidence score derived from the normalised residual, which can be used to evaluate if the point is anomalous. Following the computation of the parameter's confidence score, and in light of the fact that some attackers want to generate updates with abnormal magnitudes in order to boost the damage, a useful protection is to identify low confidence values based on a threshold δ. Once the server recognises an update $\mathbf{w}_{i,n}^t$ of the client i with confidence values less than δ, rather than altering this update to the repeat median estimation, our technique replaces it with the median of $\mathbf{w}_n^t$, as follows:

$$\mathbf{w}_{i,n}^t = \begin{cases} \mathbf{w}_{i,n}^t & \text{if } s_{i,n}^t > \delta \\ \underset{i}{\text{median}} \left\{ \mathbf{w}_{i,n}^t \right\} & \text{if } s_{i,n}^t \leq \delta \end{cases} \quad (5)$$

With the above, not only we bound the range of updates, but also improve the aggregation by introducing a robustness estimator.

2) Reputation Model: During the aggregation phase in FL, we use a subjective logic model to produce client reputation scores. The subjective logic model is a subset of probabilistic logic that depicts probability values of belief and disbelief as degrees of uncertainty [3]. In the subjective logic model, reputation score R_i^t for client i in t iteration correlates to a subjective belief in the dependability of the client's behaviour [39], as measured by the belief metric opinion τ_i^t [9]. An opinion is comprised of three elements: belief b_i^t, disbelief d_i^t and uncertainty u_i^t, with restrictions that $b_i^t + d_i^t + u_i^t = 1$ and $b_i^t, d_i^t, u_i^t \in [0, 1]$. The reputation score may be calculated as the expected value of an opinion $E(\tau_i^t)$ which can be regarded as the degree of trustworthiness in client i. As a result, the value of the client's reputation is defined as follows:

$$R_i^t = E(\tau_i^t) = b_i^t + a u_i^t \quad (6)$$

where $a \in [0, 1]$ denotes the prior probability in the absence of belief, which reflects the fraction of uncertainty that may be converted to belief. On the other side, distinct observations determined by the rectification phase in our Algorithm 2 are used to count belief, disbelief, and uncertainty opinions. The positive observation denoted by P_i^t indicates that the update $\mathbf{w}_{i,n}^t$ is accepted ($s_{i,n}^t > \delta$), whereas a negative observation denoted by N_i^t indicates that the update is rejected ($s_{i,n}^t \leq \delta$). As a consequence, the positive observations boost the client's reputation, and vice versa. To penalise the negative observations from the unreliable updates, a higher weight η is assigned to negative observations than the weight κ to positive observations with constrain $\eta + \kappa = 1$. Therefore, in Beta distribution below:

$$Beta(p|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha - 1}(1 - x)^{\beta - 1} \quad (7)$$

with the constraints $0 \leq x \leq 1$, parameters $\alpha > 0, \beta > 0$, and $x \neq 0$ if $\alpha < 1$ and $x \neq 1$ if $\beta < 1$. The parameters α and β that represent positive and negative observations respectively, can be expressed as below

$$\begin{cases} \alpha = \kappa P_i^t + W a \\ \beta = \eta N_i^t + W(1 - a) \end{cases} \quad (8)$$

where W is the non-information prior weight and the default value is 2 [3].

As a consequence, the expected value of Beta distribution, which also stands for reputation value, can be calculated as follows:

$$E(Beta(p|\alpha, \beta)) = \frac{\alpha}{\alpha + \beta} = \frac{\kappa P_i^t + W a}{\kappa P_i^t + \eta N_i^t + W} = R_i^t \quad (9)$$

Based on (6) and (9), we can derive

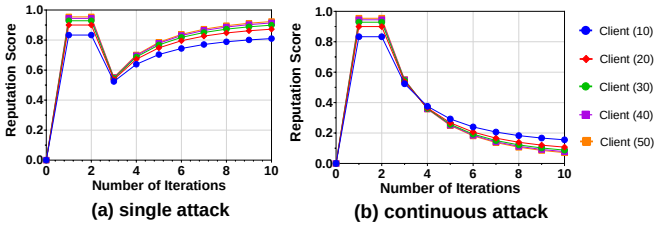


Fig. 3: The decay of reputation score in Client (X) with X model parameters when they (a) attack once at 3rd iteration and (b) attack continuously at and after the 3rd iteration.

$$\begin{cases} b_i^t = \frac{\kappa P_i^t}{\kappa P_i^t + \eta N_i^t + W} \\ d_i^t = \frac{\eta N_i^t}{\kappa P_i^t + \eta N_i^t + W} \\ u_i^t = \frac{W}{\kappa P_i^t + \eta N_i^t + W} \end{cases} \tag{10}$$

In addition, in order to take the client's historical reputation values in previous rounds into consideration, a time decay mechanism is included to lower the relevance of past performances without disregarding their influence. In other words, the reputation value from the most recent iteration contributes the most to the reputation model. We use exponential time decay in our model, as shown below:

$$\theta_{j,t} = \exp(-c(t-j)) \tag{11}$$

where $\exists c > 0$, $j \in [\tilde{s}, t]$, $\tilde{s} = \max(t-s, 0)$. We include a sliding window with a window length s that allows us to get a reputation for a certain time interval rather than the entire training procedure. We remove expired tuples with timestamps outside the window period during computation since they cannot provide meaningful information for the reputation. Hence, the final reputation score $\tilde{R}_i^t$ can be expressed as:

$$\tilde{R}_i^t = \frac{\sum_{j=\tilde{s}}^{t} \theta_{j,t} R_i^j}{\sum_{j=\tilde{s}}^{t} \theta_{j,t}} \tag{12}$$

To demonstrate how the reputation model evolves, we consider four scenarios where each client: (i) only attacks once at the same iteration, (ii) attacks continuously after launching an attack at the same iteration, (iii) only attacks once at different iteration, (iv) attacks continuously after launching an attack at different iteration. Here, clients conduct attacks as described in Section IV-A2 by utilising polluted data while training the local model, whereas the server uses our attack detection mechanism to identify these attacks.

Figure 3 displays the first two scenarios (i)-Figure 3a and (ii)-Figure 3b, respectively with Client X, who has X parameters in their local models, under single and continuous attack. In Figure 3a, all of the clients only attack once at the third iteration. When they start attacking, their reputation score plummets dramatically. In both scenarios, we observe the client who has more parameters has a larger relative decline in reputation score. This is also compatible with Corollary 1 in the Section III-C, that is, increasing the number of parameters N in the global model results in a lower error rate.

Figure 4 shows the last two scenarios (iii) and (iv) respectively with clients, who have 20 parameters in their local models, under single and continuous attack. In Figure 4a,

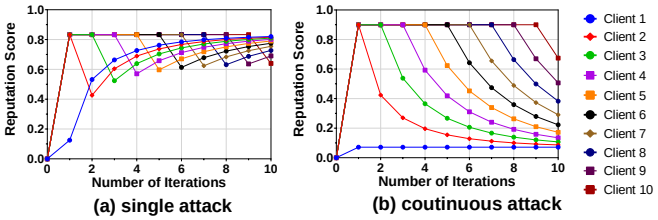


Fig. 4: The decay of reputation score in Client X with same model parameters when they (a) attack once at X iteration and (b) attack continuously after starting to attack at X iteration.

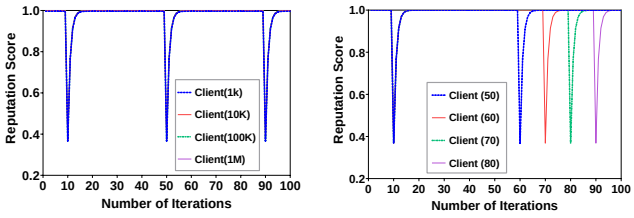


Fig. 5: The decay of reputation score in (left) Client with X model parameters when they attack at 10, 50 and 90 iteration; (right) Client X with 1 million parameters when they attack at 10 and $10 + X$ iteration.

Client X only launches an attack at X iteration. We observe that only one attack would lead to at least a 25.11% relative decrease in reputation score. In Figure 4b, Client X launches an attack at X iteration and keeps attacking in the following iterations. We observe in the end that 80% of their reputation scores are below 0.5, which is approximately half of the reputation score of honest clients, implying that the damage that they can inflict throughout the aggregation process is considerably decreased.

In addition, we consider a scenario in which an attacker spreads out the poisoning over a longer time duration, while using a higher number of model parameters. Figure 5 (left) depicts an attack over 40 iterations under different parameter sizes. Figure 5 (right) depicts an attack with 1 million parameters repeating every 50 to 80 iterations. These figures show that even if attackers spread our poisoning over multiple iterations and then try to recover their reputation score by acting benignly, our detection scheme can still identify them. This is because our attack detection and reputation schemes work in sequence. The attack detection scheme detects malicious updates without considering any reputation scores and rectifies them to mitigate damage. Then, the reputation scheme modifies the reputation scores based on the detection results. Also, attackers that employ a higher number of model parameters suffer a slightly higher reduction of reputation, which is consistent with Corollary 1.

3) Aggregation Algorithm: Algorithm 2 explains our aggregation method based on the attack detection scheme and subjective logic reputation model. First, the server sends all clients the pre-trained global model with initial parameters. Then, using their own data samples, clients train the global model locally and send the trained parameters back to the server. At this point, the server executes the attack detection scheme. In round t, if the n-th update parameter $\mathbf{w}_{i,n}^t$ from

Algorithm 2: Aggregation Algorithm

Server :

Input : $w^0 \leftarrow$ Pretrained Model
κ,η,a,W,c,$s \leftarrow$ Reputation parameters

Output: Global model M_{global} with w^T

for *Iteration* $t \leftarrow 1$ **to** T **do**
  // Broadcast global model to clients
  send(w^{t-1});
  // Wait until all updates arrive
  receive(w^t);
  // Rescale parameters by Algorithm 1
  $\bar{w}^t \leftarrow$ `Rescale`(w^t);
  for $n \leftarrow 1$ **to** N **do**
    for $i \leftarrow 1$ **to** M **do**
      // Compute parameter confidence
      $s^t_{i,n} = Eq\ 4(\bar{w}^t_{i,n})$;
      // Rectify abnormal parameters
      $w^t_{i,n} := Eq\ 5\ (s^t_{i,n}, \delta)$;
      record (P^t_i, N^t_i);
    end for
  end for
  for $i \leftarrow 1$ **to** M **do**
    // Calculate reputation score
    $\tilde{R}^t_i = Eq\ 12(P^t_i,\ N^t_i,\ \kappa,\ \eta,\ a,\ W,\ c,\ s)$;
  end for
  // Normalisation
  $\bar{R}^t \leftarrow$ `Norm`$(\tilde{R}^t)$;
  for $n \leftarrow 1$ **to** N **do**
    // Update the parameters
    $w^t_n := \sum_{i=1}^{M} \frac{\bar{R}^t_i}{\sum_{i=1}^{M} \bar{R}^t_i} w^t_{i,n}$;
  end for
  // Obtain parameters for global model
  $w^t := [w^t_1, \cdots, w^t_n]$;
end for

Client :

for *Client* $i \leftarrow 1$ **to** M **do in parallel**
  receive(w^{t-1});
  // Train local model
  $w^t_i \leftarrow w^{t-1}_i - r\frac{\partial \ell_i(w^{t-1}_i)}{\partial w}$;
  send(w^t_i);
end forpar

the client i has been rectified by the attack detection scheme in Section III-B1 to the median value, the server regards it as a negative observation, whereas no rectification represents a positive observation. Then, the server punishes the negative observation by reducing the corresponding client's reputation. Both types of observations are accumulated through all the N parameters of client i to obtain the reputation value $\tilde{R}^t_i$ in t round for client i so as to all the other clients. The server would conduct Min-Max normalisation to obtain $\bar{R}^t$ after receiving the reputation values $\tilde{R}^t$ of all the clients in t round.

After the server gets correction updates and the normalised reputation of each client, it aggregates the updates using average weighted reputation as the weights to get our global model updates for the current iteration. In this way, even over many training rounds, the attackers are still incapable of shifting parameters notably from the target direction and this ensures the quality of the resulting global model as will be demonstrated experimentally and analytically next.

C. Theoretical Guarantees

We prove the convergence of our reputation-based aggregation method. Our major results are Theorem 1 and Corollary 1, which state that convergence is guaranteed in bounded time. Regarding the performance of our algorithm in terms of metric average accuracy and convergence, we show that it is consistent with our theoretical analysis. We start by stating our assumptions, which are standard and common for such types of results, and per recent works such as [7], [20].

Assumption 1 (Smoothness)**.** *The loss functions are L-smooth, which means they are continuously differentiable and their gradients are Lipschitz-continuous with Lipschitz constant $L > 0$, whereas:*

$$\forall i \in N,\ \forall \mathbf{w}_1, \mathbf{w}_2 \in \mathbf{R}^d$$
$$\|\nabla \mathcal{L}(\mathbf{w}_1)) - \nabla \mathcal{L}(\mathbf{w}_2))\|_2 \leq L \|\mathbf{w}_1 - \mathbf{w}_2\|_2$$
$$\|\nabla \ell(\mathbf{w}_1; \mathcal{D}) - \nabla \ell(\mathbf{w}_2; \mathcal{D})\|_2 \leq L \|\mathbf{w}_1 - \mathbf{w}_2\|_2$$

Assumption 2 (Bounded Gradient)**.** *The expected square norm of gradients $\sqsupseteq$ is bounded:*

$$\forall \mathbf{w} \in \mathbf{R}^d, \exists \mathcal{G}_{\mathbf{w}} < \infty, \mathbb{E}\|\nabla \ell(\mathbf{w}; \mathcal{D})\|_2^2 \leq \mathcal{G}_{\mathbf{w}}$$

Assumption 3 (Bounded Variance)**.** *The variance of gradients $\mathbf{w}$ is bounded:*

$$\forall \mathbf{w} \in \mathbf{R}^d, \exists \mathcal{V}_{\mathbf{w}} < \infty, \mathbb{E}\|\nabla \ell(\mathbf{w}; \mathcal{D}) - \mathbb{E}(\nabla \ell(\mathbf{w}; \mathcal{D})\|_2^2 \leq \mathcal{V}_{\mathbf{w}}$$

Assumption 4 (Convexity)**.** *The loss function $\mathcal{L}(\sqsupseteq)$ are μ-strongly convex:*

$$\exists \mu > 0, \forall \mathbf{w}_1, \mathbf{w}_2 \in \mathbf{R}^d, \nabla \mathcal{L}(\mathbf{w}^*) = 0$$
$$\mathcal{L}(\mathbf{w}_1) - \mathcal{L}(\mathbf{w}_2) \geq \langle \nabla \mathcal{L}(\mathbf{w}_2), \mathbf{w}_1 - \mathbf{w}_2 \rangle + \frac{\mu}{2}\|\mathbf{w}_1 - \mathbf{w}_2\|_2^2$$

Suppose the percentage of attackers in the whole clients is p, and all the clients in the system participant every training iteration. r is the learning rate($r < \frac{1}{L}$) and $\hat{Q} = \max\{Q_i\}_{i=1}^M$. $\forall \mathbf{w} \in \mathcal{W}$, we denote

$$\mathbf{m}_i(\mathbf{w}^t) = \begin{cases} * & \text{if } i \in \text{malicious clients} \\ \nabla l_i(\mathbf{w}^t; \mathcal{D}) & \text{if } i \in \text{honest clients} \end{cases}$$

where $*$ stands for an arbitrary value from the malicious clients.

$$\mathbf{m}(\mathbf{w}^t) = \sum_{i=1}^{M} \bar{R}_i \mathbf{m}_i(\mathbf{w}^t)$$
$$s.t.\ \bar{R}_i = \frac{\tilde{R}^t_i}{\sum_{i=1}^{M} \tilde{R}^t_i},\ \sum_{i=1}^{M} \bar{R}_i = 1,\ \bar{R}_i \in (0, 1)$$

Consider the assumptions above and lemmas presented in Appendix A, we have

Theorem 1. *Under Assumptions 1, 2, 3 and 4, $\exists \epsilon > 0$ that:*

$$\sqrt{\frac{d \log(1 + \hat{Q} M L D_\epsilon)}{M(1-p)}} + C\frac{\mathcal{G}_{\mathbf{w}}}{\sqrt{\hat{Q}}} + p \leq \frac{1}{2} - \epsilon \tag{13}$$

After t rounds, Algorithm 2 converges with probability at least $1 - \xi \in \left[1 - \frac{4d}{(1+\hat{Q}MLv)^d}, 1\right)$ as

$$\left\|\mathbf{w}^t - \mathbf{w}^*\right\|_2 \leq (1 - Lr)^t \left\|\mathbf{w}^0 - \mathbf{w}^*\right\|_2 + \frac{\sqrt{N}}{L}\Delta_1 + \frac{1}{L}\Delta_2 \tag{14}$$

where

$$\Delta_1 = \frac{M\left(\varpi(M-1) + \frac{2E}{\sqrt{M\delta}}\right)}{\frac{Wa(M-1)(\kappa N+W)}{(\eta N+W)(\kappa N+Wa)} + 1}$$

$$\Delta_2 = 2\sqrt{2}\frac{1}{M\hat{Q}} + \sqrt{\frac{2}{\hat{Q}}}D_\epsilon V_w \left(\sqrt{\frac{d\log(1+\hat{Q}ML\upsilon)}{M(1-p)}} + C\frac{\mathcal{G}_w}{\sqrt{\hat{Q}}} + p\right)$$

$$D_\epsilon := \sqrt{2\pi}\exp\left(\frac{1}{2}\left(\Phi(1-\epsilon)\right)^2\right)$$

with $\Phi(\cdot)$ being the cumulative distribution function of Wald distribution.

Corollary 1. *Continuing with Theorem 1, when the iterations satisfy $t \geq \frac{1}{Lr}\log\left(\frac{L}{\sqrt{N}\Delta_1+\Delta_2}\left\|\mathbf{w}^0 - \mathbf{w}^*\right\|_2\right)$, $\exists\xi \in \left(0, \frac{4d}{(1+\hat{Q}ML\upsilon)^d}\right]$, we have:*

$$\mathbb{P}\left(\left\|\mathbf{w}^t - \mathbf{w}^*\right\|_2 \leq \frac{2\sqrt{N}}{L}\Delta_1 + \frac{2}{L}\Delta_2\right) \geq 1 - \xi$$

Remark 1. *Due to*

$$\Delta_1 := \mathcal{O}\left(\frac{\varpi}{a\kappa WN} + \frac{1}{\kappa N} + \frac{1}{\sqrt{MN\delta}}\right)$$

and

$$\Delta_2 := \mathcal{O}\left(\frac{1}{\hat{Q}} + \frac{p}{\sqrt{\hat{Q}}} + \frac{1}{\sqrt{\hat{Q}M}}\right)$$

Based on Corollary 1, we achieve an error rate:

$$\mathcal{O}\left(\frac{\varpi}{a\kappa W\sqrt{N}} + \frac{1}{\kappa\sqrt{N}} + \frac{1}{\sqrt{M\delta}} + \frac{1}{\hat{Q}} + \frac{p}{\sqrt{\hat{Q}}} + \frac{1}{\sqrt{\hat{Q}M}}\right)$$

we observe the experimental results in Figure 3 and 11 of Sections III and IV respectively, when varying the parameters of N, p, a and κ, results are consistent with this error rate.

Remark 2. *Derived from the Corollary 1 and Remark 1, there is a trade-off problem between convergence speed and error rate according to the level of reward κ and punishment η from the reputation model. This trade-off problem is mainly based on the fact that if the model penalises the bad behaviours of clients heavily, it would decrease their reputation dramatically so the model would take a longer time to converge. On the other hand, mitigating the punishment to increase the reward, would lead to an increase in the error rate.*

IV. Performance Evaluation

The objectives of our experimental evaluation are the following: (a) evaluate the performance of our aggregation method against other state-of-art robust aggregation methods, (b) benchmark it in three different scenarios, namely, no attack, label flipping attack, and backdoor attack, (c) do so using a text based real-world dataset of sensitive categories from [4] to which we will henceforth refer to as SURL, and finally (d) show that our experimental result are consistent with our previous theoretical analysis.

A. Experimental Setup

1) Datasets: The SURL dataset comes from a crowdsourcing taxonomy in the Curlie project [32], containing six categories of URLs: five sensitive categories (Health, Politics, Religion, Sexual Orientation, Ethnicity) and one for non-sensitive URLs, with a total of 442,190 webpages. The number of URLs in sensitive and non-sensitive categories are equally balanced. Each sample contains content, metadata and a class label of the webpage. For the SURL text classification task, we train a neural network with three fully connected layers and a final softmax output layer, same as in the evaluated methods [16], [20]. Furthermore, in order to fulfil the fundamental setting of an heterogeneous and unbalanced dataset for FL, we sample u_k from a Dirichlet distribution [18] with the concentration parameter $\iota = 0.9$ as in [12], which controls the imbalance level of the dataset, then assigns a $u_{k,i}$ fraction of samples in class k to client i, with the intention of generating non-IID and unbalanced data partitions. As a sanity check, we also tested our reputation scheme on a different classification task involving images and got consistent results as those we got for sensitive content (see Appendix C).

2) Threat Model: We consider the following threat model.
Attack capability: In the FL setting, the malicious clients have complete control over their local training data, training process and training hyper-parameters, e.g., the learning rate, iterations and batch size. They can pollute the training data as well as the parameters of the trained model before submitting it to the server but cannot impact the training process of other clients. We follow the common practice in the computer security field of overrating the attacker's capability rather than underrating it, so we limit our analysis to worst-case scenarios. There, an attacker has perfect knowledge about the learning algorithm, the loss function, the training data and is able to inspect the global model parameters. However, attackers would still have to train with the model published by the server, thus complying with the prescribed training scheme by FL to their local data. Furthermore, the percentage of byzantine clients p is an important factor that determines the level of success for the attack. We assume that the number of attackers is less than the number of honest clients, which is a common setting in similar methods [16], [20] to the ones we evaluate and compare our method with.
Attack strategy: We focus on two common attack strategies for sensitive context classification, namely, (i) label flipping attack [24] and (ii) backdoor attack [12]. Comparing to other attacks, for example model poisoning attack [29], [63], these two data poisoning attacks are more likely to be carried out by real users in the real world via our browser extension described in Section V, since polluting data is easier than manipulating model updates using the browser extension. Note that privacy attacks including membership inference attack [65] and property inference attack [64], are out of the scope of this paper, but form part of our ongoing and future work.

In a label flipping attack, the attacker flips the labels of training samples to a targeted label and trains the model accordingly. In our case, the attacker changes the label of "Health" to "Non-sensitive". In a backdoor attack, attackers inject a designed pattern into their local data and train these manipulated data with clean data, in order to develop a local model that learns to recognise such pattern. We realise

backdoor attacks inserting the top 10 frequent words with their frequencies for the "Health" category. Therein the backdoor targets are the labels "non-sensitive". A successful backdoor attack would acquire a global model that predicts the backdoor target label for data along with specific patterns.

For both attacks, instead of a single-shot attack where an attacker only attacks in one round during the training, we enhance the attacker by a repeated attack schedule in which an attacker submits the malicious updates in every round of the training process. Also, we evaluate a looping attack where the attackers spreads out poisoning every 30 epochs for the label flipping attack based on Figure 5. It is important to note that even if the attackers have full knowledge of our method, they would still be unable to mount smarter attacks that would try to maximise the damage caused while minimising their reputation drop. This is because the attackers are unaware and cannot compute their reputation score since the latter is computed at the server and requires input from all clients. Moreover, we allow extra training epochs for an attacker, namely, being able to train the local models with 5 more epochs as in [12].

3) Evaluated Aggregation Methods: We compare the performance of our aggregation method against the existing state-of-the-art in the area FedAvg [5], as well as against popular robust aggregation methods such as Coordinate-wise median [20], Trimmed-mean [20], FoolsGold [8], [15], Residual-based re-weighting [16], and FLTrust [60].

FedAvg is a FL aggregation method that demonstrates impressive empirical performance in non-adversarial settings [5]. Nevertheless, even a single adversarial client could control the global model in FedAvg easily [27]. This method averages local model updates of clients as a global model update weighted by the fraction of training samples size of each client compared to total training samples size. We use it as baseline evaluation to assess the performance of our method.
Median is using coordinate-wise median for aggregation. After receiving the updates in round t, the global update is set equal to the coordinate-wise median of the updates, where the median is the 1-dimensional median.
Trimmed-mean is another coordinate-wise mean aggregation technique that requires prior knowledge of the attacker fraction β, which should be less than half of the number of model parameters. For each model parameter, the server eliminates the highest and lowest β values from the updates before computing the aggregated mean with remaining values.
FoolsGold presents a strong defence against attacks in FL, based on a similarity metric. Such approach identifies attackers based on the similarity of the client updates and decreases the aggregate weights of participating parties that provide indistinguishable gradient updates frequently while keeping the weights of parties that offer distinct gradient updates. It is an effective defence for sybil attacks but it requires more iterations to converge to an acceptable accuracy.
Residual-based re-weighting weights each local model by accumulating the outcome of its residual-based parameter confidence multiplying the standard deviation of parameter based on the robust regression through all the parameters of this local model. In our reputation-based aggregation method, we implement the same re-weighting scheme IRLS [10] as residual-based aggregation, but choose the collection of reputation as the weights of clients' local models.
FLTrust establishes trust in the system by bootstrapping it via the server, instead of depending entirely on updates from clients, like the other methods do. The server obtains an initial server model trained on clean root data. Then, depending on the cosine similarity of the server model and each local model, it assigns a trust score to each client in each iteration.

4) Performance Metrics: We use the average accuracy (Avg-ACC) of the global model to evaluate the result of the aggregation defence for the poisoning attack in which attackers aim to mislead the global model during the testing phase. The accuracy is the percentage of testing examples with the correct predictions by the global model in the whole testing dataset, which is defined as:

$$\text{Avg-ACC} = \frac{\text{\# correct predictions}}{\text{\# testing samples}}$$

In addition, there is existence of targeted attacks that aim to attack a specific label while keeping the accuracy of classification on other labels unaltered. Therefore, instead of Avg-ACC, we choose the attack success rate (ASR) to measure how many of the samples that are attacked, are classified as the target label chosen by a malicious client, namely:

$$\text{ASR} = \frac{\text{\# successfully attacked samples}}{\text{\# attacked samples}}$$

A robust federated aggregation method would obtain higher Avg-ACC as well as a lower ASR under poisoning attacks. An ideal aggregation method can achieve 100% Avg-ACC and has the ASR as low as the fraction of attacked samples from the target label.

5) Evaluation Setup: For the malicious attack, we assume that 30% of the clients are malicious as in [27], which is also a common byzantine consensus threshold for resistance to failures in a typical distributed system [6]. For the server-side setting, in order to evaluate the reliability of the local model updates sent by the client to the server, we assume that the server has the ability to look into and verify the critical properties of the updates from the clients before aggregating.

Also, we only consider FL to be executed in a synchronous manner, as most existing FL defences require [7], [20], [27], [29]. For all the above aggregation methods under attack, we perform 100 iterations using the SURL dataset with a batch size of 64 and 10 clients. Furthermore, we evaluate our method for increasing numbers of clients. These settings are inline with existing state-of-the-art methods for security in FL [12], [16], [20] More details related to the training setting are presented in Appendix B.

B. Convergence and Accuracy

In Figure 6 (left), we analyse the performance of our method in the no attack scenario and compare the convergence and accuracy of our method with others during training. We show the training loss (left axis) and average accuracy (right axis) during 100 training iterations for 7 methods.

Our aggregation starts with the lowest training loss and maintains it throughout the training process. It only takes 24 iterations to achieve 82% accuracy and then converge to 82.13%, which represents a 2.7× faster converge rate than FedAvg. In comparison, Residual-based and Trimmed-mean

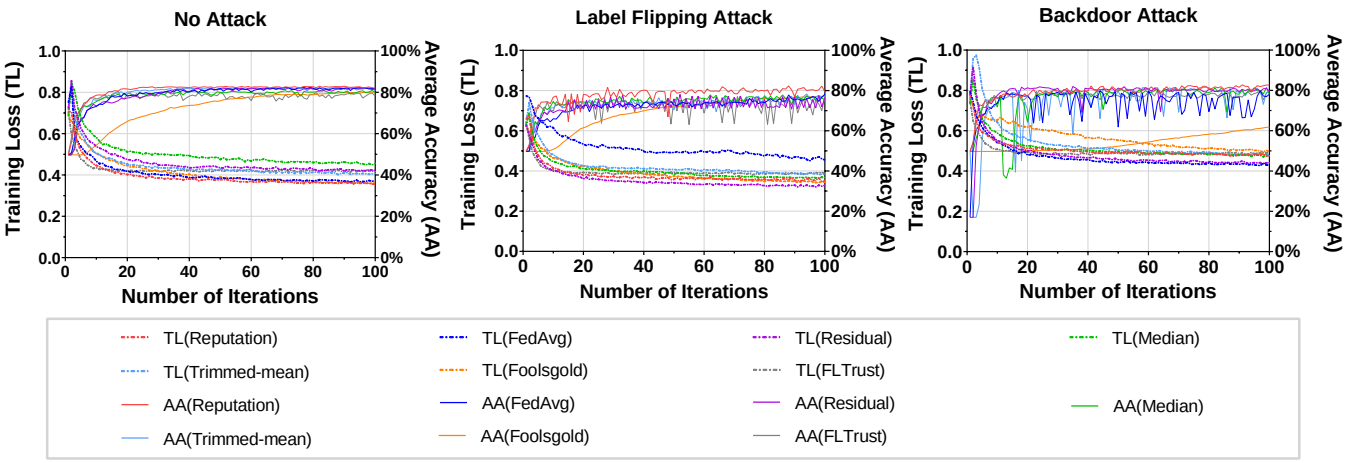


Fig. 6: Training Loss (TL) and Average Accuracy (AA) for 100 epochs of Reputation-based, FedAvg, Residual-based, Median, Trimmed-mean, FoolsGold and FLTrust methods in SURL Dataset under no attack (left) scenario, under label flipping attack (middle) and backdoor attack (right) scenarios with 30% malicious clients.

have almost identical training loss and take 52 and 47 iterations to reach 82% accuracy and practically converge to 81.79% and 80.76% respectively, which is 2.2× and 2× slower than our reputation method. Median reaches 81% at 83 rounds and after that converges to 79.94%, which amounts to a 3.6× slower converge rate than our method. Especially, Foolsgold and FLTrust are slow to converge and do not converge within 100 iterations, so our convergence rate is at least 4.2× better than FoolsGold and FLTrust. This demonstrates that our reputation model benefits from convergence speed and accuracy performance. This is because our reputation scheme assigns higher weight to more reliable clients when there is no ongoing attack, which generates more consistent updates thereby accelerating the convergence.

C. Resilience to Attacks

We begin by analysing the performance with a static percentage (30%) of attackers, and then move on to the performance with a varying percentage of attackers under label flipping and backdoor attacks.

1) Label Flipping Attack: **Static percentage of attackers:** Figure 6 (middle) shows the convergence of mentioned methods under label flipping attack. Our method converges 1.8× to 2× faster than all competing state-of-the-art methods under attack, enlarging its performance benefits compared to the no attack scenario. In addition, our method outperforms competing methods by at least 1.4% in terms of accuracy.

Varying the percentage of attackers: Here we analyse the impact on our aggregation method as the proportion of attackers increases. Figure 7 (left) shows the change of performance metrics for varying percentage of attackers for seven evaluated methods When the percentage of attackers p ranges from 10% to 50%, our method is resistant against label flipping attacks with a small loss in accuracy and a consistent attack success rate of all the methods. As p approaches 50%, FedAvg, Residual-based, Median and FLTrust defences become ineffective in mitigating the attack, and correspondingly their Avg-ACC decreases linearly. Moreover, under label flipping attack during the whole process, our reputation-based method has the highest accuracy outperforming other methods by 1% to 23.1%. At the same time it has the lowest ASR. The average ASR of other methods are at least 82.8% higher than ours.

2) Backdoor Attack: **Static percentage of attackers:** Figure 6 (right) shows the convergence of mentioned methods under backdoor attack. Same as in no attack and label flipping attack scenario, our method converges 1.6× to 2.4× faster than all competing state-of-the-art methods. In addition, our method outperforms competing methods by 3.5% to 33.6% in terms of classification accuracy.

Varied percentage of attackers: We examine the scenario in which the percentage of attackers increases. Figure 7 (right) shows the performance for the seven evaluated methods under backdoor attack when varying the percentage of attackers p from 10% to 50%. Figure 7 (right) demonstrates that under backdoor attack, our reputation-based method has a consistent accuracy throughout the process with the lowest attack success rate, whereas the average ASR of other methods is at least 72.3% higher than ours. As p changes, the ASR of the Residual-based, Median, and Foolsgold methods increase linearly. Although FLTrust has a stable ASR, it increases by a factor of 1.39 when p reaches 50%.

First, we evaluate a varying compromise rate for the label flipping and backdoor attacks using our reputation-based method. For the label-flipping attack, we vary the percentage of the flipped label poisoned by attackers from 10% to 90%. Also, for the backdoor attack, we vary the number of top frequent words inserted as the trigger pattern, from 5 to 25. The remaining settings are the same as in previous experiments. Figure 8 plots the ACC and ASR when varying the compromise rate for both attacks. Figure 8 (left) shows that when the percentage of the poisoned sample is increased, it leads to the decrease of the accuracy of the model and to a slight increase of ASR. Figure 8 (right) shows that when we increase the number of frequent words from 5 to 20, the ASR remains unaffected. When the frequent words exceed 25, the attack becomes less stealthy and thus can be more easily detected resulting in a lower ASR.

We also evaluate the performance of our method in terms of the number of participating clients. With a 30% compromise rate, we expand the number of clients from 10 to 200. Our

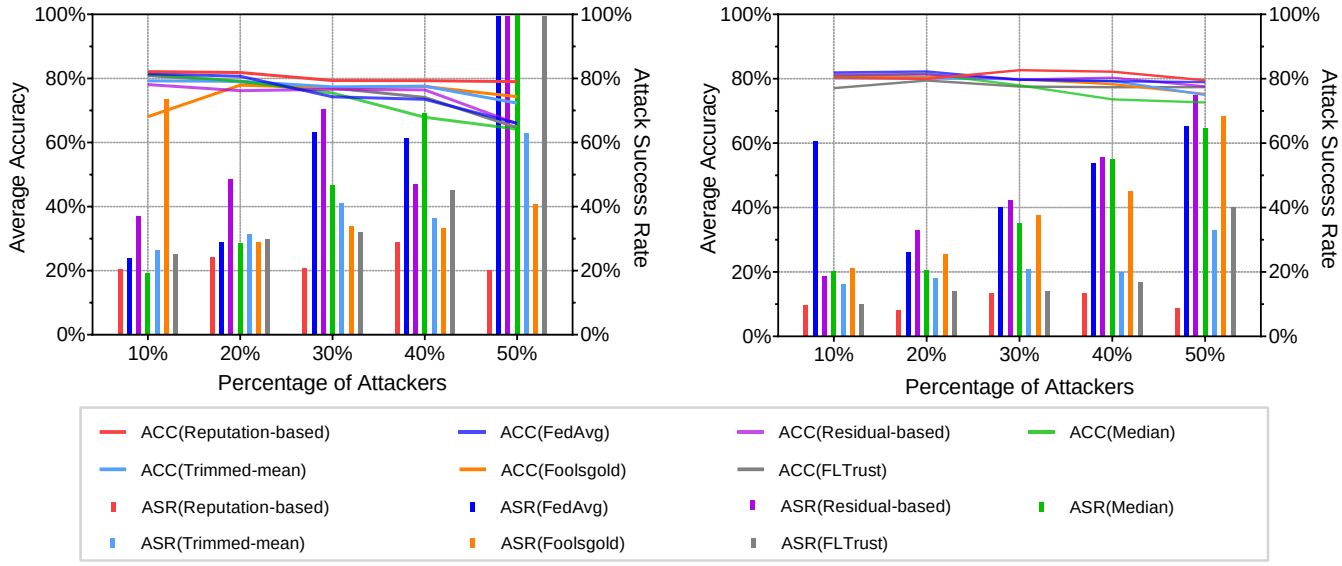


Fig. 7: Average accuracy (ACC) and attack success rate (ASR) for varying percentage of attackers from 10% to 50% under label flipping (left) and backdoor (right) attack for Reputation-based, FedAvg, Residual-based, Median, Trimmed-mean, FoolsGold and FLTrust methods in SURL Dataset.

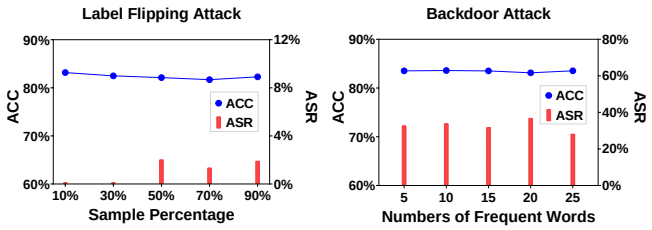


Fig. 8: ACC and ASR as we vary the percentage of flipped label from 10% to 90% (left) for label flipping attack, and the number of the frequent words as the trigger pattern from 5 to 25 for backdoor attack(right).

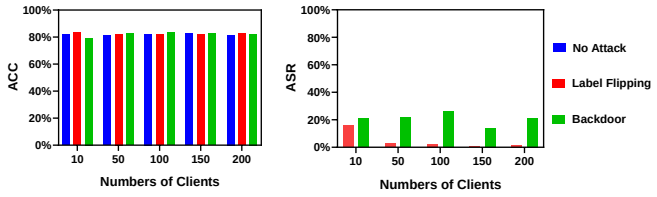


Fig. 9: Average accuracy (ACC) and attack success rate (ASR) for varying the number of clients from 10 to 200 under label flipping and backdoor attack with 30% malicious clients.

method performs consistently for a larger number of clients, as seen by the stable ACC and ASR as the number of clients grows in Figure 9.

3) Analysis of Attacks: Finally, instead of repeating the attack at every epoch, the attacker stretches poisoning across 30 epochs in our study of the looping attack. The performance of the looping attack is seen in Figure 10. As expected, the looping attack is not as effective as the repeated attack that we previously assessed. All the methods manage to defend it with low ASR, and our method still has the greatest accuracy.

4) Evaluation Results: In the no attack scenario, we observe (i) Our method converges $2\times$ to $4.2\times$ faster than all competing state-of-the-art methods. (ii) Our method is at least as good or outperforms competing methods in terms of classification accuracy. The above validates that our reputation scheme is helpful even in the no attack scenario. This is due to the fact that in our algorithm we give higher weights to the clients with high-quality updates, as illustrated in Figure 12, causing the model to converge rapidly and retain consistent accuracy. In addition, even under the two different attacks, our method:

- converges $1.6\times$ to $2.4\times$ faster than all competing state-of-the-art methods.
- provides the same or better accuracy than competing methods.
- yields the lowest ASR compared to all other methods, with the average ASR of them being at least 72.3% higher than ours.

We obtained comparable findings for the evaluation of the aforementioned methods on 100 clients, as presented in Appendix D. Furthermore, the result is consistent with the theoretical analysis: as p increases, so does the error rate.

D. Stability of Hyper-parameters

We employ four hyper-parameters in our reputation model: rewarding weight κ, prior probability a, time decay parameter c and window length s. As shown in Remark 1, c and s do not affect the performance of our model, we only consider hyper-parameters κ and a, where κ controls the reward weight to positive observations and a controls the fraction of uncertainty converted to belief. To demonstrate the impact of these two hyper-parameters of our reputation model, we grid search κ in $[0.1, 0.2, 0.3, 0.4]$ and a in $[0.1, 0.3, 0.5, 0.7, 0.9]$. The setup is the same as on SURL dataset under label flipping attack. The ultimate accuracy of stability of reputation-based aggregation

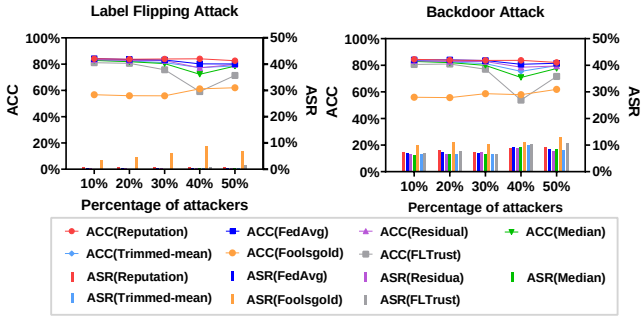
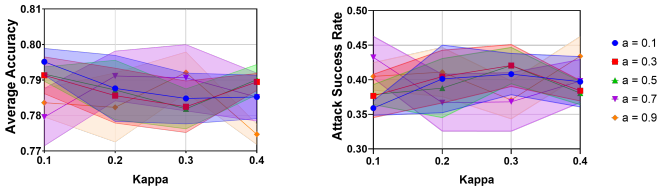


Fig. 10: Average accuracy (ACC) and attack success rate (ASR) for varying percentage of attackers from 10% to 50% under label flipping (left) and backdoor (right) attack with a looping attack in which an attacker attacks every 30 epochs.



(a) Accuracy (b) Attack Success Rate

Fig. 11: Average accuracy and attack success rate as we vary rewarding weight κ and prior probability a.

are shown in Figure 11. Note that these results are tested for the label flipping attack and they hold according to theory also for backdoor.

The result in Figure 11 demonstrates that our approach is very stable and efficient in terms of hyper-parameter selection, and it achieves a high degree of precision. Furthermore, the result is compatible with the theoretical analysis in Section III-C.

E. Comparison against a residual-based method

To demonstrate how our method improves the residual-base method by assigning the aggregation weights based on reputation, we consider a scenario with 10 clients in the FL system, 8 of which are malicious. The training lasts 10 communication rounds during which attackers carry out the backdoor attack. The remaining settings are the same as the default. Results are shown in Figure 12, in which the first two clients are benign, and the rest are malicious. We observe that for our reputation method the aggregation weights of malicious clients, which are their reputations, are rectified to 0 since the second round, demonstrating that our method is successful in eradicating their influence. On the other hand, the aggregate weights of malicious clients in residual-based methods, which are calculated by multiplying the parameter confidence by its standard deviation, are nearly similar and non-zero. This is because repeated median regression seldom yields 0 for the parameter confidence, which causes practically non-zero weights to be assigned to malicious clients by residual-based methods. To address this issue, the reputation model uses positive and negative observations that introduce rewards and punishments to assign divergent weights to clients. As a result, benign clients are given higher weights whereas malicious clients are eliminated from the aggregation.

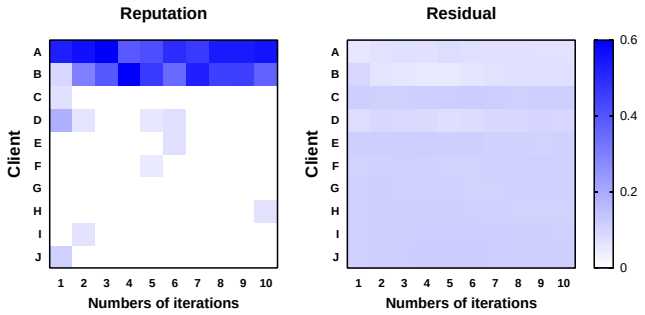


Fig. 12: The aggregation weights of clients from our reputation-based (left) and residual-based method (right) for 10 communication rounds under label flipping attack with 80% attackers.

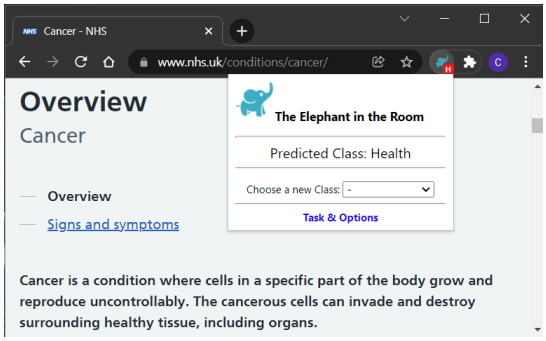


Fig. 13: *EITR* extension in action. The letter "H" inside the red frame at the bottom right of the extension's icon indicates that a health-related page has been detected.

V. The *EITR* System

In this section, we provide a high level description of our *EITR* [67] system (standing for "Elephant In the Room" of privacy). We then present some preliminary results with real users demonstrating the ability of the system to quickly learn how to classify yet unseen sensitive content, in our case COVID-19 URLs pertaining to the category Health, even in view of inaccurate user input. The system is currently being used as a research prototype to evaluate the robustness of our algorithm in a simple real-world setting. A full in depth description of the system and its performance with more users and more intricate settings, including adoption, incentives, and HCI issues, over a longer time period is the topic of our ongoing efforts and will be covered by our future work.

A. System Architecture and Implementation

The *EITR* system is based on the client-server model. The back-end server is responsible to distribute the initial classification model and the consequent updated model(s) to the clients and receive new annotations from the different clients of the system. The client is in the form of a web browser extension that is responsible to fetch and load the most recent global classification model to the users' browser from the back-end server. The loaded model can then be used to label website in real time into the 5 different sensitive topics as defined by GDPR, i.e., Religion, Health, Politics, Ethnicity and Sexual Orientation. Next, we provide more details for each part of

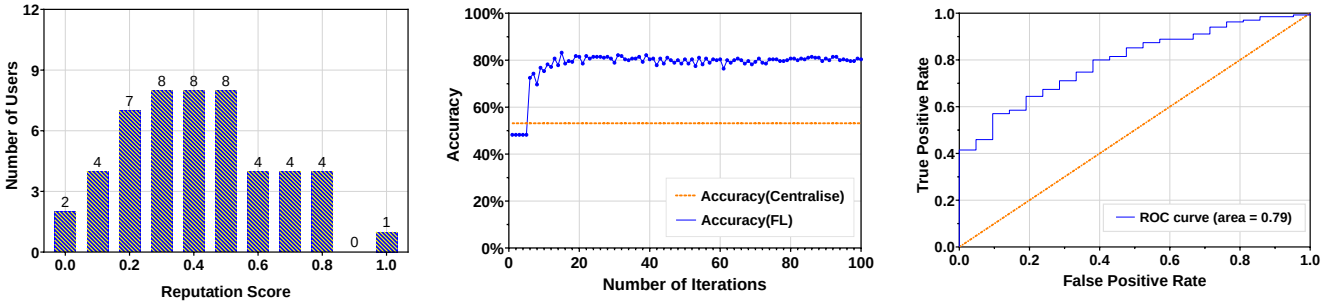


Fig. 14: Results of real-user experiment for COVID-19 related URLs with 50 users over 100 iterations. (left): the reputation score of the real users at the end of experiment, (middle): the accuracy of the centralised classifier and the global model of our reputation-based FL approach for COVID-19 URLs, and (right): the ROC curve of real-user experiment with 0.79 area under the ROC curve (AUC).

the system.

Back-end server: The back-end server is written in JavaScript using the node.js Express [40] framework. To build the initial classification model we use the dataset provided by Matic et al. [4], and the TensorFlow [41] and Keras [42] machine learning libraries. The final trained model is then exported using the TensorFlow.js [43] library in order to be able to distribute it to the system's clients (browser extension). The back-end server also includes additional functionalities such as the creation and distribution of users' tasks, i.e., a short list of URLs that the users need to visit and annotate, and an entry point that collects the resulting users annotations during the execution of the task.

Web browser extension: Currently the browser extension only supports the Google Chrome browser and is implemented in JavaScript using the Google Chrome Extension APIs [45]. To handle the classification model the extension utilises the TensorFlow.js [43] library to load, use, and update the model. The main functionality of the extension is to classify the visited website in real time and provide information to the user related to the predicted class as depicted in Figure 13. The website classification is based on the metadata (included in the website `<head>` HTML tag) and the visible text of the website. The extension also allows users to provide their input related to their agreement or disagreement with the predicted class using a drop down list as depicted in Figure 13 with the label *"Choose a new Class"*.

B. Real Users Experimental Setup

The goal of the real-user experiment is to evaluate our federated reputation-based method on real user activity (instead of systematic tests), and demonstrate that even with real users with different comprehensions of the definition of sensitive information, our method can learn new content fast and achieve higher accuracy than centralised classifiers, which is compatible with our simulation experiment.

For the setup, the participant is directed to visit the experiment website that provides the necessary information and instructions on the goal of our study, the definition of sensitive information provided by the current GDPR and how to participate in our study. In order to have access to the browser extension and the installation instructions the user must in advance give explicit consent and accept the data privacy policy. Upon successful installation of the extension, new users are asked to provide a valid email address (to contact them for the reward) and then receive their task, a list of 20 URLs, that they need to visit and provide their labels in order to successfully complete their task. The list of 20 URLs is sampled by the Dirichlet distribution with $\iota = 0.9$ for each participant from a database, which includes 300 URLs with sensitive and non-sensitive content related to COVID-19.

Ethical Considerations: We have ensured compliance with the GDPR pertaining to collecting, handling, and storing data generated by real users. To that end, we have acquired all the proper approvals from our institutions. Furthermore, the participants are directed to visit a pre-selected set of URLs selected by us to avoid collecting the actual visiting patterns of our users. In addition, the user input is only collected *if and only if* the user explicitly provide input to the drop-down list labelled *"Choose a new Class"* to avoid collecting the visiting patterns of the user accidentally while they are executing their tasks. Finally, we only use the users' email address to contact them for the reward. The mapping between the user input and their email address is based on a random identifier that is generated during the installation time of the extension.

C. Validation with Real Users

Data collection: We had 50 users participating in our experiment. In order to evaluate our reputation-based FL method using real-user data, we define a methodology to label ground truth on COVID-19 related sites.

Ground truth methodology: To set the ground truth for COVID-19 sites related to our sensitive or non-sensitive labels, we create a database of 100 websites, which we collect by searching on Google with the query "sensitive websites about COVID-19" and choose the top 100 sites returned from the query. Then, four experts in the privacy field, independently annotate them based on their professional expertise in order to achieve an agreement on whether each of those sites included sensitive or non-sensitive content.

Ground truth annotation: In order to evaluate the annotation of the 100 websites from human experts, we calculate the inter-rater agreement among them using Fleiss Kappa [61]. We obtain 0.56 of Fleiss Kappa, which is an acceptable agreement because the values of Fleiss Kappa. above 0.5 are regarded as good. Furthermore, given that COVID-19 is a controversial issue, it is difficult for humans to agree on what constitutes sensitive content relating to it. Even though, we still attain a

valid ground truth of 85 items belonging to the health sensitive category with agreement ratings of at least 0.5. Note that we also classify the above 100 websites using the centralised classifier proposed in [4] and get only 53.13% accuracy.

Result with real users: Figure 14 shows the results of accuracy and reputation score with 50 real users in the experiment. Figure 14 (left) shows that the majority of users have reputation scores falling in the intermediate range, with some having a very high reputation and a few having a very low reputation. This indicates the divergence of the user's interpretation of the sensitive information as we expect. In Figure 14 (middle) we compare the accuracy of the centralised classifier and the global model of our reputation-based FL approach for COVID-19 URLs. Despite the diversity of reputation scores of real users, our method converges as rapidly as in simulation and achieves an average accuracy of 80.36%, thereby verifying the quick convergence and high accuracy results presented in the previous sections. Figure 14 (right) shows that the ROC curve in real-user experiment yielded 0.79 AUC. Our result is acceptable in this scenario because most existing FL techniques are designed to minimise the conventional cost function and are not optimal for optimising more appropriate metrics for imbalanced data, such as AUC [62].

As we observe, with real users holding our method achieve a good performance. This means that, as new sensitive content appears and/or is defined by GDPR or new upcoming legislation, we will be able to continue training our FL model for this type of task with quick convergence and good accuracy. The empirical results in Figure 14 (middle) also shows that there is a quick convergence to the accuracy's stable value within a small number of iterations (around 30), in line with the theoretical results in Section III.

VI. Conclusion

In this paper we have shown how to use federated learning to implement a robust to poisoning attacks distributed classifier for sensitive web content. Having demonstrated the benefits of our approach in terms of convergence rate and accuracy against state-of-the-art approaches, we implemented and validated it with real users using our *EITR* browser extension. Collectively, our performance evaluation has showed that our reputation-based approach to thwarting poisoning attacks consistently converges faster than the state-of-the-art while maintaining or improving the classification accuracy.

We are currently working towards disseminating *EITR* to a larger user-base and using it to classify additional sensitive and non-sensitive types of content. This includes but it is not limited to categories defined by the users themselves for different purposes, not necessarily related to sensitive content, as well as evaluating additional attacks and threat models under our subjective-logic reputation scheme for FL. In turn, our approach can support other FL models going beyond sensitive content classification in future work.

Acknowledgements

We thank the EITR volunteers for their help. This work and dissemination efforts were supported in part by the European Commission under DataBri-X project (101070069), the TV-HGGs project (OPPORTUNITY/0916/ERCCoG/0003) co-funded by the European Regional Development Fund and the Republic of Cyprus through the Research and Innovation Foundation, and the European Research Council (ERC) Starting Grant ResolutioNet (ERC-StG-679158).

References

[1] McMahan, H., Ramage, D., Talwar, K. & Zhang, L. Learning differentially private recurrent language models. *Proceedings of ICLR*. (2017)

[2] Mammen, P. Federated Learning: Opportunities and Challenges. *ArXiv Preprint ArXiv:2101.05428*. (2021)

[3] Jøsang, A. Subjective logic. (Springer,2016)

[4] Matic, S., Iordanou, C., Smaragdakis, G. & Laoutaris, N. Identifying Sensitive URLs at Web-Scale. *Proceedings of The ACM Internet Measurement Conference*. pp. 619-633 (2020)

[5] McMahan, B., Moore, E., Ramage, D., Hampson, S. & Arcas, B. Communication-Efficient Learning of Deep Networks from Decentralized Data. *Artificial Intelligence And Statistics*. pp. 1273-1282 (2017)

[6] Castro, M., Liskov, B. Practical Byzantine Fault Tolerance. *USENIX OSDI*. **99**, 173-186 (1999)

[7] Xie, C., Koyejo, S. & Gupta, I. Zeno: Distributed Stochastic Gradient Descent with Suspicion-based Fault-tolerance. *International Conference On Machine Learning*. pp. 6893-6901 (2019)

[8] Fung, C., Yoon, C. & Beschastnikh, I. The Limitations of Federated Learning in Sybil Settings. *23rd International Symposium On Research In Attacks, Intrusions And Defenses (RAID 2020)*. pp. 301-316 (2020,10), https://www.usenix.org/conference/raid2020/presentation/fung

[9] Josang, A., Hayward, R. & Pope, S. Trust Network Analysis with Subjective Logic. *Proceedings of The Twenty-Ninth Australasian Computer Science Conference (ACSW 2006)*. pp. 85-94 (2006)

[10] Wilcox, R. Introduction to robust estimation and hypothesis testing. (Academic press,2011)

[11] Krizhevsky, A., Hinton, G. & Others Learning Multiple Layers of Features from Tiny Images. University of Toronto Technical Report, 2009.

[12] Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D. & Shmatikov, V. How to Backdoor Federated Learning. *International Conference on Artificial Intelligence And Statistics*. pp. 2938-2948 (2020)

[13] Konečny, J., McMahan, H., Yu, F., Richtárik, P., Suresh, A. & Bacon, D. Federated Learning: Strategies for Improving Communication Efficiency. *Proceedings of NIPS*. (2016)

[14] Siegel, A. Robust Regression using Repeated Medians. *Biometrika*. **69**, 242-244 (1982)

[15] Fung, C., Yoon, C. & Beschastnikh, I. Mitigating Sybils in Federated Learning Poisoning. *ArXiv Preprint ArXiv:1808.04866*. (2018)

[16] Fu, S., Xie, C., Li, B. & Chen, Q. Attack-resistant Federated Learning with Residual-based Reweighting. *AAAI Workshops: Towards Robust, Secure And Efficient Machine Learning*. (2021)

[17] Sun, Z., Kairouz, P., Suresh, A. & McMahan, H. Can you Really Backdoor Federated Learning?. *The 2nd International Workshop on Federated Learning For Data Privacy And Confidentiality, in Neural Information Processing Systems*. (2019)

[18] Minka, T. Estimating a Dirichlet Distribution. (Technical report, MIT, 2000)

[19] He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition*. pp. 770-778 (2016)

[20] Yin, D., Chen, Y., Kannan, R. & Bartlett, P. Byzantine-robust Distributed Learning: Towards Optimal Statistical Rates. *International Conference On Machine Learning*. pp. 5650-5659 (2018)

[21] Minsker, S. Geometric Median and Robust Estimation in Banach Spaces. *Bernoulli*. **21**, 2308-2335 (2015)

[22] Karimireddy, S., He, L. & Jaggi, M. Learning from History for Byzantine Robust Optimization. *International Conference On Machine Learning*. pp. 5311-5319 (2021)

[23] Bubeck, S. Convex Optimization: Algorithms and Complexity. *Foundations and Trends in Machine Learning, 8 (3-4)*. (2014)

[24] Barreno, M., Nelson, B., Joseph, A. & Tygar, J. The Security of Machine Learning. *Machine Learning*. **81**, 121-148 (2010)

[25] Deng, J., Dong, W., Socher, R., Li, L., Li, K. & Fei-Fei, L. Imagenet: A Large-scale Hierarchical Image Database. *2009 IEEE Conference On Computer Vision And Pattern Recognition*. pp. 248-255 (2009)

[26] Bhagoji, A., Chakraborty, S., Mittal, P. & Calo, S. Analyzing Federated Learning through an Adversarial Lens. *International Conference On Machine Learning*. pp. 634-643 (2019)

[27] Blanchard, P., El Mhamdi, E., Guerraoui, R. & Stainer, J. Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent. *Proceedings Of The 31st International Conference On Neural Information Processing Systems*. pp. 118-128 (2017)

[28] Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L. & Lerer, A. Automatic Differentiation in Pytorch. (2017)

[29] Fang, M., Cao, X., Jia, J. & Gong, N. Local Model Poisoning Attacks to Byzantine-robust Federated Learning. *29th USENIX Security Symposium*. pp. 1605-1622 (2020)

[30] Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H., Patel, S., Ramage, D., Segal, A. & Seth, K. Practical Secure Aggregation for Privacy-preserving Machine Learning. *Proceedings of The 2017 ACM SIGSAC Conference On Computer And Communications Security*. pp. 1175-1191 (2017)

[31] El Mahdi, E. M., Guerraoui, R., Rouault, S. The Hidden Vulnerability of Distributed Learning in Byzantium. *International Conference On Machine Learning*. pp. 3521-3530 (2018)

[32] Curlie.org Curlie - The Collector of URLs. (2019), https://curlie.org/

[33] Commission, E. Data protection in the EU, The General Data Protection Regulation (GDPR); Regulation (EU) 2016/679. (2018), https://ec.europa.eu/%0Ainfo/law/law-topic/data-protection/

[34] California, S. California Consumer Privacy Act - Assembly Bill No. 375. (2018), https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180AB375

[35] Government of Canada. Amended Act on The Personal Information Protection and Electronic Documents Act. (2018), https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/

[36] Government of Israel. Protection of privacy regulations (data security) 5777-2017. (2018), https://www.gov.il/en/Departments/legalInfo/data_%0Asecurity_regulation/

[37] European Commission. Personal Information Protection Commission, J. Amended Act on the Protection of Personal Information. (2017), https://www.ppc.go.jp/en/

[38] Australian Information Commissioner. Australian Privacy Principles guidelines; Australian Privacy Principle 5 - Notification of the collection of personal information. (2018), https://www.oaic.gov.au/agencies-and-organisations/

[39] Kang, J., Xiong, Z., Niyato, D., Xie, S. & Zhang, J. Incentive Mechanism for Reliable Federated Learning: A Joint Optimization Approach to Combining Reputation and Contract Theory. *IEEE Internet Of Things Journal*. **6**, 10700-10714 (2019)

[40] Expressjs.com Express - Fast, unopinionated, minimalist web framework for Node.js. (2021), https://expressjs.com/

[41] TensorFlow.org TensorFlow - An end-to-end open source machine learning platform.. (2021), https://www.tensorflow.org/

[42] Keras.io Keras - Simple. Flexible. Powerful. (2021), https://keras.io/

[43] Tensorflow.org TensorFlow.js is a library for machine learning in JavaScript. (2021), https://www.tensorflow.org/js

[44] Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., Kiddon, C., Konečny, J., Mazzocchi, S., McMahan, H. Towards Federated Learning at Scale: System Design. *Proceedings of MLSys*. (2019)

[45] Google Google Chrome Extension APIs. (2021), https://developer.chrome.com/docs/extensions/reference/

[46] Hard, A., Rao, K., Mathews, R., Ramaswamy, S., Beaufays, F., Augenstein, S., Eichner, H., Kiddon, C. & Ramage, D. Federated Learning for Mobile Keyboard Prediction. *ArXiv Preprint ArXiv:1811.03604*. (2018)

[47] Feng, J., Rong, C., Sun, F., Guo, D. & Li, Y. PMF: A Privacy-preserving Human Mobility Prediction Framework via Federated Learning. *Proceedings of The ACM On Interactive, Mobile, Wearable And Ubiquitous Technologies*. **4**, 1-21 (2020)

[48] Yu, T., Li, T., Sun, Y., Nanda, S., Smith, V., Sekar, V. & Seshan, S. Learning Context-aware Policies from Multiple Smart Homes via Federated Multi-task Learning. *IEEE/ACM International Conference On Internet-of-Things Design And Implementation (IoTDI)*. pp. 104-115 (2020)

[49] Huang, L., Shea, A., Qian, H., Masurkar, A., Deng, H. & Liu, D. Patient Clustering Improves Efficiency of Federated Machine Learning to Predict Mortality and Hospital Stay Time using Distributed Electronic Medical Records. *Journal Of Biomedical Informatics*. **99** pp. 103291 (2019)

[50] Brisimi, T., Chen, R., Mela, T., Olshevsky, A., Paschalidis, I. & Shi, W. Federated Learning of Predictive Models from Federated Electronic Health Records. *International Journal Of Medical Informatics*. **112** pp. 59-67 (2018)

[51] Lee, J., Sun, J., Wang, F., Wang, S., Jun, C. & Jiang, X. Privacy-preserving Patient Similarity Learning in a Federated Environment: Development and Analysis. *JMIR Medical Informatics*. **6**, e7744 (2018)

[52] Ahn, E., Kumar, A., Feng, D., Fulham, M. & Kim, J. Unsupervised Deep Transfer Feature Learning for Medical Image Classification. *IEEE International Symposium On Biomedical Imaging (ISBI 2019)*. pp. 1915-1918 (2019)

[53] Lin, B., He, C., Zeng, Z., Wang, H., Huang, Y., Soltanolkotabi, M., Ren, X. & Avestimehr, S. FedNLP: Benchmarking Federated Learning Methods for Natural Language Processing Tasks. *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. (2022)

[54] Iordanou, C., Soriente, C., Sirivianos, M. & Laoutaris, N. Who is Fiddling with Prices? Building and Deploying a Watchdog Service for e-commerce. *Proceedings of ACM SIGCOMM* (2017)

[55] Iordanou, C., Kourtellis, N., Carrascosa, J., Soriente, C., Cuevas, R. & Laoutaris, N. Beyond Content Analysis: Detecting Targeted Ads via Distributed Counting. *Proceedings Of The 15th International Conference On Emerging Networking Experiments And Technologies*. pp. 110-122 (2019)

[56] Su, D., Cao, J., Li, N., Bertino, E. & Jin, H. Differentially Private k-means Clustering. *Proceedings Of The Sixth ACM Conference On Data And Application Security And Privacy*. pp. 26-37 (2016)

[57] Cormode, G. & Muthukrishnan, S. An Improved Data Stream Summary: the Count-Min Sketch and its Applications. *Journal Of Algorithms*. **55**, 58-75 (2005)

[58] Salton, G. & Buckley, C. Term-weighting Approaches in Automatic Text Retrieval. *Information Processing and Management*. **24**, 513-523 (1988)

[59] Lopuhaa, H. & Rousseeuw, P. Breakdown points of Affine Equivariant Estimators of Multivariate Location and Covariance Matrices. *The Annals Of Statistics*. pp. 229-248 (1991)

[60] Cao, X., Fang, M., Liu, J. & Gong, N. FLTrust: Byzantine-robust Federated Learning via Trust Bootstrapping. *Proceedings Of NDSS*. (2021)

[61] Fleiss, J. Measuring Nominal Scale Agreement among Many Raters. *Psychological Bulletin*. **76**, 378 (1971)

[62] Yuan, Z., Guo, Z., Xu, Y., Ying, Y. & Yang, T. Federated Deep AUC Maximization for Hetergeneous Data with a Constant Communication Complexity. *International Conference On Machine Learning*. pp. 12219-12229 (2021)

[63] Shejwalkar, V. & Houmansadr, A. Manipulating the Byzantine: Optimizing Model Poisoning Attacks and Defenses for Federated Learning. *NDSS*. (2021)

[64] Melis, L., Song, C., De Cristofaro, E. & Shmatikov, V. Exploiting Unintended Feature Leakage in Collaborative Learning. *2019 IEEE Symposium On Security And Privacy (SP)*. pp. 691-706 (2019)

[65] Naseri, M., Hayes, J. & De Cristofaro, E. Local and Central Differential Privacy forRobustness and Privacy in Federated Learning. *In Proceedings of NDSS*. (2022)

[66] Rieger, P., Nguyen, T., Miettinen, M. & Sadeghi, A. DeepSight: Mitigating Backdoor Attacks in Federated Learning Through Deep Model Inspection. *In Proceedings of NDSS*. (2022)

[67] EITR System, https://eitr-experiment.networks.imdea.org (2022)

APPENDIX

A. Proofs

The following are the lemmas we utilise in the proof of Theorem 1.

Lemma 1. *From Assumption 1 and 4, $\mathcal{L}(w)$ is L-smooth and μ-strongly convex. Then $\forall w_1, w_2 \in \mathcal{W}$, one has*

$$\langle \nabla\mathcal{L}(\mathbf{w}_1) - \nabla\mathcal{L}(\mathbf{w}_2), \mathbf{w}_1 - \mathbf{w}_2\rangle \geq \frac{L\mu}{L+\mu}\left\|\mathbf{w}_1 - \mathbf{w}_2\right\|_2^2 + \frac{1}{L+\mu}\left\|\nabla\mathcal{L}(\mathbf{w}_1) - \nabla\mathcal{L}(\mathbf{w}_2)\right\|_2^2 \quad (15)$$

Lemma 2. *The difference between $\mathbf{m}(\mathbf{w})$ and $\nabla\mathcal{L}(\mathbf{w})$ is bounded in every iteration:*

$$\left\|\mathbf{m}(\mathbf{w}) - \nabla\mathcal{L}(\mathbf{w})\right\|_2 \leq \left\|\mathbf{m}_0(\mathbf{w}) - \nabla\mathcal{L}(\mathbf{w})\right\|_2 + \sqrt{N}\Delta_1 \quad (16)$$

where:

$$\Delta_1 = \frac{M(\varpi(M-1) + \frac{2E}{\sqrt{M}\delta})}{\frac{Wa(M-1)(\kappa N+W)}{(\eta N+W)(\kappa N+Wa)} + 1}$$

$$E = \sup\left\{\frac{37\sqrt{2}\lambda(M+4)}{25(M-1)}\underset{i}{\text{median}}\left\{|w_{i,n}^t - \hat{B}_n x_{i,n}^t - \hat{A}_n|\right\}\right\}$$

and

$$\mathbf{m}_0(\mathbf{w}) := \underset{i}{\text{median}}\left\{\mathbf{m}_i(\mathbf{w})\right\}$$

1) Proof of Lemma 1: Let $g(\mathbf{w}) = \mathcal{L}(\mathbf{w}) - \frac{\varsigma}{2}\left\|\mathbf{w}\right\|_2^2$. Base on the assumption 4, we have $g(\mathbf{w})$ is $(L-\varsigma)$-strongly convex. from [23] 3.6, we have

$$\langle \nabla\mathcal{L}(\mathbf{w}_1) - \nabla\mathcal{L}(\mathbf{w}_2), \mathbf{w}_1 - \mathbf{w}_2\rangle \geq \frac{1}{L}\left\|\nabla\mathcal{L}(\mathbf{w}_1) - \nabla\mathcal{L}(\mathbf{w}_2)\right\|_2^2 \quad (17)$$

Hence,

$$\langle \nabla g(\mathbf{w}_1) - \nabla g(\mathbf{w}_2), \mathbf{w}_1 - \mathbf{w}_2\rangle \geq \frac{1}{L-\varsigma}\left\|\nabla g(\mathbf{w}_1) - \nabla g(\mathbf{w}_2)\right\|_2^2 \quad (18)$$

Now We have

$$\begin{aligned}&\langle \nabla\left(\mathcal{L}(\mathbf{w}_1) - \tfrac{\varsigma}{2}\left\|\mathbf{w}_1\right\|_2^2\right) - \nabla\left(\mathcal{L}(\mathbf{w}_2) - \tfrac{\varsigma}{2}\left\|\mathbf{w}_2\right\|_2^2\right), \mathbf{w}_1 - \mathbf{w}_2\rangle \\ &\geq \frac{1}{L+\mu}\left\|\nabla\left(\mathcal{L}(\mathbf{w_1}) - \tfrac{\varsigma}{2}\left\|\mathbf{w}_1\right\|_2^2\right) - \nabla\left(\mathcal{L}(\mathbf{w}_2) - \tfrac{\varsigma}{2}\left\|\mathbf{w}_2\right\|_2^2\right)\right\|_2^2\end{aligned} \quad (19)$$

And therefore

$$\begin{aligned}&\langle \nabla\mathcal{L}(\mathbf{w}_1) - \nabla\mathcal{L}(\mathbf{w}_2), \mathbf{w}_1 - \mathbf{w}_2\rangle - \langle \varsigma\mathbf{w}_1 - \varsigma\mathbf{w}_2, \mathbf{w}_1 - \mathbf{w}_2\rangle \\ &\geq \frac{1}{L-\varsigma}\left\|(\nabla\mathcal{L}(\mathbf{w}_1) - \nabla\mathcal{L}(\mathbf{w}_2)) - (\varsigma\mathbf{w}_1 - \varsigma\mathbf{w}_2)\right\|_2^2\end{aligned} \quad (20)$$

Refer to Assumption 1, we obtain

$$\begin{aligned}&\langle \nabla\mathcal{L}(\mathbf{w}_1) - \nabla\mathcal{L}(\mathbf{w}_2), \mathbf{w}_1 - \mathbf{w}_2\rangle \geq \frac{L\varsigma}{L-\varsigma}\left\|\mathbf{w}_1 - \mathbf{w}_2\right\|_2^2 \\ &- \frac{2\varsigma}{L-\varsigma}\langle \nabla\mathcal{L}(\mathbf{w}_1) - \nabla\mathcal{L}(\mathbf{w}_2), \mathbf{w}_1 - \mathbf{w}_2\rangle \\ &+ \frac{1}{L-\varsigma}\left\|\nabla\mathcal{L}(\mathbf{w}_1) - \nabla\mathcal{L}(\mathbf{w}_2)\right\|_2^2 \\ &\geq -\frac{L\varsigma}{L-\varsigma}\left\|\mathbf{w}_1 - \mathbf{w}_2\right\|_2^2 + \frac{1}{L-\varsigma}\left\|\nabla\mathcal{L}(\mathbf{w}_1) - \nabla\mathcal{L}(\mathbf{w}_2)\right\|_2^2\end{aligned} \quad (21)$$

Let $\varsigma = -\mu$, then we conclude the proof of Lemma 1.

2) Proof of Lemma 2: We have the following equation:

$$\begin{aligned}\left\|\mathbf{m}(\mathbf{w}) - \nabla\mathcal{L}(\mathbf{w})\right\|_2 \leq &\left\|\mathbf{m}(\mathbf{w}) - \mathbf{m}_0(\mathbf{w})\right\|_2 \\ &+ \left\|\mathbf{m}_0(\mathbf{w}) - \nabla\mathcal{L}(\mathbf{w})\right\|_2\end{aligned} \quad (22)$$

from [16] inequality 18, we know $\forall i, n, \exists E > 0$

$$\sup|e_{i,n}| \leq \frac{E}{\sqrt{M}\delta}$$

Where

$$E = \sup\left\{\frac{37\sqrt{2}\lambda(M+4)}{25(M-1)}\underset{i}{\text{median}}\left\{|w_{i,n}^t - \hat{B}_n x_{i,n}^t - \hat{A}_n|\right\}\right\}$$

and the dimension of $\mathbf{w}$ is N. Hence the distance between the two aggregation functions satisfies

$$\left\|\mathbf{m}(\mathbf{w}) - \mathbf{m}_0(\mathbf{w})\right\|_2 \leq \sqrt{N}\left\|\sum_{i=1}^{M}\bar{R}_i\left(\hat{B}_i\left(M-1\right) + \frac{2E}{\sqrt{M}\delta}\right)\right\|_2 \quad (23)$$

Based on Equation 12

$$\frac{s\exp(-cs)}{\sum_{j=0}^{s}\exp(-cj)}\cdot\frac{Wa}{\eta N+W} \leq \tilde{R}_i^t \quad (24)$$

$$\tilde{R}_i^t \leq \frac{s}{\sum_{j=0}^{s}\exp(-cj)}\cdot\frac{\kappa N+Wa}{\kappa N+W} \quad (25)$$

so we have

$$\bar{R}_i = \frac{\tilde{R}_i^t}{\sum_{i=1}^{M}\tilde{R}_i^t} \leq \frac{1}{\frac{Wa(M-1)(\kappa N+W)}{(\eta N+W)(\kappa N+Wa)} + 1} \quad (26)$$

Due to our Aggregation Algorithm

$$\hat{B}_n = \underset{i}{\text{median}}\left(\underset{i\neq j}{\text{median}}\frac{w_{j,n} - w_{i,n}}{x_{j,n} - x_{i,n}}\right) \leq \varpi \quad (27)$$

Therefore, we have

$$\begin{aligned}\left\|\sum_{i=1}^{M}\bar{R}_i\left(\hat{B}_i\left(M-1\right) + \frac{2E}{\sqrt{M}\delta}\right)\right\|_2 \leq &\frac{M(\varpi(M-1) + \frac{2E}{\sqrt{M}\delta})}{\frac{Wa(M-1)(\kappa N+W)}{(\eta N+W)(\kappa N+Wa)} + 1} \\ &= \Delta_1\end{aligned} \quad (28)$$

Hence, we proof Lemma 2.

3) Proof of Theorem 1: We first consider a general problem of robust estimation of a one dimensional random variable. Suppose that there are M clients, and p percentage of them are malicious and own adversarial data. In t iteration, we have:

$$\begin{aligned}&\left\|\mathbf{w}^t - \mathbf{w}^*\right\|_2 = \left\|(\mathbf{w}^{t-1} - r\mathbf{m}(\mathbf{w}^{t-1}) - \mathbf{w}^*\right\|_2 \\ &\leq \underbrace{\left\|\mathbf{w}^{t-1} - r\nabla\mathcal{L}(\mathbf{w}^{t-1}) - \mathbf{w}^*\right\|_2}_{A} \\ &+ \underbrace{r\left\|\mathbf{m}(\mathbf{w}^{t-1}) - \nabla\mathcal{L}(\mathbf{w}^{t-1})\right\|_2}_{B}\end{aligned} \quad (29)$$

We bound part A first. We have

$$\begin{aligned}&\left\|\mathbf{w}^{t-1} - r\nabla\mathcal{L}(\mathbf{w}^{t-1}) - \mathbf{w}^*\right\|_2^2 = \left\|\mathbf{w}^{t-1} - \mathbf{w}^*\right\|_2^2 \\ &+ r^2\left\|\nabla\mathcal{L}(\mathbf{w}^{t-1})\right\|_2^2 - 2r\left\langle\nabla\mathcal{L}(\mathbf{w}^{t-1}), \mathbf{w}^{t-1} - \mathbf{w}^*\right\rangle\end{aligned} \quad (30)$$

Under the Assumption 4 and Lemma 1, we have

$$\left\langle \nabla \mathcal{L}(\mathbf{w}^{t-1}), \mathbf{w}^{t-1} - \mathbf{w}^* \right\rangle \geq \frac{L\mu}{L+\mu} \left\| \mathbf{w}^{t-1} - \mathbf{w}^* \right\|_2^2 + \frac{1}{L+\mu} \left\| \nabla \mathcal{L}(\mathbf{w}^{t-1}) \right\|_2^2 \quad (31)$$

Then we combine inequalities 31 to equation 30

$$\left\| \mathbf{w}^{t-1} - r\nabla \mathcal{L}(\mathbf{w}^{t-1}) - \mathbf{w}^* \right\|_2^2 \leq (1 - 2r\frac{L\mu}{L+\mu}) \left\| \mathbf{w}^{t-1} - \mathbf{w}^* \right\|_2^2 + (r^2 - \frac{2r}{L+\mu}) \left\| \nabla \mathcal{L}(\mathbf{w}^{t-1}) \right\|_2^2 \quad (32)$$

From Assumption 1, we can derive:

$$\left\| \nabla \mathcal{L}(\mathbf{w}^{t-1}) - \nabla \mathcal{L}(\mathbf{w}^*) \right\|_2^2 \leq L^2 \left\| \mathbf{w}^{t-1} - \mathbf{w}^* \right\|_2^2 \quad (33)$$

Combining inequalities 32 and 33, we have:

$$\left\| \mathbf{w}^{t-1} - r\nabla \mathcal{L}(\mathbf{w}^{t-1}) - \mathbf{w}^* \right\|_2^2 \leq (1 - Lr)^2 \left\| \mathbf{w}^{t-1} - \mathbf{w}^* \right\|_2^2 \quad (34)$$

Let $r < \frac{1}{L}$, we have

$$\left\| \mathbf{w}^{t-1} - r\nabla \mathcal{L}(\mathbf{w}^{t-1}) - \mathbf{w}^* \right\|_2 \leq (1 - Lr) \left\| \mathbf{w}^{t-1} - \mathbf{w}^* \right\|_2 \quad (35)$$

Then we turn to bound part B. Based on Lemma 2, we know:

$$\left\| \mathbf{m}(\mathbf{w}) - \nabla \mathcal{L}(\mathbf{w}) \right\|_2 \leq \left\| \mathbf{m}_0(\mathbf{w}) - \nabla \mathcal{L}(\mathbf{w}) \right\|_2 + \sqrt{N}\Delta_1 \quad (36)$$

Assume Assumption 1, 2, 3 and 4 holds, and $\exists \epsilon$ fulfills inequality 13. Based on Lemma 1 in [20], with the probability $1 - \xi \geq 1 - \frac{4d}{\left(1+\hat{Q}ML\upsilon\right)^a}$, we have

$$\left\| \mathbf{m}_0(\mathbf{w}) - \nabla \mathcal{L}(\mathbf{w}) \right\|_2 \leq \sqrt{\frac{2}{\hat{Q}}} D_\epsilon V_w \left(\sqrt{\frac{d \log(1 + \hat{Q}ML\upsilon)}{M(1-p)}} + C\frac{\mathcal{G}_w}{\sqrt{\hat{Q}}} + p \right) + 2\sqrt{2}\frac{1}{M\hat{Q}} = \Delta_2 \quad (37)$$

where $C = 0.4748$. After obtaining the bound of part A and B, now we have

$$\left\| \mathbf{w}^t - \mathbf{w}^* \right\|_2 \leq \underbrace{(1 - Lr) \left\| \mathbf{w}^{t-1} - \mathbf{w}^* \right\|_2}_{Bound\ A} + \underbrace{r\left(\sqrt{N}\Delta_1 + \Delta_2\right)}_{Bound\ B} \quad (38)$$

Hence, we can prove Theorem 1 through iterations using the formula of a finite geometric series,

$$\begin{aligned} \left\| \mathbf{w}^t - \mathbf{w}^* \right\|_2 &\leq (1 - Lr)^t \left\| \mathbf{w}^0 - \mathbf{w}^* \right\|_2 + \frac{1 - (1 - Lr)^t}{Lr} r\left(\sqrt{N}\Delta_1 + \Delta_2\right) \\ &\leq (1 - Lr)^t \left\| \mathbf{w}^0 - \mathbf{w}^* \right\|_2 + \frac{1}{L}\left(\sqrt{N}\Delta_1 + \Delta_2\right) \end{aligned} \quad (39)$$

B. Experimental Setting

Our simulation experiments are implemented with Pytorch framework [28] on the cloud computing platform Google Colaboratory Pro (Colab Pro) with access to Nvidia K80s, T4s, P4s and P100s with 25 GB of Random Access Memory. Table II shows the default setting in our experiments.

TABLE II: Default experimental settings

Explanation	Notation	Default Setting
prior probability	a	0.5
non-information prior weight	W	2
weight for positive observation	κ	0.3
time decay parameter	c	0.5
window length	s	10
confidence threshold	δ	0.1
value range	ϖ	2
Objective Function	$\mathcal{L}(\cdot)$	Negative Log-likelihood Loss
Learning rate	r	0.01
Batch size per client		64
The number of local iterations		10
The number of total iterations		100

C. Supplementary dataset for experiment

CIFAR-10 is a supplementary dataset assessing the robustness of our reputation scheme in image datasets to poisoning attacks. The CIFAR-10 dataset is a 32×32 colour image dataset that includes ten classes with a total number of 50 thousand images for training and 10 thousand images for testing. Here we use ResNet-18 [19] model pertaining to ImageNet [25] with 20 iterations. In CIFAR-10 dataset, attackers switch the label of "cat" images to the "dog".

Figure 15 shows that under label flipping and backdoor attack during the whole process, our reputation-based method has the highest accuracy outperforming other methods and the lowest ASR, excluding Foolsgold in CIFAR-10, yielding a result that is similar to the one in SURL.

D. Extra Experimental Result

We evaluate the proposed defence for a varying number of clients from 10 to 200 in the SURL dataset. Here, we analyze the performance of the aforementioned methods for 100 clients. The results are presented in Figure 16 and Figure 17 that correspond to the average accuracy (ACC) and attach success rate (ASR), respectively.

In the no attack scenario, we observe that our method converges between $1.7\times$ to $7.1\times$ faster than all competing state-of-the-art methods with at least as good performance (or outperforms) compared with competing methods in terms of classification accuracy, see Figure 16 (left). In addition, even under the two different attacks, our method: (i) converges between $1.6\times$ to $3.6\times$ faster than all competing state-of-the-art methods, (ii) provides the same or better accuracy than competing methods, and (iii) yields the lowest ASR compared to all other methods.

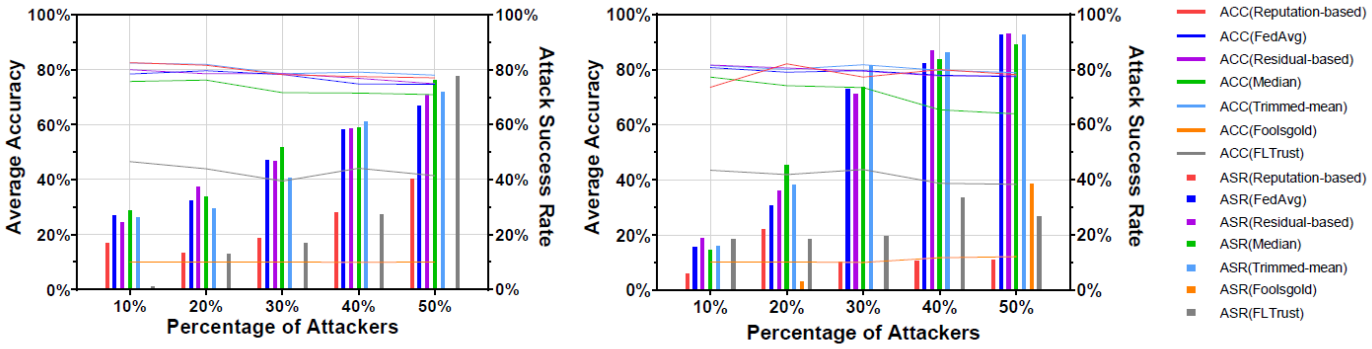


Fig. 15: Average accuracy (ACC) and attack success rate (ASR) for varying percentage of attackers from 10% to 50% under label flipping (left) and backdoor (right) attack for Reputation, FedAvg, Median, Residual-based, Median, Trimmed-Median, FoolsGold and FLTruts in CIFAR-10 dataset.

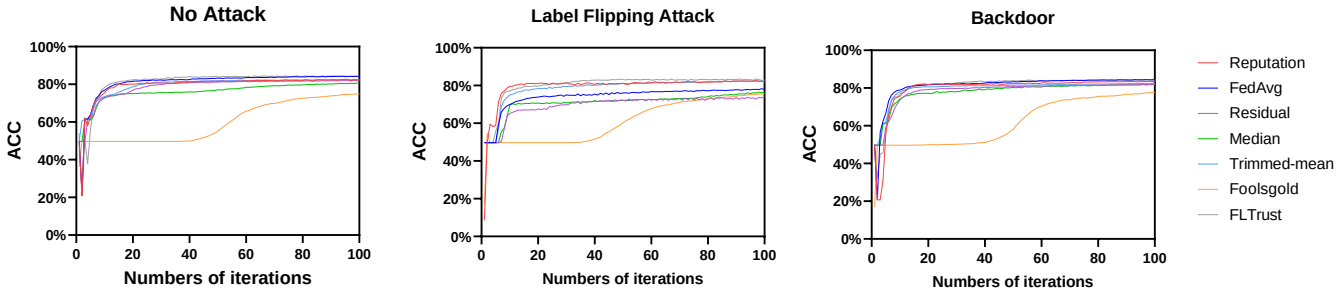


Fig. 16: Average accuracy (ACC) with no attack (left) and varying percentage of attackers from 10% to 50% under label flipping (middle) and backdoor (right) attack for Reputation, FedAvg, Residual-based, Median, Trimmed-mean, Foolsgold, FLTrust in SURL dataset.

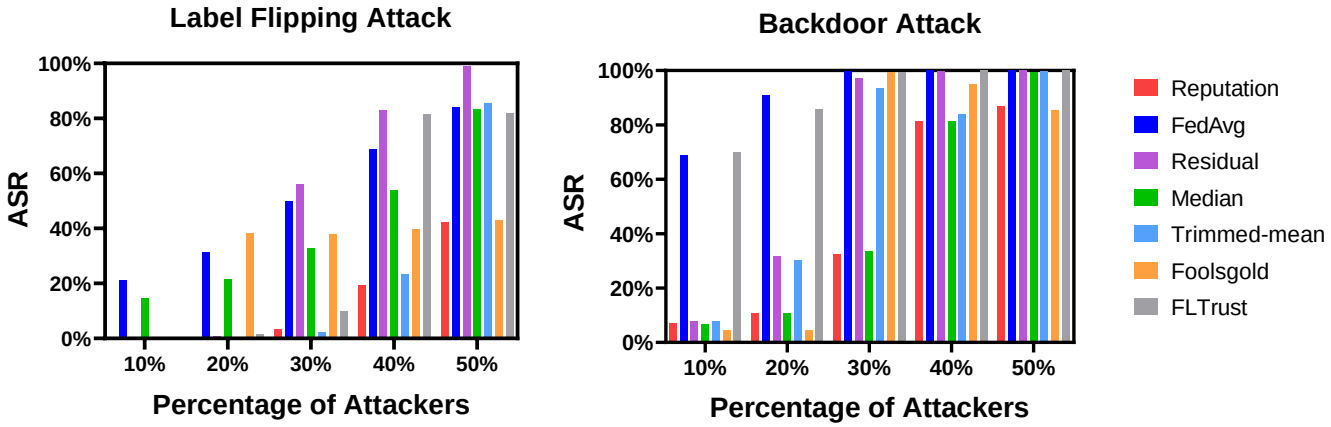


Fig. 17: Attack success rate (ASR) for varying percentage of attackers from 10% to 50% under label flipping (left) and backdoor (right) attack for Reputation, FedAvg, Residual-based, Median, Trimmed-mean, Foolsgold, FLTrust in SURL dataset.

Anexo 4: PriPrune: Quantifying and Preserving Privacy in Pruned Federated Learning

PriPrune: Quantifying and Preserving Privacy in Pruned Federated Learning

TIANYUE CHU*, IMDEA Networks Institute, Madrid (Leganes), Spain and Universidad Carlos III de Madrid, Madrid (Leganes), Spain
MENGWEI YANG*, University of California Irvine, Irvine, United States
NIKOLAOS LAOUTARIS, IMDEA Networks Institute, Madrid (Leganes), Spain
ATHINA MARKOPOULOU, University of California Irvine, Irvine, United States

Model pruning has been proposed as a technique for reducing the size and complexity of Federated learning (FL) models. By making local models coarser, pruning is intuitively expected to improve protection against privacy attacks. However, the level of this expected privacy protection has not been previously characterized, or optimized jointly with utility.

In this paper, we first characterize the privacy offered by pruning. We establish information-theoretic upper bounds on the information leakage from pruned FL and we experimentally validate them under state-of-the-art privacy attacks across different FL pruning schemes. Second, we introduce *PriPrune*– a privacy-aware algorithm for pruning in FL. *PriPrune* uses defense pruning masks, which can be applied locally after *any* pruning algorithm, and adapts the defense pruning rate to jointly optimize privacy and accuracy. Another key idea in the design of *PriPrune* is *Pseudo-Pruning*: it undergoes defense pruning within the local model and only sends the pruned model to the server; while the weights pruned out by defense mask are withheld locally for future local training rather than being removed. We show that *PriPrune* significantly improves the privacy-accuracy tradeoff compared to state-of-the-art pruned FL schemes. For example, on the FEMNIST dataset, *PriPrune* improves the privacy of *PruneFL* by 45.5% without reducing accuracy.

CCS Concepts: • **Security and privacy** → **Distributed systems security**; • **Computing methodologies** → **Machine learning**; • **Theory of computation** → *Theory and algorithms for application domains.*

Additional Key Words and Phrases: Federated Learning, Privacy, Model Pruning

1 Introduction

Federated Learning (FL) has emerged as the predominant paradigm for distributed machine learning across a multitude of user devices [18, 26]. It is known to have several benefits in terms of reducing the communication, computation and storage costs for the users, training better global models, and raising the bar for privacy by not sharing the local data. FL allows users to train models locally on their (user) devices without revealing their local data but instead collaborate by sharing only model updates that can be combined to build a global model through a central server. A typical FL scenario involves numerous users and resource-constrained devices [23] making it challenging to perform resource-intensive tasks like training Deep Neural Networks (DNNs) on them.

*Both authors contributed equally to the paper.

Authors' Contact Information: Tianyue Chu, IMDEA Networks Institute, Madrid (Leganes), Spain and Universidad Carlos III de Madrid, Madrid (Leganes), Spain; e-mail: tianyue.chu@imdea.org; Mengwei Yang, University of California Irvine, Irvine, California, United States; e-mail: mengwey@uci.edu; Nikolaos Laoutaris, IMDEA Networks Institute, Madrid (Leganes), Spain; e-mail: nikolaos.laoutaris@imdea.org; Athina Markopoulou, University of California Irvine, Irvine, California, United States; e-mail: athina@uci.edu.

ACM 2376-3647/2024/11-ART
https://doi.org/10.1145/3702241

In parallel, significant efforts have been put toward optimizing sparse DNNs to create lightweight models suitable for training on edge devices [6, 14, 24]. Model pruning, which involves the removal of a certain percentage of parameters, has gained significant attention [13, 20, 28]. Its primary objective is to derive a sparse DNNs model that decreases the computational requirements and enhances efficiency without compromising accuracy [10]. Several model pruning techniques have been developed for FL to optimize model sparsity under given constraints on the processing and communication capabilities of devices, especially under heterogeneous clients, [3, 17].

It is intuitively expected that by reducing the number of weights of FL models, the pruning method should, as a side-effect, also improve their resistance to reconstruction attacks. These attacks aim at using these weights to reverse-engineer the model and obtain information about the input data; a.k.a. gradient inversion attacks launched by the server. Prior work on model pruning focuses solely on the scalability/accuracy trade-off and thus neither considers nor quantifies the privacy impact of pruning on FL. To the best of our knowledge, our paper is the first to consider privacy in the context of FL model pruning, and to address the following key questions: **Q1.** *Can we quantify privacy in FL model pruning?* **Q2.** *How can we further optimize pruning for privacy, and jointly with accuracy?* To address these questions, we make the following two contributions.

First, **we theoretically and empirically quantify privacy leakage in model pruning.** We derive information-theoretic upper bounds on the amount of information revealed about any single user's dataset via model updates, in *any* pruned FL scheme. This is the first theoretical characterization of privacy leakage in pruned FL models. We also conduct a comprehensive empirical evaluation that quantifies the actual amount of privacy leakage of six different pruning schemes, considering several state-of-the-art privacy attacks. Within the family of gradient inversion attacks (including Deep Leakage from Gradients [45] and Gradient Inversion (GI) [11]), we also design a novel privacy attack (referred to as Sparse Gradient Inversion, or SGI), specifically tailored to exploit vulnerabilities inherent in FL model pruning. This evaluation combined with our theoretical analysis provides valuable insights into the choices and parameters that affect the privacy provided by pruning.

Building upon these insights, we make our second contribution: we design ***PriPrune*– a privacy-preserving pruning mechanism.** *PriPrune* defends against gradient inversion attacks in pruned FL, by performing additional local (defense) pruning, after any (base) FL pruning scheme. *PriPrune* applies a personalized defense mask and adapts the defense pruning rate, so as to jointly optimize model accuracy *and* privacy, using back-propagation augmented with Gumbel Softmax Sampling. Another key idea in the design of *PriPrune* is what we refer to as *Pseudo-Pruning*: it entails pruning with the defense mask within the local model and transmitting the pruned model to the server, thereby enhancing privacy. However, the weights pruned out by the defense mask in the local model are not discarded; instead, they are retained locally for subsequent rounds of local training, thereby preserving model accuracy. We evaluate *PriPrune* via comprehensive experiments across various FL pruning schemes, a state-of-the-art attack and several benchmark datasets. We show that it achieves a significantly better privacy-accuracy tradeoff than privacy-unaware pruned FL baselines. For example, on the FEMNIST dataset, *PriPrune* improves the privacy *PruneFL* [17] by 45.5% without imposing any accuracy loss.

2 Related Work

Model Pruning in FL. State-of-the-art research for model pruning in FL utilizes two main approaches: one involves the pruning mask being decided on the server side and the client only using this mask without changing it, and the other involves the collaborative selection of the pruning mask by both the user and the server side. *PruneFL* [17] follows the first approach., aiming to reduce the size and complexity of FL models without compromising accuracy. It achieves this by removing less important weights and connections in each layer of the neural network based on their magnitude and importance scores. However, the pruning mask is only decided by the server without considering user-side local representations. *LotteryFL* [22] belongs to the second approach,

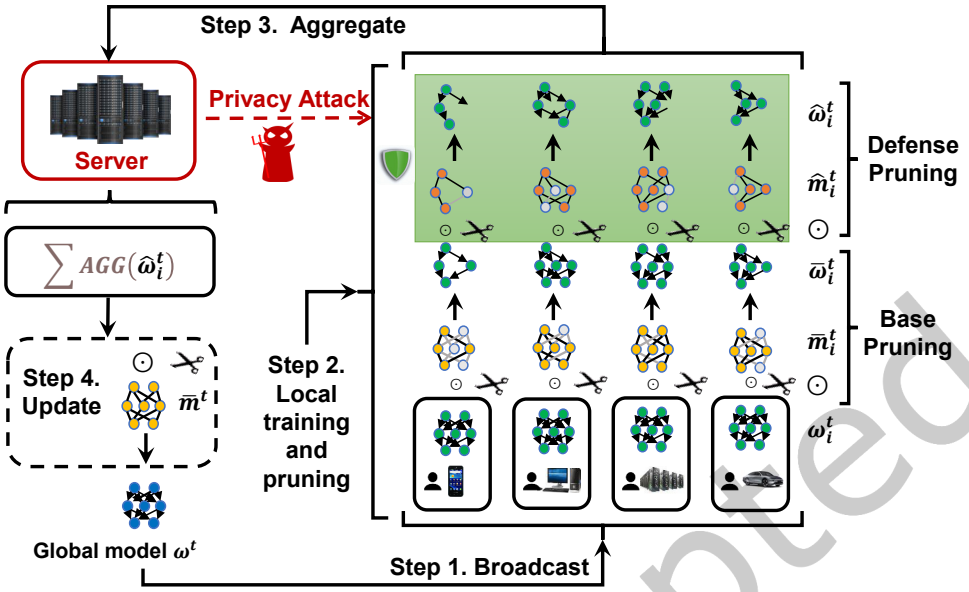


Fig. 1. FL with model pruning, under a privacy attack by an honest-but-curious server. Step 1: the server broadcasts the global model with weight w^t to users. Step 2: After local training, user i employs base pruning with pruning mask $\bar{m}_i^t$, resulting in a pruned model with weights $\bar{w}_i^t$. The user then applies a defense pruning technique, yielding a pruned model with weights $\hat{w}_i^t$. Step 3, the server aggregates the model updates and launches a gradient inversion attack upon receiving $\hat{w}_i^t$. Step 4: the server applies global pruning with pruning mask $\bar{m}^t$, and updates the global model accordingly. Notice *Base* pruning can be *any* FL pruning scheme. *Defense* pruning is introduced, in this paper, specifically to protect against privacy attacks.

introducing a method that allows users to maintain local representations by selecting a subset of the global network using personalized masks. However, the system of sparse models it produces performs well only on local datasets. *FedDST* [3] involves dynamic sparse training on both the client and server sides, extracting and training sparse sub-networks. Other recent state-of-the-art studies have introduced various gradient-based one-shot pruning algorithms at initialization. *Snip* [21] leverages connection sensitivity to retain critical connections and remove less significant ones, aiming to limit the loss due to pruning. On the other hand, *GraSP* [38] focuses on preserving the gradient flow through the network by scoring weights based on the Hessian-gradient product. *SynFlow* [34], being a data-agnostic pruning algorithm, emphasizes preserving the total flow of synaptic strengths through the neural network to achieve Maximal Critical Compression during initialization. More importantly, none of them considered privacy implications in their design, as *PriPrune* does.

Reconstruction Attacks in FL. In terms of privacy leakage, communicating gradients throughout the training process in federated learning can reveal sensitive information about the participants. *Deep leakage from gradients* (DLG) [45] demonstrated that sharing the gradients can leak private training data, including images and text. *Improved DLG* (iDLG) [44], presented an analytical procedure to extract the ground-truth labels from the shared gradients and showed the advantages of iDLG over DLG. *Inverting Gradients* [11] introduced cosine similarity as a cost function in their reconstruction attacks and employed total variation loss $\mathcal{L}_{\mathcal{TV}}$ as an image prior. *Inverting Gradients* [11] and *GradInversion* [41] demonstrated the capability to reconstruct high-resolution images with increased batch sizes. The reconstruction quality of the attack was improved in *Reconstruction From Obfuscated*

Gradient (ROG) [42]. ROG [42] proposed conducting the reconstruction optimization in low dimensional space and trained a neural network as a postprocessing module.

While these privacy attacks have demonstrated the privacy vulnerabilities of standard federated learning, none of them has considered pruned federated learning. Therefore, we do not currently know how effective existing attacks are against the sparser models from pruned FL.

Privacy-Preserving Techniques in FL. FL is particularly vulnerable to the aforementioned inverting-the-gradient type of reconstruction attacks, since it relies on gradient updates to operate. There exists a large body of literature on privacy-enhancing methods that protect against this type of attacks in FL. The adversary is typically an "honest-but-curious" server, *i.e.,* a server that participates correctly in the FL protocol, but tries to infer training data of individual clients based on their gradient updates. One family of defense approaches is based on differential privacy (DP) [7, 8, 12, 36, 40, 43], which can be applied locally (at the clients), centrally (at the server), or in a distributed way. In the local DP model for FL, each client adds noise to the local update before sharing with the server, which makes it difficult for the server to invert the true gradient and infer the client's training data. However, local DP is known to significantly degrade utility, in this case model accuracy. Another defense approach applied to FL is Secure Aggregation (SecAgg) [4, 32, 33, 37, 39]. SecAgg ensures that individual updates are encrypted (thus preventing the server from inverting the gradients of individual users), while their aggregate can still be decrypted (thus enabling the server to train a good model). SecAgg is powerful but has computational overhead, although significant advances have been done recently.

Both DP and SecAgg are important but orthogonal to the underlying FL and can be combined with a number of FL protocols, including all the pruning protocols we consider as base in this study. Our proposed *PriPrune* can also be combined with DP and/or SecAgg. For instance, after applying *PriPrune* to base pruning, clients can add local DP noise to the pruned updates before sending them to the server, and/or encrypt them to enable SecAgg following standard protocols. We focus on pruning itself and the privacy-utility tradeoff that can be achieved, *before* adding orthogonal defense techniques. Future work could explore adding DP and/or SecAgg, which are out of the scope of this paper.

Recent works like *FedMap* [15] and *Fed-LTP* [30] discuss privacy specifically in the context of pruned FL. FedMap, which appeared after this submission, enhances privacy by training models from scratch and by avoiding the sharing of detailed pruning masks. However, it does not quantify the privacy leakage that may still occur from the pruned models. *FedMap* can be considered as a base pruning technique, and *PriPrune* can still be applied to protect it. *Fed-LTP*, on the other hand, combines Lottery Ticket Pruning with Zero-concentrated Differential Privacy (zCDP) to enhance privacy. Integrating zCDP, or any type of DP, with our proposed *PriPrune* is orthogonal to our core contribution, as discussed above, and can be explored as part of future research.

3 Problem Setup

Federated Learning (FL) with Pruning. We consider a general FL system with multiple users and one server, employing *any* model pruning scheme on the users and/or the server. We refer to pruning at this stage as *base* pruning with base pruning rate $\bar{p}$, and it can be any state-of-the-art pruning FL scheme [3, 17]. This is depicted in Fig. 1 and formalized in Algorithm 1.

The goal is to train a global model $F(\mathcal{D}, \boldsymbol{w})$ through FL. At the beginning of round t, user i with the local data $\mathcal{D}_i$ and labels y_i has local model weights $\boldsymbol{w}_i^t$; the server broadcasts the most recent global model $F(\mathcal{D}, \boldsymbol{w}^t)$ to all participating users. Then the round proceeds as follows:

Algorithm 1 FL with Pruning under Privacy Attack

Given: T number of global training rounds; N number of users indexed by i; E number of local epochs; the server aims to reconstruct the local data of the target user.
Server executes:
Initialize $\boldsymbol{w}^0$
for $t \leftarrow 1$ **to** T **do**
  for each user $i \in N$ **in parallel do**
    $\hat{\boldsymbol{w}}_i^t \leftarrow \mathit{UserUpdate}(\boldsymbol{w}^{t-1}, \boldsymbol{m}^{t-1}, i)$
    if user i is targeted **then**
      $\nabla\hat{\boldsymbol{w}}_i^t \leftarrow \hat{\boldsymbol{w}}_i^t - \hat{\boldsymbol{w}}_i^{t-1}$
      Privacy Attack on $(F(\mathcal{D}_i; \hat{\boldsymbol{w}}_i^t), \hat{\boldsymbol{w}}_i^t, \nabla\hat{\boldsymbol{w}}_i^t)$
    end if
  end for
  $\boldsymbol{w}^t \leftarrow \left(\sum_{i=1}^{I} \frac{\|\mathcal{D}_i\|}{\|\mathcal{D}\|} \hat{\boldsymbol{w}}_i^t\right) \odot \bar{\boldsymbol{m}}^t$;
end for
function $\mathit{UserUpdate}(\boldsymbol{w}^{t-1}, \boldsymbol{m}^{t-1}, i)$:
for local epoch $e \leftarrow 1$ **to** E **do**
  Local Training and Base Pruning: $\bar{\boldsymbol{w}}_i^t \leftarrow$ Eq. (1)
  Defense Pruning: $\hat{\boldsymbol{w}}_i^t \leftarrow \bar{\boldsymbol{w}}_i^t \odot \hat{\boldsymbol{m}}_i^t$
end for
Return $\hat{\boldsymbol{w}}_i^t$ to server

Local Training and Base Pruning: Each user i utilizes base pruning strategy $\mathbb{P}_i(\cdot)$ with base pruning masks $\bar{\boldsymbol{m}}_i^t$ on its trained local model $F(\mathcal{D}_i, \boldsymbol{w}_i^t)$:

$$\bar{\boldsymbol{w}}_i^t := \left(\boldsymbol{w}^t - r\frac{\partial \ell_i\left(F(\mathcal{D}_i, \boldsymbol{w}^t \odot \bar{\boldsymbol{m}}^t), y_i\right)}{\partial \boldsymbol{w}^t}\right) \odot \bar{\boldsymbol{m}}_i^t \tag{1}$$

Where $\bar{\boldsymbol{m}}_i^t = \mathbb{P}_i(\boldsymbol{w}_i^t)$, r is the learning rate and $\odot$ is the Hadamard product. The *base* pruning mask $\bar{\boldsymbol{m}}_i^t$ exhibits variability across different base pruning schemes, reflecting the diverse pruning strategy $\mathbb{P}_i(\cdot)$ employed.
Server: The server receives model updates $\hat{\boldsymbol{w}}_i^t$ from all participating users and aggregates to update the global model, as it is typical in FL. We do not consider secure aggregation [33], differential privacy [27], or other defenses known in the literature.[1]

Threat Model. We consider an *honest-but-curious* server, which correctly follows the FL protocol but uses model updates to infer private information, by launching privacy attacks against target user(s), as shown in Algorithm 1. In our case, the server has full visibility of all masked local models and, upon receiving an update from a target user, it attempts to reconstruct the user's training data. We consider reconstruction attacks that invert gradients, including Deep-Leakage-from-Gradients (DLG) [45], Gradient Inversion (GI) [11], and a custom attack SGI described in Section 4.2. We focus on single-round attacks, *i.e.,* attacks that invert gradients observed at a single round[2].

[1]We consider those defenses as out-of-the-scope for this paper, since they are orthogonal to pruning. They can be implemented on top/combined with our pruning-based defense methods.
[2]To the best of our knowledge, there are currently no known *practical* multi-round inverting-the-gradient types of attacks that combine information from multiple rounds. There is theoretical analysis of upper *bounds* of multi-round information leakage, including [9, 32] and we derive our own multi-round leakage in Section 4.1.2. Designing a practical multi-round attack is orthogonal to our contribution, which focuses on quantifying the leakage and designing an add-on defense, for *existing* attacks.

Defense via Pruning: In this paper, we propose to defend against privacy attacks, via pruning itself. One could specifically design an entire pruning scheme not only for accuracy and efficiency but also for privacy, and the defense would then be optimized jointly with the particular base pruning scheme *e.g.,* PruneFL [17] and FedDST [3].

In this paper, we take a modular and universal approach: we consider *any* given state-of-the-art base pruning scheme (with rate $\bar{p}_i$ and mask $\bar{\boldsymbol{m}}_i^t$) without any modification. Then, we introduce additional local *defense* pruning, with mask $\hat{\boldsymbol{m}}_i^t$ and pruning rate $\hat{p}_i$, specifically designed for protecting each user i from privacy attacks. User i eventually sends back to the server weights after applying both base ($\bar{\boldsymbol{m}}_i^t$) and defense ($\hat{\boldsymbol{m}}_i^t$) pruning: $\hat{\boldsymbol{w}}_i^t = \bar{\boldsymbol{m}}_i^t \odot \hat{\boldsymbol{m}}_i^t \odot \boldsymbol{w}_i^t$. This is depicted in Fig. 1 and Algorithm 1 (see Line 15 and 16), and elaborated upon on Fig. 7. Given a base pruning ($\bar{\boldsymbol{m}}_i^t$), there are many possible defense pruning strategies ($\hat{\boldsymbol{m}}_i^t$). In Section 5.2, we propose *PriPrune* for $\hat{\boldsymbol{m}}_i^t$, to optimize the accuracy-vs-privacy tradeoff.

4 Quantifying Privacy in FL Model Pruning

In this section, to address question Q1, we provide a theoretical and empirical quantification of privacy leakage in FL model pruning. We aim to assess the server's ability to infer information about an individual user's local dataset $\mathcal{D}_i$. We employ Mutual Information (MI) as the metric for quantifying privacy leakage:

$$\begin{aligned} \mathbf{I}(X;Y) &= \int_{\mathcal{X}\times\mathcal{Y}} log \frac{d\mathbb{P}_{X,Y}(x,y)}{d\mathbb{P}_X(x)\otimes\mathbb{P}_Y(y)} d\mathbb{P}_{X,Y}(x,y) \\ &= \mathbf{H}(X) + \mathbf{H}(Y) - \mathbf{H}(X,Y) \end{aligned} \tag{2}$$

Where $(X;Y)$ is a pair of random variables with values over the space $\mathcal{X}\times\mathcal{Y}$. Their joint distribution is $\mathbb{P}_{X,Y}(x,y)$ and the marginal distributions are $\mathbb{P}_X(x)$ and $\mathbb{P}_Y(y)$.

This concept, rooted in information theory and Shannon entropy, captures the interdependence between two random variables. It is very powerful for privacy analysis, given its applicability across various domains and tasks [1, 2]. This applicability extends across various datasets, pruning schemes, and attack scenarios, as recently demonstrated in the quantification of privacy within the context of FL aggregation [9].

4.1 Theoretical Quantification

We seek to quantify how much information the server can infer about the private data $\mathcal{D}_i$ of user i over T global training rounds, based on the pruned updates $\left\{\boldsymbol{w}_i^t \odot \bar{\boldsymbol{m}}_i^t\right\}_{t\in[T]}$ submitted by user i, via:

$$\mathbf{I}_i = \mathbf{I}\left(\mathcal{D}_i; \left\{\boldsymbol{w}_i^t \odot \bar{\boldsymbol{m}}_i^t\right\}_{t\in[T]}\right) \tag{3}$$

where $\mathbf{I}$ represents mutual information between $\mathcal{D}_i$ and $\left\{\boldsymbol{w}_i^t \odot \bar{\boldsymbol{m}}_i^t\right\}_{t\in[T]}$, $\boldsymbol{w}_i^t \in \mathbb{R}^d$ is the local model weights of user i at round t.

This quantification is based on two mild assumptions:

Assumption 1. *The random vector $\hat{g}_i\left(\boldsymbol{w}_i^t, b\right)$ has non-singular covariance matrix Σ_i and mean 0.*

Here, $\hat{g}_i\left(\boldsymbol{w}_i^t, b\right) \in \mathbb{R}^{d^*}$ is the largest sub-vector of the $g_i\left(\boldsymbol{w}_i^t, b\right)$, where d^* is the rank of $g_i\left(\boldsymbol{w}_i^t, b\right)$, $d^* \leq d$, d is the model size, b denotes the size of the random samples. $g_i(\boldsymbol{w}_i^t, b)$ denotes the stochastic estimate of the gradient of the local loss function for user i, computed based on a random sample b, and $\mathbf{B}_i^t$ is a data batch of size B sampled uniformly at random from it local dataset $\mathcal{D}_i$.

Assumption 2. *$\bar{p}_i$ the pruning rate of user i remains unchanged throughout the training process.*

Building upon Equation 2, Equation 3 can be written as:

$$\mathbf{I}_i \leq \sum_{t=1}^{T} \left(\mathbf{I} \left(\mathrm{x}_i^t; \boldsymbol{w}_i^t \odot \bar{\boldsymbol{m}}_i^t \middle| \left\{ \boldsymbol{w}_i^k \odot \bar{\boldsymbol{m}}_i^k \right\}_{k \in [t-1]} \right) \right) \tag{4}$$

Where $\mathrm{x}_i^t = \frac{\partial \ell_i(\tilde{\boldsymbol{w}}_i^{t-1}; \mathcal{D}_i)}{\partial \boldsymbol{w}} = \frac{1}{B} \sum_{b \in \mathbf{B}_i^t} g_i\left(\boldsymbol{w}_i^t, b\right)$.

To that end, we first characterize the privacy leakage for a *single round t*:

$$\mathbf{I} \left(\mathrm{x}_i^t; \boldsymbol{w}_i^t \odot \bar{\boldsymbol{m}}_i^t \middle| \left\{ \boldsymbol{w}_i^k \odot \bar{\boldsymbol{m}}_i^k \right\}_{k \in [t-1]} \right) \tag{5}$$

Under these two mild assumptions, we obtain our main theoretical result; the proof is deferred to Appendix A.1.

4.1.1 Upper Bound for Privacy Leakage in a Single Round. We now provide an upper bound for Equation 5, i.e., for the privacy leakage in a single round.

Theorem 4.1 (Single Round Leakage). *Under Assumption 1 and 2, we have an upper bound of* $\mathbf{I}_i^t$ *for a single round:*

$$\mathbf{I}_i^t \leq 1 - \frac{\bar{p}_i - 1}{2 \ln 2} + 2 \log \frac{1}{B} + 2\Delta + d^* \log (2\pi e) \tag{6}$$

where $\Delta = \log \left| \det \left(\Sigma_i^{-\frac{1}{2}} \right) \right|$.

Proof. Based on Equation 2, we have

$$\mathbf{I} \left(\mathrm{x}_i^t; \boldsymbol{w}_i^t \odot \bar{\boldsymbol{m}}_i^t \middle| \left\{ \boldsymbol{w}_i^k \odot \bar{\boldsymbol{m}}_i^k \right\}_{k \in [t-1]} \right) \leq \underbrace{\mathbf{H} \left(\boldsymbol{w}_i^t \odot \bar{\boldsymbol{m}}_i^t \middle| \left\{ \boldsymbol{w}_i^k \odot \bar{\boldsymbol{m}}_i^k \right\}_{k \in [t-1]} \right)}_{\mathbf{A}} + \underbrace{\mathbf{H} \left(\mathrm{x}_i^t \middle| \left\{ \boldsymbol{w}_i^k \odot \bar{\boldsymbol{m}}_i^k \right\}_{k \in [t-1]} \right)}_{\mathbf{B}} \tag{7}$$

For part **A**, based on the chain rule of entropy, we have

$$\mathbf{A} = \mathbf{H} \left(\boldsymbol{w}_i^t \odot \bar{\boldsymbol{m}}_i^t \middle| \left\{ \boldsymbol{w}_i^k \odot \bar{\boldsymbol{m}}_i^k \right\}_{k \in [t-1]} \right) + \mathbf{H} \left(\boldsymbol{w}_i^t, \bar{\boldsymbol{m}}_i^t \middle| \left\{ \boldsymbol{w}_i^k \odot \bar{\boldsymbol{m}}_i^k \right\}_{k \in [t]} \right) \geq 0 \tag{8}$$

Due to $\mathbf{H} \left(\boldsymbol{w}_i^t \odot \bar{\boldsymbol{m}}_i^t \middle| \boldsymbol{w}_i^t, \bar{\boldsymbol{m}}_i^t, \left\{ \boldsymbol{w}_i^k \odot \bar{\boldsymbol{m}}_i^k \right\}_{k \in [t-1]} \right) = 0$, and $\mathbf{H} \left(\boldsymbol{w}_i^t, \bar{\boldsymbol{m}}_i^t \middle| \left\{ \boldsymbol{w}_i^k \odot \bar{\boldsymbol{m}}_i^k \right\}_{k \in [t]} \right) \geq 0$, we have

$$\mathbf{H} \left(\boldsymbol{w}_i^t \odot \bar{\boldsymbol{m}}_i^t \middle| \left\{ \boldsymbol{w}_i^k \odot \bar{\boldsymbol{m}}_i^k \right\}_{k \in [t-1]} \right) \leq \underbrace{\mathbf{H} \left(\mathrm{x}_i^t \middle| \left\{ \boldsymbol{w}_i^k \odot \bar{\boldsymbol{m}}_i^k \right\}_{k \in [t-1]} \right)}_{\mathbf{B}} + 1 - \frac{\bar{p}_i - 1}{2 \ln 2} \tag{9}$$

The proof is shown in the Appendix A.1.

$$\mathbf{B} = \mathbf{H} \left(\frac{1}{B} \sum_{b \in \mathbf{B}_i^t} g_i\left(\boldsymbol{w}_i^t, b\right) \middle| \left\{ \boldsymbol{w}_i^k \odot \bar{\boldsymbol{m}}_i^k \right\}_{k \in [t-1]} \right) \tag{10}$$

Let $X = \sum_{b \in \mathbf{B}_i^t} g_i\left(\boldsymbol{w}_i^t, b\right)$, $Z = \left\{ \boldsymbol{w}_i^k \odot \bar{\boldsymbol{m}}_i^k \right\}_{k \in [t-1]}$, $Y = \frac{1}{B} X$. $f_{X,Z}(x, z)$ and $f_{Y,Z}(y, z)$ are the corresponding joint probability density functions.

Then Equation 10 can be written as follows:

$$\mathbf{H}\left(Y|Z\right) = \underbrace{\mathbf{H}\left(\sum_{b\in\mathbf{B}_i^t} g_i\left(\boldsymbol{w}_i^t, b\right)\middle|\left\{\boldsymbol{w}_i^k \odot \bar{\boldsymbol{m}}_i^k\right\}_{k\in[t-1]}\right)}_{\mathbf{C}} + \log\frac{1}{B} \tag{11}$$

The proof is shown in the Appendix A.1.

In recent theoretical results for analyzing the behaviour of SGD, the SGD vector is approximated by a distribution with independent components or by a multivariate Gaussian vector [46]. Based on the ZCA whitening transformation and Assumption 1, we have $\hat{g}_i\left(\boldsymbol{w}_i^t, b\right) = \Sigma_i^{-\frac{1}{2}}\boldsymbol{z}$, where $\boldsymbol{z}$ has zero mean and $\mathbb{I}_{d^*}$ convariance matrix.

For the part **C**, we set $\boldsymbol{v} = \sum_{b\in\mathbf{B}_i^t}\boldsymbol{z}$ and $\boldsymbol{u} = \Sigma_i^{-\frac{1}{2}}\boldsymbol{v}$, then we have:

$$\begin{aligned}\mathbf{C} &= \mathbf{H}\left(\sum_{b\in\mathbf{B}_i^t}\hat{g}_i\left(\boldsymbol{w}_i^t, b\right)\middle|\left\{\boldsymbol{w}_i^k \odot \bar{\boldsymbol{m}}_i^k\right\}_{k\in[t-1]}\right) = \mathbf{H}\left(\Sigma_i^{-\frac{1}{2}}\sum_{b\in\mathbf{B}_i^t}\boldsymbol{z}\middle|\left\{\boldsymbol{w}_i^k \odot \bar{\boldsymbol{m}}_i^k\right\}_{k\in[t-1]}\right) = -\int \mathbb{P}_{\mathcal{U}}(\boldsymbol{u})\log\mathbb{P}_{\mathcal{U}}(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u} \\ &\overset{(a)}{=} -\int \frac{\mathbb{P}_{\mathcal{V}}(\boldsymbol{v})}{\left|\det\left(\Sigma_i^{-\frac{1}{2}}\right)\right|}\log\left(\frac{\mathbb{P}_{\mathcal{V}}(\boldsymbol{v})}{\left|\det\left(\Sigma_i^{-\frac{1}{2}}\right)\right|}\right)\left|\det\left(\Sigma_i^{-\frac{1}{2}}\right)\right|\mathrm{d}\boldsymbol{v} = \log\left|\det\left(\Sigma_i^{-\frac{1}{2}}\right)\right| + \underbrace{\mathbf{H}\left(\sum_{b\in\mathbf{B}_i^t}\boldsymbol{z}\middle|\left\{\boldsymbol{w}_i^k \odot \bar{\boldsymbol{m}}_i^k\right\}_{k\in[t-1]}\right)}_{\mathbf{D}}\end{aligned} \tag{12}$$

where (a) is based on the transformation of random vectors. Based on the Maximum Entropy Upper Bound [35], which the continuous distribution X with prescribed variance $\mathcal{V}(\mathrm{X})$ maximizing the entropy is the Gaussian distribution of same variance. Since the entropy of a multivariate Gaussian distribution with mean μ and covariance $K_{i,j}$ is

$$\mathbf{h}(\mathcal{N}_n(\mu, K)) = \frac{1}{2}\log\left(2\pi e\right)^n|K|$$

where $|K|$ denotes the determinant of K. Hence, the maximum entropy upper bound for part **D** is

$$\mathbf{D} \le \frac{1}{2}\log\left(2\pi e\right)^{d^*}|\mathbb{I}_{d^*}| = \frac{d^*}{2}\log\left(2\pi e\right) \tag{13}$$

Therefore, by combining Equation 12 and Equation 13, the upper bound for part **B** is

$$\mathbf{B} \le \log\frac{1}{B} + \log\left|\det\left(\Sigma_i^{-\frac{1}{2}}\right)\right| + \frac{d^*}{2}\log\left(2\pi e\right) \tag{14}$$

By adding parts **A** and **B**, we establish the upper bound for privacy leakage for a single round as in Theorem 4.1.

4.1.2 Upper Bound for Privacy Leakage across Multiple Rounds. Using Equation 4, we can also obtain an upper bound for Equation 3, i.e., how much information the aggregated model over T global training rounds could leak about the private data:

Corollary 4.2 (Multiple Rounds Leakage). *Continuing with Theorem 4.1, we can upper bound* $\mathbf{I}_i$ *after* T *global training rounds as follows:*

$$\mathbf{I}_i \le T\left(1 - \frac{\bar{p}_i - 1}{2\ln 2} + 2\log\frac{1}{B} + 2\Delta + d^*\log\left(2\pi e\right)\right) \tag{15}$$

Algorithm 2 Sparse Gradient Inversion (SGI) Attack

Input: $F(\mathcal{D}_i;\hat{\boldsymbol{w}}_i^t)$: model at round t from targeted user i; learning rate η for inverting gradient optimizer; S: max iterations for attack; τ: regularization term for cosine loss in inverting gradient attack; $\boldsymbol{m}_{rec}$:the recovery mask generated by server; d:the model size
Output: reconstructed training data $(\mathcal{D}_i, y_i)$ at round t
$$\boldsymbol{m}_{rec}[j]_{j\in d} = \begin{cases} 1 & \text{if } \hat{\boldsymbol{w}}_i^t[j] \neq 0 \\ 0 & \text{if } \hat{\boldsymbol{w}}_i^t[j] = 0 \end{cases}$$
$\nabla\hat{\boldsymbol{w}}_i^t \leftarrow \left(\hat{\boldsymbol{w}}_i^t - \hat{\boldsymbol{w}}_i^{t-1}\right) \odot \boldsymbol{m}_{rec}$
Initialize $\mathcal{D}_0' \leftarrow \mathcal{N}(0,1)$, $y_0' \leftarrow Randint(0, max(y))$
for $s \leftarrow 0$ **to** $S-1$ **do**
  $\nabla\boldsymbol{w}_s' \leftarrow \partial\ell(F(\mathcal{D}_s', \hat{\boldsymbol{w}}_i^t), y_s')/\partial\hat{\boldsymbol{w}}_i^t$
  $\mathcal{L}_s' \leftarrow 1 - \frac{\nabla\hat{\boldsymbol{w}}_i^t \cdot \nabla\boldsymbol{w}_s'}{\|\nabla\hat{\boldsymbol{w}}_i^t\|\|\nabla\boldsymbol{w}_s'\|} + \tau$
  $\mathcal{D}_{s+1}' \leftarrow \mathcal{D}_s' - \eta\nabla_{\mathcal{D}_s'}\mathcal{L}_s'$, $y_{s+1}' \leftarrow y_s' - \eta\nabla_{y_s'}\mathcal{L}_s'$
end for
Return $\mathcal{D}_S', y_S'$

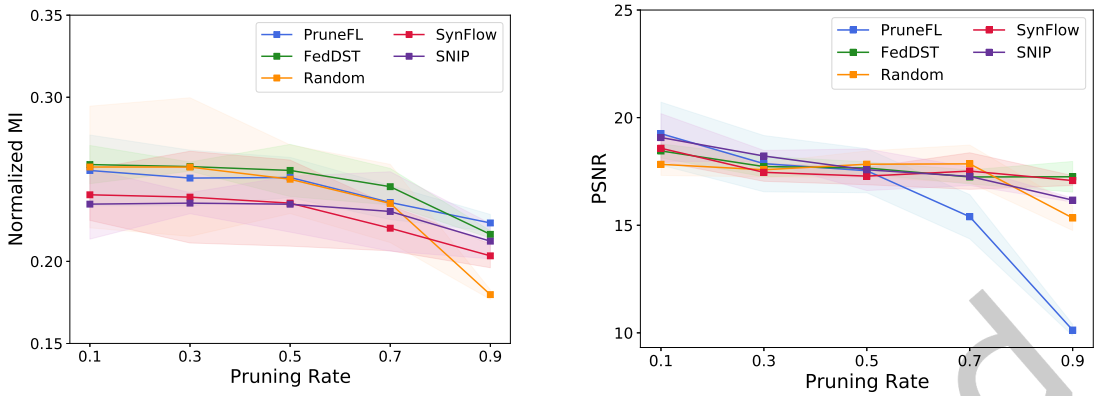
This shows that increasing the number of global training rounds (T) leads to a proportional rise in the upper bound of information leakage from the user's local training model.

Importance of the Upper Bounds: The derived upper bounds are important for several reasons.

- They provide a theoretical upper limit to the amount of information that can potentially be leaked from model updates in pruned models. To the best of our knowledge, this has not been previously characterized for pruned models.
- Because they are information-theoretical in nature, these bounds are universally applicable. They apply to *any* learning task, *any* data distribution, *any* pruning scheme, and *any* privacy attack. The attacker is the server that sees the pruned updates $\{\boldsymbol{w}_i^t \odot \bar{\boldsymbol{m}}_i^t\}_{t\in[T]}$ from client i, and infers information about the local training Data $\mathcal{D}_i$ of that client. Regardless of the exact reconstruction strategy, the inferred information is at most the mutual information $\mathbf{I}_i$ between the update and the local data, quantified in Equation 3. The derivation of the bounds assumes the existence of a pruning mask but does not depend on how this mask was generated, *e.g.,* whether through magnitude-based pruning, gradient-based pruning, or any other pruning method.
- Theorem 1 expresses the bound in terms of important parameters (e.g., the pruning rate p, model size d, etc), which provides insights into the choices and parameters that affect the privacy provided by pruning in practical algorithms. However, the bounds may or may not be tight compared to the actual information leakage, depending on the real-world scenario. In fact, these bounds are necessarily conservative since they are universal. This is why, in the next section, we complement the theoretical analysis with experimental evaluations in specific settings (i.e., considering specific base pruning, attack and defense schemes, specific data sets, and learning tasks).

4.2 Empirical Quantification

We perform an experimental evaluation to quantify the exact amount of privacy loss in concrete settings. This allows us to compare the theoretical bounds to experimental results and show that they are qualitatively aligned. It also allows us to obtain insights (in Section 5.1) that inform the design of defense pruning (in Section 5.2). More evaluation results are provided in Section 6.



(a) Normalized Mutual Information (NMI).

(b) Peak Signal-to-Noise Ratio (PSNR).

Fig. 2. Impact of varying pruning rate on privacy leakage using key metrics: Normalized Mutual Information (NMI) displayed on (a) and Peak Signal-to-Noise Ratio (PSNR) shown on (b) under a Sparse Gradient Inversion attack (SGI) in FEMNIST for five pruning methods under varying the base pruning rate.

4.2.1 Privacy Metrics. To quantify the extent of privacy leakage, we use the Normalized Mutual Information (NMI) between the training data and the reconstructed data. For two vectors $U, V \in \mathcal{R}^N$, NMI is computed as follows. The entropy of U is calculated by $H(U) = -\sum_{i=1}^{|U|} P(i)\log(P(i))$, where $P(i) = |U_i|/N$. The Mutual Information between U and V is given by $I(U;V) = \sum_{i=1}^{|U|}\sum_{j=1}^{|V|} \frac{|U_i \cap V_j|}{N} \log\left(\frac{N|U_i \cap V_j|}{|U_i||V_j|}\right)$. NMI is then measured as $\frac{I(U;V)}{\text{mean}(H(U),H(V))}$.

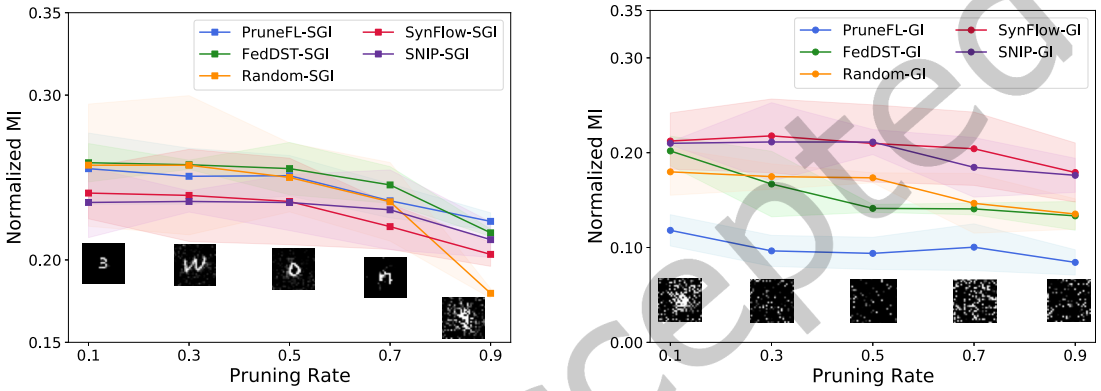
For image data, prior research [9] has shown that NMI aligns well with the Peak Signal-to-Noise Ratio (PSNR). This is also the case here: we see an agreement between NMI and PSNR with increasing pruning rates in Fig. 2 and Fig. A1, both showing a similar decreasing trend. In the rest of the paper, we use both NMI and PSNR as metrics to quantify the success of a reconstruction attack. Intuitively, the higher the NMI, the higher the privacy leakage, the more successful the reconstruction attack. The higher the PSNR, the closer the original and reconstructed images and the more successful the attack. Hence, Fig. 2 and Fig. A1 compare the NMI and PSNR metrics and show that that increasing pruning rate reduces the success of the privacy attack.

4.2.2 Reconstruction Attack Algorithms. We first implement the classic Gradient Inversion (GI) attack [11] from the DLG attack family. However, due to the unique sparsity patterns introduced by model pruning in FL, we identify the potential for further optimizing this attack to specifically exploit the sparsity. Consequently, we introduce an advanced attack tailored for the context of model pruning in FL, the Sparse Gradient Inversion (SGI) attack. This attack is designed to recover the pruning mask within the pruned model. Moreover, its adaptability allows seamless integration with existing privacy attacks in FL, thereby amplifying their effectiveness.

The SGI attack is provided in Algorithm 2. In each iteration t, the SGI algorithm proceeds as follows: (i) Based on the user i' local model weights $\boldsymbol{w}_i^t$, the server attempts to recover the pruning mask of user i (Line 2). The recovered mask is then integrated into the calculation of gradient, allowing the server to obtain a gradient $\nabla\hat{\boldsymbol{w}}_i^t$ that closely approximates the true gradients of user i; (ii) The SGI attack randomly initializes a set of dummy data, comprising dummy inputs $\mathcal{D}_0'$ and dummy labels y_0'. (iii) After the dummy gradient $\boldsymbol{w}'$ is acquired, the server then updates the dummy data in the direction that minimizes the cosine distance between the dummy gradient and the gradient $\nabla\hat{\boldsymbol{w}}_i^t$.

Method	Pruning Criteria	Pruning Cycles	Pruning at Server	Pruning at Client
SNIP [21]	gradient-based	one shot	✓	✗
SynFlow [34]	gradient-based	one shot	✓	✗
Random Pruning	magnitude-based	iterative	✓	✓
FedDST [3]	magnitude-based	iterative	✓	✓
PruneFL [17]	magnitude-based	iterative	✓	✓

Table 1. Comparison for Evaluated Methods

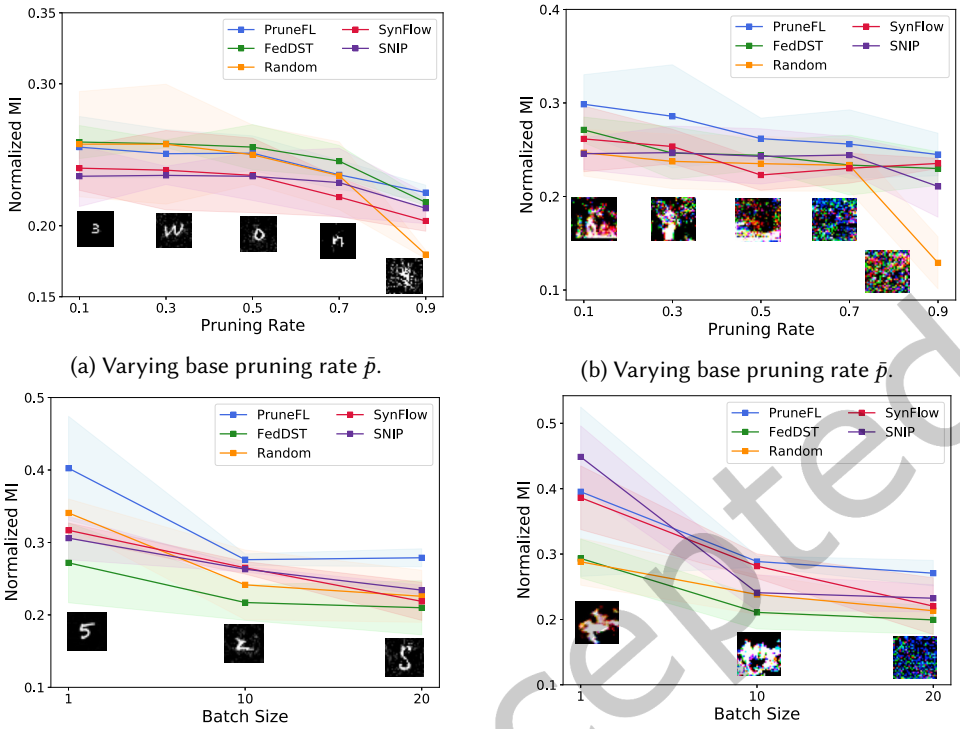


(a) SGI attack. (b) GI attack.

Fig. 3. Comparison between the SGI (Sparse Gradient Inversion) attack and the GI (Gradient Inversion) attack on the FEMNIST dataset for five pruning methods under varying the base pruning rate. In addition to the NMI metric, we show the corresponding recovered images, which show the degradation with an increasing pruning rate.

4.2.3 Base Pruning Schemes under Attack. We subject a set of well-established FL pruning methods to the privacy attack to assess their vulnerabilities and effectiveness in protecting user privacy. The pruning methods under evaluation include Random pruning, Snip [21], SynFlow [34], FedDST [3], and PruneFL [17], all of which are discussed in Related Work (Section 2). We present an overview of the pruning criteria, cycles, and schedules employed by the six evaluated methods. These specifics are elaborated in Table 1. By scrutinizing the table, we discern notable trends: three of the methods execute pruning exclusively at the server side, while the remaining methods engage in pruning activities at both the server and client sides. The latter approaches adopt an iterative pruning strategy throughout the course of the FL process.

We compare the effectiveness of the Sparse Gradient Inversion (SGI) attack with the Gradient Inversion attack under varying the base pruning rate. Our evaluation is based on the normalized mutual information (NMI) metric. Specifically, we analyze the NMI achieved by each attack for different pruning methods. Figure 3 demonstrates that the SGI attack consistently outperforms the Gradient Inversion attack by inferring a larger amount of information (as captured by NMI) at each pruning rate level, across all tested pruning methods. This is because the SGI attack exploits the sparsity inherent to model pruning, ultimately allowing better reconstruction attacks. In the rest of the paper, we use the SGI attack as a baseline, with a learning rate of 0.01 and 10,000 attack iterations.



(a) Varying base pruning rate $\bar{p}$.

(b) Varying base pruning rate $\bar{p}$.

(c) Varying the batch size B.

(d) Varying the batch size B.

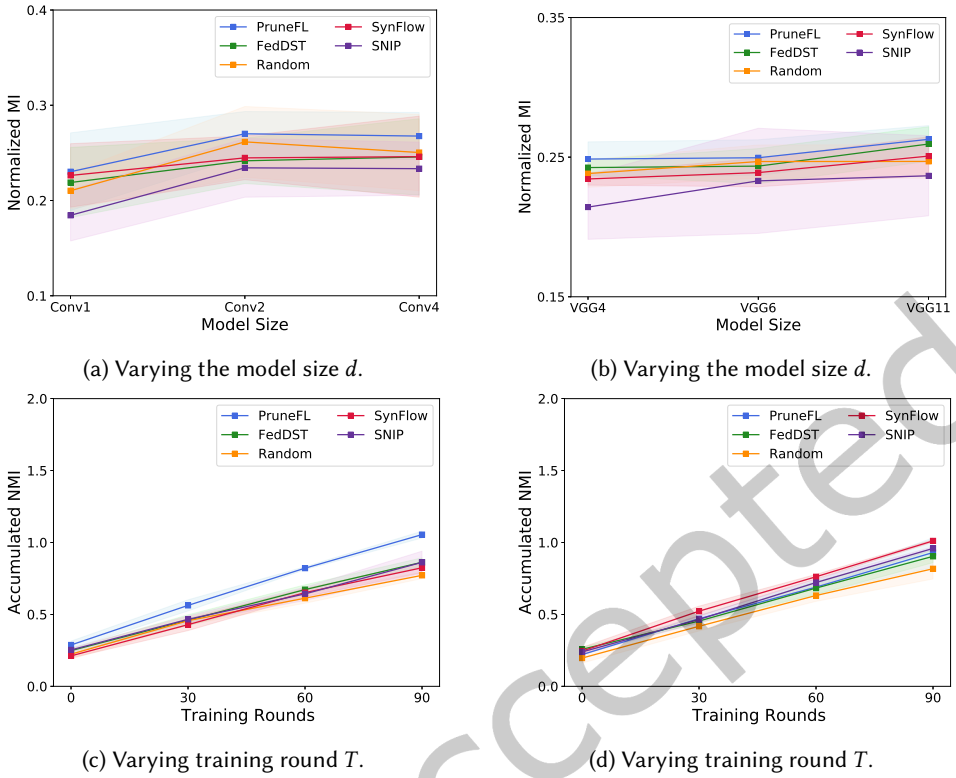
Fig. 4. Impact of varying base pruning rate $\bar{p}$ and batch size B on privacy leakage. We consider the Normalized Mutual Information (NMI) as privacy metric. We launch a Sparse Gradient Inversion (SGI) attack against 5 base pruning methods (Random, SNIP, SynFlow, FedDST, and PruneFL), over two datasets (FEMNIST in the First Column and CIFAR-10 in the Second Column).

4.2.4 Attack Performance. We launch the SGI attack on the aforementioned FL pruning schemes, on the FEMNIST and CIFAR10 datasets, to assess how FL parameters influence privacy leakage. The experimental results in Fig. 4 and Fig. 5 demonstrate that the impact of the parameters $\bar{p}$, B, d, and T qualitatively align with our theoretical findings.

Impact of Pruning Rate ($\bar{p}$). Fig. 4a and Fig. 4b illustrate that a higher pruning rate generally leads to a decrease in NMI. This is intuitively expected as pruning removes information and makes it more difficult to attack the original model. The effect is clearer for large pruning rates (*e.g.,* above 0.5). Image reconstruction is clearer in FEMNIST than CIFAR, consistently with prior work.

Impact of Batch Size (B). Fig. 4c and Fig. 4d show increasing the batch size B contributes to reducing information leakage from the local training model. Larger batch sizes enable more data to be added during model training, which can help obscure single data and increase privacy protection.

Impact of Model Size (d). We consider 3 architecture of the models for FEMNIST dataset: Conv-1, Conv-2, Conv-4, featuring 1, 2, 4 convolutional layers, along with 2 linear layers. These models encompass 626720, 6601504 and 13077392 parameters, respectively. We consider 3 different model architectures for CIFAR 10 dataset: VGG-4, VGG-6, and VGG-11. These models consist of 1 convolutional layer and 2 linear layers, 4 convolutional layers



(a) Varying the model size d.

(b) Varying the model size d.

(c) Varying training round T.

(d) Varying training round T.

Fig. 5. Impact of varying model size d and training rounds T on privacy leakage using Normalized Mutual Information (NMI) under a Sparse Gradient Inversion (SGI) attack for 5 base pruning methods (Random, SNIP, SynFlow, FedDST, and PruneFL), over two datasets (FEMNIST in the First Column and CIFAR-10 in the Second Column).

and 2 linear layers, and 8 convolutional layers with 3 linear layers, respectively. These models encompass 268464, 624048 and 9747136 parameters, respectively. Figure 5a and Figure 5b depicts the relationship between the model size (d) and the corresponding information leakage. The x-axis in the figure represents the number of layers in each model. Fig. 5a and Fig. 5b show the upper bounds on information leakage increase as d^* increases. Interestingly, the impact of model size d on information leakage is not linear. In the case of over-parameterised models, some parameters may not contribute to the model's performance or information retention. Therefore, increasing the size of such models may not proportionally affect information leakage.

Impact of Global Training Rounds (T). Fig. 5c and Fig. 5d show that increasing the global training rounds (T) leads to a proportional rise in the upper bound of information leakage. As the training progresses over multiple rounds, the user's model updates are repeatedly exposed to the central server. This repeated exposure increases the risk of potential memorization of private training information by the server.

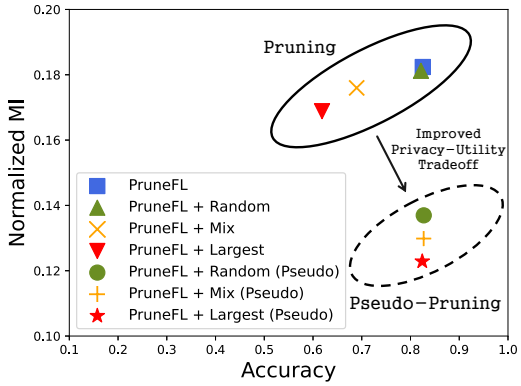


Fig. 6. Tradeoff of privacy vs. accuracy with three defense strategies: *Largest* (red), *Random* (green), and *Mix* (yellow), and their enhanced versions incorporating *Pseudo-Pruning* (*Pseudo*). *Pseudo* could improve the overall tradeoff by elevating model accuracy and *Largest* offers the highest level of privacy protection among all defense methods.

5 Design of Defense Pruning

5.1 Insights from Privacy Quantification

As shown above, pruning designed only with model accuracy and model size in mind, does not sufficiently protect against privacy attacks. Next, we describe three insights gained in the process of privacy quantification that inform our design of privacy-aware FL model pruning (*PriPrune* described in Section 5.2).

Insight 1: Largest Weights Matter. Pruning weights with large gradients improves privacy the most, but also hurts model accuracy the most. The intuition is that (base) pruning criteria, as used *e.g.,* in *PruneFL* and *FedDST*, avoid pruning model weights with the largest gradients, in order to preserve the most valuable information for the model. These gradients can then be exploited by gradient inversion attacks. Fig. 4a and Fig. 4b confirm empirically that, as more weights with large gradients are pruned, the attack is less successful.

This implies that the defense should strategically prune certain weights with large gradients, in addition to those pruned by the base pruning scheme. We explored three defense strategies for weights to prune: (1) weights with the top-k largest gradients (denoted as *Largest*), (2) random weight pruning (denoted as *Random*), and (3) a hybrid approach combining the *Largest* and *Random* (denoted as *Mix*). We compared the three strategies for the same defense pruning rate $\hat{p}_i$ (set 0.3) on top of the base method (*PruneFL*) and details are provided in Appendix C.1. The results in Fig. 6, within the Pruning oval, show the tradeoff achieved between privacy and accuracy: the "Largest" method achieves the best privacy and the worst model accuracy. This is because pruning weights with large gradients, removes valuable information, thus improving privacy but also significantly reducing model accuracy. Therefore, to improve this tradeoff and preserve privacy without harming accuracy, we need to introduce an additional mechanism.

Insight 2: *Pseudo-Pruning*. Users can prune weights with large magnitude gradients when they communicate with the server (which helps privacy) *and* still keep these weights for their local training (which maintains accuracy).

We refer to this idea as *Pseudo-Pruning*. and illustrate it in Fig. 7. Before sending the weights $\bar{\boldsymbol{w}}_i^t$ to the server user i conducts *Pseudo-Pruning* with defense mask $\hat{\boldsymbol{m}}_i^t$. The weights pruned out by defense pruning with large gradients $\tilde{\boldsymbol{w}}_i^t$ $(= \bar{\boldsymbol{w}}_i^t \odot \tilde{\boldsymbol{m}}_i^t)$ are withheld locally, thus preserving accuracy. Meanwhile, the weights remaining after the defense pruning without large gradients, $\hat{\boldsymbol{w}}_i^t$ $(= \bar{\boldsymbol{w}}_i^t \odot \hat{\boldsymbol{m}}_i^t)$, are transmitted to the server, thus preserving privacy.

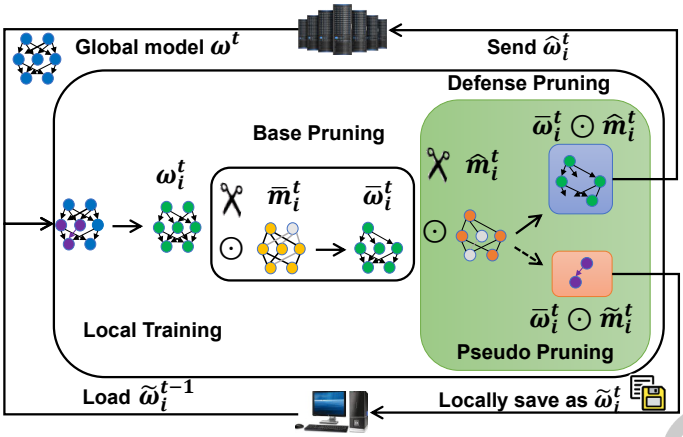


Fig. 7. Illustration of local pruning at user i with *Pseudo-Pruning* defense: First user applies base pruning with base mask $\bar{m}_i$ to derive a pruned model$\bar{w}_i^t$. Then, defense pruning is conducted with mask $\hat{m}_i$. Defense pruning is implemented as *Pseudo-Pruning*, where the remaining weights $\hat{w}_i^t$ ($= \bar{w}_i^t \odot \hat{m}_i^t$) are sent to the server; while the pruned weights $\tilde{w}_i^t$ ($= \bar{w}_i^t \odot \tilde{m}_i^t$) are locally saved for future local training. Notably, these weights pruned by defense are not actually pruned, hence the term "Pseudo-Pruning."

Here, $\tilde{m}$ is the bit-wise complement matrix of $\hat{m}$. In the next round, after user i receives the global model w^{t+1}, the locally saved weights $\tilde{w}_i^t$ are loaded into the global model, serving as the local initial model.

We combine *Pseudo-Pruning* with each of the three strategies (*Largest*, *Random*, and *Mix*) in Insight 1, and we show the evaluation results within the *Pseudo-Pruning* dashed oval in Fig. 6. Fig. 6 shows that the model accuracy is notably better with *Pseudo-Pruning* than real pruning, with the same mask. Furthermore, combining the two insights into *Largest (Pseudo)* offers the most effective privacy guarantee among all defenses discussed above. Hence, we adopt *Largest (Pseudo)* as our defense strategy. Additional evaluation details are available in Appendix C. Still, a remaining shortcoming of the defense considered so far, is that their pruning rate is manually selected and remains fixed.

Insight 3: Adapt.

The defense pruning rate ($\hat{p}_i$) should adapt to the model weights, to jointly optimize privacy and model accuracy. A fixed defense rate is unable to capture the evolving dynamics of the training process. The defense strategy should adapt to the changing dynamics of the training process by employing an adaptive defense pruning rate, enabling a more effective optimization of the tradeoff between privacy and accuracy.

5.2 The *PriPrune* Mechanism

Combining all aforementioned insights, we introduce the defense strategy, ***PriPrune***, which dynamically adapts the *Pseudo-Pruning* defense mask $\hat{m}$ to jointly optimize accuracy and privacy throughout the FL training process.

Loss Function. The objective of user i is to minimize a loss function that combines privacy and accuracy:

$$\mathcal{L}_{local} = \lambda_{acc}\mathcal{L}_{acc} + \lambda_{pri}\mathcal{L}_{pri} + \lambda_{sha} \sum_{l \in L, j \in d_l} \alpha_{i_{(l,j)}}^t \tag{16}$$

Here, $\mathcal{L}_{acc}$ represents the local model training loss, $\mathcal{L}_{pri}$ represents the local privacy loss, and $\sum_{l \in L, j \in d_l} \alpha_{i_{(l,j)}}^t$ serves as the privacy regularization term, where $\alpha_{i_{(l,j)}}^t$ represents the probability of user i not sharing the j-th parameter

in l-th layer with the server at round t. L represents the total number of layers in the model, and d_l refers to the number of parameters in the l-th layer. Additionally, λ_{acc}, λ_{pri}, and λ_{sha} act as weights for accuracy loss, privacy loss, and the privacy regularization term, respectively.

Next, we explain the terms of the loss function of the user in Equation 16 and their rationale. First, the accuracy loss is given by: $\mathcal{L}_{acc} = \mathcal{L}\left(\mathcal{D}_i; \hat{\boldsymbol{w}}_i^t | \boldsymbol{w}_{i,init}^t\right)$. The initial local weight is computed as: $\boldsymbol{w}_{i,init}^t = \boldsymbol{w}^t \odot \hat{\boldsymbol{m}}_i^{t-1} + \bar{\boldsymbol{w}}_i^{t-1} \odot \tilde{\boldsymbol{m}}_i^{t-1}$, where $\tilde{\boldsymbol{m}}$ denotes the bit-wise complement matrix of $\hat{\boldsymbol{m}}$. It is the composition of locally saved weights with global weights through the utilization of the defense mask $\hat{\boldsymbol{m}}$.

Then, the privacy loss $\mathcal{L}_{pri}$ is given by:

$$\mathcal{L}_{pri} = \sum_{l \in L, j \in d_l} -\frac{N_l}{\sum\limits_{l \in L} N_l} \frac{|\boldsymbol{g}_{l,j}|}{\sum\limits_{j \in d_l} |\boldsymbol{g}_{l,j}|} \log \boldsymbol{\alpha}_{i_{(l,j)}}^t \tag{17}$$

N_l indicates the number of weights in l-th layer, $|\boldsymbol{g}_{l,j}|$ represents the magnitude of the gradient for the j-th parameter in l-th layer. The privacy loss is designed to discourage sharing weights with large gradients inspired by Insight 1, aiming to mitigate information leakage. Therefore, we assign higher weights to parameters with larger gradients in the loss function. Through this process, during the optimization of the privacy loss, the algorithm updates a higher $\boldsymbol{\alpha}_i^t$ associated with the larger weight term. This higher $\boldsymbol{\alpha}_i^t$ signifies a higher probability of not sharing the parameters with larger gradients with the server, thus promoting the sharing of less valuable information with the server. Moreover, we also distribute constraints to l-th layer based on its parameter count N_l, in proportion to the total model weights, which ensures uniform layer-wise sparsity.

The term $\sum\limits_{l \in L, j \in d_l} \boldsymbol{\alpha}_{i_{(l,j)}}^t$ introduces a privacy penalty based on the magnitude of $\boldsymbol{\alpha}_i^t$. The optimization process aims to minimize the sum of $\boldsymbol{\alpha}_i^t$, resulting in a decrease in the overall probability of not sharing parameters. This, in turn, leads to an increase in the total probability of sharing parameters. Thus, while the privacy loss discourages the sharing of parameters with larger gradients with the server, this term encourages the sharing of parameters with smaller gradients, thus preserving accuracy while maintaining privacy.

Optimization. In *PriPrune*, we optimize the defense mask $\hat{\boldsymbol{m}}$ and model weights $\bar{\boldsymbol{w}}_i$ jointly through gradient descent based on our designed loss function. To overcome the discrete and non-differentiable nature of the defense mask $\hat{\boldsymbol{m}}$, we employ Gumbel-Softmax sampling [16, 25] to substitute the original non-differentiable sample with a differentiable sample. Specifically, $\hat{\boldsymbol{m}}$ is parameterized by a distribution vector $\boldsymbol{\pi}_{l,j} = [\boldsymbol{\alpha}_{i_{(l,j)}}^t, 1 - \boldsymbol{\alpha}_{i_{(l,j)}}^t]$. $\boldsymbol{v}_{l,j}$ represents the soft *Pseudo-Pruning* decision for the j-th parameter in layer l. The reparameterization trick is used for differentiable training:

$$\boldsymbol{v}_{l,j}(k) = \frac{\exp\left((\log \boldsymbol{\pi}_{l,j}(k) + G_{l,j}(k))/\tau\right)}{\sum\limits_{k \in \{0,1\}} \exp\left((\log \boldsymbol{\pi}_{l,j}(k) + G_{l,j}(k))/\tau\right)} \tag{18}$$

where $G_{l,j} = -\log\left(-\log U_{l,j}\right)$ is Gumbel distribution with $U_{l,j}$ sampled from a uniform i.i.d. distribution $\mathrm{Unif}(0, 1)$, τ is the temperature of the softmax and $k \in \{0, 1\}$. We apply the hard sample trick introduced by PyTorch document [29] to acquire the defense mask $\hat{\boldsymbol{m}}$ and the main trick is to do $y_{hard} - f_{StopGradient}(y_{soft}) + y_{soft}$.

PriPrune. The *PriPrune* algorithm is descibed in Algorithm 3, which updates the *UserUpdate* function of Algorithm 1. As depicted in Algorithm 3, in each FL round, after user i receives the latest global model $\boldsymbol{w}_i^t$ from the server, they sample their defense mask $\hat{\boldsymbol{m}}_i$ using Equation 18. The defense mask $\hat{\boldsymbol{m}}_i$ is then employed to combine the global model $\boldsymbol{w}^t$ and the user's locally saved $\tilde{\boldsymbol{w}}_i^{t-1}$ to generate the local initial model $\boldsymbol{w}_{i,init}^t$. Following this composition step, the user initiates local training based on Equation 16, leading to updates in $\bar{\boldsymbol{w}}_i^t$, $\tilde{\boldsymbol{w}}_i^t$, and $\boldsymbol{\alpha}_i^t$.

Algorithm 3 *PriPrune*

function *UserUpdate* $(\boldsymbol{w}^t, \boldsymbol{m}^t, i)$*:*
Load $\tilde{\boldsymbol{w}}_i^t$ *from local storage of user* i
Initialize $\boldsymbol{\alpha}_i^t = \boldsymbol{\alpha}_i^{t-1}$
for *local epoch* $e \in E$ ***do***
  Sample defense mask $\hat{\boldsymbol{m}}_i^t$ *by Equation 18 with* $\boldsymbol{\alpha}_i^t$
  Combine local and global model by:

$$\boldsymbol{w}_i^t[j]_{j\in d} = \begin{cases} \tilde{\boldsymbol{w}}_i^t[j] & \text{if } \hat{\boldsymbol{m}}_i^t[j] = 0 \\ \boldsymbol{w}_i^t[j] & \text{if } \hat{\boldsymbol{m}}_i^t[j] = 1 \end{cases}$$

  Base Pruning: $\bar{\boldsymbol{w}}_i^t \leftarrow \boldsymbol{w}_i^t \odot \bar{\boldsymbol{m}}_i^t$
  $\mathcal{L}_{local} \leftarrow Equation16.$
  Model parameter update:

$$\bar{\boldsymbol{w}}_i^t \leftarrow \bar{\boldsymbol{w}}_i^t - \eta \nabla \mathcal{L}_{local}(\bar{\boldsymbol{w}}_i^t, \boldsymbol{\alpha}_i^t)$$

  Defense mask parameter update:

$$\boldsymbol{\alpha}_i^t \leftarrow \boldsymbol{\alpha}_i^t - \eta \nabla \mathcal{L}_{local}(\bar{\boldsymbol{w}}_i^t, \boldsymbol{\alpha}_i^t)$$

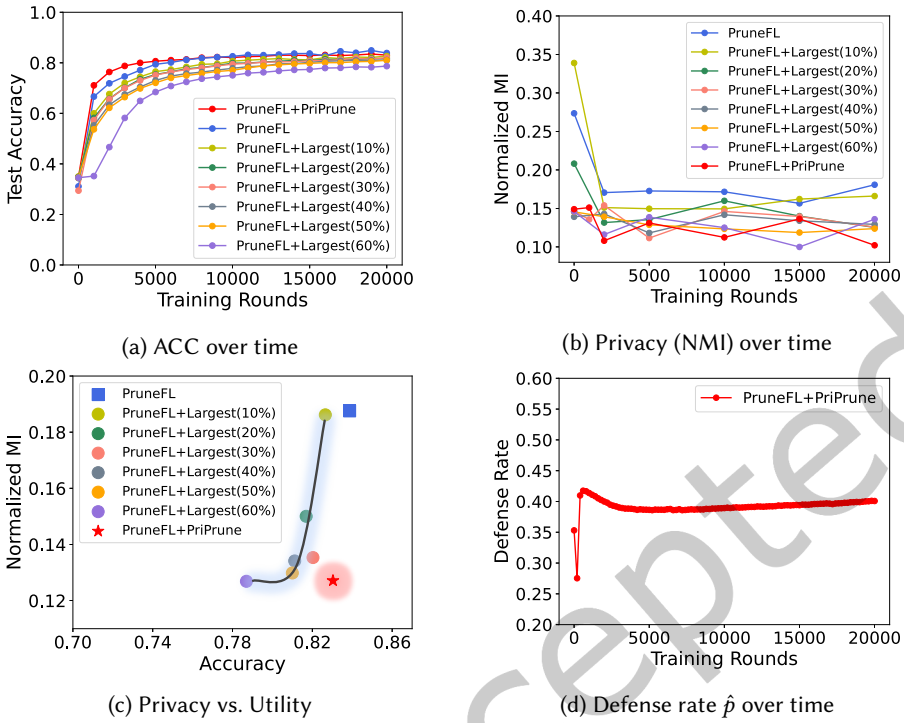
end for
$\hat{\boldsymbol{m}}_i^t \leftarrow argmax(\boldsymbol{\alpha}_i^t, 1 - \boldsymbol{\alpha}_i^t)$
Defense Pruning: $\hat{\boldsymbol{w}}_i^t \leftarrow \bar{\boldsymbol{w}}_i^t \odot \hat{\boldsymbol{m}}_i^t, \tilde{\boldsymbol{w}}_i^t \leftarrow \bar{\boldsymbol{w}}_i^t \odot \tilde{\boldsymbol{m}}_i^t$
Return $\hat{\boldsymbol{w}}_i^t$ *to server, locally save* $\tilde{\boldsymbol{w}}_i^t$

Then user i conducts defense pruning, retaining the non-shared weights $\tilde{\boldsymbol{w}}_i^t$ locally, and transmits the shared weights $\hat{\boldsymbol{w}}_i^t$ to the server.

6 Experimental Evaluation

6.1 Experimental Setup

Datasets and Models. We evaluate models Conv-2 [5] and VGG-11 [31]) along with their corresponding datasets FEMNIST [5] and CIFAR-10 [19]), which are commonly employed in FL studies.
Implementation details. In each round of training, we randomly select 10 clients from a pool of 193 users for the FEMNIST and 100 clients for the CIFAR-10. The training process involves 20,000 iterations, with a batch size of 20 for the FEMNIST and CIFAR-10. We employ 1 local training and initialize the learning rate at 0.25 with the base pruning rate set to 0.3. The details regarding the data split, resource and configurations are provided in Appendix B.1.
Performance Metrics. To assess the effectiveness of the evaluated pruning methods, we employ the privacy metrics and utility metrics. For privacy assessment, we use the Normalized Mutual Information (NMI) between the training data and the reconstructed data as the privacy metric. Higher values of NMI indicate a higher risk of privacy leakage. Additionally, the Peak Signal-to-Noise Ratio (PSNR) is employed as a well-established metric for image quality evaluation. The comparison of NMI and PSNR can be found in Fig. 2. See Appendix B.2 for additional evaluation results. The utility metric relies on model accuracy (ACC) to gauge model performance. Ideally, we want improved privacy, without significantly compromising accuracy.
Hyper-parameter Selection. We conducted a search for our main parameter, the defense rate $\hat{p}_i$, from 0.1 to 0.6, as depicted in Fig. 8c. With regards to our hyperparameters in loss functions, we search λ_{acc} in the list of [1, 5, 10], λ_{pri} in the list of [1, 10, 15] and λ_{sha} in the list of [2e-06, 2e-05, 2e-04]. Detailed results are deferred to Appendix C.3.



(a) ACC over time

(b) Privacy (NMI) over time

(c) Privacy vs. Utility

(d) Defense rate $\hat{p}$ over time

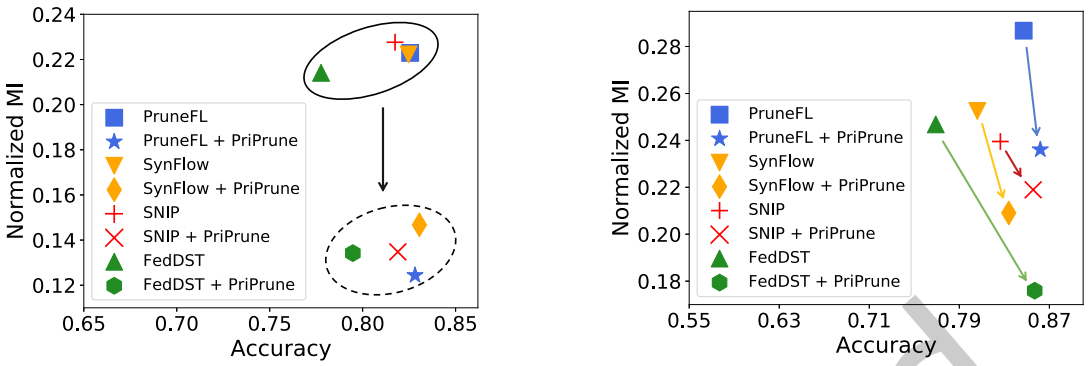
Fig. 8. Performance evaluation of *PriPrune* with adaptive defense rate and *Largest (Pseudo)* with 6 fixed defense rates, as defense strategies, integrated with PruneFL (as the base pruning method), conducted over 20,000 training rounds across the FEMNIST dataset. (a) Comparison of ACC over training rounds under these different defense strategies. (b) Comparison of Privacy (NMI) over training rounds under these different defense strategies. (c) The Privacy-Utility tradeoff, measured in terms of NMI and ACC, respectively.(d) Change in the adaptive defense rate of *PriPrune* over training rounds.

6.2 Performance of *PriPrune*

Recall that in Section 5.1, Insight 1, we initially assessed three defense strategies atop the base pruning method, namely: *Random*, *Largest*, *Random and Mix*. The results in Figure 6 indicate that while these defense strategies enhance privacy, they also degrade utility. Consequently, in Insight 2, we introduce *Pseudo-Pruning* to optimize this tradeoff. We evaluate the aforementioned three defense strategies both with and without *Pseudo-Pruning*, revealing that their integration with *Pseudo-Pruning* leads to improved accuracy and enhanced privacy. Notably, the *Largest + Pseudo-Pruning* strategy achieves the optimal balance between privacy and utility, thus emerging as our preferred defense mechanism. In Insight 3, we tackle the challenge of a fixed defense rate by proposing *PriPrune*, which incorporates *Largest + Pseudo-Pruning* (*Largest (Pseudo)*) with an adaptive defense rate. The evaluation of the privacy improvement achieved by each individual insight, in Section 5.1, essentially provides an ablation study. In this section, we conduct comprehensive experiments to meticulously evaluate the performance of the entire *PriPrune* across various pruning mechanisms and datasets.

First, we present a performance comparison between the *Largest (Pseudo)* with 6 fixed defense rates and *PriPrune* with the adaptive defense rate, both are integrated with PruneFL (as the base pruning method).

Comparable Utility. Fig. 8a illustrates how the utility metrics, ACC, change over FL rounds. It shows the convergence curve of *PriPrune*, with its model accuracy closely aligned with that of PruneFL. This alignment



(a) *PriPrune* in FEMNIST (b) *PriPrune* in CIFAR-10

Fig. 9. Comparison of tradeoffs between privacy-vs-accuracy among four base pruning mechanisms, both with and without the integration of *PriPrune*, across the FEMNIST and CIFAR-10 datasets.

indicates that *PriPrune* maintains utility comparable to PruneFL. However, for the *Largest (Pseudo)* defense, increasing its fixed defense rate from 10% to 60% results in decreased utility, rendering it unable to match the ACC achieved by PruneFL.

Enhanced Privacy. Figure 8b illustrates the change in the privacy metric NMI across FL rounds. It shows that *PriPrune* consistently maintains a low NMI throughout all FL training rounds, significantly outperforming PruneFL. Moreover, its NMI is comparable to that of *Largest (Pseudo)* with a fixed defense rate of 60%.

Optimal Tradeoff. Figure 8c illustrates the comparison of the privacy and utility tradeoffs among the mentioned defenses. It shows that increasing the defense rate of *Largest (Pseudo)* enhances privacy protection but at the expense of model accuracy. Conversely, *PriPrune* achieves the optimal tradeoff between privacy and accuracy. This is attributed to the adaptive nature of *PriPrune*'s defense rate $\hat{p}$, which adjusts to the model weights and is optimized jointly for privacy and model accuracy. The dynamic change of the defense rate of *PriPrune* is depicted in Fig. 8d.

Communication Efficiency and Training Speed. We conduct experiments to compare communication and training in PruneFL with and without *PriPrune*; details are deferred to Appendix C.2. When *PriPrune* is applied, the number of model parameters is reduced by 32.20% ± 0.91% for FEMNIST and by 15.51% ± 0.17% for CIFAR-10, thus significantly reducing the communication cost between clients and the server. This benefit is achieved because of *Pseudo-Pruning*, which retains pseudo-pruned weights locally, but transmits fewer model updates. When *PriPrune* is applied, the training time is increased by 0.18s ± 0.01s for FEMNIST and by 0.50s ± 0.01s for CIFAR-10, this is a negligible increase in the training speed (due to updating the defense pruning rate).

We have demonstrated that *PriPrune* stands out as the best defense solution among other strategies. Subsequently, we present the privacy performance of *PriPrune* when integrated with various state-of-the-art base pruning schemes in FL, such as SNIP, SynFlow, FedDST, and PruneFL, across both the FEMNIST and CIFAR-10 datasets.

***PriPrune* with State-of-the-Art Pruned FL.** Figure 9 shows the results of using *PriPrune* for defense after a number of state-of-the-art base pruning schemes. Across all base pruning methods, this integration consistently enhances privacy while preserving accuracy. This highlight *PriPrune*'s effectiveness and versatility in achieving an optimal tradeoff between privacy and accuracy in diverse scenarios. This outcome is primarily due to the nature of *PriPrune*, which not only actually prunes but also selectively retains certain critical weights for local training (through *Pseudo-Pruning*). This mechanism allows the model to maintain or even improve accuracy while

reducing the information leakage that could be exploited in privacy attacks, as also shown in Fig.6. Specifically, *PriPrune* focuses on *Pseudo-Pruning* weights with large gradients that are most likely to leak sensitive information and keep them for local training. This selective pruning can lead to a model that is more robust against privacy attacks without a significant loss in accuracy. Additionally, the adaptive pruning rate in *PriPrune* is optimized for the tradeoff between privacy and accuracy, allowing the model to balance these two objectives effectively.

	FEMNIST		CIFAR-10	
Methods	Base	Base + *PriPrune*	Base	Base + *PriPrune*
PruneFL	0.22 ± 0.008	**0.12** ± 0.007	0.29 ± 0.009	**0.24** ± 0.009
FedDST	0.21 ± 0.025	**0.13** ± 0.007	0.25 ± 0.008	**0.18** ± 0.006
SynFlow	0.22 ± 0.009	**0.15** ± 0.005	0.25 ± 0.012	**0.21** ± 0.018
SNIP	0.23 ± 0.012	**0.13** ± 0.014	0.24 ± 0.004	**0.22** ± 0.008

Table 2. Comparison of privacy levels (NMI) between Base and Base + *PriPrune* for FL pruning methods. A lower NMI indicates a reduced risk of privacy leakage.

While Fig. 9 compares the tradeoffs between privacy-vs-accuracy, Table 2 further elaborates on the privacy improvement achieved by using *PriPrune* as defense after any base scheme, reporting results over three experiments. Compared to the base approach alone, adding *PriPrune* as defense consistently reduces NMI, thus improving privacy. For instance, on the FEMNIST dataset, *PriPrune* improves the privacy of *PruneFL* by 45.5% without compromising accuracy; on the CIFAR-10 dataset, *PriPrune* improves privacy of *FedDST* by 28% while maintaining accuracy.

7 Conclusion

In this paper, we revisit federated learning with model pruning, from the point of view of privacy. First, we quantify the information leakage and privacy gain offered for model pruning in FL, both theoretically and experimentally. Inspired by insights obtained from our extensive privacy quantification, we design *PriPrune*– an adaptive privacy-preserving local pruning mechanism in FL, that is jointly optimized for privacy and model performance. Our proposed mechanism achieves the best tradeoff between privacy and accuracy across diverse pruning methods and datasets under privacy attacks. One direction for our future work is to combine our pruning-based defense with classic, orthogonal defenses in FL such as differential privacy and secure aggregation, to enhance privacy further.

Acknowledgment

Tianyue Chu and Nikolaos Laoutaris were supported by the MLEDGE project (REGAGE22e00052829516), funded by the Ministry of Economic Affairs and Digital Transformation and the European Union-NextGenerationEU/PRTR. Mengwei Yang and Athina Markopoulou were supported by NSF awards 1956393 and 1900654, and a UC Noyce Foundation gift. This work started during a long-term visit of Tianyue to the ProperData project, also supported by NSF Award 1956393

References

[1] Philip Bachman, R Devon Hjelm, and William Buchwalter. 2019. Learning Representations by Maximizing Mutual Information Across Views. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc., Vancouver Convention Centre, Canada. https://proceedings.neurips.cc/paper_files/paper/2019/file/ddf354219aac374f1d40b7e760ee5bb7-Paper.pdf

[2] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. 2018. Mutual Information Neural Estimation. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, Stockholmsmässan, Stockholm Sweden, 531–540. https://proceedings.mlr.press/v80/belghazi18a.html

[3] Sameer Bibikar, Haris Vikalo, Zhangyang Wang, and Xiaohan Chen. 2022. Federated Dynamic Sparse Training: Computing Less, Communicating Less, Yet Learning Better. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 6 (6 2022), 6080–6088. https://doi.org/10.1609/aaai.v36i6.20555

[4] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. 2017. Practical Secure Aggregation for Privacy-Preserving Machine Learning. In *CCS '17: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (Dallas, Texas, USA) *(CCS '17)*. Association for Computing Machinery, New York, NY, USA, 1175–1191. https://doi.org/10.1145/3133956.3133982

[5] Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečnỳ, H Brendan McMahan, Virginia Smith, and Ameet Talwalkar. 2018. Leaf: A Benchmark for Federated Settings.

[6] Tianlong Chen, Jonathan Frankle, Shiyu Chang, Sijia Liu, Yang Zhang, Zhangyang Wang, and Michael Carbin. 2020. The lottery ticket hypothesis for pre-trained bert networks. *Advances in neural information processing systems(NeurIPS)* 33 (2020), 15834–15846.

[7] Min Du, Ruoxi Jia, and Dawn Song. 2020. Robust anomaly detection and backdoor attack detection via differential privacy. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, Addis Ababa, Ethiopia. https://openreview.net/forum?id=SJx0q1rtvS

[8] Cynthia Dwork. 2006. Differential privacy. In *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II 33*. Springer, Berlin, Heidelberg, 1–12.

[9] Ahmed Roushdy Elkordy, Jiang Zhang, Yahya H. Ezzeldin, Konstantinos Psounis, and Salman Avestimehr. 2023. How Much Privacy Does Federated Learning with Secure Aggregation Guarantee? *Proceedings on Privacy Enhancing Technologies (PoPETs)* 2023, 1 (2023), 510–526. https://doi.org/10.56553/POPETS-2023-0030

[10] Jonathan Frankle and Michael Carbin. 2019. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, New Orleans, LA, USA. https://openreview.net/forum?id=rJl-b3RcF7

[11] Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. 2020. Inverting gradients-how easy is it to break privacy in federated learning? *Advances in Neural Information Processing Systems(NeurIPS)* 33 (2020), 16937–16947.

[12] Antonious Girgis, Deepesh Data, Suhas Diggavi, Peter Kairouz, and Ananda Theertha Suresh. 2021. Shuffled Model of Differential Privacy in Federated Learning. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 130)*, Arindam Banerjee and Kenji Fukumizu (Eds.). PMLR, Virtual Event, 2521–2529. https://proceedings.mlr.press/v130/girgis21a.html

[13] Song Han, Huizi Mao, and William J. Dally. 2016. Deep Compression: Compressing Deep Neural Network with Pruning, Trained Quantization and Huffman Coding. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). OpenReview.net, San Juan, Puerto Rico. http://arxiv.org/abs/1510.00149

[14] Song Han, Jeff Pool, John Tran, and William J. Dally. 2015. Learning both weights and connections for efficient neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1* (Montreal, Canada) *(NIPS'15)*. MIT Press, Cambridge, MA, USA, 1135–1143.

[15] Alexander Herzog, Robbie Southam, Ioannis Mavromatis, and Aftab Khan. 2024. FedMap: Iterative Magnitude-Based Pruning for Communication-Efficient Federated Learning.

[16] Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical Reparameterization with Gumbel-Softmax. In *International Conference on Learning Representations*. OpenReview.net, Toulon, France. https://openreview.net/forum?id=rkE3y85ee

[17] Yuang Jiang, Shiqiang Wang, Víctor Valls, Bong Jun Ko, Wei-Han Lee, Kin K. Leung, and Leandros Tassiulas. 2023. Model Pruning Enables Efficient Federated Learning on Edge Devices. *IEEE Trans. Neural Networks Learn. Syst.* 34, 12 (2023), 10374–10386. https://doi.org/10.1109/TNNLS.2022.3166101

[18] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. 2021. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning* 14, 1–2 (2021), 1–210.

[19] Alex Krizhevsky, Geoffrey Hinton, et al. 2009. *Learning Multiple Layers of Features from Tiny Images*. Technical Report. University of Toronto, Toronto, ON, Canada. https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf

[20] Yann LeCun, John S. Denker, and Sara A. Solla. 1989. Optimal Brain Damage. In *Advances in Neural Information Processing Systems 2, [NIPS Conference, Denver, Colorado, USA, November 27-30, 1989]*, David S. Touretzky (Ed.). Morgan Kaufmann, Denver, Colorado, USA, 598–605. http://papers.nips.cc/paper/250-optimal-brain-damage

[21] Namhoon Lee, Thalaiyasingam Ajanthan, and Philip H. S. Torr. 2019. Snip: single-Shot Network Pruning based on Connection sensitivity. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, Louisiana, USA, May 6-9, 2019*. OpenReview.net, New Orleans, Louisiana. https://openreview.net/forum?id=B1VZqjAcYX

[22] Ang Li, Jingwei Sun, Binghui Wang, Lin Duan, Sicheng Li, Yiran Chen, and Hai Li. 2020. Lotteryfl: Personalized and communication-efficient federated learning with lottery ticket hypothesis on non-iid datasets. arXiv:2008.03371

[23] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. 2020. Federated Learning: Challenges, Methods, and Future Directions. *IEEE Signal Processing Magazine* 37, 3 (2020), 50–60. https://doi.org/10.1109/MSP.2020.2975749

[24] Xiaolong Ma, Geng Yuan, Xuan Shen, Tianlong Chen, Xuxi Chen, Xiaohan Chen, Ning Liu, Minghai Qin, Sijia Liu, Zhangyang Wang, et al. 2021. Sanity checks for lottery tickets: Does your winning ticket really win the jackpot? *Advances in Neural Information Processing Systems(NeurIPS)* 34 (2021), 12749–12760.

[25] Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. 2017. The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, Toulon, France. https://openreview.net/forum?id=S1jE5L5gl

[26] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA (Proceedings of Machine Learning Research, Vol. 54)*, Aarti Singh and Xiaojin (Jerry) Zhu (Eds.). PMLR, Fort Lauderdale, FL, USA, 1273–1282. http://proceedings.mlr.press/v54/mcmahan17a.html

[27] H. Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. 2018. Learning Differentially Private Recurrent Language Models. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, Vancouver, BC, Canada. https://openreview.net/forum?id=BJ0hF1Z0b

[28] Jun-Hyung Park, Yeachan Kim, Junho Kim, Joon-Young Choi, and SangKeun Lee. 2023. Dynamic Structure Pruning for Compressing CNNs. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, Brian Williams, Yiling Chen, and Jennifer Neville (Eds.). AAAI Press, Washington, DC, USA, 9408–9416. https://doi.org/10.1609/AAAI.V37I8.26127

[29] Pytorch. n.d.. GUMBEL SOFTMAX In Pytorch documentation. https://pytorch.org/docs/stable/generated/torch.nn.functional.gumbel_softmax.html. [Online].

[30] Yifan Shi, Kang Wei, Li Shen, Jun Li, Xueqian Wang, Bo Yuan, and Song Guo. 2024. Efficient Federated Learning With Enhanced Privacy via Lottery Ticket Pruning in Edge Computing. *IEEE Transactions on Mobile Computing* 23, 10 (Feb. 2024), 9946–9958. https://doi.org/10.1109/TMC.2024.3370967

[31] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). OpenReview.net, San Diego, CA, USA. http://arxiv.org/abs/1409.1556

[32] Jinhyun So, Ramy E. Ali, Basak Güler, Jiantao Jiao, and Amir Salman Avestimehr. 2023. Securing Secure Aggregation: Mitigating Multi-Round Privacy Leakage in Federated Learning. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI Washington, DC, USA, February 7-14, 2023*, Brian Williams, Yiling Chen, and Jennifer Neville (Eds.). AAAI Press, Washington, DC, USA, 9864–9873. https://doi.org/10.1609/AAAI.V37I8.26177

[33] Jinhyun So, Chaoyang He, Chien-Sheng Yang, Songze Li, Qian Yu, Ramy E Ali, Basak Guler, and Salman Avestimehr. 2022. Lightsecagg: a lightweight and versatile design for secure aggregation in federated learning. *Proceedings of Machine Learning and Systems* 4 (2022), 694–720.

[34] Hidenori Tanaka, Daniel Kunin, Daniel L. K. Yamins, and Surya Ganguli. 2020. Pruning neural networks without any data by iteratively conserving synaptic flow. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '20)*. Curran Associates Inc., Red Hook, NY, USA, Article 535, 13 pages.

[35] MTCAJ Thomas and A Thomas Joy. 2006. Elements of information theory.

[36] Aleksei Triastcyn and Boi Faltings. 2019. Federated Learning with Bayesian Differential Privacy. In *2019 IEEE International Conference on Big Data (IEEE BigData)*, Chaitanya K. Baru, Jun Huan, Latifur Khan, Xiaohua Hu, Ronay Ak, Yuanyuan Tian, Roger S. Barga, Carlo Zaniolo, Kisung Lee, and Yanfang (Fanny) Ye (Eds.). IEEE, Los Angeles, CA, USA, 2587–2596. https://doi.org/10.1109/BIGDATA47090.2019.9005465

[37] Georgia Tsaloli, Bei Liang, Carlo Brunetta, Gustavo Banegas, and Aikaterini Mitrokotsa. 2021. sf DEVA: Decentralized, Verifiable Secure Aggregation for Privacy-Preserving Learning. In *Information Security - 24th International Conference, ISC 2021, Virtual Event, November 10-12, 2021, Proceedings (Lecture Notes in Computer Science, Vol. 13118)*, Joseph K. Liu, Sokratis K. Katsikas, Weizhi Meng, Willy Susilo, and Rolly Intan (Eds.). Springer, Virtual Event, 296–319. https://doi.org/10.1007/978-3-030-91356-4_16

[38] Chaoqi Wang, Guodong Zhang, and Roger B. Grosse. 2020. Picking Winning Tickets Before Training by Preserving Gradient Flow. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, Addis Ababa, Ethiopia. https://openreview.net/forum?id=SkgsACVKPH

[39] Danye Wu, Miao Pan, Zhiwei Xu, Yujun Zhang, and Zhu Han. 2020. Towards Efficient Secure Aggregation for Model Update in Federated Learning. In *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*. IEEE, Taipei, Taiwan, 1–6. https://doi.org/10.1109/GLOBECOM42002.2020.9347960

[40] Xiyuan Yang, Wenke Huang, and Mang Ye. 2023. Dynamic personalized federated learning with adaptive differential privacy. *Advances in Neural Information Processing Systems* 36 (2023), 72181–72192.

[41] Hongxu Yin, Arun Mallya, Arash Vahdat, Jose M. Alvarez, Jan Kautz, and Pavlo Molchanov. 2021. See through Gradients: Image Batch Recovery via GradInversion. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Los Alamitos, CA, USA, 16332–16341. https://doi.org/10.1109/CVPR46437.2021.01607

[42] Kai Yue, Richeng Jin, Chau-Wai Wong, Dror Baron, and Huaiyu Dai. 2023. Gradient Obfuscation Gives a False Sense of Security in Federated Learning. In *32nd USENIX Security Symposium (USENIX Security 23)*. USENIX Association, Anaheim, CA, 6381–6398. https://www.usenix.org/conference/usenixsecurity23/presentation/yue

[43] Xinwei Zhang, Xiangyi Chen, Mingyi Hong, Zhiwei Steven Wu, and Jinfeng Yi. 2022. Understanding clipping for federated learning: Convergence and client-level differential privacy. In *International Conference on Machine Learning, ICML 2022*, Vol. 162. PMLR, Baltimore, Maryland, USA, 26048–26067.

[44] Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. 2020. idlg: Improved deep leakage from gradients. arXiv:2001.02610 [cs.LG]

[45] Ligeng Zhu, Zhijian Liu, and Song Han. 2019. *Deep leakage from gradients*. Curran Associates Inc., Red Hook, NY, USA, 11.

[46] Zhanxing Zhu, Jingfeng Wu, Bing Yu, Lei Wu, and Jinwen Ma. 2019. The Anisotropic Noise in Stochastic Gradient Descent: Its Behavior of Escaping from Sharp Minima and Regularization Effects. *Proceedings of the 36th International Conference on Machine Learning* 97 (09–15 Jun 2019), 7654–7663. https://proceedings.mlr.press/v97/zhu19e.html

Appendix

This appendix extends the main paper, providing supplemental materials, including additional details and results, which could not be included in the main paper, due to lack of space.

A Proof of Theorem

This section provides the proof of Theorem 4.1 in Section 4.1.

A.1 Proof

Equation 2 is the definition of mutual information, which is:

$$\begin{aligned}\mathbf{I}(X;Y) &= \int_{X\times\mathcal{Y}} log\frac{d\mathbb{P}_{X,Y}(x,y)}{d\mathbb{P}_X(x)\otimes\mathbb{P}_Y(y)}d\mathbb{P}_{X,Y}(x,y)\\ &= \mathbf{H}(X) - \mathbf{H}(X|Y)\\ &= \mathbf{H}(X) + \mathbf{H}(Y) - \mathbf{H}(X,Y)\end{aligned}$$

Hence, Equation 3 can be written as:

$$
\begin{aligned}
\mathbf{I}\left(\mathcal{D}_i;\left\{\boldsymbol{w}_i^t\odot\bar{\boldsymbol{m}}_i^t\right\}_{t\in[T]}\right) &\overset{(a)}{\leq} \sum_{t=1}^{T}\mathbf{I}\left(\mathcal{D}_i;\boldsymbol{w}_i^t\odot\bar{\boldsymbol{m}}_i^t\middle|\left\{\boldsymbol{w}_i^k\odot\bar{\boldsymbol{m}}_i^k\right\}_{k\in[t-1]}\right)\\
&\overset{(b)}{=} \sum_{t=1}^{T}\left(\mathbf{I}\left(\boldsymbol{w}_i^t;\boldsymbol{w}_i^t\odot\bar{\boldsymbol{m}}_i^t\middle|\left\{\boldsymbol{w}_i^k\odot\bar{\boldsymbol{m}}_i^k\right\}_{k\in[t-1]}\right)\right)+\sum_{t=1}^{T}\left(\mathbf{I}\left(\mathcal{D}_i;\boldsymbol{w}_i^t\odot\bar{\boldsymbol{m}}_i^t\middle|\boldsymbol{w}_i^t,\left\{\boldsymbol{w}_i^k\odot\bar{\boldsymbol{m}}_i^k\right\}_{k\in[t-1]}\right)\right)\\
&\quad-\sum_{t=1}^{T}\left(\mathbf{I}\left(\mathcal{D}_i;\boldsymbol{w}_i^t\middle|\left\{\boldsymbol{w}_i^k\odot\bar{\boldsymbol{m}}_i^k\right\}_{k\in[t]}\right)\right)\\
&\overset{(c)}{\leq} \sum_{t=1}^{T}\left(\mathbf{I}\left(\boldsymbol{w}_i^t;\boldsymbol{w}_i^t\odot\bar{\boldsymbol{m}}_i^t\middle|\left\{\boldsymbol{w}_i^k\odot\bar{\boldsymbol{m}}_i^k\right\}_{k\in[t-1]}\right)\right)\\
&= \sum_{t=1}^{T}\left(\mathbf{H}\left(\boldsymbol{w}_i^t\middle|\left\{\boldsymbol{w}_i^k\odot\bar{\boldsymbol{m}}_i^k\right\}_{k\in[t-1]}\right)\right)+\sum_{t=1}^{T}\left(\mathbf{H}\left(\boldsymbol{w}_i^t\odot\bar{\boldsymbol{m}}_i^t\middle|\left\{\boldsymbol{w}_i^k\odot\bar{\boldsymbol{m}}_i^k\right\}_{k\in[t-1]}\right)\right)\\
&\quad-\sum_{t=1}^{T}\left(\mathbf{H}\left(\boldsymbol{w}_i^t,\boldsymbol{w}_i^t\odot\bar{\boldsymbol{m}}_i^t\middle|\left\{\boldsymbol{w}_i^k\odot\bar{\boldsymbol{m}}_i^k\right\}_{k\in[t-1]}\right)\right)\\
&\overset{(d)}{=} \sum_{t=1}^{T}\left(\mathbf{I}\left(\mathrm{x}_i^t;\boldsymbol{w}_i^t\odot\bar{\boldsymbol{m}}_i^t\middle|\left\{\boldsymbol{w}_i^k\odot\bar{\boldsymbol{m}}_i^k\right\}_{k\in[t-1]}\right)\right)
\end{aligned}
\tag{19}
$$

Where (a) and (b) are from the chain rule; (c) is from data processing inequality $\mathcal{D}_i\rightarrow\boldsymbol{w}_i^t\rightarrow\boldsymbol{w}_i^t\odot\bar{m}_i^t$ that $\mathbf{I}\left(\mathcal{D}_i;\boldsymbol{w}_i^t\odot\bar{m}_i^t\middle|\boldsymbol{w}_i^t,\left\{\boldsymbol{w}_i^k\odot\bar{m}_i^k\right\}_{k\in[t-1]}\right)=0$; (d) is from zero conditional entropy and invariant of mutual information, where

$$
\mathrm{x}_i^t=\frac{\partial\ell_i(\bar{\boldsymbol{w}}_i^{t-1};\mathcal{D}_i)}{\partial\boldsymbol{w}}=\frac{1}{B}\sum_{b\in\mathbf{B}_i^t}g_i\left(\boldsymbol{w}_i^t,b\right)
\tag{20}
$$

b denotes the size of the random samples.

Therefore, the privacy leakage for a single round t is Equation 5:

$$
\mathbf{I}\left(\mathrm{x}_i^t;\boldsymbol{w}_i^t\odot\bar{\boldsymbol{m}}_i^t\middle|\left\{\boldsymbol{w}_i^k\odot\bar{\boldsymbol{m}}_i^k\right\}_{k\in[t-1]}\right)
$$

We first prove the upper bound for Equation 5.

Based on Equation 2, we have

$$
\begin{aligned}
\mathbf{I}\left(\mathrm{x}_i^t;\boldsymbol{w}_i^t\odot\bar{\boldsymbol{m}}_i^t\middle|\left\{\boldsymbol{w}_i^k\odot\bar{\boldsymbol{m}}_i^k\right\}_{k\in[t-1]}\right) &= \mathbf{H}\left(\mathrm{x}_i^t\middle|\left\{\boldsymbol{w}_i^k\odot\bar{\boldsymbol{m}}_i^k\right\}_{k\in[t-1]}\right)+\mathbf{H}\left(\boldsymbol{w}_i^t\odot\bar{\boldsymbol{m}}_i^t\middle|\left\{\boldsymbol{w}_i^k\odot\bar{\boldsymbol{m}}_i^k\right\}_{k\in[t-1]}\right)\\
&\quad-\mathbf{H}\left(\mathrm{x}_i^t,\boldsymbol{w}_i^t\odot\bar{\boldsymbol{m}}_i^t\middle|\left\{\boldsymbol{w}_i^k\odot\bar{\boldsymbol{m}}_i^k\right\}_{k\in[t-1]}\right)\\
&\leq \underbrace{\mathbf{H}\left(\boldsymbol{w}_i^t\odot\bar{\boldsymbol{m}}_i^t\middle|\left\{\boldsymbol{w}_i^k\odot\bar{\boldsymbol{m}}_i^k\right\}_{k\in[t-1]}\right)}_{\mathbf{A}}+\underbrace{\mathbf{H}\left(\mathrm{x}_i^t\middle|\left\{\boldsymbol{w}_i^k\odot\bar{\boldsymbol{m}}_i^k\right\}_{k\in[t-1]}\right)}_{\mathbf{B}}
\end{aligned}
$$

For part **A**, based on the chain rule of entropy, we have

$$\begin{aligned} 0 \leq \mathbf{H}\left(\boldsymbol{w}_i^t, \bar{\boldsymbol{m}}_i^t, \boldsymbol{w}_i^t \odot \bar{\boldsymbol{m}}_i^t \middle| \left\{\boldsymbol{w}_i^k \odot \bar{\boldsymbol{m}}_i^k\right\}_{k\in[t-1]}\right) &= \mathbf{H}\left(\boldsymbol{w}_i^t, \bar{\boldsymbol{m}}_i^t \middle| \left\{\boldsymbol{w}_i^k \odot \bar{\boldsymbol{m}}_i^k\right\}_{k\in[t-1]}\right) + \mathbf{H}\left(\boldsymbol{w}_i^t \odot \bar{\boldsymbol{m}}_i^t \middle| \boldsymbol{w}_i^t, \bar{\boldsymbol{m}}_i^t, \left\{\boldsymbol{w}_i^k \odot \bar{\boldsymbol{m}}_i^k\right\}_{k\in[t-1]}\right) \\ &= \mathbf{H}\left(\boldsymbol{w}_i^t \odot \bar{\boldsymbol{m}}_i^t \middle| \left\{\boldsymbol{w}_i^k \odot \bar{\boldsymbol{m}}_i^k\right\}_{k\in[t-1]}\right) + \mathbf{H}\left(\boldsymbol{w}_i^t, \bar{\boldsymbol{m}}_i^t \middle| \left\{\boldsymbol{w}_i^k \odot \bar{\boldsymbol{m}}_i^k\right\}_{k\in[t]}\right) \end{aligned} \tag{21}$$

Due to $\mathbf{H}\left(\boldsymbol{w}_i^t \odot \bar{\boldsymbol{m}}_i^t \middle| \boldsymbol{w}_i^t, \bar{\boldsymbol{m}}_i^t, \left\{\boldsymbol{w}_i^k \odot \bar{\boldsymbol{m}}_i^k\right\}_{k\in[t-1]}\right) = 0$, and $\mathbf{H}\left(\boldsymbol{w}_i^t, \bar{\boldsymbol{m}}_i^t \middle| \left\{\boldsymbol{w}_i^k \odot \bar{\boldsymbol{m}}_i^k\right\}_{k\in[t]}\right) \geq 0$, we have

$$\begin{aligned} \mathbf{H}\left(\boldsymbol{w}_i^t \odot \bar{\boldsymbol{m}}_i^t \middle| \left\{\boldsymbol{w}_i^k \odot \bar{\boldsymbol{m}}_i^k\right\}_{k\in[t-1]}\right) &\leq \mathbf{H}\left(\boldsymbol{w}_i^t, \bar{\boldsymbol{m}}_i^t \middle| \left\{\boldsymbol{w}_i^k \odot \bar{\boldsymbol{m}}_i^k\right\}_{k\in[t-1]}\right) \\ &\overset{(a)}{\leq} \mathbf{H}\left(\boldsymbol{w}_i^t \middle| \left\{\boldsymbol{w}_i^k \odot \bar{\boldsymbol{m}}_i^k\right\}_{k\in[t-1]}\right) + \mathbf{H}\left(\bar{\boldsymbol{m}}_i^t \middle| \left\{\boldsymbol{w}_i^k \odot \bar{\boldsymbol{m}}_i^k\right\}_{k\in[t-1]}\right) \\ &\overset{(b)}{=} \mathbf{H}\left(\mathrm{x}_i^t \middle| \left\{\boldsymbol{w}_i^k \odot \bar{\boldsymbol{m}}_i^k\right\}_{k\in[t-1]}\right) + 1 - \frac{1}{2\ln 2}\sum_{n=1}^{\infty}\frac{(2\bar{p}_i - 1)^{2n}}{n(2n-1)} \\ &\leq \underbrace{\mathbf{H}\left(\mathrm{x}_i^t \middle| \left\{\boldsymbol{w}_i^k \odot \bar{\boldsymbol{m}}_i^k\right\}_{k\in[t-1]}\right)}_{\mathbf{B}} + 1 - \frac{\bar{p}_i - 1}{2\ln 2} \end{aligned}$$

(a) is from conditioning reduces entropy; (b) is from zero conditional entropy and the Taylor series of the binary entropy function in a neighbourhood of 0.5 with the base pruning rate $\bar{p}_i$.

For part **B**,

$$\mathbf{H}\left(\mathrm{x}_i^t \middle| \left\{\boldsymbol{w}_i^k \odot \bar{\boldsymbol{m}}_i^k\right\}_{k\in[t-1]}\right) = \mathbf{H}\left(\frac{1}{B}\sum_{b\in\mathbf{B}_i^t} g_i\left(\boldsymbol{w}_i^t, b\right) \middle| \left\{\boldsymbol{w}_i^k \odot \bar{\boldsymbol{m}}_i^k\right\}_{k\in[t-1]}\right)$$

Let $X = \sum_{b\in\mathbf{B}_i^t} g_i\left(\boldsymbol{w}_i^t, b\right)$, $Z = \left\{\boldsymbol{w}_i^k \odot \bar{\boldsymbol{m}}_i^k\right\}_{k\in[t-1]}$, $Y = \frac{1}{B}X$. $f_{X,Z}(x,z)$ and $f_{Y,Z}(y,z)$ are the corresponding joint probability density functions.

Then part **B** can be written as

$$\begin{aligned} \mathbf{H}\left(Y|Z\right) &= -\int_{\mathcal{Y},\mathcal{Z}} f_{Y,Z}\left(y,z\right)\log\frac{f_{Y,Z}(y,z)}{f_Z(z)}\,\mathrm{d}y\,\mathrm{d}z = -\int_{\mathcal{X},\mathcal{Z}} |B|\, f_{X,Z}\left(By,z\right)\log|B|\,\frac{f_{X,Z}(By,z)}{f_Z(z)}\,\mathrm{d}y\,\mathrm{d}z \\ &= -\int_{\mathcal{X},\mathcal{Z}} f_{X,Z}\left(x,z\right)\log\frac{f_{X,Z}(x,z)}{f_Z(z)}\,\mathrm{d}x\,\mathrm{d}z + log\frac{1}{|B|} = \underbrace{\mathbf{H}\left(\sum_{b\in\mathbf{B}_i^t} g_i\left(\boldsymbol{w}_i^t, b\right) \middle| \left\{\boldsymbol{w}_i^k \odot \bar{\boldsymbol{m}}_i^k\right\}_{k\in[t-1]}\right)}_{\mathbf{C}} + \log\frac{1}{B} \end{aligned}$$

In recent theoretical results for analyzing the behaviour of SGD, they approximate the SGD vector by a distribution with independent components or by a multivariate Gaussian vector.

We define $\hat{g}_i\left(\boldsymbol{w}_i^t, b\right) \in \mathbb{R}^{d^*}$ is the largest sub-vector of the $g_i\left(\boldsymbol{w}_i^t, b\right)$ with non-singular convariance matrices, where $d^* \leq d$. Based on the ZCA whitening transformation and Assumption 1, we have $\hat{g}_i\left(\boldsymbol{w}_i^t, b\right) = \Sigma_i^{-\frac{1}{2}}\boldsymbol{z}$, where $\boldsymbol{z}$ has zero mean and $\mathbb{I}_{d^*}$ convariance matrix.

Then for the part **C**, we set $\boldsymbol{v} = \sum_{b\in\mathbf{B}_i^t} \boldsymbol{z}$ and $\boldsymbol{u} = \Sigma_i^{-\frac{1}{2}}\boldsymbol{v}$, then we have:

$$\mathbf{C} = \mathbf{H}\left(\sum_{b\in\mathbf{B}_i^t} \hat{g}_i\left(\boldsymbol{w}_i^t, b\right)\middle|\left\{\boldsymbol{w}_i^k \odot \bar{\boldsymbol{m}}_i^k\right\}_{k\in[t-1]}\right) = \mathbf{H}\left(\Sigma_i^{-\frac{1}{2}}\sum_{b\in\mathbf{B}_i^t}\boldsymbol{z}\middle|\left\{\boldsymbol{w}_i^k \odot \bar{\boldsymbol{m}}_i^k\right\}_{k\in[t-1]}\right) = -\int \mathbb{P}_{\mathcal{U}}(\boldsymbol{u})\log\mathbb{P}_{\mathcal{U}}(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u}$$

$$\overset{(a)}{=} -\int \frac{\mathbb{P}_{\mathcal{V}}(\boldsymbol{v})}{\left|\det\left(\Sigma_i^{-\frac{1}{2}}\right)\right|}\log\left(\frac{\mathbb{P}_{\mathcal{V}}(\boldsymbol{v})}{\left|\det\left(\Sigma_i^{-\frac{1}{2}}\right)\right|}\right)\left|\det\left(\Sigma_i^{-\frac{1}{2}}\right)\right|\mathrm{d}\boldsymbol{v} = \log\left|\det\left(\Sigma_i^{-\frac{1}{2}}\right)\right| + \underbrace{\mathbf{H}\left(\sum_{b\in\mathbf{B}_i^t}\boldsymbol{z}\middle|\left\{\boldsymbol{w}_i^k \odot \bar{\boldsymbol{m}}_i^k\right\}_{k\in[t-1]}\right)}_{\mathbf{D}} \tag{22}$$

(a) is based on the transformation of random vectors. Based on the Maximum Entropy Upper Bound [35], which the continuous distribution X with prescribed variance $\mathcal{V}(\mathrm{X})$ maximizing the entropy is the Gaussian distribution of same variance. Since the entropy of a multivariate Gaussian distribution with mean μ and covariance $K_{i,j}$ is

$$\mathbf{h}(\mathcal{N}_n(\mu, K)) = \frac{1}{2}\log\left(2\pi e\right)^n |K|$$

where $|K|$ denotes the determinant of K. Hence, the maximum entropy upper bound for part **D** is

$$\mathbf{H}\left(\sum_{b\in\mathbf{B}_i^t}\boldsymbol{z}\middle|\left\{\boldsymbol{w}_i^k \odot \bar{\boldsymbol{m}}_i^k\right\}_{k\in[t-1]}\right) \le \frac{1}{2}\log\left(2\pi e\right)^{d^*}|\mathbb{I}_{d^*}| = \frac{d^*}{2}\log\left(2\pi e\right) \tag{23}$$

Therefore, by combining Equation 22 and Equation 23, the upper bound for part **B** is

$$\mathbf{H}\left(\mathrm{x}_i^t\middle|\left\{\boldsymbol{w}_i^k \odot \bar{\boldsymbol{m}}_i^k\right\}_{k\in[t-1]}\right) \le \log\frac{1}{B} + \log\left|\det\left(\Sigma_i^{-\frac{1}{2}}\right)\right| + \frac{d^*}{2}\log\left(2\pi e\right) \tag{24}$$

After summing part **A** and part **B**, we derive the upper bound for the privacy leakage for the single round, which is Equation 5:

$$\mathbf{I}\left(\mathrm{x}_i^t; \boldsymbol{w}_i^t \odot \bar{\boldsymbol{m}}_i^t\middle|\left\{\boldsymbol{w}_i^k \odot \bar{\boldsymbol{m}}_i^k\right\}_{k\in[t-1]}\right) \le 1 - \frac{\bar{p}_i - 1}{2\ln 2} + 2\log\frac{1}{B} + 2\log\left|\det\left(\Sigma_i^{-\frac{1}{2}}\right)\right| + d^*\log\left(2\pi e\right)$$

Based on Equation 4, we have the upper bound for objective function Equation 3, standing how much information the aggregated model over T global training rounds could leak about the private data,

$$\mathbf{I}_i \le T - \frac{T(\bar{p}_i - 1)}{2\ln 2} + 2T\log\frac{1}{B} + 2T\log\left|\det\left(\Sigma_i^{-\frac{1}{2}}\right)\right| + d^*T\log\left(2\pi e\right)$$

For PruneFL, the upper bound for the single-round leakage is

$$\begin{aligned}\mathbf{I}_i^t &\le 1 + \frac{\bar{\mathcal{P}} \cup \mathcal{A}}{2d\ln 2} + 2\log\frac{1}{B} + 2\log\left|\det\left(\Sigma_i^{-\frac{1}{2}}\right)\right| + d^*\log\left(2\pi e\right)\\ &\le 1 + \frac{\bar{\mathcal{P}}}{2d\ln 2} + 2\log\frac{1}{B} + 2\log\left|\det\left(\Sigma_i^{-\frac{1}{2}}\right)\right| + d^*\log\left(2\pi e\right)\end{aligned} \tag{25}$$

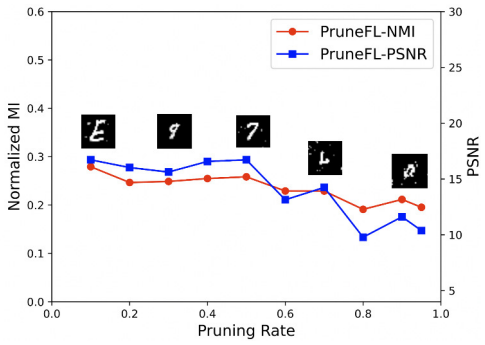


Fig. A1. Impact of varying pruning rate in FEMNIST on privacy leakage using key metrics: Normalized Mutual Information (NMI) displayed on the left y-axis and Peak Signal-to-Noise Ratio (PSNR) shown on the right y-axis under a Gradient Inversion attack for PruneFL.

B Privacy Attacks

B.1 Extra Implementation Details

Our algorithms are implemented by Pytorch and we implement experiments on two NVIDIA RTX A5000 and two Xeon Silver 4316. Across all the examined methods, we employ 1 local training and initialize the learning rate to 0.25 and The baseline pruning rate is set at 0.3 for all datasets.

Data Partitioning. Our data partitioning methodology is aligned with the FEMNIST dataset setting, utilizing its inherent 193-user partition. As for the CIFAR-10 dataset, we divided it into 100 equal-sized, non-overlapping users, following the methodology used in PruneFL [17].

B.2 Evaluation on both NMI and PSNR Metrics

To quantify the extent of privacy leakage, we utilized the Normalized Mutual Information (NMI) metric, which has been demonstrated in prior research to align well with the Peak Signal-to-Noise Ratio (PSNR) [9] as an effective measure of privacy leakage.

We examine the relationship between NMI and PSNR. The observed trend, depicted in Fig. A1, illustrates a consistent decrease in both NMI and PSNR metrics with higher pruning rates. This result is consistent with previous research, confirming the reliability of NMI as a measure of privacy leakage, which is comparable to the widely accepted PSNR metric.

B.3 Image Reconstructions under Privacy Attacks

Figure A2 shows the recovered images under the SGI (Sparse Gradient Inversion) attack in the FEMNIST dataset, for different base pruning methods and base pruning rates.

C Defense Details

C.1 Results of Insights

Within the scope of the three defense methods in the Insights from Evaluation Section, specific configurations have been established. These settings are detailed below. For all three defense methods, the total pruning rate ($\hat{p}$) is uniformly set at 0.3, mirroring the original PruneFL's pruning rate.

For each defense method:

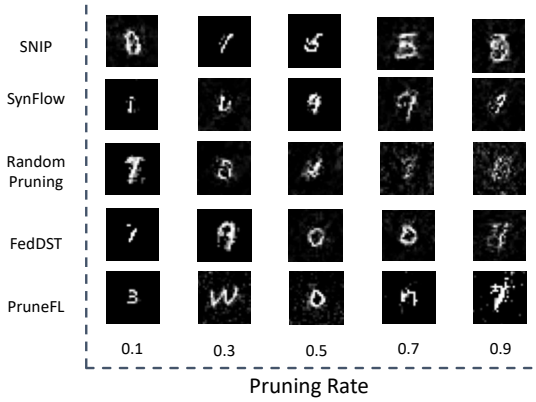
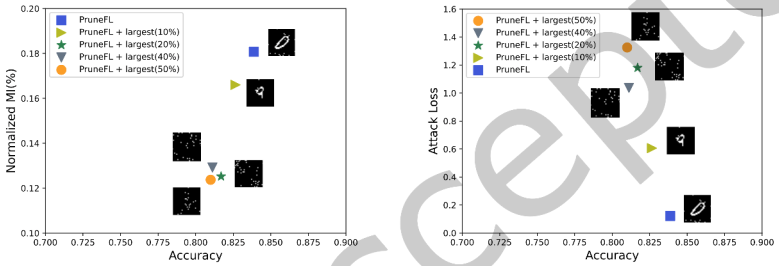


Fig. A2. Impact of varying pruning rate on image reconstruction in FEMNIST using different base pruning methods.



(a) ACC vs NMI (b) ACC vs Attack Loss

Fig. A3. Comparison of Privacy-Accuracy Trade-Off with Varying Defense Rate (0% to 50%) in Largest Method based on PruneFL using NMI and Attack Loss Privacy Metrics

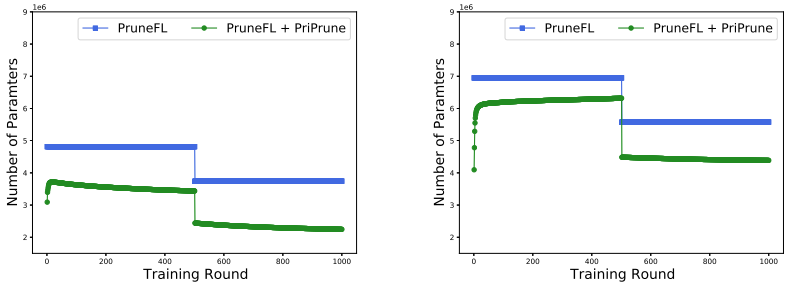
- PruneFL + Largest: The defense strategy involves pruning based on the weights with top-k largest gradients, with the defense pruning rate set to $\hat{p}_{largest} = 0.3$.
- PruneFL + Random: The defense strategy involves random pruning, with the defense pruning rate ($\hat{p}$) set to $\hat{p}_{random} = 0.3$.
- PruneFL + Mix: This is a hybrid approach that combines both the largest gradient-based pruning and random pruning. The defense pruning rate is determined as the sum of $\hat{p}_{largest} = 0.15$ and $\hat{p}_{random} = 0.15$.

Figure A3 shows Privacy-Accuracy Trade-Off with Varying Defense Rate (0% to 50%) in Largest Method based on PruneFL using NMI and Attack Loss Privacy Metrics.

The outcomes are indicative of a notable trend: as more weights associated with larger gradients are pruned, privacy is enhanced at the expense of accuracy. This substantiates our **Insight 1**: Largest Weights Matter. Pruning weights with large gradients improves privacy the most but also hurts model accuracy the most.

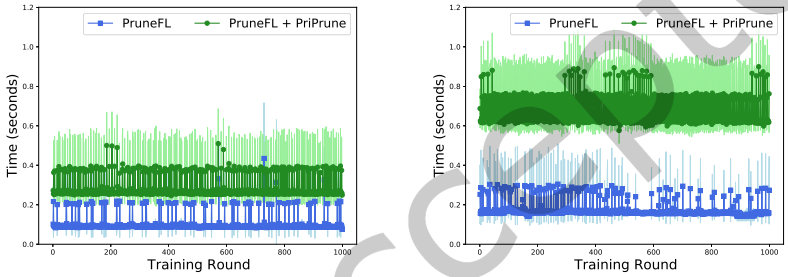
C.2 Communication Efficiency and Training Speed

Figure A4 demonstrates the communication cost in terms of the number of model parameters across 1000 iterations of FL. In *PriPrune*, we employ *Pseudo-Pruning*, where the pseudo-pruned weights are retained locally for subsequent training rounds. Thus, *PriPrune* directly reduces communication costs by transmitting fewer



(a) Conv-2 in FEMNIST (b) VGG-11 in CIFAR-10

Fig. A4. Comparison of the communication costs of PruneFL with and without *PriPrune* during the training process with (a) representing FL training using the Conv-2 model on the FEMNIST dataset and (b) the VGG-11 model on the CIFAR-10 dataset.



(a) Training speed in FEMNIST (b) Training speed in CIFAR-10

Fig. A5. Comparison of the training speed of PruneFL with and without *PriPrune* during the training process with (a) representing FL training on the FEMNIST dataset and (b) on the CIFAR-10 dataset.

model parameters. Figure A4 shows a significant reduction in the number of model parameters communicated between clients and the server with the application of *PriPrune*.

Figure A5 displays the training speed access through the total time of clients' local training across 1000 iterations of FL. *PriPrune* involves additional steps of updating the defense pruning rate, which is computationally lightweight. Figure A5 demonstrates that *PriPrune* introduces only a minor overhead and the additional steps brought by *PriPrune* has a minimal impact on training speed.

C.3 Hyper-parameter Selection

Table A1 has searched λ_{acc} in the list of [1, 5, 10], λ_{pri} in the list of [1, 10, 15] and λ_{sha} in the list of [2e-06, 2e-05, 2e-04], which showing the comparisons of privacy levels during the hyper-parameter search. As shown in Table A1, when we increase the λ_{pri}, the NMI value becomes lower, indicating better privacy protection. We explore different combinations of hyper-parameters and finally utilize the best set of values that could achieve the best privacy-accuracy tradeoff.

For different pruning methods and different datasets, we are adopting different combinations of hyper-parameters so as to achieve better privacy-utility tradeoff, as detailed in Table A2.

	$\lambda_{pri} = 1$			$\lambda_{pri} = 10$			$\lambda_{pri} = 15$		
Parameter	$\lambda_{sha} = 2e-04$	$\lambda_{sha} = 2e-05$	$\lambda_{sha} = 2e-06$	$\lambda_{sha} = 2e-04$	$\lambda_{sha} = 2e-05$	$\lambda_{sha} = 2e-06$	$\lambda_{sha} = 2e-04$	$\lambda_{sha} = 2e-05$	$\lambda_{sha} = 2e-06$
$\lambda_{acc} = 1$	0.166	0.171	0.156	0.138	0.153	0.154	0.138	0.159	0.151
$\lambda_{acc} = 5$	0.154	0.158	0.136	0.125	0.121	0.123	0.135	0.123	0.130
$\lambda_{acc} = 10$	0.148	0.142	0.140	0.132	0.126	0.118	0.135	0.132	0.140

Table A1. Comparison of privacy levels (NMI) during hyper-parameter selection: with regards to our hyperparameters in loss functions, we search λ_{acc} in the list of [1, 5, 10], λ_{pri} in the list of [1, 10, 15] and λ_{sha} in the list of [2e-06, 2e-05, 2e-04].

Dataset	Method	λ_{pri}	λ_{sha}	λ_{acc}
FEMNIST	PruneFL+*PriPrune*	15	2×10^{-5}	5
	Synflow+*PriPrune*	1	2×10^{-6}	10
	SNIP+*PriPrune*	1	2×10^{-6}	10
	FedDST+*PriPrune*	8	2×10^{-5}	10
CIFAR-10	PruneFL+*PriPrune*	10	2×10^{-5}	5
	Synflow+*PriPrune*	1	2×10^{-6}	5
	SNIP+*PriPrune*	1	2×10^{-6}	10
	FedDST+*PriPrune*	10	2×10^{-5}	10

Table A2. Parameter settings for λ_{pri}, λ_{sha}, and λ_{acc} in FEMNIST and CIFAR-10 datasets.

Received 29 April 2024; revised 26 August 2024; accepted 7 October 2024

Anexo 5: FedQV: Leveraging Quadratic Voting in Federated Learning

FEDQV: LEVERAGING QUADRATIC VOTING IN FEDERATED LEARNING

Tianyue Chu
IMDEA Networks Institute
Universidad Carlos III de Madrid

Nikolaos Laoutaris
IMDEA Networks Institute

ABSTRACT

Federated Learning (FL) permits different parties to collaboratively train a global model without disclosing their respective local labels. A crucial step of FL, that of aggregating local models to produce the global one, shares many similarities with public decision-making, and elections in particular. In that context, a major weakness of FL, namely its vulnerability to poisoning attacks, can be interpreted as a consequence of the *one person one vote* (henceforth *1p1v*) principle that underpins most contemporary aggregation rules.

In this paper, we introduce FEDQV, a novel aggregation algorithm built upon the *quadratic voting* scheme, recently proposed as a better alternative to *1p1v*-based elections. Our theoretical analysis establishes that FEDQV is a truthful mechanism in which bidding according to one's true valuation is a dominant strategy that achieves a convergence rate matching that of state-of-the-art methods. Furthermore, our empirical analysis using multiple real-world datasets validates the superior performance of FEDQV against poisoning attacks. It also shows that combining FEDQV with unequal voting "budgets" according to a reputation score increases its performance benefits even further. Finally, we show that FEDQV can be easily combined with Byzantine-robust privacy-preserving mechanisms to enhance its robustness against both poisoning and privacy attacks.

1 Introduction

Federated Learning (FL) has emerged as a promising privacy-preserving paradigm for conducting distributed collaborative model training across parties unwilling to disclose their local data. Parties collaborate to train a global model by submitting their local models to a server, which applies a specific aggregation rule to generate a new version of the global model to be sent back to parties. The process of agreeing on a common global model in Federated Learning shares many similarities with public decision-making and elections in particular. Indeed, the weights of local model updates from each party can be seen as votes of preference that affect the global model resulting from an aggregation rule applied at the centralised server of an FL group.

FEDAVG [1] has been the "de facto" aggregation rule used in FL tasks such as Google's emoji and next-word prediction for mobile device keyboards [2, 3]. In FEDAVG the global model is produced from a simple weighted averaging of local updates with weights that represent the amount of data that each party has used for its training.

The problem. Recent work [4] has shown that FEDAVG is vulnerable to poisoning attacks, as even a single attacker can degrade the global model by sharing faulty local updates of sufficiently large weight. Such attacks become possible because FEDAVG treats all local data points equally. In essence, the aggregation rule, when seen at the granularity of individual training data, resembles the *one person one vote (1p1v)* election rule of modern democratic elections. In this context, the server distributes votes (weights) to a party in accordance with the amount of its training data, which may be regarded as its population. This, however, may confer an unjust advantage to malicious parties who may have, or falsely claim to have, large training datasets.

Our approach. To address this issue, we propose a robust aggregation rule inspired by elections based on *Quadratic Voting* [5] (henceforth QV). In QV, each party is given a voting budget that can be spent on different rounds of voting for proposals. Within a particular vote, an individual has to decide the number of "credit voices" to commit, whose square root is what impacts the corresponding outcome of the vote, hence the name quadratic voting. The number of "credit voices" is determined by the individual's voting preference for the proposal. QV has been proposed as a means to break out from the tyranny-of-the-majority vs. subsidising-the-minority dilemma of election systems [6]. Its formal

analysis [7] under a game theoretic price-taking model, has shown that QV outperforms *1p1v* in terms of efficiency and robustness. Importantly, it has the unique capacity to deter collusion attacks by effectively taxing extreme behaviours.

Our contributions. In this paper, we propose FEDQV, a novel FL aggregation scheme that draws inspiration from QV. Our objective is to mitigate the ability of malicious peers to impose disproportional damage on the global model – a vulnerability inherent in FEDAVG that applies the *1p1v* principle at the granularity of individual votes. By addressing this issue, FEDQV serves as a superior alternative to FEDAVG, offering increased robustness against poisoning attacks while retaining compatibility with privacy-guaranteed mechanisms to defend against privacy attacks, as will be demonstrated later.

Our primary contribution involves integrating QV principles into FL, augmented by the implementation of multiple defensive layers, to establish the truthful mechanism, FEDQV. We begin with the incorporation of quadratic computation from QV into the FL setting. This incorporation restricts the ability of malicious peers to inflict high damages by taxing their aggregation weights more than linear. However, this direct incorporation alone falls short of capturing each party's voting preference, a crucial aspect of QV. To capture this preference, we require parties to submit the similarity of their local model with the previous round's global model, which serves as a measure of the parties' preference. This modification enables the integration of QV principles into the FL system. Then in response to potential malicious attempts from peers, we introduce FEDQV, a truthfulness mechanism alongside our application of QV to FL. This mechanism employs a masked voting rule on the server side to conceal the voting calculation process from parties. Additionally, it incorporates outlier detection and a limited voting budget, information that is exclusive to the server. These defence measures collectively act as a deterrent against potential poisoning attacks and untruthful strategic behaviour, ensuring the overall robustness of the FEDQV system in the face of adversarial threats.

To further enhance resilience against poisoning attacks, we extend FEDQV to use adaptive voting budgets. In election-related applications, QV allocates equal budgets to all voters, reflecting the democratic principle of equal rights. However, in our adaptation of QV for FL, it makes sense to allocate more votes to benign peers and limit the influence of malicious ones by assigning them smaller voting budgets. We achieve this by employing unequal budgets, which are tied to a reputation score for each peer, as discussed in Section 5.7.

Our next contribution is the implementation of various state-of-the-art Byzantine fault tolerance techniques on top of FEDQV and demonstrating that FEDQV serves as a complementary defence that further boosts the robustness of existing defence techniques.

Put differently, FEDQV is an enhancer of existing byzantine fault tolerance techniques, not a competitor – implementing these defences atop FEDQV consistently yields superior results compared to implementing them atop FEDAVG.

To also defend against privacy attacks, we design FEDQV such that it can be easily combined with existing privacy-guaranteeing mechanisms to thwart *inference and reconstruction attacks* [8, 9, 10]. Specifically, we showcase the compatibility of FEDQV by implementing SECAGG[11] on top of the FEDQV framework. We then demonstrate experimentally the effectiveness of SECAGG within the FEDQV framework, focusing on its ability to withstand powerful privacy attacks.

Our final contribution comprises an extensive theoretical analysis in order to: 1) establish convergence guarantees, and 2) prove the truthfulness of our method.

Our findings. Using a combined theoretical and experimental evaluation, we show that:

- FEDQV is a truthful mechanism and is theoretically and empirically compatible with FEDAVG in terms of accuracy and convergence under attack and no-attack scenarios.
- FEDQV consistently outperforms FEDAVG under various state-of-the-art poisoning attacks, especially for local model poisoning attacks improving the robustness to such attacks by a factor of at least $4\times$.
- The combination of FEDQV with a reputation model to assign unequal credit voice budgets to parties according to their respective reputations, improves robustness against poisoning attacks by at least 26% compared to the baseline FEDQV that uses equal budgets.
- We show that integrating FEDQV with established Byzantine-robust FL defences, including Multi-Krum [4], Trimmed-Mean [12], and Reputation [13], results in substantial enhancements in accuracy and reductions in the attack success rate (ASR) under state-of-the-art attacks when compared to the original defence methods. Specifically, integrating FEDQV into Multi-Krum results in a twofold improvement in accuracy under two untargeted attack scenarios, along with an average enhancement of 26.72% in accuracy and a notable average reduction of 70% in ASR under two targeted attack scenarios.

- We demonstrate the compatibility of FEDQV with SECAGG, and our empirical evaluation confirms that this integration enhances resistance against two privacy attacks.

2 Background

2.1 Election Mechanisms in FL

Election mechanisms are widely used in distributed systems for choosing a coordinator from a collection of processes [14, 15]. Likewise, there exist works that explore the value of the election mechanism for the aggregation step of FL. Plurality voting is employed in *FedVote* [16] for weighting the local updates, and *FedVoting* [17] for treating the validation results as votes to decide the optimal model. Also in [18], the authors propose two forms of election coding, random Bernoulli codes and deterministic algebraic codes, for discovering majority opinions for the aggregation step. *DETOX* [19] proposes a hierarchical aggregation step based on majority votes upon groups of updates. Finally, *DRACO* [20] and *ByzShield* [21] also employ majority voting to fend off attacks against the aggregation step. All the aforementioned election mechanisms suffer from the tyranny of the majority problem in election systems [22]. In FL, this means that if attackers manage to control the majority of votes, then via poisoning their tyranny will manifest itself as a degradation of the accuracy of the FL model used by the minority.

To address these limitations, QV is proposed as a solution that combines simplicity, practicality, and efficiency under relatively broad conditions. QV considers a quadratic vote pricing rule, inspired by economic theory, under which voters can purchase votes at ever-increasing prices within a predetermined voting budget. The advantages of QV over *1p1v* have a rigorous theoretical basis, which of course applies also to the use of QV in FL. For any type of symmetric Bayes-Nash equilibrium, the price-taking assumption approximately holds for all voters, as a result, the expected inefficiency of QV is bounded by constant [23]. This theoretical analysis [24, 25] combined with strong empirical validation, both at the laboratory [26] and on the field [27], suggest that QV is near-perfectly efficient and more robust than *1p1v* which, as already explained, forms the basis of contemporary FL aggregation mechanisms. The advantages of QV can also be observed from the viewpoint of collusion, which is generally deterred either by unilateral deviation incentives or by the reactions of non-participants [7].

2.2 FL Aggregation Against Poisoning and Privacy Attacks

There exist several Byzantine-robust FL aggregation methods for mitigating poisoning attacks either by leveraging statistic-based outlier detection techniques [4, 12, 28, 13] or by utilising auxiliary labelled data collected by the aggregation server in order to verify the correctness of the received gradients [29, 30]. Both approaches, though, require examining the properties of the updates of individual parties, which can jeopardise their privacy due to inference [10] and reconstruction attacks [8, 9] mounted by an honest but curious aggregation server.

To fight against privacy attacks, Secure aggregation (SA) protocols have been proposed as potential countermeasures [31]. Secure aggregation can be achieved using four main privacy-enhancing technologies: differential privacy [32], trusted-execution environment (TEE) [33], secure shuffling under anonymity assumptions [34], and cryptography. Among those, secure aggregation based on cryptography is the most widely studied [11, 35, 36]. Although crypto-based secure aggregation provides strong security guarantees compared to alternatives, i.e., differential privacy, it also suffers from high computational and communication overhead. Consequently, these crypto-based secure aggregations are primarily established in the FEDAVG framework due to the impracticality of integrating them into Byzantine-robust FL aggregation methods, owing to the computational complexity in these aggregations.

In response to this limitation, FEDQV emerges as a promising solution, offering enhanced resilience against poisoning attacks while inheriting the simplicity of FEDAVG. Notably, this simplicity facilitates FEDQV's integration with crypto-based secure aggregations, thereby bolstering defences against privacy attacks while also exhibiting superior resilience against poisoning attacks compared to FEDAVG. Although a few FL aggregation approaches [37, 38, 39] can be adapted to incorporate secure aggregation, they still rely on majority voting as the aggregation scheme, which can be integrated with FEDQV to enhance its robustness against poisoning attack.

3 FEDQV: Quadratic Voting in FL

3.1 Federated Learning Setting

Consider an FL system involving N parties and a central server. During training round t, a subset of parties $\mathcal{S}^t$ is selected to participate in the training task. Each party i has the local dataset $\mathcal{D}_i$ with $|\mathcal{D}_i|$ samples (voters), drawn from

non-independent and non-identically (Non-IID) distribution $\mathcal{X}_i(\mu_i,\, \sigma_i^2)$. The goal of using FL is to learn a global model for the server. Given the loss function $\ell(\boldsymbol{w}; \mathcal{D})$, the objective function of FL can be described as

$$\mathcal{L}(\boldsymbol{w}) = \mathbb{E}_{\mathcal{D}\sim\mathcal{X}}\left[\ell(\boldsymbol{w}; \mathcal{D})\right]$$

Therefore, the task becomes:

$$\boldsymbol{w}^* = \underset{\boldsymbol{w}\in\mathbb{R}^d}{\arg\min}\, \mathcal{L}(\boldsymbol{w})$$

To find the optimal $\boldsymbol{w}^*$, Stochastic Gradient Descent (SGD) is employed to optimise the objective function. Let T be the total number of every part's SGD, E be the local iterations between two communication rounds, and thus $\frac{T}{E}$ is the number of communication rounds.

The FL model training process entails several rounds of communication between the parties and the server, including broadcasting, local training, and aggregation, as demonstrated in Algorithm 1. The global model is first initialised at a random state by the server and, then, in round t the following steps are taken:

Broadcast: The server first randomly selects a subset of parties $\mathcal{S}^t$ ($|\mathcal{S}^t| = \mathcal{C} \geq 1$) and then distributes the latest global model as the global proposal to all chosen parties. **Local Training**: The selected party i performs local computation based on the global proposal and its local dataset $\mathcal{D}_i$, and sends the update $\boldsymbol{w}_i^t$ back to the server:

$$\boldsymbol{w}_i^t \leftarrow \boldsymbol{w}_i^{t-1} - r_{t-1}\frac{\partial \ell_i(\boldsymbol{w}_i^{t-1}; \mathcal{D}_i)}{\partial \boldsymbol{w}}$$

where r_{t-1} is the learning rate.

Aggregation: The server receives the model updates from all participating parties and aggregates them to update the global model.

For aggregation rule, FEDAVG uses the fraction of the local training sample size of each party over the total training samples as the weight of a party:

$$\boldsymbol{w}^{t+1} = \frac{1}{\bigcup_{i\in\mathcal{S}^t}|\mathcal{D}_i|}\sum_{i\in\mathcal{S}^t}|\mathcal{D}_i|\cdot\boldsymbol{w}_i^t$$

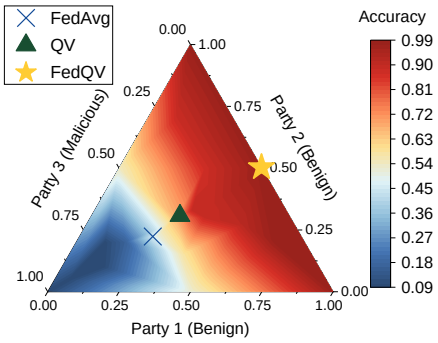
Similar to *1p1v*, each sample here represents a single voter, and since party i possesses $|\mathcal{D}_i|$ samples, it is able to cast $|\mathcal{D}_i|$ votes for its local model during the aggregation. Hence, the global proposal is a combination of all parties' local proposals weighted by their votes.

3.2 Preliminary Results

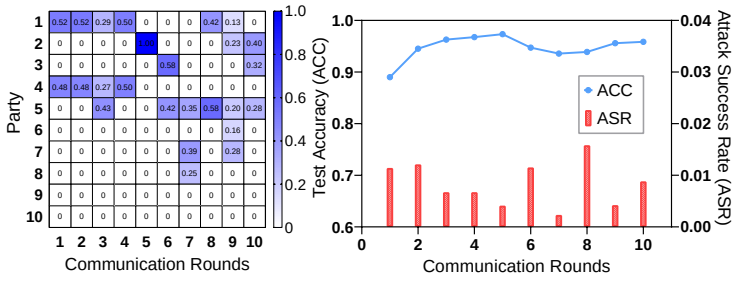
To show the rationale behind the algorithm design, we consider a toy use case involving two benign parties and one malicious party in FL. These parties claim possession of training datasets with the dataset sizes set to $\{1, 1, 2\}$, respectively. Importantly, the malicious party asserts a larger training dataset than the benign ones. These claims are reported to the central server, which lacks access to the raw data of the participating parties for verification purposes.

First, we introduce quadratic computation from QV into the FL. The quadratic computation involves taxing the aggregation weights with the square root. Consequently, while the aggregation weights in FEDAVG remain $\{1, 1, 2\}$, the corresponding weights in QV are adjusted to $\left\{1, 1, \sqrt{2}\right\}$. We conduct a 10-round training of a multi-layer CNN on the MNIST dataset, during which the malicious party executes a backdoor attack under the same settings as discussed in Section 5. The test accuracy is colour-coded and presented in Figure 1a, where the vertices of the triangle correspond to different parties, and their positions inside the triangle reflect their respective aggregation weights. As expected, the compromised accuracy in FEDAVG arises from the presence of a malicious party that possesses or falsely claims to possess a larger dataset. Compared to FEDAVG, QV, with the adjusted weights, exhibits superior accuracy. This suggests that the quadratic computation in QV can enhance performance by restraining the influence of attackers within FEDAVG by taxing their aggregation weights more than linear.

However, the direct incorporation of quadratic computation alone proves insufficient in capturing the voting preferences of each party, a critical aspect of QV. In QV, parties with pronounced disagreements with a proposal express their dissent by casting more votes to reject it. To capture each party's voting preference into the FL system, we refine the integration by requiring parties to submit the similarity of their local model with the previous round's global model. The similarity score serves as a measure of agreement, where higher scores indicate stronger alignment between the local model (proposal) and the global model (proposal). Consequently, parties with higher similarity scores will cast fewer votes in this round, reflecting their reduced need for adjustment to the existing global model. Conversely, parties showing greater dissimilarity ($1 - s_i^t$) between their local proposal and the global one cast more votes. This strategic



(a) Aggregation weights (position within the triangle) and corresponding test accuracy (colour-coded) with three parties (two benign and one malicious). FEDAVG is located at the bottom left ; QV is positioned around the centre ; FEDQV is situated along the right triangle side.



(b) FEDQV aggregation weights for each communication round, involving 10 parties (on the left). The first 6 parties are benign, while the subsequent 7 are malicious. The presentation is complemented by the corresponding metrics (on the right), including test accuracy (ACC) and attack success rate (ASR) of the global model. This scenario is observed during 10 communication rounds in the MNIST dataset under a Backdoor attack.

allocation empowers parties with substantial disagreements to wield more significant influence over the impending global proposal $\boldsymbol{w}^t$, thereby making it a closer alignment with their local proposal $\boldsymbol{w}_i^t$.

Based on these similarity scores, the voting calculation is conducted to determine the votes, serving as the aggregation weights. Rather than having parties directly calculate their votes, which could lead to malicious attempts, we introduce the FEDQV mechanism. This approach adopts a more secure approach by delegating the vote allocation task to the server. It employs a masked voting rule to obscure the vote calculation process from the parties, thereby preventing them from knowing the exact votes they have cast for aggregation. This confidentiality measure is crucial to maintaining the integrity and impartiality of the voting process. Built upon the lack of awareness regarding the voting calculation process among parties, the FEDQV system incorporates two additional defensive mechanisms: a limited voting budget and outlier detection, which are known only by the server. Given that parties are unaware of their budget and the possibility of their updates being detected as abnormal, attempts by malicious parties to manipulate their updates and overcast their votes entail a dual risk. There is a potential for exclusion from the aggregation process through outlier detection or unknowingly reaching their voting budget limit, resulting in the assignment of no votes. These protections serve as a deterrent against potential poisoning attacks, ensuring the overall robustness of the FEDQV system in the face of adversarial threats. Details of the design of FEDQV are presented in the next subsection.

Returning to our previous toy example, the implementation of FEDQV leads to the allocation of weights $\{1, 1, 0\}$, as depicted in Figure 1a. This weight distribution effectively excludes the malicious party from the aggregation process. Consequently, this exclusion contributes to the improved accuracy of the resulting global model. To further illustrate how FEDQV assigns aggregation weights to different parties during training, consider another toy case with 10 parties in the FL system, where 4 of them are attackers. The settings remain the same as the first toy case, and the results are presented in Figure 1b. In the left of Figure 1b, the first six parties are benign, and the rest are malicious. Notably, during the 10 communication rounds, only two attackers (parties 7 and 8) successfully participated in the aggregation, with party 7 participating twice and party 8 once, while most others did not contribute, resulting in an aggregation weight of 0 for them. This outcome demonstrates FEDQV's effectiveness in expelling malicious parties. Even when a round includes two malicious parties, the decrease in test accuracy is limited, and the Attack Success Rate (ASR) remains low (<2%) as shown in the right of Figure 1b. This demonstration underscores FEDQV's capability to mitigate the influence of malicious parties, thus preventing significant damage to the global model.

3.3 FEDQV Design

We use QV in FL to overcome the drawback of *1p1v*, which improves the robustness of aggregation in comparison to FEDAVG without compromising any efficiency. Figure 2 provides an overview of FEDQV algorithm, which comprises two key components: *similarity computation* executed on the party side to capture the voting preference of each party and *voting scheme* managed on the server side to deter potential poisoning attacks and untruthful strategic behaviour of parties. We detail each component in the following subsections.

Similarity Computation: In round t, based on the server instructions, party i ($i \in \mathcal{S}^t$) trains its local model $\boldsymbol{w}_i^t$, which can be regarded as its local proposal. Following the local training phase, party i computes a similarity score s_i^t utilising

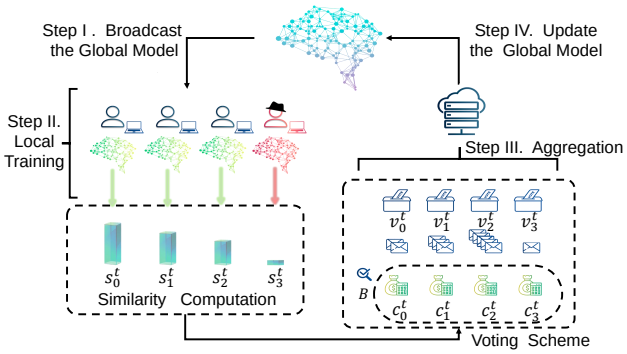


Figure 2: Overview of FEDQV algorithm.

cosine similarity, quantifying the alignment between its locally trained model $\boldsymbol{w}_i^t$ and the previous global model $\boldsymbol{w}^{t-1}$.

$$s_i^t = S_{\cos}\left(\boldsymbol{w}_i^t, \boldsymbol{w}^{t-1}\right) = \frac{\left\langle \boldsymbol{w}_i^t, \boldsymbol{w}^{t-1} \right\rangle}{\|\boldsymbol{w}_i^t\| \cdot \|\boldsymbol{w}^{t-1}\|}$$

Notably, the cosine similarity function can be adapted to different similarity metrics, such as L2 distance, to better suit specific tasks. In this context, a higher s_i^t value indicates a stronger agreement with the previous global model (proposal). Once selected parties finish training, they send their updates $\boldsymbol{w}_i^t$ to the server, with the message $\langle s_i^t \rangle$.

It's noteworthy that similarity calculation can also be launched on the server side. However, this approach presents certain risks: (i) it exposes the system to privacy attacks initiated by the server, and (ii) it may produce distorted similarity scores due to the utilisation of regularisation and privacy-preservation methods [40] at the party side. Given our primary objective of comparing FEDQV with FEDAVG, where weights are also computed at the party side, we choose to conduct the calculation on the party side within FEDQV. Additionally, to counteract potential malicious attempts launched from the party side, we have the option to introduce defence mechanisms on the server side, such as other Byzantine-robust aggregations, as detailed in Section 5.8. This approach makes it more challenging for attackers to mount successful attacks compared to FEDAVG.

Voting Scheme (Server Side): Upon receiving the updates and messages from selected parties, the server proceeds with the following steps:

- The server normalises the similarity scores using Min-Max scaling to obtain $\bar{s}_i^t$ as:

$$\bar{s}_i^t = \text{norm}\left(s_i^t, \{s_i^t\}_{i \in [\mathcal{S}^t]}\right) \tag{1}$$

  Following normalisation, parties no longer being aware of their exact similarity scores $\bar{s}_i^t$.

- The server employs a penalty mechanism for abnormal similarity scores, specifically when $\bar{s}_i^t$ falls below θ or exceeds $1-\theta$. Here, θ represents the similarity threshold and is established to capture scores that are excessively small or large, indicating potential anomalies. In response to such abnormality, the server enforces a penalty by reducing the budgets B_i allocated to the respective party as:

$$B_i = \max\left(0, B_i + \ln \bar{s}_i^t - 1\right) \tag{2}$$

- The server calculates the voice credit c_i^t for party i utilising the masked voting rule $\mathcal{H}$ as:

$$c_i^t = \mathcal{H}(\bar{s}_i^t) = \left(-\ln \bar{s}_i^t + 1\right) \mathbb{1}_{\theta < \bar{s}_i^t < 1-\theta} \tag{3}$$

  Here, the voice credits signify the price party i is required to pay in round t for its local proposal. Parties with higher similarity scores, indicating stronger agreement with the global proposal, are allocated fewer credit votes from the server. Conversely, parties showing greater dissimilarity $(1 - s_i^t)$ between their local proposal and the global one receive an increased allocation of c_i^t. This mechanism empowers parties with substantial disagreements to exert greater influence over the forthcoming global proposal $\boldsymbol{w}^t$. Notably, parties with abnormal similarity scores are excluded from participating in the aggregation, with receiving zero voice credits.

- The server checks the budget B_i for each party and computes their final votes v_i^t as:

$$v_i^t = \sqrt{\min\left(c_i^t, \max\left(0, B_i\right)\right)} \tag{4}$$

Algorithm 1: FEDQV

Input : $\boldsymbol{w}^0 \leftarrow$ random initialisation; B, $\theta \leftarrow$ FEDQV parameters

Server :

for *Iteration* $t \leftarrow 1$ **to** $\frac{T}{E}$ **do**
  Broadcast $\boldsymbol{w}^{t-1}$ to randomly selected set of parties $\mathcal{S}^t$ ($|\mathcal{S}^t| = \mathcal{C} \geq 1$);
  Receive *the local updates* $(\boldsymbol{w}_i^t, s_i^t)$ from selected parties ($i \in \mathcal{S}^t$);
  Normalisation: $\bar{s}_i^t \leftarrow$ Eq.(1), $i \in \mathcal{S}^t$;
  for $i \leftarrow 1$ **to** N **do in parallel**
    if $\bar{s}_i^t \leq \theta$ *or* $\bar{s}_i^t \geq 1 - \theta$ **then**
      Update $B_i \leftarrow \max(0, B_i + \ln \bar{s}_i^t - 1)$
    Credit voice $c_i^t \leftarrow$ Equation 3, *Vote* $v_i^t \leftarrow$ Equation 4;
    Budget $B_i \leftarrow \max\left(0, B_i - (v_i^t)^2\right)$ `// Update the budget`
  end forpar
  return $\boldsymbol{w}_n^t \leftarrow \sum_{i=1}^{N} \frac{v_i^t}{\sum_{i=1}^{M} v_i^t} \boldsymbol{w}_{i,n}^t$
end for

Party :

for *Party* $i \in \mathcal{S}^t$ **do in parallel**
  Receive *the global model* $\boldsymbol{w}^{t-1}$;
  for *local epoch* $e \leftarrow$ *1* ***to*** E **do**
    $\boldsymbol{w}_i^t \leftarrow \boldsymbol{w}_i^{t-1} - r_{t-1} \frac{\partial \ell_i(\boldsymbol{w}_i^{t-1}; \mathcal{D}_i)}{\partial \boldsymbol{w}}$ `// Local training`
  end for
  Calculate *the similarity score*: $s_i^t \leftarrow \frac{\langle \boldsymbol{w}_i^t, \boldsymbol{w}^{t-1} \rangle}{\|\boldsymbol{w}_i^t\| \cdot \|\boldsymbol{w}^{t-1}\|}$;
  Send $(\boldsymbol{w}_i^t, s_i^t)$
end forpar

- The server updates the budget as:

$$B_i = \max(0, B_i - \left(v_i^t\right)^2) \tag{5}$$

Thus, the server determines the weight (v_i^t) of party i for aggregation and generates the updated global model $\boldsymbol{w}^t$, reflecting the collective opinion of all selected parties (voters). Algorithm 1 summarises all these steps of FEDQV.

Malicious Party: In cases where malicious parties attempt to manipulate the similarity scores, their capabilities are restricted by:

(a)*No knowledge of the voting process.* Only the server possesses knowledge of each party's remaining budget and the number of actual votes cast in the current round. The lack of knowledge about the voting calculation process that contains outlier detection, coupled with the parties' unawareness of their limited voting budget, exposes malicious parties to the risk of exclusion from the aggregation process.

(b)*Punitive Measures*: FEDQV, with its masked voting rule that contains outlier detection and limited budget, empower the system to penalise and potentially remove malicious parties who attempt to conduct poisoning attacks;

(c)*Limited Influence*: Even if a manipulated similarity score is accepted by the server, the influence the malicious party can exert is inherently constrained due to the nature of QV and the limited budgets, minimising the potential damage.

Benefits of FEDQV:

- **Truthful Mechanism.** FEDQV is a *truthful* mechanism [41] as we prove in Theorem 4.17. This means that this mechanism compels the parties, even malicious ones, to tell the truth about their votes (weights) for aggregation, rather than any possible lie. This truthfulness is reinforced by the aforementioned several defence layers.
- **Ease of Integration and Compatibility.** FEDQV is highly adaptable and can be seamlessly integrated into Byzantine-robust FL defence schemes with minimal adjustments, specifically by modifying the aggregation weight calculation while leaving other algorithm components unchanged. This integration is demonstrated

Algorithm 2: FEDQV with Adaptive Budget

Input : $\boldsymbol{w}_i^t, c_i^t, B_i^t \leftarrow$ FEDQV; κ,a,W,M,λ,$\delta \leftarrow$ *Reputation model* parameters

for $i \in \mathcal{S}^t$ **do**
  for $j \leftarrow 1$ **to** M **do**
    Subjective Observations $(P_i^t, Q_i^t) :=$ `IRLS` $(w_{i,j}^t, \delta)$;
  end for
  Reputation Score $R_i^t :=$ `Rep` $(P_i^t, Q_i^t, \kappa, a, W)$
  Budget $B_i^t \leftarrow R_i^t \mathbb{1}_{\lambda \leq R_i^t} + B_i^t$, *Credit voice* $c_i^t \leftarrow (R_i^t + c_i^t)\mathbb{1}_{\lambda \leq R_i^t}$
end for

in Section 5.8. Furthermore, similar to FEDAVG, FEDQV boasts efficient communication and simplicity, rendering it compatible with various mechanisms employed in FL. It can effortlessly incorporate the regularisation, sparsification, and privacy modules, encompassing techniques such as clipping [40], gradient compression [42], differential privacy [43], and secure aggregation [11].

3.4 FEDQV with Adaptive Budgets

In democratic elections, all individuals are typically granted equal voting rights, entailing an equal voting budget. In FL, however, it often makes sense to give malicious parties fewer votes than honest ones. Thus to improve the robustness of standard FEDQV, we combine it with the reputation model in [13] to assign an unequal budget based on the reputation score of parties in each round t. Specifically, if a party's reputation score R^t surpasses a predefined threshold λ, we increase their budget, and vice versa. We present a summary of this combination in Algorithm 2, with a detailed explanation, expanding on the well-established components from the original paper. We provide empirical evidence in Section 5.7 showcasing the substantial performance improvements achieved by the enhanced version of FEDQV featuring an adaptive budget.

Here we present a concise elucidation of key components of the Algorithm 2 as followings:

- **IRLS** (Iteratively Reweighted Least Squares): IRLS serves as an optimisation technique employed to solve specific regression problems. Within [13], IRLS is utilised to compute the Subjective Observations of participating clients based on their parameter's confidence score, which is calculated using the repeated-median regression technique.
- **Subjective Observations**: Positive observations denoted by P_i^t signify acceptance of an update, while negative observations denoted by Q_i^t indicate rejection. Consequently, positive observations enhance a party's reputation, and negative ones have the opposite effect.
- **Reputation Score Calculation**: The reputation score of a party is determined using a subjective logic model, formulated as follows:

$$R_i^t = \frac{\kappa P_i^t + Wa}{\kappa P_i^t + \eta Q_i^t + W}$$

Regarding the integration of the reputation model, our objective is to demonstrate how combining FEDQV with the reputation model enables the allocation of unequal budgets, thereby enhancing the robustness of standard FEDQV. This integration's adaptability extends beyond a single reputation model, allowing customisation to suit various needs. The example presented in the paper serves to showcase the concept's viability.

4 Theoretical Analysis

In this section, we establish the convergence guarantees and truthfulness properties of FEDQV. Our first result is Theorem 4.9 that states FEDQV converges to the global optimal solution at a rate of $\mathcal{O}(\frac{1}{T})$, comparable to the convergence rate exhibited by FEDAVG. Theorem 4.10 extends this statement, affirming that FEDQV converges to a near-optimal solution in the presence of malicious parties, with the resulting performance gap determined by the percentage of malicious parties. The empirical convergence performance of our algorithm, as gauged by metric test accuracy and training loss, is consistent with our theoretical analysis, as elaborated in the following section. Our final result, Theorem 4.17, establishes that FEDQV is a truthful mechanism, wherein honesty emerges as the dominant strategy within this framework. Fully detailed proofs are in Appendix A.

4.1 Convergence

We start by stating our assumptions, which are standard and common for such types of analysis and per recent works such as [12, 28, 44, 30, 13, 45].

Assumption 4.1. The loss functions are L-smooth, which means they are continuously differentiable and their gradients are Lipschitz-continuous with Lipschitz constant $L > 0$, whereas:

$$\forall i \in N,\ \forall \boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{R}^d,\ \|\nabla \mathcal{L}(\boldsymbol{w}_1)) - \nabla \mathcal{L}(\boldsymbol{w}_2))\|_2 \leq L \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2$$
$$\|\nabla \ell(\boldsymbol{w}_1; \mathcal{D}) - \nabla \ell(\boldsymbol{w}_2; \mathcal{D})\|_2 \leq L \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2$$

Assumption 4.2. The loss function $\ell(\boldsymbol{w}_i, D)$ are μ-strongly convex:

$$\exists \mu > 0, \forall \boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{R}^d, \nabla \ell(\boldsymbol{w}^*; \mathcal{D}) = 0, \nabla \mathcal{L}(\boldsymbol{w}^*) = 0$$
$$2\left(\mathcal{L}(\boldsymbol{w}_1) - \mathcal{L}(\boldsymbol{w}_2)\right) \geq 2\left\langle \nabla \mathcal{L}(\boldsymbol{w}_2), \boldsymbol{w}_1 - \boldsymbol{w}_2 \right\rangle + \mu \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2^2$$
$$2\left(\ell(\boldsymbol{w}_1; \mathcal{D}) - l(\boldsymbol{w}_2; \mathcal{D})\right) \geq 2\left\langle \nabla \ell(\boldsymbol{w}_2; \mathcal{D}), \boldsymbol{w}_1 - \boldsymbol{w}_2 \right\rangle + \mu \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2^2$$

Assumption 4.3. The expected square norm of gradients $\boldsymbol{w}$ is bounded:

$$\forall \boldsymbol{w} \in \mathbb{R}^d, \exists \mathcal{G}_{\boldsymbol{w}} < \infty, \mathbb{E}\|\nabla \ell(\boldsymbol{w}; \mathcal{D})\|_2^2 \leq \mathcal{G}_{\boldsymbol{w}}^2$$

Assumption 4.4. The variance of gradients $\boldsymbol{w}$ is bounded:

$$\forall \boldsymbol{w} \in \mathbb{R}^d, \exists \mathcal{V}_{\boldsymbol{w}} < \infty, \mathbb{E}\|\nabla \ell(\boldsymbol{w}; \mathcal{D}) - \mathbb{E}(\nabla \ell(\boldsymbol{w}; \mathcal{D})\|_2^2 \leq \mathcal{V}_{\boldsymbol{w}}$$

The lemmas we utilise in the proof of Theorem 4.9 and Theorem 4.10, are presented below.

Lemma 4.5. *From Assumption 4.1 and 4.2, $\mathcal{L}(\boldsymbol{w})$ is L-smooth and μ-strongly convex. Then $\forall \boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{R}^d$, one has*

$$\langle \nabla \ell(\boldsymbol{w}_1) - \nabla \ell(\boldsymbol{w}_2), \boldsymbol{w}_1 - \boldsymbol{w}_2 \rangle \geq \frac{L\mu}{L+\mu} \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2^2 + \frac{1}{L+\mu} \|\nabla \ell(\boldsymbol{w}_1) - \nabla \ell(\boldsymbol{w}_2)\|_2^2$$

Lemma 4.6. *Assume Assumption 4.1, Assumption 4.2 and Lemma 4.5 hold, Let $p_i^t = \frac{1}{C} \mathbb{1}_{i \in \mathcal{S}^t}$, we have*

$$\left\|\boldsymbol{w}^{t-1} - r\nabla \mathcal{L}(\boldsymbol{w}^{t-1}) - \boldsymbol{w}^*\right\|_2^2 \leq \sum_{i=1}^{N} p_i^{t-1} \left\|\boldsymbol{w}^{t-1} - \boldsymbol{w}_i^{t-1}\right\|_2^2 + \left(r^2\left(1 + L^2\right) - \frac{2rL\mu + 1}{L+\mu}\right) \left\|\boldsymbol{w}^{t-1} - \boldsymbol{w}^*\right\|_2^2 \quad (6)$$

Lemma 4.7. *Assume Assumption 4.3 holds, it follows that*

$$\mathbb{E} \sum_{i=1}^{N} p_i^{t-1} \left\|\boldsymbol{w}^{t-1} - \boldsymbol{w}_i^{t-1}\right\|_2^2 \leq (E-1)^2 r^2 \mathcal{G}_{\boldsymbol{w}}^2$$

Lemma 4.8. *Assume Assumption 4.4 holds, according to our Algorithm 1, it follows that*

$$\mathbb{E}\left\|\mathcal{F}(\boldsymbol{w}^{t-1}) - \nabla \mathcal{L}(\boldsymbol{w}^{t-1})\right\|_2^2 \leq (1 - 2\theta)\, \mathcal{C} \mathcal{V}_{\boldsymbol{w}} \sqrt{B}$$

Where

$$\mathcal{F}(\boldsymbol{w}^{t-1}) = \sum_{i \in \mathcal{S}^{t-1}} p_i^{t-1} \nabla \ell(\boldsymbol{w}_i^{t-1}; \mathcal{D}_i^{t-1})$$

Under these four mild and standard assumptions, along with the support of Lemmas, we have:

Theorem 4.9. ***(Secure Convergence Without Attack)*** *Under Assumptions 4.1, 4.2, 4.3 and 4.4, and $m = 0$. Choose $\alpha = \frac{L+\mu}{\mu L}$ and $\beta = 2\frac{(L+1)(L+\mu)}{\mu L}$, then* FEDQV *satisfies*

$$\mathbb{E}\,\mathcal{L}(\boldsymbol{w}^T) - \mathcal{L}(\boldsymbol{w}^*) \leq \frac{L}{2\varphi + T}\left(\varphi\, \mathbb{E}\left\|\boldsymbol{w}^0 - \boldsymbol{w}^*\right\|_2^2 + \frac{\alpha^2}{2}\Delta\right) \quad (7)$$

Where

$$\Delta = (E-1)^2\, \mathcal{G}_{\boldsymbol{w}}^2 + (1 - 2\theta)\, \mathcal{C} \mathcal{V}_{\boldsymbol{w}} \sqrt{B},\ \varphi = \alpha\,(L+1)$$

Suppose the percentage of attackers in the whole parties is m, we denote

$$\mathcal{M}_i(\boldsymbol{w}_i^t) = \begin{cases} * & \text{if } i \in \text{malicious parties} \\ \nabla \ell(\boldsymbol{w}_i^t; \mathcal{D}_i^t) & \text{if } i \in \text{honest parties} \end{cases}$$

Where $*$ stands for an arbitrary value from the malicious parties. Then we have:

Theorem 4.10. ***(Robust Convergence Under Attack)*** *Under Assumptions 4.1, 4.2, 4.3 and 4.4, Choose* $\alpha = \frac{L+\mu}{\mu L}$ *and* $\beta = 2\frac{(L+1)(L+\mu)}{\mu L}$*, then* FEDQV *satisfies*

$$\mathbb{E}\,\mathcal{L}(\boldsymbol{w}^T) - \mathcal{L}(\boldsymbol{w}^*) \leq \frac{L + 2Lr_{T-1}\varpi}{2\varphi + T}\left(\varphi\,\mathbb{E}\left\|\boldsymbol{w}^0 - \boldsymbol{w}^*\right\|_2^2 + \frac{\alpha^2}{2}\Delta\right) + \frac{L\varpi^2}{2} \tag{8}$$

Where

$$\Delta = (E-1)^2\,\mathcal{G}_{\boldsymbol{w}}^2 + (1-2\theta)\,\mathcal{CV}_{\boldsymbol{w}}\sqrt{B},\ \varphi = \alpha\,(L+1)\,,\ \varpi = mN\mathcal{G}_{\boldsymbol{w}} r_{T-1}\sqrt{4+6\theta-\theta^2}$$

Remark 4.11. According to Theorem 4.9, FEDQV obtains a secure convergence rate of $\mathcal{O}(\frac{1}{T})$ in the absence of malicious parties, which is comparable to the convergence rate of FEDAVG [46].

Remark 4.12. According to Theorem 4.10, FEDQV converge to a near-optimal solution in the presence of malicious parties, with the resulting performance gap determined by the percentage of malicious parties.

Remark 4.13. The error rate exhibits dependence on the budget B, the similarity threshold θ, and the percentage of malicious parties m. It is noteworthy that a larger budget allocation, a reduction in the similarity threshold, or an augmentation in the proportion of malicious parties induce more pronounced disparities in model updates, consequently resulting in an elevated error rate. The impact of these hyperparameters is shown in Figure 6b in Section 5.10.

4.2 Truthfulness

The FEDQV mechanism belongs to a single-parameter domain, where the real parameter votes v_i play a direct role in determining the eligibility of party i for aggregation. The mechanism is normalised according to game theory principles [41], where for every v_i and v_{-i} such that $f(v_i, v_{-i}) \notin W_i$, $p_i(v_i, v_{-i}) = 0$. Here, v_{-i} represents votes cast by all parties except for i, W_i is the subset of participants in aggregation, f represents the voting scheme outcome, and p_i is the payment function with $p_i(v_i, v_{-i}) = v_i^2$ in FEDQV. The upcoming definition of truthfulness and the following lemmas contribute to the proof of Theorem 4.17 in alignment with monotone and critical value concepts in game theory [41].

Definition 4.14. A mechanism $(f, p_1, ..., p_n)$ is called truthfulness if for every party i, we denote $a = f(v_i, v_{-i})$ and $a' = f(v_i^{'}, v_{-i})$ as the outcome of the voting, then $v_i(a) - p_i(v_i, v_{-i}) \geq v_i(a^{'}) - p_i(v_i^{'}, v_{-i})$, where $v_i(a)$ denotes the gain of party i if the outcome of the voting is a.

Here, $v_i(a) - p_i(v_i, v_{-i})$ is the utility of party i, indicating the gain from voting ($v_i(a)$) minus its cost ($p_i(v_i, v\_i)$). Intuitively this implies that party i with v_i would prefer "telling the truth" v_i to the server rather than any possible "lie" $v_i^{'}$ since this gives him higher (in the weak sense) utility.

Lemma 4.15. *f is monotone: : $\forall v_{-i}$ and $\forall v_i^{'} > v_i$, if $f(v_i, v_{-i}) \in W_i$, then $f(v_i^{'}, v_{-i}) \in W_i$.*

Lemma 4.16. *In* FEDQV*, $\forall i, v_i, v_{-i}$ that $f(v_i, v_{-i}) \in W_i$, we have that $p_i(v_i, v_{-i}) = \Phi_i(v_{-i})$, where Φ_i is the critical value of a monotone function f on a single parameter domain that $\Phi_i(v_{-i}) = \sup_{v_i : f(v_i, v_{-i}) \notin W_i} v_i$.*

Based on Lemma 4.15 and Lemma 4.16, we establish the following theorem:

Theorem 4.17. FEDQV *is incentive compatible (truthful).*

Remark 4.18. Regarding the concept of truthfulness, it theoretically ensures that being honest is the dominant strategy since providing manipulated similarity scores may lead to penalties and removal from the system due to the masked voting rule $\mathcal{H}$ and limited budget B. This is an integral part of the nature of QV embedded within our FEDQV framework.

5 Experiments

The objectives of our experimental evaluation are the following: (a) To assess the performance of our aggregation method relative to FEDAVG.(b) To benchmark our method across 10 distinct state-of-the-art poisoning attack scenarios.(c) To validate the alignment of our experimental findings with our prior theoretical analyses. (d) To demonstrate the ease of integration and compatibility of our method within Byzantine-robust FL defence schemes and privacy-preserving mechanisms.

5.1 Experimental setting

Datasets and global models. We implement the typical FL setting where each party owns its local data and transmits/receives information to/from the central server. To demonstrate the generality of our method, we train different global models on different datasets. We use four popular benchmark datasets: MNIST [47], Fashion-MNIST [48], FEMNIST [49] and CIFAR10 [50]. We consider a multi-layer CNN same as in [40], consisting of 2 convolutional layers and 2 fully connected layers for MNIST, Fashion-MNIST and FEMNIST, and the ResNet18 [51] for CIFAR10.

Non-IID setting. In order to fulfil the setting of a heterogeneous and unbalanced dataset for FL, we sample from a Dirichlet distribution with the concentration parameter $\iota = 0.9$ as the Non-IID degree as in [52, 53], with the intention of generating non-IID and unbalanced data partitions. Moreover, we have examined the performance across varying levels of non-IID data in Section 5.9.

Parameter Settings. The server selects 10 ($\mathcal{C}$) out of 100 (N) parties to participate in each communication round and train the global models for 100 communication rounds($\frac{T}{E}$), where the local training epoch E equals 5. We set the model hyper-parameters budget B and the similarity threshold θ to 30 and 0.2 respectively based on the hyper-parameter searching. All additional settings are provided in the Appendix B.1.

5.2 Thread Model

We assume we the parties are *malicious* while the server is *honest-but-curious*. Within this context, parties may deliberately submit false or manipulated local model updates to the central server. Their objectives could range from injecting biases and compromising the model's performance to extracting sensitive information from the aggregated global model. The server correctly follows the FL protocol steps but remains curious to discover any private information. In our case, the server has full visibility of all local models and launches privacy attacks upon receiving an update from a target party to reconstruct users' training data.

Furthermore, the percentage of malicious party m is an important factor in determining the success of the poisoning attack. In our analysis, we assume that the number of malicious parties is less than the number of honest parties, a common setting in such types of analysis and recent works [4, 12, 13].

5.3 Evaluated Poisoning Attacks

Our paper addresses three distinct attack schemes:

- **Data poisoning**: Attackers submit the true similarity score based on their poisoned updates, including **Labelflip Attack** [54], **Gaussian Attack** [55], **Backdoor** [56], **Scaling Attack** [52], **Neurotoxin** [57].
- **Model poisoning**: Attackers submit the true similarity score based on their clean updates and poison their model, including: **Krum Attack** [54], **Trim Attack** [54], and Aggregation-agnostic attacks: **Min-Max** and **Min-Sum** [58]
- **QV-tailored Attack**: Attackers submit both poisoned similarity score and the local model to exploit vulnerabilities in FEDQV. This dual-pronged attack, termed **QV-Adaptive**, presents a heightened challenge for FEDQV.

Here are the details of the aforementioned attacks. We begin with data poisoning attacks:

Labelflip Attack [54]: In the Label-Flip scenario, all the labels of the training data for the malicious clients are set to zero. This scenario simulates a directed attack, with the goal to disproportionally bias the jointly trained model towards one specific class. This is a data poisoning attack that does not require knowledge of the training data distribution. Under this attack, the malicious parties train with clean data but with flipped labels. Specifically, we flip a label k as $K - k - 1$, where K is the total class number.

Gaussian Attack [55]: This attack forges local model updates via Gaussian distribution on the malicious parties. malicious parties forge local model updates via Gaussian distribution.

Backdoor Attack [56] Malicious parties inject specific backdoor triggers into the training data and modify their labels to the attacker-chosen target label. Specifically, we use the same backdoor pattern trigger and attacker-chosen target label as in [59] as our trigger and set the attacker-chosen target label as 5. The backdoor can be introduced into a model by an attacker who poisons the training data with specially crafted inputs. A backdoor transformation applied to any input causes the model to mis-classify it to an attacker-chosen label The pattern must be applied by the attacker during local training, by modifying the digital image.

Scaling attack [52] The malicious parties generate poisoned local model updates by backdoor attack and only launch this attack during the last communication round after scaling these updates by a factor of N.

Neurotoxin attack [57] In this attack, the adversary starts by downloading the gradient from the previous round and employs it to approximate the benign gradient for the upcoming round. The attacker identifies the top-k% coordinates of the benign gradient and treats them as the constraint set. Over several epochs of Projected Gradient Descent (PGD), the attacker computes gradient updates on the manipulated dataset and projects this gradient onto the constraint set, which consists of the bottom-k% coordinates of the observed benign gradient. PGD is employed to approach the optimal solution within the span of the bottom-k% coordinates. We adopt the original parameter setting from the paper, where k is set to 0.1.

Next, we move to model poisoning attacks:

Krum Attack [54]: Malicious parties craft poisoned local model updates opposite from benign ones, and enable them to circumvent the defence of Krum [4].

Trim Attack [54] The poisoned local model updates constructed by malicious parties are optimised for evading the Trim-mean and Median [12].

Min-Max Attack [58]In order to ensure that the malicious gradients closely align with the benign gradients within the clique, attackers strategically compute the malicious gradient. This computation is carried out to limit the maximum distance of the malicious gradient from any other gradient, which is constrained by the maximum distance observed between any two benign gradients.

Min-Sum [58] The Min-Sum attack enforces an upper bound on the sum of squared distances between the malicious gradient and all the benign gradients. This upper bound is determined by the sum of squared distances between any one benign gradient and the rest of the benign gradients.

Finally, we introduce the QV-tailored attack:

QV-Adaptive attack Tailored for FEDQV, this attack leverages the Aggregation-agnostic optimisations [58] within the LMP framework [54]. This attack manipulates both the similarity score and the local model, following the procedure below:

1. The malicious party i generates benign updates $\boldsymbol{w}_i^t$ using clean data $\mathcal{D}_i$ in round t and calculates the corresponding similarity score;
2. malicious parties (with counts of m) collectively normalise all the similarity scores and employ the Aggregation-agnostic Min-Max optimisation to select the optimal similarity score. This optimisation objective aims to increase the likelihood of the score being accepted by the server.
3. the adaptive attack focuses on local model poisoning to optimise the following problem:

$$\max \nu \tag{9}$$

$$\text{s.t.} \quad \boldsymbol{w}_{i\in m}^{t'} = \text{FedQV}(\boldsymbol{w}_1^t, \boldsymbol{w}_2^t, \ldots, \boldsymbol{w}_m^t) \tag{10}$$

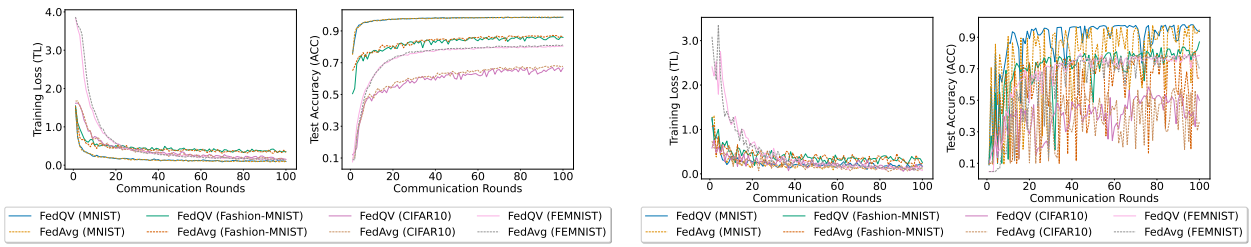
$$\boldsymbol{w}_{i\in m}^{t'} = \boldsymbol{w}_i^t - \nu \hat{d} \tag{11}$$

Here, $\hat{d}$ represents a column vector encompassing the estimated changing directions of all global model parameters. The variables $\boldsymbol{w}_{i\in m}^t$ and $\boldsymbol{w}_{i\in m}^{t'}$ correspond to the local model before and after the attack. The parameter ν denotes the extent of the attack's impact on the model.

It is noteworthy that Labelflip, Gaussian, Krum, Trim, Min-Max, Min-Sum and QV Adaptive attacks are untargeted attacks, whereas, Backdoor, Scaling and Neurotoxin attacks are targeted attacks. We confine our analysis to the worst-case scenario in which the attackers submit the poisoned updates in every round of the training process for all attack strategies with the exception of the Scaling attack.

5.4 Performance Metrics

We use the average test accuracy (ACC) of the global model to evaluate the result of the aggregation defence for poisoning attacks, in which attackers aim to mislead the global model during the testing phase. ACC is the percentage of testing examples with the correct predictions by the global model in the whole testing dataset, which is defined as ACC = (# correct predictions)/(# testing samples). In addition, there are targeted attacks that aim to attack a specific label while keeping the accuracy of classification on other labels unaltered. Therefore, besides ACC, we choose the attack success rate (ASR) to measure how many of the samples that are attacked, are classified as the target label chosen



(a) Without attack.

(b) Under labelflip attack with 40% malicious parties.

Figure 3: Convergence comparison of FEDAVG and FEDQV in Training Loss (TL) and Accuracy (ACC) over 100 epochs across four benchmark datasets, both under no attack and under attack scenarios.

by malicious parties. A robust federated aggregation method would obtain a higher Avg-ACC as well as a lower ASR under poisoning attacks. An ideal aggregation method can achieve 100% Avg-ACC and has the ASR as low as the fraction of attacked samples from the target label.

5.5 Convergence

We evaluate the convergence of FEDAVG and FEDQV across the aforementioned four datasets under both benign and adversarial conditions. The training loss (TL) and accuracy (ACC) of global models trained using FEDQV and FEDAVG are depicted in Figure 3a without any attack and in Figure 3a under labelflip attacks with 40% malicious parties. In the absence of Byzantine attacks, the convergence of the global model trained with FEDQV matches that of FEDAVG across all datasets, consistent with Theorem 4.9. However, under attack, FEDAVG struggles to converge due to its susceptibility to poisoning attacks, as shown in Figure 3a. While FEDQV also exhibits slower convergence under attack, it eventually converges and outperforms FEDAVG. This outcome aligns with Theorem 4.10, indicating that FEDQV achieves convergence to a near-optimal solution even in the presence of malicious parties, with the extent of performance improvement determined by the percentage of malicious parties.

5.6 Defence against Poisoning Attacks

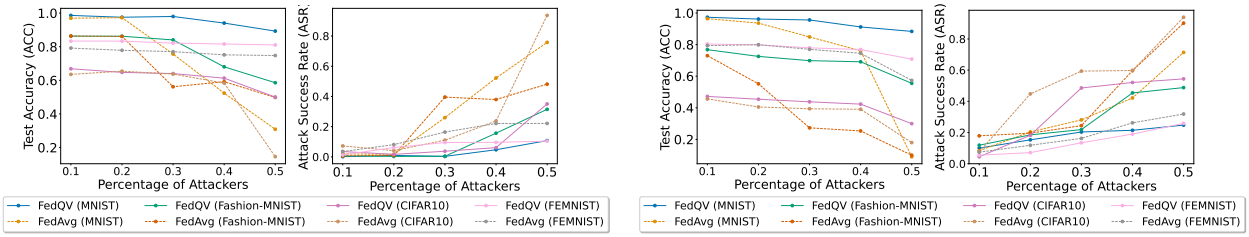
Static percentage of attackers: We present ACC and ASR of global models trained using both FEDAVG and FEDQV under the 10 aforementioned poisoning attacks across all four datasets in Table 1. The experiments involve 30% malicious parties, same as in previous studies such as [4, 13], which is also a common byzantine consensus threshold for resistance to failures in a typical distributed system [60]. In data poisoning attacks, the results consistently demonstrate that FEDQV outperforms FEDAVG, achieving the highest ACC with the smallest standard error. When considering targeted attacks, FEDQV again stands out, displaying the highest ACC along with the lowest ASR when compared to FEDAVG. In the context model poisoning attacks, FEDQV outperforms FEDAVG, except for the QV-Adaptive attack, which is tailored for FEDQV. Especially for local model poisoning attacks: Trim and Krum attacks, FEDQV outperforms FEDAVG by at least *4 times* in terms of accuracy. Setting this observation as the alternative hypothesis H_1 and using the Wilcoxon signed-rank test, we can reject the null hypothesis H_0 at a confidence level of 1% in favour of H_1, proving the robustness of FEDQV.

Varying the percentage of attackers: Then we examine the performance of our method as the proportion of attackers increases. Figure 4 shows the changes in performance metrics for varying percentages of attackers for both FEDQV and FEDAVG under the backdoor attack (depicted in Figure 4a) and Neurotoxin attack (shown in Figure 4b) for both FEDQV and FEDAVG. Across scenarios where the percentage of attackers m ranges from 10% to 50%, FEDQV consistently outperforms FEDAVG in terms of both ACC and ASR under both attacks, even in scenarios where half the parties are malicious. Notably, the Neurotoxin attack exhibits a more pronounced impact on ACC reduction and ASR increase in both FEDQV and FEDAVG compared to the backdoor attack as the percentage of attackers varies.

Unlike untargeted attacks, strong targeted attacks can be mounted with just a single attacker and data poisoning [61]. To investigate the behaviour of FEDQV in scenarios with finer gradations, we also evaluate it with small, realistic percentages of attackers, same as in [61], in Table 2 and Appendix Table A2. The results indicate that under these small, realistic percentages of attackers, the performance of FEDQV remains consistent with its behaviour under larger attack scenarios consistently outperforming FEDAVG. This outcome underscores the robustness of FEDQV across varying threat levels.

	MNIST		Fashion-MNIST		CIFAR10		FEMNIST	
	FEDAVG	FEDQV	FEDAVG	FEDQV	FEDAVG	FEDQV	FEDAVG	FEDQV
Data Poison								
Labelflip	98.81±0.03	98.54±0.05	**86.70±0.02**	85.22±0.05	66.88±0.48	67.36±0.22	74.92±2.55	**78.42±0.65**
Gaussian	9.68±0.41	10.49±0.46	10.00±0.00	**27.38±17.38**	15.29±0.57	**19.76±3.66**	4.64±0.13	4.83±0.25
Backdoor								
ACC	37.38±19.82	**98.30±0.15**	74.27±9.12	**78.40±3.95**	59.85±2.18	60.65±1.72	49.78±22.38	**75.20±3.96**
ASR	68.49±22.00	**0.19±0.07**	14.58±12.53	**7.05±6.35**	18.20±5.27	**3.21±1.30**	30.88±7.52	**28.26±9.57**
Scaling								
ACC	10.33±0.05	11.16±0.88	10.22±0.09	11.27±0.99	10.00±0.00	**28.55±18.55**	26.30±21.55	**64.80±1.38**
ASR	99.94±0.06	98.96±1.04	99.74±0.10	98.21±1.45	100.00±0.00	67.66±32.34	0.47±0.08	0.56±0.06
Neurotoxin								
ACC	81.17±15.39	**95.73±1.45**	70.00±7.85	**79.58±1.60**	22.40±7.16	**45.40±3.22**	47.29±18.07	**79.99±0.70**
ASR	23.19±2.25	**18.11±1.67**	20.65±2.21	18.12±4.16	51.63±1.03	57.42±1.91	40.42±4.35	**9.00±1.29**
Model Poison								
Krum	10.57±0.39	**97.96±0.14**	10.00±0.00	**79.43±0.86**	10.00±0.00	**53.27±1.12**	5.20±0.22	**51.86±3.06**
Trim	10.04±0.16	**98.36±0.11**	10.00±0.00	**84.45±0.70**	10.00±0.00	**57.33±2.34**	5.09±0.33	**52.19±4.52**
Min-Max	35.00±25.38	**85.32±6.45**	10.00±0.00	**67.25±7.44**	10.00±0.00	**19.07±6.97**	56.37±13.67	**72.58±2.11**
Min-Sum	96.69±0.94	95.97±0.59	10.88±0.87	**83.93±0.81**	17.40±4.27	**43.94±3.56**	52.56±23.91	**72.36±1.61**
QV-Adaptive	**71.43±22.67**	56.94±23.95	35.92±4.60	**62.13±11.25**	10.00±0.00	11.14±1.14	22.08±18.72	**43.78±20.72**

Table 1: Comparison of FEDQV and FEDAVG on four benchmark datasets under 10 attack scenarios with 30% malicious parties. The best results are highlighted in bold.



(a) Backdoor attack

(b) Neurotoxin attack

Figure 4: Performance comparison of FEDQV and FEDAVG in terms of Accuracy (ACC) and Attack Success Rate (ASR) over 100 epochs across four benchmark datasets under two targeted attacks, with varying percentages of attackers m from 10% to 50%.

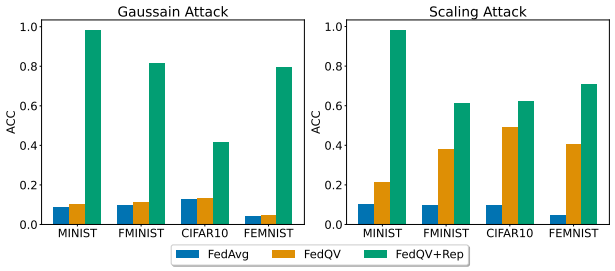
However, we notice that none of these methods yields satisfactory accuracy results for Gaussian and Scaling attacks. To address this, we present the enhanced version of FEDQV with an adaptive budget assigned according to a reputation model.

5.7 Adaptive Budget

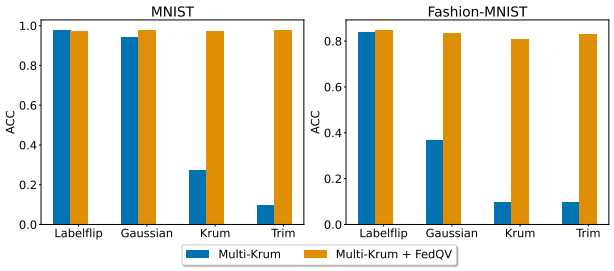
To enhance the robustness of FEDQV, we integrate it with the reputation model proposed in [13] to allocate unequal budgets based on the reputation scores of parties in each round t, as detailed in Section 3.4. The performance of FEDQV with an adaptive budget (referred to as FEDQV+REP) is compared with FEDAVG and FEDQV during the Gaussian and Scaling attacks, with the percentage of attackers increased to 50%. As depicted in Figure 5a, the combination of FEDQV and the reputation model significantly enhances resistance against Gaussian and Scaling attacks by at least a factor of 26%. This observation, formulated as the alternative hypothesis H_1, is supported by the Wilcoxon signed-rank test, where the null hypothesis H_0 can be rejected at a confidence level of 1% in favour of H_1. This integration effectively enhances the robustness of FEDQV, rendering it successful in defending against the two attacks that it previously failed to counter.

5.8 Integration with Byzantine-robust Aggregation

Our objective is *not to position FedQV in competition with existing defence techniques but rather to demonstrate that FedQV can act as a complementary approach to advanced defences.* Hence, FEDQV can be seamlessly integrated into Byzantine-robust defence to enhance the overall defence performance, by adapting the weight calculation process. We illustrate this with examples using Muilt-Krum [4], Trim-mean [12] and Reputation [13].



(a) Performance comparison of FEDAVG, FEDQV, and FEDQV+REP, in terms of ACC for 100 epochs in four benchmark datasets under 2 attack scenarios with 50% malicious parties.



(b) Test accuracy (ACC) for 100 epochs of both Krum alone and Multi-Krum + FEDQV on two benchmark datasets under 4 untargeted attack scenarios with 30% malicious parties.

Figure 5: Performance comparison of FEDAVG, FEDQV, and FEDQV+REP (FEDQV with adaptive budgets based on reputation model). Performance comparison of Krum alone and Multi-Krum integrated with FEDQV.

	Multi-Krum	Multi-Krum + FedQV	Trimmed-Mean	Trimmed-Mean + FedQV	Rep	Rep + FedQV
Neurotoxin						
1%	78.99±1.03/1.05±0.01	80.61±0.66/0.86±0.17	85.23±1.77/0.59±0.13	84.75±0.84/0.45±0.06	80.99±1.15/0.84±0.32	**85.82±0.55/0.38±0.06**
5%	76.21±0.73/3.29±0.92	80.32±1.07/**1.34±0.34**	85.15±0.38/0.87±0.17	85.09±0.97/0.73±0.05	80.48±1.17/1.62±0.11	**84.12±0.38/1.35±0.40**
10%	72.79±1.02/21.73±7.07	**77.41±1.36/16.30±3.01**	85.13±0.70/2.32±0.27	84.68±0.73/1.45±0.21	80.86±0.86/1.30±0.10	**83.78±0.09/0.66±0.04**
Min-Max						
10%	71.62±4.48	**79.48±0.82**	75.41 ±0.77	78.46±0.67	72.66±1.34	**76.98±1.48**
30%	52.29±0.34	**58.42±4.92**	59.62 ±1.20	60.24±6.81	54.94±0.82	**58.02±0.71**
50%	10.28±0.28	**22.95±9.94**	9.47±0.42	10.64±0.63	11.23±1.00	13.64±2.58
QV-Adaptive						
10%	52.98±1.78	**73.35±3.44**	83.17±1.85	**85.55±0.33**	12.60±2.36	**41.55±19.78**
30%	34.93±14.60	**55.07±11.40**	29.14±19.14	**42.17±25.95**	12.44±2.44	**38.44±14.32**
50%	10.20±0.20	12.24±1.12	10.00±0.00	**13.35±3.37**	10.00±0.00	10.55 ±0.44

Table 2: Comparison of Multi-Krum, Trimmed-Mean, Reputation, and their integration with FEDQV under three State-of-the-Art attacks on FEMNIST dataset. The best results are in bold. The results of targeted attacks are in the form of "ACC / ASR".

Table 2 presents a comparative analysis illustrating the performance of Multi-Krum, Trimmed-Mean, and Reputation methods both individually and when integrated with FEDQV under three state-of-the-art attacks conducted on the FEMNIST dataset. The results highlight that the integration of FEDQV with these defense mechanisms consistently yields superior performance characterised by higher ACC and lower ASR compared to their standalone counterparts.

Especially, in the context of Multi-Krum, as illustrated in Figure 5b, the integration of FEDQV (referred to as Multi-Krum + FEDQV) results in a significant enhancement in accuracy under 4 attack scenarios. Notably, under local poisoning attacks such as the Krum attack and Trim attack, the accuracy improves by a factor of 2.5. The results presented in Table 6a elucidate that the incorporation of FEDQV into Multi-Krum yields an average ACC enhancement of 26.72% and a significant average reduction of 70% in the ASR under two targeted attack scenarios: backdoor and scale attack. Setting this observation as the alternative hypothesis H_1 and Employing the Wilcoxon signed-rank test, we reject the null hypothesis H_0 at a confidence level of 1% in favour of H_1.

These findings underscore the potential of FEDQV *as a promising complementary method* to augment the performance of existing defence mechanisms.

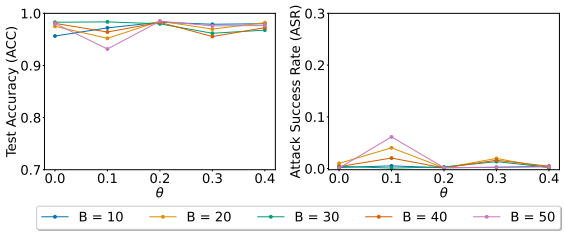
5.9 Non-IID Degree

To fulfil the fundamental setting of a heterogeneous and unbalanced dataset for FL, our experimental setup integrates datasets exhibiting non-IID properties, with a non-IID degree (ι) of 0.9, consistent with the settings outlined in previous studies such as [52, 13]. The ι represents the concentration parameter utilised in sampling from a Dirichlet distribution [62], which governs the degree of dataset imbalance. This parameter choice aims to generate non-IID and unbalanced data partitions conducive to our experimental objectives.

Moreover, we have examined the performance of FEDQV and FEDAVG across varying degrees of non-IID data, ranging from 0.1 to 0.9, as depicted in Appendix Table A3. These results demonstrate that as the non-IID degree increases among the parties, the performance of the global model declines. Notably, FEDQV consistently maintains a superior performance compared to FEDAVG, even when confronted with different degrees of data heterogeneity under attack conditions.

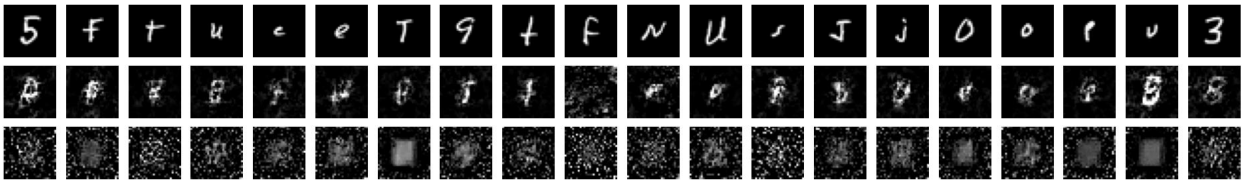
	MNIST		Fashion-MNIST	
	Multi-Krum	+ FEDQV	Multi-Krum	+ FEDQV
Backdoor				
ACC	70.20±9.99	**89.96±1.85**	33.24±13.24	**70.89±3.17**
ASR	32.03±11.20	**9.59±2.28**	68.87±17.77	**9.72±4.50**
Scaling				
ACC	68.35±16.76	**96.55±0.41**	59.43±14.22	**82.48±0.24**
ASR	33.65±19.15	**0.41±0.06**	33.64±19.08	**0.91±0.18**

(a) Comparison of Multi-Krum and Multi-Krum + FEDQV under two targeted attacks with 30% malicious parties in MNIST and Fashion-MNIST dataset. The best results are in bold.

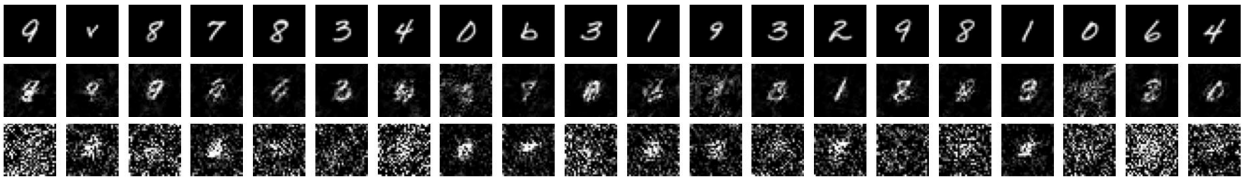


(b) Impact of Hyperparameters B and θ of FEDQV on ACC and ASR under a backdoor attack scenario with 30% malicious parties, using the MNIST dataset.

Figure 6: Comparison of Multi-Krum and Multi-Krum + FEDQV under two targeted attacks, and examining the impact of hyperparameters.



(a) The Deep Leakage from Gradients (DLG) Attack.



(b) The Gradient Invention (GI) Attack.

Figure 7: Visual Comparison of Images Reconstructed by Privacy Attacks (DLG and GI) on FEDQV with and without SECAGG. The first row displays the original private training images from the targeted client. The second row illustrates reconstructed images resulting from the privacy attacks (DLG and GI) on FEDQV without SECAGG. In the third row, reconstructed images from the privacy attacks (DLG and GI) on FEDQV with SECAGG are presented. The use of SECAGG is observed to contribute significantly to mitigating information leakage from the images.

5.10 Impact of Hyperparameters

As noted, Theorem 4.10 provides general guidelines for hyperparameters tuning. As shown in Remark 4.13, the error rate is influenced by B and θ. To demonstrate the impact of these two hyper-parameters, we conduct a grid search over B in $[10, 20, 30, 40, 50]$ and θ in $[0.1, 0.2, 0.3, 0.4, 0.5]$. The setup is the same as on the MNIST dataset under the backdoor attack with 30% malicious parties. Figure 6b illustrates the stability of the metrics ACC and ASR when varying the hyperparameters B and θ. Specifically, as B increases, there is a discernible decline in ACC accompanied by an increase in ASR, attributable to the augmented voting budget allotted to malicious parties. Conversely, an increase in θ yields an elevation in ACC coupled with a reduction in ASR, as a higher threshold implies stricter acceptance criteria for abnormal similarity scores. The optimal values of B and θ are determined to be 30 and 0.2, respectively.

We can see from Theorem 4.10, that the number of malicious devices m will affect the algorithm, and more malicious devices can lead to increased damage. However, this does mean the server needs to know the number of malicious devices to do the fine-tuning. We agree that determining optimal parameters can be challenging, especially in the absence of complete knowledge about the FL system. A better tuning is possible if more information is available. For specific tasks, more information can indeed be collected from which practical parameter sets can be extracted either via exhaustive search or via simpler online algorithms using trial and error. We will add this to our future work and consider it when we study particular domain-specific problems using our method.

6 Defence against Privacy Attacks

To illustrate the compatibility of FEDQV with existing privacy-guaranteeing mechanisms, we integrate SECAGG[11] into the FEDQV framework. Subsequently, we assess the framework's resilience against privacy attacks, specifically the Deep Leakage from Gradients (DLG)[8] and Gradient Inversion (GI) [9] attacks, with and without the incorporation of SECAGG. Figure 7a provides visualisations of the reconstructed images by the DLG attack. In the absence of SECAGG, the DLG attack can successfully recover approximately 8 out of 20 private images from the party. However, upon integrating SECAGG, the DLG attack fails to recover any of the party's images. Figure **??** presents visualisations of the reconstructed images by the GI attack. Without SECAGG, the GI attack successfully recovers half of the private images from the party. Yet, with SECAGG, the GI attack is unable to recover any party images, although it may vaguely leak the overall shape of the image. These results highlight the efficacy of SECAGG in strengthening FEDQV against privacy attacks, with the GI attack exhibiting a higher level of attack potency compared to the DLG attack. More details regarding privacy attacks and implementations are provided in Appendix C.

7 Conclusion

In this paper, we have proposed FEDQV, a novel aggregation scheme for FL based on quadratic voting instead of *1p1v*, which is the underlying principle that makes the currently employed FEDAVG vulnerable to poisoning attacks. The proposed method aggregates local models based upon the votes from a truthful mechanism employed in FEDQV. The efficiency of the proposed method has been comprehensively analysed from both a theoretical and an experimental point of view. Collectively, our performance evaluation has shown that FEDQV achieves superior performance than FEDAVG in defending against various poisoning attacks. Moreover, FEDQV is a reusable module that can be integrated with reputation models to assign unequal voting budgets, incorporate Byzantine-robust techniques, and employ privacy-preserving mechanisms. This versatility provides robustness against both poisoning and privacy attacks, positioning FEDQV as a promising complement to existing aggregation in FL.

Acknowledgement

Tianyue Chu and Nikolaos Laoutaris were supported by the MLEDGE project (REGAGE22e00052829516), funded by the Ministry of Economic Affairs and Digital Transformation and the European Union NextGenerationEU/PRTR.

References

[1] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[2] Swaroop Ramaswamy, Rajiv Mathews, Kanishka Rao, and Françoise Beaufays. Federated learning for emoji prediction in a mobile keyboard. *arXiv preprint arXiv:1906.04329*, 2019.

[3] Andrew Hard, Kanishka Rao, Rajiv Mathews, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. Federated learning for mobile keyboard prediction. *CoRR*, abs/1811.03604, 2018.

[4] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 119–129, Long Beach, CA, USA, 2017. Curran Associates, Inc.

[5] Steven P Lalley and E Glen Weyl. Quadratic voting: How mechanism design can radicalize democracy. In *AEA Papers and Proceedings*, volume 108, pages 33–37, 2018.

[6] Eric A Posner and E Glen Weyl. Voting squared: Quadratic voting in democratic politics. *Vand. L. Rev.*, 68:441, 2015.

[7] E Glen Weyl. The robustness of quadratic voting. *Public choice*, 172(1):75–107, 2017.

[8] Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, Vancouver Convention Center, Vancouver CANADA, 2019. Curran Associates, Inc.

[9] Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients-how easy is it to break privacy in federated learning? *Advances in Neural Information Processing Systems*, 33:16937–16947, 2020.

[10] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE symposium on security and privacy (SP)*, pages 691–706. IEEE, 2019.

[11] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS '17, page 1175–1191, New York, NY, USA, 2017. Association for Computing Machinery.

[12] Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International Conference on Machine Learning*, pages 5650–5659. PMLR, 2018.

[13] Tianyue Chu, Álvaro García-Recuero, Costas Iordanou, Georgios Smaragdakis, and Nikolaos Laoutaris. Securing federated sensitive topic classification against poisoning attacks. In *30th Annual Network and Distributed System Security Symposium, NDSS 2023*, San Diego, California, USA, 2023. The Internet Society.

[14] Hector Garcia-Molina. Elections in a distributed computing system. *IEEE transactions on Computers*, 31(01):48–59, 1982.

[15] Mack W Alford, Jean-Pierre Ansart, Günter Hommel, Leslie Lamport, Barbara Liskov, Geoff P Mullery, and Fred B Schneider. *Distributed systems: methods and tools for specification. An advanced course*. Springer, Verlag, 1985.

[16] Kai Yue, Richeng Jin, Chau-Wai Wong, and Huaiyu Dai. Federated learning via plurality vote. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–14, 2022.

[17] Yinghao Liu, Zipei Fan, Xuan Song, and Ryosuke Shibasaki. Fedvoting: A cross-silo boosting tree construction method for privacy-preserving long-term human mobility prediction. *Sensors*, 21(24):8282, 2021.

[18] Jy-yong Sohn, Dong-Jun Han, Beongjun Choi, and Jaekyun Moon. Election coding for distributed learning: Protecting signsgd against byzantine attacks. *Advances in Neural Information Processing Systems*, 33:14615–14625, 2020.

[19] Shashank Rajput, Hongyi Wang, Zachary Charles, and Dimitris Papailiopoulos. Detox: A redundancy-based framework for faster and more robust gradient aggregation. *Advances in Neural Information Processing Systems*, 32, 2019.

[20] Lingjiao Chen, Hongyi Wang, Zachary Charles, and Dimitris Papailiopoulos. Draco: Byzantine-resilient distributed training via redundant gradients. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 902–911, Stockholmsmässan, Stockholm, Sweden, 2018. PMLR.

[21] Konstantinos Konstantinidis and Aditya Ramamoorthy. Byzshield: An efficient and robust system for distributed training. *Proceedings of Machine Learning and Systems*, 3:812–828, 2021.

[22] Giovanni Sartori. *The theory of democracy revisited*, volume 2. NJ, 1987.

[23] Steven P Lalley, E Glen Weyl, et al. Quadratic voting. *Available at SSRN*, 2016.

[24] Bharat Chandar and E Glen Weyl. Quadratic voting in finite populations. *Available at SSRN: https://ssrn.com/abstract=2571026 or http://dx.doi.org/10.2139/ssrn.2571026*, 2019.

[25] Nicolaus Tideman and Florenz Plassmann. Efficient collective decision-making, marginal cost pricing, and quadratic voting. *Public Choice*, 172(1):45–73, 2017.

[26] Alessandra Casella and Luis Sanchez. Storable votes and quadratic voting. an experiment on four california propositions. Technical report, National Bureau of Economic Research, 2019.

[27] David Quarfoot, Douglas von Kohorn, Kevin Slavin, Rory Sutherland, David Goldstein, and Ellen Konar. Quadratic voting in the wild: real people, real votes. *Public Choice*, 172(1):283–303, 2017.

[28] Cong Xie, Sanmi Koyejo, and Indranil Gupta. Zeno: Distributed stochastic gradient descent with suspicion-based fault-tolerance. In *International Conference on Machine Learning*, pages 6893–6901. PMLR, 2019.

[29] Hanxi Guo, Hao Wang, Tao Song, Yang Hua, Zhangcheng Lv, Xiulang Jin, Zhengui Xue, Ruhui Ma, and Haibing Guan. Siren: Byzantine-robust federated learning via proactive alarming. In *Proceedings of the ACM Symposium on Cloud Computing*, SoCC '21, page 47–60, New York, NY, USA, 2021. Association for Computing Machinery.

[30] Xiaoyu Cao, Minghong Fang, Jia Liu, and Neil Gong. Fltrust: Byzantine-robust federated learning via trust bootstrapping. In *28th Annual Network and Distributed System Security Symposium, NDSS 2021, February 21-25, 2021*, virtually, 2021. The Internet Society.

[31] Mohamad Mansouri, Melek Önen, Wafa Ben Jaballah, and Mauro Conti. Sok: Secure aggregation based on cryptographic schemes for federated learning. *Proc. Priv. Enhancing Technol.*, 2023.

[32] Slawomir Goryczka, Li Xiong, and Vaidy S. Sunderam. Secure multiparty aggregation with differential privacy: a comparative study. In *EDBT/ICDT Conferences, EDBT/ICDT*, New York, NY, USA, 2013. Association for Computing Machinery.

[33] Lingchen Zhao, Jianlin Jiang, Bo Feng, Qian Wang, Chao Shen, and Qi Li. SEAR: secure and efficient aggregation for byzantine-robust federated learning. *IEEE Trans. Dependable Secur. Comput.*, 2022.

[34] Yuval Ishai, Eyal Kushilevitz, Rafail Ostrovsky, and Amit Sahai. Cryptography from anonymity. In *Annual IEEE Symposium on Foundations of Computer Science (FOCS*. IEEE Computer Society, 2006.

[35] Georgia Tsaloli, Bei Liang, Carlo Brunetta, Gustavo Banegas, and Aikaterini Mitrokotsa. sfDEVA: decentralized, verifiable secure aggregation for privacy-preserving learning. In *Information Security -International Conference, ISC*, Lecture Notes in Computer Science. Springer, 2021.

[36] Danye Wu, Miao Pan, Zhiwei Xu, Yujun Zhang, and Zhu Han. Towards efficient secure aggregation for model update in federated learning. In *IEEE Global Communications Conference, GLOBECOM*. IEEE, 2020.

[37] Xu Ma, Yuqing Zhou, Laihua Wang, and Meixia Miao. Privacy-preserving byzantine-robust federated learning. *Computer Standards & Interfaces*, 80:103561, 2022.

[38] Jinhyun So, Başak Güler, and A Salman Avestimehr. Byzantine-resilient secure federated learning. *IEEE Journal on Selected Areas in Communications*, 39(7):2168–2181, 2020.

[39] Felix Marx, Thomas Schneider, Ajith Suresh, Tobias Wehrle, Christian Weinert, and Hossein Yalame. Hyfl: A hybrid approach for private federated learning. *arXiv preprint arXiv:2302.09904*, 2023.

[40] H Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. In *International Conference on Learning Representations*, 2018.

[41] Liad Blumrosen and Noam Nisan. *Algorithmic game theory*. Springer, Berlin, Heidelberg, 2007.

[42] Felix Sattler, Simon Wiedemann, Klaus-Robert Müller, and Wojciech Samek. Sparse binary compression: Towards distributed deep learning with minimal communication. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019.

[43] Cynthia Dwork. Differential privacy. In *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II 33*, pages 1–12, Berlin, Heidelberg, 2006. Springer.

[44] Hao Yu, Sen Yang, and Shenghuo Zhu. Parallel restarted sgd with faster convergence and less communication: Demystifying why model averaging works for deep learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33,01, pages 5693–5700, 2019.

[45] X. Cao, J. Jia, Z. Zhang, and N. Gong. Fedrecover: Recovering from poisoning attacks in federated learning using historical information. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 1366–1383, Los Alamitos, CA, USA, may 2023. IEEE Computer Society.

[46] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. In *International Conference on Learning Representations(ICLR)*, 2020.

[47] Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

[48] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[49] Sebastian Caldas, Peter Wu, Tian Li, Jakub Konečný, H. Brendan McMahan, Virginia Smith, and Ameet Talwalkar. LEAF: A benchmark for federated settings. *CoRR*, abs/1812.01097, 2018.

[50] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[51] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, pages 770–778, Las Vegas, NV, USA, 2016. IEEE Computer Society.

[52] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2938–2948. PMLR, 26–28 Aug 2020.

[53] Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *CoRR*, abs/1909.06335, 2019.

[54] Minghong Fang, Xiaoyu Cao, Jinyuan Jia, and Neil Gong. Local model poisoning attacks to byzantine-robust federated learning. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 1605–1622, Virtual, 2020. USENIX Association.

[55] Bo Zhao, Peng Sun, Tao Wang, and Keyu Jiang. Fedinv: Byzantine-robust federated learning by inversing local model updates. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 36(8), 9171–9179, 2022.

[56] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019.

[57] Zhengming Zhang, Ashwinee Panda, Linyue Song, Yaoqing Yang, Michael Mahoney, Prateek Mittal, Ramchandran Kannan, and Joseph Gonzalez. Neurotoxin: Durable backdoors in federated learning. In *International Conference on Machine Learning (ICML)*, pages 26429–26446. PMLR, 2022.

[58] Virat Shejwalkar and Amir Houmansadr. Manipulating the byzantine: Optimizing model poisoning attacks and defenses for federated learning. In *28th Annual Network and Distributed System Security Symposium, NDSS 2021, virtually, February 21-25, 2021*. The Internet Society, 2021.

[59] Eugene Bagdasaryan and Vitaly Shmatikov. Blind backdoors in deep learning models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 1505–1521, virtually, 2021. USENIX Association.

[60] Miguel Castro and Barbara Liskov. Practical byzantine fault tolerance. In *OsDI*, pages 173–186, New York, NY, United States, 1999. Association for Computing Machinery.

[61] Virat Shejwalkar, Amir Houmansadr, Peter Kairouz, and Daniel Ramage. Back to the drawing board: A critical evaluation of poisoning attacks on production federated learning. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1354–1371. IEEE, 2022.

[62] Thomas Minka. Estimating a dirichlet distribution, 2000.

[63] Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

[64] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS 2017 Workshop on Autodiff*, 2017.

Appendix

A Proof of Theoretical Analysis

A.1 Proof of Theorem 4.9 and Theorem 4.10

A.1.1 Proof of Lemmas

Lemmas 4.5, Lemmas 4.6, Lemmas 4.7 and Lemmas 4.8 are all the lemmas we utilise during the proof of Theorem 4.9, and we prove them in that order. Notice, Lemmas 4.5 are used in the proof Lemmas 4.6, and Theorem 4.9 is proved using Lemmas 4.6, Lemmas 4.7 and Lemmas 4.8 in order.

Proof of Lemma 4.5

Proof. Let $g(\boldsymbol{w}) = \ell(\boldsymbol{w}) - \frac{\varsigma}{2}\|\boldsymbol{w}\|_2^2$. Base on the Assumption 4.2, we have $g(\boldsymbol{w})$ is $(L-\varsigma)$-strongly convex. from [63] Equation 3.6, we have

$$\langle \nabla\ell(\boldsymbol{w}_1) - \nabla\ell(\boldsymbol{w}_2), \boldsymbol{w}_1 - \boldsymbol{w}_2\rangle \geq \frac{1}{L}\|\nabla\ell(\boldsymbol{w}_1) - \nabla\ell(\boldsymbol{w}_2)\|_2^2 \tag{12}$$

Hence,

$$\langle \nabla g(\boldsymbol{w}_1) - \nabla g(\boldsymbol{w}_2), \boldsymbol{w}_1 - \boldsymbol{w}_2\rangle \geq \frac{1}{L-\varsigma}\|\nabla g(\boldsymbol{w}_1) - \nabla g(\boldsymbol{w}_2)\|_2^2 \tag{13}$$

Now We have

$$\begin{aligned}\langle \nabla \left(\ell(\boldsymbol{w}_1) - \frac{\varsigma}{2}\left\|\boldsymbol{w}_1\right\|_2^2\right) - \nabla \left(\ell(\boldsymbol{w}_2) - \frac{\varsigma}{2}\left\|\boldsymbol{w}_2\right\|_2^2\right), \boldsymbol{w}_1 - \boldsymbol{w}_2\rangle \\ \geq \frac{1}{L+\mu}\left\|\nabla \left(\ell(\mathbf{w_1}) - \frac{\varsigma}{2}\left\|\boldsymbol{w}_1\right\|_2^2\right) - \nabla \left(\ell(\boldsymbol{w}_2) - \frac{\varsigma}{2}\left\|\boldsymbol{w}_2\right\|_2^2\right)\right\|_2^2 \end{aligned} \tag{14}$$

And therefore

$$\begin{aligned}\langle \nabla\ell(\boldsymbol{w}_1) - \nabla\ell(\boldsymbol{w}_2), \boldsymbol{w}_1 - \boldsymbol{w}_2\rangle - \langle \varsigma\boldsymbol{w}_1 - \varsigma\boldsymbol{w}_2, \boldsymbol{w}_1 - \boldsymbol{w}_2\rangle \\ \geq \frac{1}{L-\varsigma}\left\|(\nabla\ell(\boldsymbol{w}_1) - \nabla\ell(\boldsymbol{w}_2)) - (\varsigma\boldsymbol{w}_1 - \varsigma\boldsymbol{w}_2)\right\|_2^2 \end{aligned} \tag{15}$$

Refer to Assumption 4.1, we obtain

$$\begin{aligned}\langle \nabla\ell(\boldsymbol{w}_1) - \nabla\ell(\boldsymbol{w}_2), \boldsymbol{w}_1 - \boldsymbol{w}_2\rangle &\geq \frac{L\varsigma}{L-\varsigma}\left\|\boldsymbol{w}_1 - \boldsymbol{w}_2\right\|_2^2 - \frac{2\varsigma}{L-\varsigma}\langle \nabla\ell(\boldsymbol{w}_1) - \nabla\ell(\boldsymbol{w}_2), \boldsymbol{w}_1 - \boldsymbol{w}_2\rangle \\ &+ \frac{1}{L-\varsigma}\left\|\nabla\ell(\boldsymbol{w}_1) - \nabla\ell(\boldsymbol{w}_2)\right\|_2^2 \\ &\geq -\frac{L\varsigma}{L-\varsigma}\left\|\boldsymbol{w}_1 - \boldsymbol{w}_2\right\|_2^2 + \frac{1}{L-\varsigma}\left\|\nabla\ell(\boldsymbol{w}_1) - \nabla\ell(\boldsymbol{w}_2)\right\|_2^2 \end{aligned} \tag{16}$$

Let $\varsigma = -\mu$, then we conclude the proof of Lemma 4.5.

□

Proof of Lemma 4.6

Proof. We have

$$\left\|\boldsymbol{w}^{t-1} - r_{t-1}\nabla\mathcal{L}(\boldsymbol{w}^{t-1}) - \boldsymbol{w}^*\right\|_2^2 = \left\|\boldsymbol{w}^{t-1} - \boldsymbol{w}^*\right\|_2^2 \underbrace{-2r_{t-1}\left\langle \nabla\mathcal{L}(\boldsymbol{w}^{t-1}), \boldsymbol{w}^{t-1} - \boldsymbol{w}^*\right\rangle}_{\mathbf{A1}} + \underbrace{r_{t-1}^2\left\|\nabla\mathcal{L}(\boldsymbol{w}^{t-1})\right\|_2^2}_{\mathbf{A2}} \tag{17}$$

For part $\mathbf{A1}$ under the Assumption 4.2, Lemma 4.5 and Maclaurin inequality, we have

$$\begin{aligned}\mathbf{A1} &= -2r_{t-1}\sum_{i=1}^{N} p_i^{t-1}\left\langle \nabla\ell(\boldsymbol{w}_i^{t-1}), \boldsymbol{w}^{t-1} - \boldsymbol{w}^*\right\rangle \\ &= -2r_{t-1}\sum_{i=1}^{N} p_i^{t-1}\left(\left\langle \nabla\ell(\boldsymbol{w}_i^{t-1}), \boldsymbol{w}^{t-1} - \boldsymbol{w}_i^{t-1}\right\rangle\right) \\ &- 2r_{t-1}\sum_{i=1}^{N} p_i^{t-1}\left(\left\langle \nabla\ell(\boldsymbol{w}_i^{t-1}), \boldsymbol{w}_i^{t-1} - \boldsymbol{w}^*\right\rangle\right) \\ &\leq \sum_{i=1}^{N} p_i^{t-1}\left(r_{t-1}^2\left\|\nabla\ell(\boldsymbol{w}_i^{t-1})\right\|_2^2 + \left\|\boldsymbol{w}^{t-1} - \boldsymbol{w}_i^{t-1}\right\|_2^2\right) - \\ &2r_{t-1}\sum_{i=1}^{N} p_i^{t-1}\left(\frac{1}{L+\mu}\left\|\nabla\ell(\boldsymbol{w}_i^{t-1})\right\|_2^2 + \frac{L\mu}{L+\mu}\left\|\boldsymbol{w}_i^{t-1} - \boldsymbol{w}^*\right\|_2^2\right) \\ &= \left(r_{t-1}^2 - \frac{1}{L+\mu}\right)\sum_{i=1}^{N} p_i^{t-1}\left(\left\|\nabla\ell(\boldsymbol{w}_i^{t-1})\right\|_2^2\right) \\ &+ \sum_{i=1}^{N} p_i^{t-1}\left\|\boldsymbol{w}^{t-1} - \boldsymbol{w}_i^{t-1}\right\|_2^2 - \frac{2r_{t-1}L\mu}{L+\mu}\left\|\boldsymbol{w}^{t-1} - \boldsymbol{w}^*\right\|_2^2 \end{aligned}$$

From Assumption 4.1 and Jensen inequality, we can derive:

$$\left\|\nabla\ell(\boldsymbol{w}_i^{t-1}) - \nabla\ell(\boldsymbol{w}^*)\right\|_2^2 \leq L^2 \left\|\boldsymbol{w}_i^{t-1} - \boldsymbol{w}^*\right\|_2^2 \tag{18}$$

Hence for $\mathbf{A1}$, by Jensen inequality and Equation 18, we have

$$\begin{aligned}
\mathbf{A1} &\leq \left(r_{t-1}^2 - \frac{1}{L+\mu}\right) \sum_{i=1}^N p_i^{t-1} \left(\left\|\nabla\ell(\boldsymbol{w}_i^{t-1})\right\|_2^2\right) \\
&+ \sum_{i=1}^N p_i^{t-1} \left\|\boldsymbol{w}^{t-1} - \boldsymbol{w}_i^{t-1}\right\|_2^2 - \frac{2r_{t-1}L\mu}{L+\mu} \left\|\boldsymbol{w}^{t-1} - \boldsymbol{w}^*\right\|_2^2 \\
&\leq \left(r_{t-1}^2 - \frac{1}{L+\mu}\right) \sum_{i=1}^N p_i^{t-1} \left\|\boldsymbol{w}_i^{t-1} - \boldsymbol{w}^*\right\|_2^2 \\
&+ \sum_{i=1}^N p_i^{t-1} \left\|\boldsymbol{w}^{t-1} - \boldsymbol{w}_i^{t-1}\right\|_2^2 - \frac{2r_{t-1}L\mu}{L+\mu} \left\|\boldsymbol{w}^{t-1} - \boldsymbol{w}^*\right\|_2^2 \\
&\leq \left(r_{t-1}^2 - \frac{2r_{t-1}L\mu + 1}{L+\mu}\right) \left\|\boldsymbol{w}^{t-1} - \boldsymbol{w}^*\right\|_2^2 \\
&+ \sum_{i=1}^N p_i^{t-1} \left\|\boldsymbol{w}^{t-1} - \boldsymbol{w}_i^{t-1}\right\|_2^2
\end{aligned}$$

Similar for $\mathbf{A2}$, we have

$$\begin{aligned}
\mathbf{A2} &= r_{t-1}^2 \left\|\sum_{i=1}^N p_i^{t-1} \nabla\ell(\boldsymbol{w}_i^{t-1})\right\|_2^2 \leq r_{t-1}^2 \sum_{i=1}^N p_i^{t-1} \left\|\nabla\ell(\boldsymbol{w}_i^{t-1})\right\|_2^2 \\
&\leq r_{t-1}^2 L^2 \sum_{i=1}^N p_i^{t-1} \left\|\boldsymbol{w}_i^{t-1} - \boldsymbol{w}^*\right\|_2^2 = r_{t-1}^2 L^2 \left\|\boldsymbol{w}^{t-1} - \boldsymbol{w}^*\right\|_2^2
\end{aligned}$$

Then we combine results of $\mathbf{A1}$ and $\mathbf{A2}$ for Equation 17, it follows that

$$\begin{aligned}
\left\|\boldsymbol{w}^{t-1} - r_{t-1}\nabla\mathcal{L}(\boldsymbol{w}^{t-1}) - \boldsymbol{w}^*\right\|_2^2 \leq &\left(r_{t-1}^2\left(1+L^2\right) - \frac{2r_{t-1}L\mu + 1}{L+\mu}\right) \left\|\boldsymbol{w}^{t-1} - \boldsymbol{w}^*\right\|_2^2 \\
&+ \sum_{i=1}^N p_i^{t-1} \left\|\boldsymbol{w}^{t-1} - \boldsymbol{w}_i^{t-1}\right\|_2^2
\end{aligned} \tag{19}$$

□

Proof of Lemma 4.7

Proof. For each E step FL necessitates a communication. As a result, for any $t-1 \geq 0$, $\exists t^* \leq t-1$ that $t - t^* \leq E, t^* \in T$, accordingly $\forall i, j \in \mathcal{S}^{t^*}, \boldsymbol{w}_i^{t^*} = \boldsymbol{w}_j^{t^*} = \boldsymbol{w}^{t^*}$. Then, based on $\mathbb{E}\left\|\mathbf{X} - \mathbb{E}\,\mathbf{X}\right\|_2^2 \leq \mathbb{E}\left\|\mathbf{X}\right\|_2^2$, Jensen inequality and Assumption 4.3, we have

$$
\begin{aligned}
\mathbb{E}\sum_{i=1}^{N} p_i^{t-1}\left\|\boldsymbol{w}^{t-1}-\boldsymbol{w}_i^{t-1}\right\|_2^2 &= \mathbb{E}_{\mathcal{S}^{t^*}}\sum_{i\in\mathcal{S}^{t^*}} p_i^{t-1}\left\|\left(\boldsymbol{w}_i^{t-1}-\boldsymbol{w}^{t^*}\right)-\left(\boldsymbol{w}^{t-1}-\boldsymbol{w}^{t^*}\right)\right\|_2^2 \\
&= \mathbb{E}_{\mathcal{S}^{t^*}}\left[\mathbb{E}_{\mathcal{S}^{t^*}}\left\|\left(\boldsymbol{w}_i^{t-1}-\boldsymbol{w}^{t^*}\right)-\mathbb{E}_{\mathcal{S}^{t^*}}\left[\boldsymbol{w}_i^{t-1}-\boldsymbol{w}^{t^*}\right]\right\|_2^2\right) \\
&\le \mathbb{E}_{\mathcal{S}^{t^*}}\left[\mathbb{E}_{\mathcal{S}^{t^*}}\left\|\left(\boldsymbol{w}_i^{t-1}-\boldsymbol{w}^{t^*}\right)\right\|_2^2\right] \\
&= \mathbb{E}_{\mathcal{S}^{t^*}}\sum_{i\in\mathcal{S}^{t^*}} p_i^{t-1}\left\|\boldsymbol{w}_i^{t-1}-\boldsymbol{w}^{t^*}\right\|_2^2 \\
&= \mathbb{E}_{\mathcal{S}^{t^*}}\sum_{i\in\mathcal{S}^{t^*}} p_i^{t-1}\left\|\sum_{t=t^*}^{t-2}\nabla\ell(\boldsymbol{w}_i^{t-1},D_i^{t-1})\right\|_2^2 \\
&\le \sum_{i\in\mathcal{S}^{t^*}} p_i^{t-1}\,\mathbb{E}_{\mathcal{S}^{t^*}}\left(t-1-t^*\right)\sum_{t=t^*}^{t-2} r_{t-1}^2\left\|\nabla\ell(\boldsymbol{w}_i^{t-1},D_i^{t-1})\right\|_2^2 \\
&\le \sum_{i\in\mathcal{S}^{t^*}} p_i^{t-1}\left(E-1\right)\sum_{t=t^*}^{t-2} r_{t-1}^2\left\|\nabla\ell(\boldsymbol{w}_i^{t-1},D_i^{t-1})\right\|_2^2 \\
&\le \sum_{i\in\mathcal{S}^{t^*}} p_i^{t-1}\left(E-1\right)\sum_{t=t^*}^{t-2} r_{t-1}^2\mathcal{G}_{\boldsymbol{w}}^2 \\
&\le \sum_{i\in\mathcal{S}^{t^*}} p_i^{t-1}\left(E-1\right)^2 r_{t-1}^2\mathcal{G}_{\boldsymbol{w}}^2 \\
&\le \left(E-1\right)^2 r_{t-1}^2\mathcal{G}_{\boldsymbol{w}}^2
\end{aligned}
\tag{20}
$$

□

Proof of Lemma 4.8

Proof. Due to Assumption 4.4 and Algorithm 1, we have

$$
\begin{aligned}
\mathbb{E}\left\|\mathcal{F}(\boldsymbol{w}^{t-1})-\nabla\mathcal{L}(\boldsymbol{w}^{t-1})\right\|_2^2 = \mathrm{Var}\left(\mathcal{F}(\boldsymbol{w}^{t-1})\right) &= \mathbb{E}_{\mathcal{S}^{t-1}}\left\|\sum_{i\in\mathcal{S}^{t-1}} p_i^{t-1}\left(\nabla\ell(\boldsymbol{w}_i^{t-1};\mathcal{D}_i^{t-1})-\nabla\ell(\boldsymbol{w}_i^{t-1})\right)\right\|_2^2 \\
&= \sum_{i\in\mathcal{S}^{t-1}}\left(p_i^{t-1}\right)^2\mathbb{E}\left\|\nabla\ell(\boldsymbol{w}_i^{t-1};\mathcal{D}_i^{t-1})-\nabla\ell(\boldsymbol{w}_i^{t-1})\right\|_2^2 \\
&\le \sum_{i\in\mathcal{S}^{t-1}}\left(p_i^{t-1}\right)^2\mathcal{V}_{\boldsymbol{w}} \le \mathcal{V}_{\boldsymbol{w}}\sum_{i\in\mathcal{S}^{t-1}}\left(\frac{v_i^{t-1}}{\sum_{i\in\mathcal{S}^{t-1}} v_i^{t-1}}\right)^2 \\
&\le \mathcal{V}_{\boldsymbol{w}}\frac{\sum_{i\in\mathcal{S}^{t-1}}\left(v_i^{t-1}\right)^2}{\left(\sum_{i\in\mathcal{S}^{t-1}} v_i^{t-1}\right)^2} \le \mathcal{V}_{\boldsymbol{w}}\frac{\sum_{i\in\mathcal{S}^{t-1}}\left(v_i^{t-1}\right)^2}{\sum_{i\in\mathcal{S}^{t-1}} v_i^{t-1}} \\
&\le \mathcal{V}_{\boldsymbol{w}}\sum_{i\in\mathcal{S}^{t-1}} v_i^{t-1} \le \left(1-2\theta\right)qN\mathcal{V}_{\boldsymbol{w}}\sqrt{B}
\end{aligned}
\tag{21}
$$

□

A.1.2 Proof of Theorem 4.9

Proof. In t round, due to $m = 0$, we have:

$$\begin{aligned}\left\|\boldsymbol{w}^t - \boldsymbol{w}^*\right\|_2^2 &= \left\|\boldsymbol{w}^{t-1} - r_{t-1}\mathcal{M}(\boldsymbol{w}^{t-1}) - \boldsymbol{w}^*\right\|_2^2 = \left\|\boldsymbol{w}^{t-1} - r_{t-1}\mathcal{F}(\boldsymbol{w}^{t-1}) - \boldsymbol{w}^*\right\|_2^2 \\ &= \underbrace{\left\|\boldsymbol{w}^{t-1} - r_{t-1}\nabla\mathcal{L}(\boldsymbol{w}^{t-1}) - \boldsymbol{w}^*\right\|_2^2}_{\mathbf{A}} + \underbrace{r_{t-1}^2\left\|\mathcal{F}(\boldsymbol{w}^{t-1}) - \nabla\mathcal{L}(\boldsymbol{w}^{t-1})\right\|_2^2}_{\mathbf{B}} \\ &+ \underbrace{2r_{t-1}\left\langle \boldsymbol{w}^{t-1} - r_{t-1}\nabla\mathcal{L}(\boldsymbol{w}^{t-1}) - \boldsymbol{w}^*, \mathcal{F}(\boldsymbol{w}^{t-1}) - \nabla\mathcal{L}(\boldsymbol{w}^{t-1})\right\rangle}_{\mathbf{C}}\end{aligned} \tag{22}$$

Where

$$\mathcal{M}(\boldsymbol{w}^{t-1}) = \sum_{i\in\mathcal{S}^{t-1}} p_i^{t-1}\mathcal{M}_i(\boldsymbol{w}_i^{t-1})$$

Note that $\mathbb{E}\,\mathbf{C} = 0$. For the expectation of A, from Lemma 4.6 and Lemma 4.7, it follows that

$$\begin{aligned}\mathbb{E}[\mathbf{A}] &= \mathbb{E}\left\|\boldsymbol{w}^{t-1} - r_{t-1}\nabla\mathcal{L}(\boldsymbol{w}^{t-1}) - \boldsymbol{w}^*\right\|_2^2 \\ &\leq \left(r_{t-1}^2\left(1+L^2\right) - \frac{2r_{t-1}L\mu+1}{L+\mu}\right)\left\|\boldsymbol{w}^{t-1} - \boldsymbol{w}^*\right\|_2^2 \\ &+ \left(E-1\right)^2 r_{t-1}^2\mathcal{G}_{\boldsymbol{w}}^2\end{aligned} \tag{23}$$

We use Lemma 4.8 to bound $\mathbf{B}$, we have

$$\mathbb{E}[\mathbf{B}] \leq r_{t-1}^2\left(1-2\theta\right)qN\mathcal{V}_{\boldsymbol{w}}\sqrt{B} \tag{24}$$

Hence, we have

$$\begin{aligned}\mathbb{E}\left\|\boldsymbol{w}^t - \boldsymbol{w}^*\right\|_2^2 \leq r_{t-1}^2\left(1+L^2\right)\mathbb{E}\left\|\boldsymbol{w}^{t-1} - \boldsymbol{w}^*\right\|_2^2 \\ - \frac{2r_{t-1}L\mu+1}{L+\mu}\mathbb{E}\left\|\boldsymbol{w}^{t-1} - \boldsymbol{w}^*\right\|_2^2 + r_{t-1}^2\Delta\end{aligned} \tag{25}$$

where

$$\Delta = \left(E-1\right)^2\mathcal{G}_{\boldsymbol{w}}^2 + \left(1-2\theta\right)qN\mathcal{V}_{\boldsymbol{w}}\sqrt{B}$$

For the learning rate r_t, $\exists\alpha > \frac{L+\mu}{2\mu L}, \exists\beta > 0$, such that $r_t = \frac{\alpha}{\beta+t} \leq \frac{1}{L+1}$. We use mathematical induction to prove the following statement:

Proposition: $\forall t \in \mathbb{N}, \mathbb{E}\left\|\boldsymbol{w}^t - \boldsymbol{w}^*\right\|_2^2 \leq \frac{\gamma}{\beta+t}$, where $\gamma = \max\left\{\frac{(L+\mu)\alpha^2\Delta}{2\alpha\mu L - L - \mu}, \beta\,\mathbb{E}\left\|\boldsymbol{w}^0 - \boldsymbol{w}^*\right\|_2^2\right\}$.

Let $P(t)$ be the statement $\mathbb{E}\left\|\boldsymbol{w}^t - \boldsymbol{w}^*\right\|_2^2 \leq \frac{\gamma}{\beta+t}$, we give a proof by induction on t.

Base case: The statement $P(0)$ holds for $t = 0$:

$$\mathbb{E}\left\|\boldsymbol{w}^0 - \boldsymbol{w}^*\right\|_2^2 \leq \frac{\gamma}{\beta}$$

Inductive step: Assume the induction hypothesis that for a particular j, the single case $t = j$ holds, meaning $P(j)$ is true:

$$\mathbb{E}\left\|\boldsymbol{w}^j - \boldsymbol{w}^*\right\|_2^2 \leq \frac{\gamma}{\beta+j}$$

It follows that:

$$\begin{aligned}\mathbb{E}\left\|\boldsymbol{w}^{j+1} - \boldsymbol{w}^*\right\|_2^2 &\leq \left(r_t^2\left(1+L^2\right) - \frac{2r_tL\mu+1}{L+\mu}\right)\mathbb{E}\left\|\boldsymbol{w}^j - \boldsymbol{w}^*\right\|_2^2 + r_t^2\Delta \\ &\leq \left(1 - \frac{2L\mu\alpha}{(L+\mu)(\beta+j)}\right)\frac{\gamma}{\beta+j} + \left(\frac{\alpha}{\beta+j}\right)^2\Delta \\ &= \left[\frac{\alpha^2\Delta}{(\beta+j)^2} - \frac{2\alpha\mu L - L - \mu}{(\beta+j)^2(L+\mu)}\gamma\right] + \frac{\beta+j-1}{(\beta+j)^2}\gamma \\ &\leq \frac{\gamma}{\beta+j+1}\end{aligned}$$

Therefore, the statement $P(j+1)$ also holds true, establishing the inductive step. Since both the base case and the inductive step have been proved as true, by mathematical induction the statement $P(t)$ holds for $\forall t \in \mathbb{N}$.

We choose $\alpha = \frac{L+\mu}{\mu L}$ and $\beta = 2\frac{(L+1)(L+\mu)}{\mu L}$, and we have

$$\begin{aligned}\gamma &= \max\left\{\frac{(L+\mu)\alpha^2\Delta}{2\alpha\mu L - L - \mu}, \beta\,\mathbb{E}\left\|\boldsymbol{w}^0 - \boldsymbol{w}^*\right\|_2^2\right\} \\ &\leq \frac{(L+\mu)\alpha^2\Delta}{2\alpha\mu L - L - \mu} + \beta\,\mathbb{E}\left\|\boldsymbol{w}^0 - \boldsymbol{w}^*\right\|_2^2 \\ &= \alpha^2\Delta + 2(L+1)\alpha\,\mathbb{E}\left\|\boldsymbol{w}^0 - \boldsymbol{w}^*\right\|_2^2\end{aligned}$$

Then based on Assumption 4.1 and Taylor expansion, we have the quadratic upper-bound of $\mathcal{L}(\cdot)$:

$$\mathcal{L}(\boldsymbol{w}_1) - \mathcal{L}(\boldsymbol{w}_2) \leq (\boldsymbol{w}_1 - \boldsymbol{w}_2)^T \nabla\mathcal{L}(\boldsymbol{w}_2) + \frac{L}{2}\left\|\boldsymbol{w}_1 - \boldsymbol{w}_2\right\|_2^2$$

It follows that

$$\begin{aligned}\mathbb{E}\,\mathcal{L}(\boldsymbol{w}^T) - \mathcal{L}(\boldsymbol{w}^*) &\leq \frac{L}{2}\,\mathbb{E}\left\|\boldsymbol{w}^T - \boldsymbol{w}^*\right\|_2^2 \leq \frac{\gamma L}{2(\beta+T)} \\ &\leq \frac{L}{2\alpha(L+1)+T}\left(\frac{\alpha^2}{2}\Delta + \alpha(L+1)\,\mathbb{E}\left\|\boldsymbol{w}^0 - \boldsymbol{w}^*\right\|_2^2\right) \\ &= \frac{L}{2\varphi+T}\left(\varphi\,\mathbb{E}\left\|\boldsymbol{w}^0 - \boldsymbol{w}^*\right\|_2^2 + \frac{\alpha^2}{2}\Delta\right)\end{aligned}$$

Where

$$\Delta = (E-1)^2\,\mathcal{G}_{\boldsymbol{w}}^2 + (1-2\theta)\,\mathcal{C}\mathcal{V}_{\boldsymbol{w}}\sqrt{B},\ \varphi = \alpha\,(L+1)$$

□

A.1.3 Proof of Theorem 4.10

Proof. In the t round, we have:

$$\begin{aligned}\left\|\boldsymbol{w}^t - \boldsymbol{w}^*\right\|_2^2 &= \left\|\boldsymbol{w}^{t-1} - r_{t-1}\mathcal{M}(\boldsymbol{w}^{t-1}) - \boldsymbol{w}^*\right\|_2^2 \\ &= \left\|\boldsymbol{w}^{t-1} - r_{t-1}\mathcal{F}(\boldsymbol{w}^{t-1}) - \boldsymbol{w}^* + r_{t-1}\mathcal{F}(\boldsymbol{w}^{t-1}) - r_{t-1}\mathcal{M}(\boldsymbol{w}^{t-1})\right\|_2^2 \\ &= \underbrace{\left\|\boldsymbol{w}^{t-1} - r_{t-1}\mathcal{F}(\boldsymbol{w}^{t-1}) - \boldsymbol{w}^*\right\|_2^2}_{\mathbf{A}} + \underbrace{r_{t-1}^2\left\|\mathcal{F}(\boldsymbol{w}^{t-1}) - \mathcal{M}(\boldsymbol{w}^{t-1})\right\|_2^2}_{\mathbf{B}} \\ &+ \underbrace{2r_{t-1}\left\langle\boldsymbol{w}^{t-1} - r_{t-1}\mathcal{F}(\boldsymbol{w}^{t-1}) - \boldsymbol{w}^*, \mathcal{F}(\boldsymbol{w}^{t-1}) - \mathcal{M}(\boldsymbol{w}^{t-1})\right\rangle}_{\mathbf{C}}\end{aligned} \tag{26}$$

Where

$$\mathcal{M}(\boldsymbol{w}^{t-1}) = \sum_{i\in\mathcal{S}^{t-1}} p_i^{t-1}\mathcal{M}_i(\boldsymbol{w}_i^{t-1})$$

For the expectation of A, from Theorem 4.9, it follows that

$$\mathbb{E}[\mathbf{A}] \leq \frac{1}{2\varphi+t}\left(2\varphi\,\mathbb{E}\left\|\boldsymbol{w}^0 - \boldsymbol{w}^*\right\|_2^2 + \alpha^2\Delta\right) \tag{27}$$

For $\mathbf{B}$, we have

$$\begin{aligned}\mathbb{E}[\mathbf{B}] &= r_{t-1}^2 \left\| \sum_{i\in\mathcal{S}^{t-1}} p_i^{t-1} \nabla \ell(\boldsymbol{w}_i^{t-1}) - \sum_{i\in\mathcal{S}^{t-1}} p_i^{t-1} \mathcal{M}_i(\boldsymbol{w}_i^{t-1}) \right\|_2^2 \\ &= r_{t-1}^2 \left\| \sum_{i\in\mathcal{S}^{t-1}} p_i^{t-1} \left(\nabla \ell(\boldsymbol{w}_i^{t-1}) - \mathcal{M}_i(\boldsymbol{w}_i^{t-1}) \right) \right\|_2^2 \\ &\leq r_{t-1}^2 \left\| \sum_{i\in mN} p_i^{t-1} \left(\nabla \ell(\boldsymbol{w}_i^{t-1}) - \mathcal{M}_i(\boldsymbol{w}_i^{t-1}) \right) \right\|_2^2 \end{aligned} \tag{28}$$

Where m is the percentage of the malicious parties.

Due to Equation 3, we have

$$\theta \leq \frac{\left\langle \nabla \ell(\boldsymbol{w}_i^{t-1}), \mathcal{M}_i(\boldsymbol{w}_i^{t-1}) \right\rangle}{\left\| \nabla \ell(\boldsymbol{w}_i^{t-1} \right\| \cdot \left\| \mathcal{M}_i(\boldsymbol{w}_i^{t-1}) \right\|} \leq 1 - \theta \tag{29}$$

Thus,

$$\theta \left\| \nabla \ell(\boldsymbol{w}_i^{t-1}) \right\| \left\| \mathcal{M}_i(\boldsymbol{w}_i^{t-1}) \right\| \leq \left\langle \nabla \ell(\boldsymbol{w}_i^{t-1}), \mathcal{M}_i(\boldsymbol{w}_i^{t-1}) \right\rangle \leq (1-\theta) \left\| \nabla \ell(\boldsymbol{w}_i^{t-1}) \right\| \left\| \mathcal{M}_i(\boldsymbol{w}_i^{t-1}) \right\| \tag{30}$$

Due to this, we have

$$\begin{aligned} &\left\| \nabla \ell(\boldsymbol{w}_i^{t-1}) \right\|_2^2 - 2(1-\theta) \left\| \nabla \ell(\boldsymbol{w}_i^{t-1} \right\| \left\| \mathcal{M}_i(\boldsymbol{w}_i^{t-1}) \right\| + \left\| \mathcal{M}_i(\boldsymbol{w}_i^{t-1}) \right\|_2^2 \\ &\leq \left\| \nabla \ell(\boldsymbol{w}_i^{t-1}) - \mathcal{M}_i(\boldsymbol{w}_i^{t-1}) \right\|_2^2 \\ &\leq \left\| \nabla \ell(\boldsymbol{w}_i^{t-1}) \right\|_2^2 - 2\theta \left\| \nabla \ell(\boldsymbol{w}_i^{t-1}) \right\| \left\| \mathcal{M}_i(\boldsymbol{w}_i^{t-1}) \right\| + \left\| \mathcal{M}_i(\boldsymbol{w}_i^{t-1}) \right\|_2^2 \end{aligned} \tag{31}$$

Hence we have

$$\begin{aligned} &\theta(2-\theta) \left\| \nabla \ell(\boldsymbol{w}_i^{t-1}) \right\|_2^2 + \left\| (1-\theta) \left\| \nabla \ell(\boldsymbol{w}_i^{t-1}) \right\| - \left\| \mathcal{M}_i(\boldsymbol{w}_i^{t-1}) \right\| \right\|_2^2 \\ &\leq \left\| \nabla \ell(\boldsymbol{w}_i^{t-1}) - \mathcal{M}_i(\boldsymbol{w}_i^{t-1}) \right\|_2^2 \\ &\leq (1-\theta^2) \left\| \nabla \ell(\boldsymbol{w}_i^{t-1}) \right\|_2^2 + \left\| \theta \left\| \nabla \ell(\boldsymbol{w}_i^{t-1}) \right\| - \left\| \mathcal{M}_i(\boldsymbol{w}_i^{t-1}) \right\| \right\|_2^2 \end{aligned} \tag{32}$$

Hence,

$$\theta(2-\theta) \left\| \nabla \ell(\boldsymbol{w}_i^{t-1}) \right\|_2^2 \leq \left\| \nabla \ell(\boldsymbol{w}_i^{t-1}) - \mathcal{M}_i(\boldsymbol{w}_i^{t-1}) \right\|_2^2 \tag{33}$$

Due to the Triangle Inequality, we have

$$\sqrt{\theta(2-\theta)} \left\| \nabla \ell(\boldsymbol{w}_i^{t-1}) \right\| \leq \left\| \nabla \ell(\boldsymbol{w}_i^{t-1}) - \mathcal{M}_i(\boldsymbol{w}_i^{t-1}) \right\| \leq \left\| \nabla \ell(\boldsymbol{w}_i^{t-1}) \right\| + \left\| \mathcal{M}_i(\boldsymbol{w}_i^{t-1}) \right\| \tag{34}$$

It follows that:

$$\left(\sqrt{\theta(2-\theta)} - 1 \right) \left\| \nabla \ell(\boldsymbol{w}_i^{t-1}) \right\| \leq \left\| \mathcal{M}_i(\boldsymbol{w}_i^{t-1}) \right\| \tag{35}$$

By incorporating Equation 32 and leveraging the AM-GM inequality, we can derive the following expression

$$\begin{aligned} \left\| \nabla \ell(\boldsymbol{w}_i^{t-1}) - \mathcal{M}_i(\boldsymbol{w}_i^{t-1}) \right\|_2^2 &\leq (1-\theta^2) \left\| \nabla \ell(\boldsymbol{w}_i^{t-1}) \right\|_2^2 + \left\| \theta \left\| \nabla \ell(\boldsymbol{w}_i^{t-1}) \right\| - \left\| \mathcal{M}_i(\boldsymbol{w}_i^{t-1}) \right\| \right\|_2^2 \\ &\leq \left(1 - \theta^2 + \left(1 + \theta + \sqrt{\theta(2-\theta)} \right)^2 \right) \left\| \nabla \ell(\boldsymbol{w}_i^{t-1}) \right\|_2^2 \\ &\leq \left(4 + 6\theta - \theta^2 \right) \left\| \nabla \ell(\boldsymbol{w}_i^{t-1}) \right\|_2^2 \end{aligned} \tag{36}$$

Therefore,

$$\mathbb{E}[\mathbf{B}] \leq r_{t-1}^2 \left\| \sum_{i\in mN} p_i^{t-1} \left(\sqrt{4 + 6\theta - \theta^2} \left\| \nabla \ell(\boldsymbol{w}_i^{t-1}) \right\| \right) \right\|_2^2 \leq \left(4 + 6\theta - \theta^2 \right) m^2 N^2 r_{t-1}^2 \mathcal{G}_{\boldsymbol{w}}^2 \tag{37}$$

Hence for $\mathbf{C}$, we have

$$\mathbb{E}[\mathbf{C}] \leq \frac{2mN\mathcal{G}_{\boldsymbol{w}} r_{t-1}^2 \sqrt{4+6\theta-\theta^2}}{2\varphi+t} \left(2\varphi\, \mathbb{E}\left\|\boldsymbol{w}^0-\boldsymbol{w}^*\right\|_2^2+\alpha^2\Delta\right) \tag{38}$$

Then based on Assumption 4.1 and Taylor expansion, we have the quadratic upper-bound of $\mathcal{L}(\cdot)$:

$$\mathcal{L}(\boldsymbol{w}_1)-\mathcal{L}(\boldsymbol{w}_2) \leq (\boldsymbol{w}_1-\boldsymbol{w}_2)^T \nabla\mathcal{L}(\boldsymbol{w}_2)+\frac{L}{2}\left\|\boldsymbol{w}_1-\boldsymbol{w}_2\right\|_2^2$$

It follows that

$$\begin{aligned}\mathbb{E}\,\mathcal{L}(\boldsymbol{w}^T)-\mathcal{L}(\boldsymbol{w}^*) &\leq \frac{L}{2}\,\mathbb{E}\left\|\boldsymbol{w}^T-\boldsymbol{w}^*\right\|_2^2 \\ &\leq \frac{L+2Lr_{T-1}\varpi}{2\varphi+T}\left(\varphi\,\mathbb{E}\left\|\boldsymbol{w}^0-\boldsymbol{w}^*\right\|_2^2+\frac{\alpha^2}{2}\Delta\right)+\frac{L\varpi^2}{2}\end{aligned}$$

Where $\varphi=\alpha\,(L+1)$, $\varpi=mN\mathcal{G}_{\boldsymbol{w}} r_{T-1}\sqrt{4+6\theta-\theta^2}$ □

A.2 Proof of Theorem 4.17

A.2.1 Lemmas

Lemma A.1. *f is monotone: : $\forall v_{-i}$ and $\forall v_i^{'} > v_i$, if $f(v_i, v_{-i}) \in W_i$, then $f(v_i^{'}, v_{-i}) \in W_i$.*

Lemma A.2. *In* FEDQV, *$\forall i, v_i, v_{-i}$ that $f(v_i, v_{-i}) \in W_i$, we have that $p_i(v_i, v_{-i}) = \Phi_i(v_{-i})$, where Φ_i is the critical value of a monotone function f on a single parameter domain that $\Phi_i(v_{-i}) = \sup_{v_i: f(v_i, v_{-i}) \notin W_i} v_i$.*

A.2.2 Proof of Lemmas

Proof of Lemmas 4.15

Proof. $\forall v_{-i}$ and $\forall v_i^{'} > v_i$, based on the voting scheme, if the party i who submit s_i join the aggregation with v_i, which means $f(v_i, v_{-i}) \in W_i$, then this party can also submit $\forall s_i^{'} < s_i$ that lead to $v_i^{'} > v_i$, and still join the aggregation. In other words, $f(v_i^{'}, v_{-i}) \in W_i$. Thus, f is monotone. □

Proof of Lemmas 4.16

Proof. The number of parties is $\mathcal{C}$ in each round. In voting scheme that follows Equation 3, the parties whose $s_i \leq \theta$ and $s_i \geq 1-\theta$ pay 0 credits voice. After Equation 4, the parties with 0 credit voice or 0 budget gain 0 vote. Assuming there are the top $k(k < \mathcal{C})$ parties in ranking whose payments are $c_{j\in k}$ ($c_{j\in k} > 0$). Notice in FEDQV, the payment function $p_i(v_i, v_{-i}) = c_i = v_i^2$.

$\forall j \in k$, if party j pays $c_j^{'} > p_j(v_j, v_{-j}) = \Phi_i(v_{-i}) = \sup_{v_i: f(v_i, v_{-i}) \notin W_i} v_i$, it will still remain in top k and join the aggregation. On the other hand, if party j pays $c_j^{'} < p_j(v_j, v_{-j}) = \Phi_i(v_{-i})$, then it will be replaced by the party $k+1$ in the ranking, and party j will not be able to join the aggregation regardless of whether party $k+1$ joins or not. As a result, in order to participate in the aggregation, the parties need to pay critical value, that is, $\forall i, v_i, v_{-i}$ that $f(v_i, v_{-i}) \in W_i$, we have that $p_i(v_i, v_{-i}) = \Phi_i(v_{-i})$ □

A.2.3 Proof of Theorem 4.17

Proof. According to Theorem 9.36 [41]: a normalised mechanism on a single parameter domain is incentive compatible(truthful) if and only if:
(i) The selection rule is monotone.
(ii) For every party i participants in the aggregation ($v_i > 0$) pays the critical value $\Phi_i(v_{-i}) = \sup_{v_i: f(v_i, v_{-i}) \notin W_i} v_i$.
The first condition (i) and the second one (ii) are proofed in Lemma 4.15 and Lemma 4.16 respectively. Thus, the proposed scheme FEDQV is incentive-compatible (truthful). □

B More Experimental Results

B.1 More Experimental Details

Platform Configurations. Our simulation experiments are implemented with Pytorch framework [64] on the cloud computing platform Google Colaboratory Pro (Colab Pro) with access to Nvidia K80s, T4s, P4s and P100s with 25 GB of Random Access Memory.

Table A1 shows the default setting in our experiments.

Table A1: Default experimental settings

Explanation	Notation	Default Setting
Budget	B	25
Similarity threshold	θ	0.1
The number of parties	N	100
The fraction of selected parties	$\mathcal{C}$	10
The number of total steps	T	500
The number of local epochs	E	5
Learning rate	r	0.01
Local batch size	b	10
Loss function	$\mathcal{L}(\cdot)$	Cross-entropy
Repeating times		3

B.2 Extra Experimental Results Under Small Percentages of Attackers.

To investigate FEDQV's behaviour in scenarios with finer gradations, we evaluated it with small, realistic percentages of attackers. Specifically, we examined the performance of Trimmed-Mean and Trimmed-Mean Integrated with FEDQV under two attacks, including Backdoor and QV-Adaptive, with 1%, 5%, and 10% attackers. The results

	Trimmed-Mean	Trimmed-Mean-QV
Backdoor	ACC(%)/ASR(%)	ACC(%)/ASR(%)
1%	84.99/0.57	85.67/0.52
5%	84.83/0.93	85.66/0.46
10%	85.45/2.27	85.06/1.79
Qv-adaptive	ACC(%)	ACC(%)
1%	84.13	86.38
5%	83.88	85.51
10%	79.49	85.74
30%	10.00	73.95
50%	10.00	10.00

Table A2: Comparison of Trimmed-Mean and Trimmed-Mean Integrated with FedQV Methods under Targeted Attacks (Backdoor and QV-Adaptive) Across Varying Percentages of Malicious Parties.

in Table A2 indicate that even under these small, realistic percentages of attackers, Trimmed-Mean integrated with FEDQV consistently outperforms Trimmed-Mean alone. The small percentage of attackers did not significantly impact the performance of the models, with almost unaffected accuracy (ACC) and a small attack success rate (ASR). This outcome underscores that integration with FEDQV enhances the robustness of the original defence mechanism across varying threat levels.

B.2.1 Under Different Non-IID Degree.

we have examined the performance of FEDQV and FEDAVG across varying degrees of non-IID data, ranging from 0.1 to 0.9, as depicted in Table A3. These results demonstrate that as the non-IID degree increases among the parties, the performance of the global model declines. Notably,FEDQV consistently maintains a superior performance compared to FEDAVG, even when confronted with different degrees of data heterogeneity under attack conditions.

Non-IID		0.1	0.3	0.5	0.7	0.9
FedQV	ACC(%)	84.94	86.01	83.88	81.37	75.96
	ASR(%)	3.39	4.55	17.64	20.59	24.18
FedAvg	ACC(%)	81.27	81.1	82.44	80.77	65.68
	ASR(%)	3.37	13.39	20.84	22.99	60.35

Table A3: Comparison of Accuracy (ACC) and Attack Success Rate (ASR) for FedQV and FedAvg under Backdoor Attack over 100 epochs with varying Non-IID Degrees on Fashion-MNIST Dataset.

C More Details of Defence against privacy attacks

C.1 Privacy Attack Algorithms

We mount the Deep Leakage from Gradients (DLG) attack [8] and the Gradient Inversion (GI) attack [9] on FEDQV. These attacks aim to generate dummy data and corresponding labels by leveraging a gradient-matching objective. The detailed algorithms are outlined in Algorithm 3.

Algorithm 3: DLG and GI Attacks

Input: $F(\mathcal{D}_i; \boldsymbol{w}_i^t)$: model at round t from targeted user i; learning rate η for inverting gradient optimiser; S: max iterations for attack; τ: regularisation term for cosine loss in inverting gradient attack; d:the model size

Output: reconstructed training data $(\mathcal{D}_i, y_i)$ at round t

Initialise $\mathcal{D}'_0 \leftarrow \mathcal{N}(0,1)$, $y'_0 \leftarrow Randint(0, max(y))$

for *s*$\leftarrow$ *0,1, ...* $S-1$ **do**

  $\nabla \boldsymbol{w}'_s \leftarrow \partial \ell(F(\mathcal{D}'_s, \boldsymbol{w}_i^t), y'_s)/\partial \boldsymbol{w}_i^t$

  switch *Case* **do**

    case *DLG attack* **do** $\mathcal{L}'_s \leftarrow \| \nabla \boldsymbol{w}'_s - \nabla \boldsymbol{w}_i^t \|^2$

    case *GI attack* **do** $\mathcal{L}'_s \leftarrow 1 - \frac{\nabla \boldsymbol{w}_i^t \cdot \nabla \boldsymbol{w}'_s}{\|\nabla \boldsymbol{w}_i^t\| \|\nabla \boldsymbol{w}'_s\|} + \tau$

  $\mathcal{D}'_{s+1} \leftarrow \mathcal{D}'_s - \eta \nabla_{\mathcal{D}'_s} \mathcal{L}'_s$, $y'_{s+1} \leftarrow y'_s - \eta \nabla_{y'_s} \mathcal{L}'_s$

return $\mathcal{D}'_S, y'_S$

C.2 Implementing SECAGG in FEDQV

Our approach is in line with other securely aggregated FL designs consisting of three main stages: 1) *Setup*; 2) *Generation and Protect*; and 3) *Aggregate*. Our integration of secure aggregation to FEDQV is depicted in Figure A1. We further discuss its details below:

Stage 0 (*Setup*). In this stage, the server and the partis compute SECAGG.Setup and initialise FEDQV by having the server send values of the protocol's parameters to each party, i.e. $\boldsymbol{w}^{t-1}$.

Stage 1 (*Generation and Protect*). After the setup and initialisation stage, each party first computes its local model $\boldsymbol{w}_i^t$ as in FEDQV and computes its s_i^t. Later, it transmits the message $\langle |\mathcal{D}_i|, s_i^t \rangle$. Upon receiving $\langle |\mathcal{D}_i|, s_i^t \rangle$, the server computes v_i^t as in Equation 4 and transmits v_i^t to the corresponding party. Then party i first computes $\boldsymbol{w}_i^t \cdot v_i^t$ and protects its input $\boldsymbol{w}_i^t \cdot v_i^t$ by treating it an input to the SECAGG.Protect as $[\boldsymbol{w}_i^t] = \boldsymbol{w}_i^t \cdot v_i^t + \mathsf{k}_i + \mathsf{b}_i$. Party i sends $[\boldsymbol{w}_i^t]$ to the server.

Stage 3 (*Aggregate*). Upon receiving $[\boldsymbol{w}_i^t]$, the server first calls SECAGG.Aggregate as described previously which returns the sum of all $\boldsymbol{w}_i^t$. The server requires one more step to finalise by multiplying the sum retrieved by $\frac{1}{\sum_{i=1}^N v_i^t}$, as it is the sole entity possessing complete knowledge of all v_i^t.

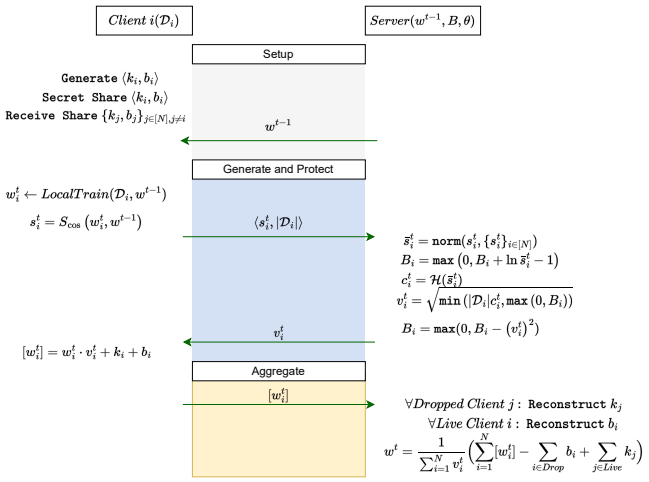


Figure A1: Overview of implementation of the SECAGG protocol in FEDQV.

Anexo 6: FreqyWM - Frequency Watermarking for the New Data Economy

FreqyWM: Frequency Watermarking for the New Data Economy

Devriş İşler[1,2], Elisa Cabana[3§], Alvaro Garcia-Recuero[4§], Georgia Koutrika[5], and Nikolaos Laoutaris[1]
[1]*IMDEA Networks Institute,* [2]*UC3M,* [3]*CUNEF University,* [4]*FUNDITEC,* [5]*Athena Research Center*

***Abstract*—We present a novel technique for modulating the appearance frequency of a few tokens within a dataset for encoding an invisible watermark that can be used to protect ownership rights upon data. We develop optimal as well as fast heuristic algorithms for creating and verifying such watermarks. We also demonstrate the robustness of our technique against various attacks and derive analytical bounds for the false positive probability of erroneously "detecting" a watermark on a dataset that does not carry it. Our technique is applicable to both single dimensional and multidimensional datasets, is independent of token type, allows for a fine control of the introduced distortion, and can be used in a variety of use cases that involve buying and selling data in contemporary data marketplaces.**

***Index Terms*—Intellectual property, digital rights management, watermarking, ownership rights, data economy**

I. Introduction

Data-driven decision making powered by Machine Learning (ML) algorithms is changing how society and the economy work. ML is driving up the demand for data in what has been called the fourth industrial revolution. To satisfy this demand, several data marketplaces (DMs),which are mediation platforms aiming to connect the two primary stakeholders of the data value chain, namely the data providers/sellers and the data buyers [1], have appeared in the last few years.

The problems: Unfortunately, as with all digital assets, being able to copy/store/transmit datasets with close to zero cost makes creating illegal copies very easy. Even worse, unlike media content and software, the issue of ownership is less obvious when it comes to datasets. Any movie, song, e-book, or software can usually be attributed to a director, musician, author, or company, respectively, but this is hard to do for large datasets. These large datasets in data economy are traded in a wholesale manner that involves large numbers of tuples/rows. Consider an anonymised mobility dataset logging the movement of people in a city. Such a dataset may have been produced by collecting GPS readings from the smartphones of individuals using a map application, or it may be deduced by analysing cell phone traces [2] or Call Description Records (CDRs) maintained by mobile operators. Deployment of advanced privacy enhancing technologies (PETs) such as multiparty computation [3], (fully) homomorphic encryption [4], functional encryption [5], and trusted execution environments [6] can protect data from leaking in the first place and allow (pre-agreed) computations on data without hampering the functioning of the data-driven economy, e.g., private set computation [7], encrypted databases [8], secure computation [9], secure data aggregation [10], and verifiable databases [11]. However, most such approaches face serious scalability challenges that hamper their deployment in real-world use-cases. An alternative to deploying PETs solutions, is to rely on purely legal tools and terms and conditions to protect data ownership in the context of the new data economy [12]. In fact, most DMs do exactly that – trade plaintext versions of entire data [1, 13, 14] assuming that the different parties will abide to pre-agreed terms and conditions. With weak to nonexistent ownership guarantees by technical means, it is difficult to imagine that the data economy will ever flourish and reach its projected potential [15]. Indeed, any sold copy of a dataset can be 'pirated' by a buyer-turned-seller that can then resell the same dataset in a DM thereby undercutting the rightful owner and rendering its investment useless.

Watermarking is a well-known technique for protecting ownership upon copying and unauthorized distribution, initially proposed for protecting digital media [16, 17] and software [18]. Watermarking techniques for datasets [19, 20, 21] and machine learning models [22] have been proposed recently. Watermarking generally consists of two algorithms: *generation* (or embedding) and *detection*. The generation allows an owner to embed an invisible (or visible) watermark into their data using a high entropy (watermarking) secret and produces a watermarked version of the data introducing tolerable distortion without degrading the data utility. During the detection algorithm, the owner proves its ownership on the suspected data (even if it is modified) using the same watermarking secret generated during the watermark generation. If the result of the detection is 1 (or *accept*), the owner can use it to prove their ownership on the (suspected) watermarked data. A watermarking scheme is assumed to be secure against the guess attack (where an attacker tries to expose the watermarking secret) and robust against (un)intentional alterations/-modifications (i.e., a watermark should be still detectable even under attacks such as [20, 23, 24, 25, 26, 27, 28]).

Limitations of existing watermarking techniques: Watermarking techniques, depending on the nature of their application, may have very different objectives, e.g., numerical database watermarking controlling the distortion on mean and standard deviation [21], reversible watermarking allowing owners to reconstruct the original data [29], watermarking text datasets preserving the meaning of a text [30] and/or the frequencies of the words [31], categorical watermarking preserving the (predefined) categories (e.g., gender) of a

§Work done while the author was affiliated with IMDEA Networks Institute.

dataset [32]. All these solutions focus on a specific data type in a specific domain [23, 33]. Another limitation of theirs relates to the level of control they offer to the user in terms of controlling the distortion introduced upon the original data due to the watermark. There are, for example, techniques that maintain the mean and the standard deviation of a numerical field [20, 34, 35] but, as we will show later, this can lead to arbitrary large distortion between the original and the watermarked data when considering the entire distribution of values that goes beyond the mean and the standard deviation. To address these limitations, *we introduce a novel watermarking technique that can be implemented over a wide range of data types and structures* (with some constraints that will be explained later) while *giving the data owner very precise control over the introduced distortion*.

A novel watermarking technique for data: In this paper, we present a novel *Frequency Watermarking* technique, henceforth *FreqyWM*, [1] for hiding a secret within a dataset in a manner that makes the said secret indistinguishable from the data that it protects. The main idea behind *FreqyWM* is *to modify slightly the appearance frequency of existing tokens* within a dataset in order *to create a secret in the form of a complex relationship* between the frequencies of different tokens. By making this relationship complex enough, we can reduce the probability that it appeared by chance close to zero. Therefore, by revealing knowledge of such secret relationship, a party can claim ownership over a dataset because the only practical way of knowing such a secret is to have inserted it in the data in the first place. A token may be a word, a database record, a URL, or any repeating value within a structured or semi-structured commercial dataset. Our secret is created by first selecting a number of token pairs. Then, for each pair, we slightly modulate the frequency counts of its tokens in order to make their difference yield zero under modulo N arithmetic. This can be easily done by adding or removing some instances of one, the other, or both tokens. By increasing the number of selected pairs we can make our watermark more resistant to attacks, as well as less likely to have appeared by chance.

FreqyWM can achieve several things. First and foremost, by revealing knowledge of the secret encoded by the watermark, a data seller can prove rightful ownership of a dataset to a third party, such as a DM. This can be used to distinguish a rightful owner from a pirate that may attempt to monetize a pirated dataset in a DM. If the DM, or the rightful owner detects such an event, the dataset can be removed and the pirate be banned. This would mimic what web-sites like YouTube do to protect copyrighted content. Detecting the presence of pirated copies can be achieved using content similarity [36], locality sensitive hashing [37, 38] and even hashing similarity [39] that go beyond the scope of watermarking.

In addition to proving ownership, our watermarking technique can also reveal who may have leaked (copied/pirated) a dataset in the first place. A dataset seller or a DM may create a different watermark for every buyer and in addition to encoding it into the data, store also a description of it in some immutable index (e.g., a blockchain). Then, if an unauthorized copy of the dataset is found at a latter point, the culprit can be identified by looking up its watermark against this index.

[1]Freqy pronounced as *freaky*.

Our major contributions are as follows:

• Our first contribution is the idea of using the appearance frequency of tokens to encode invisible watermarks upon datasets traded in DMs. We establish a family of such watermarks using frequency pairs and modulo arithmetic and prove that creating an optimal *FreqyWM* reduces to solving a Maximum Weighted Matching (MWM) problem [40, 41] combined with a polynomial special version of the 0/1 Knapsack problem [42] involving items of equal value but different weights.

• We extend frequency-based watermarking to make it resilient against a series of attacks. In particular, we protect our technique against a *Guess Attack* attempting to identify our watermarked pairs and secrets to impersonate the rightful owner. We make such an attack computationally hard by introducing a high-entropy secret while generating the watermark. We also protect against a *Re-watermarking Attack* mounted by having a pirate inject its own watermark upon an already watermarked dataset, and then present the former as a false proof of ownership. We thwart such an attack by describing a simple protocol capable of ordering chronologically multiple watermarks that may be carried by different versions of the same dataset. We protect against a *Destroy Attack* attempting to destroy our watermark by changing the frequency of different tokens in the dataset. By relaxing our modulo arithmetic rule used during the verification of a particular watermark pair, as well as the percentage of pairs to be detected before the entire watermark is verified (accepted), we oblige the attacker to effectively also destroy the actual data in the process of destroying the watermark. Finally, we show that our technique is robust to a *Sampling Attack* in which the attacker attempts to pirate only a random sample of the watermarked data.

• Our final contribution is an extensive performance evaluation study aiming to explain the impact of the main parameters of *FreqyWM* on major performance metrics under different attack scenarios using synthetic and real world datasets.

The main **findings** of our evaluation are as follows:

• We show that as long as there exists sufficient variation in the frequencies of different tokens, *FreqyWM* can encode robust watermarks with minimal distortion on the initial data. Our technique does not apply to uniform token appearance frequencies, because in this case there does not exist sufficient gap between different frequencies for encoding a watermark.

• Regarding the false positive probability, i.e., "detecting" a watermark on a dataset that does not carry it, our analytical bounds (in the form of closed form expressions) show that it quickly goes to zero as we increase the number of pairs.

• We demonstrate that a Guess Attack has negligible probability of success, thereby making it impossible for almost all practical cases. On the up side, the rightful owner or any party, that is given the watermarking secret for verifying the watermark, can do that very fast in linear time complexity.

• Regarding Sampling Attacks, we show that with the exception of very small samples, our detection algorithm is capable of detecting our watermark. Achieving this requires using the relaxed detection algorithm that trades robustness to attacks with false positives. For example, on a sample of 20% and with thresholds that impose tiny false positive probability, the detection probability exceeds 90%.
• In terms of Destroy Attacks, we show that a watermark that imposes (costs) a tiny 0.0002% distortion on the original data, remains detectable even under attacks that add random noise that imposes a 90% modification.
• Compared to existing solutions from the literature [30, 35] that are applicable only to numerical data and preserve only the mean value of the watermarked data, *FreqyWM* allows a data owner to control the exact amount of distortion introduced by the watermark in terms of cosine or other similarity metrics which, under [30, 35] may become unbounded. For example, a *FreqyWM* watermark that imposes only 0.0002% distortion in terms of cosine similarity, is stronger than watermarks from [35] and [30] that impose 46.72% and 4% distortion, respectively under the same metric.

II. Related Work

Database watermarking is the closest type of watermarking to our work. There are of course other types of watermarking and fingerprinting (when an owner generates a unique watermark for each intended party, e.g., buyers/data marketplaces), for example, for sequential [43] and genomic datasets [44]. However, as they focus on specialized types of data, we do not go into more details about them. Survey papers such as [23, 33, 45, 46] compare database watermarking techniques in terms of verifiability, distortion, supported data types, and other aspects. Many of these solutions are applicable only to numerical data and thus cannot be applied to a range of commercial datasets, e.g., to web-browsing click-streams.

The first known watermarking technique for relational data is a *numerical database watermarking* approach [20]. The watermark information is normally embedded in the Least Significant Bit (LSB) of features of relational databases to minimize distortion. Other numerical database watermarking solutions introduce distortion by considering the statistics of numeric values [34, 35, 47, 48, 49]. The proposed solutions in [20, 35] focus on keeping the change at minimum (i.e., median and standard deviation). *However*, numerical database watermarking unfortunately cannot be applied to datasets composed of string and numerical values (e.g., CDRs, web-browsing history) that we handle in our work.

Distortion-free database watermarking schemes have also been proposed [50, 51] that introduce fake tuples or columns in the original database. The fake tuples or columns are created based on a watermark secret by computing a secret function which makes watermarking visible and easy to remove. *However, an attacker can remove the watermark with minimum computational power*, making these approaches inapplicable to our case. *Reversible watermarking* allows owners to reconstruct the original data used for watermarking on the top of watermark verification [29, 30, 49, 52, 53, 54, 55, 56, 57]. They have similar properties as other relational watermarking techniques (e.g., private key based, robust, introducing distortion).

Categorical watermarking [32] is another watermarking approach that replaces tokens in a dataset with another token in the same category. However, this causes an undesired distortion and requires predefined categories (e.g., gender, clothing size) in the data. Consequently, its applicability on datasets consisting of different data types is limited. *Text watermarking* [30, 31] is for text files where it changes a token (e.g., by replacing a word with another similar word) trying to preserve the meaning of a text [30] and/or the frequencies of the words [31]. However, assume the dataset is a list of URLs visited by the owner, then this (insecure) change/replacement may invalidate a token (e.g., causing an invalid URL).

In the context of datasets in our use case, while prior works try to minimize the amount of distortion on median, average, or first moments of the distribution of a feature, the owner can limit the exact distortion between the original and the watermarked dataset as reflected by distance metrics that capture the shape of the entire distribution of a feature. Our results in Section IV-D have shown that the latter can deviate arbitrarily if an owner tries to control only the first most important moments.

III. Frequency-based Watermarking

In this section we provide an overview of *FreqyWM* and the notations used throughout the paper in Table I.

D_o	The original data to watermark.
D_w	Watermarked (data) version of D_o.
tk_i	ith token.
f_i^o	Frequency of ith token in D_o.
f_i^w	Frequency of ith token in D_w.
R	A high entropy secret.
L_{wm}	A list of chosen token pairs for watermarking.
L_{sc}	A list of secrets required for watermark detection.
L_e	A list of eligible token pairs for watermarking.
k	Threshold for detecting a watermark.
t	Threshold to accept a pair as watermarked.
b	A budget threshold for distortion that watermarking can introduce.

TABLE I: Notation.

Running Example. To provide the intuition behind our watermarking approach, assume a scenario where an owner holds a real click-stream dataset consisting of visited URLs (e.g., the dataset by [58]). Such datasets are desired by modern data analytic-based applications [59] where their frequency histograms (e.g., the number of clicks/visits, popularity of likes in social networks) are used as an essential source of information. For instance, assuming the appearance frequencies (histogram) visualized in Figure 1 via a tabular form, the most frequent token is `youtube.com`, the second one is `facebook.com`, and so forth. After watermarking, it is important that the *ranking* of the tokens based on the frequency shall not change while the frequency appearances can be modified. For instance `youtube.com` shall be the most frequent URL (token) visited in the watermarked dataset. Another important distortion metric on the histogram is *similarity*. It

is important that owners shall have control over the change in similarity. Since the similarity metric can be varied depending on the application that a dataset will be used, owners can assign a *budget* to determine the minimum similarity desired on the frequency distribution after watermarking. Based on the above, we derive two natural constraints on the data utility to allow an owner to control distortion, without limiting watermarking to a specific data type:

• *Ranking Constraint:* Watermarking should preserve the ranking of token frequency distribution (histogram). Preservation of ranking does not of course imply that frequencies of individual tokens need to remain intact. • *Similarity Constraint:* The similarity between original and watermarked frequency distributions (histograms) should not be any less than $(100 - b)\%$ where b is a budget. Input b is determined by the owner to keep distortion due to watermark generation within b. [2]

To satisfy these constraints and overcome the shortcoming of existing watermarking techniques, we introduce a new *private-key based* watermarking scheme, *FreqyWM*, that is *blind* (does not require the original data), *primary-key free* (does not need attributes that uniquely specify a tuple in a relation in a dataset), *robust*, and *secure* against *guess*, *sampling*, *destroy*, and *false-claim* attacks with a high utility and a good trade-off between the complexity of the transformation and algorithmic efficiency of the solution.

A. Overview of our Approach

FreqyWM consists of two main algorithms: the watermark generation algorithm, $WMGenerate$, and the watermark detection algorithm, $WMDetect$. $WMGenerate$ generates watermarked data based on a *budget* b capturing how much the watermarked data may differ from the original one, e.g., in terms of cosine (or other) similarity metrics of their corresponding token frequency distributions. By calling $WMGenerate$, the owner creates a watermarked version of their data consisting of tokens such that ownership can be proved. $WMDetect$ detects if a suspected dataset holds the watermark of the owner using the owner secrets produced by $WMGenerate$ and two thresholds (k and t). If $WMDetect$ outputs *accept/verified*, this evidence would prove that the owner can claim ownership of the watermark and thus the data. By nature, $WMDetect$ can be computed as many times as desired in private while it can be computed *only once* in public, because it would mean that the potential data owner shall reveal the secret leading to such watermark to the public (or whomever must verify it, e.g., a judging third party). As part of our future work, we are also looking at public verifiability without revealing the private key (Section VII).

We describe the general idea behind *FreqyWM*, illustrated in Figure 1. We use our running example. Of course, our technique is general and can be applied to any repeating token beyond just URLs, as we explain in Section IV-C.

Watermark Generation. Assume that the data owner holding a list of URLs visited creates a dataset D_o using the *domain* of each URL in the list as a token and sets a budget b for the similarity constraint. $WMGenerate$ has the following steps:

• *Histogram Generation.* Since *FreqyWM* aims to preserve the appearance frequency of tokens, it first creates a histogram of the original dataset D_o such that it sorts all unique tokens in descending order of their frequency (e.g., YouTube is the most visited, Facebook is the second, and so on).

• *Generation of Eligible Tokens. FreqyWM* cannot modify the frequencies randomly because of the ranking constraint. Therefore, it identifies a list L_e of eligible pairs of tokens that are candidates to be watermarked using some secret R.

• *Optimal Selection*. With the identification of eligible pairs, *FreqyWM* ensures that the ranking is preserved after watermarking. However, the similarity constraint is yet to be satisfied. To keep the similarity at least at $(100 - b)\%$, *FreqyWM* selects pairs of tokens from eligible pairs for watermarking, denoted by L_{wm}, based on the budget constraint b. For this purpose, *FreqyWM* benefits from solving two well-known problems: *Maximum Weight Matching* (MWM) and *Equally Valued 0/1 Knapsack problem* (QKP). To do so, eligible pairs are converted to a graph representation where vertices represent a token, and an edge represents a pair. *FreqyWM* applies Maximum Weight Matching to the graph representation (discussed in detail later). By applying MWM, *FreqyWM* selects the pairs from eligible pairs requiring the minimum change in total; however, it does not necessarily mean that the similarity between the original histogram and watermarked histogram will be at least $(100-b)\%$. To choose another set of pairs satisfying the Ranking Constraint from the pairs derived after MWM, an *Equally Valued 0/1 Knapsack problem* needs to be solved. The more the token pairs are selected to watermark, the more robust *FreqyWM* is, since the number of tokens to attack (e.g., remove/identify) increases. To fulfill the budget b, QKP selects a maximum number of pairs such that the similarity between the original frequency histogram and the watermarked one is *at least* $(100 - b)\%$.

• *Frequency Modification*. Until now, *FreqyWM* determines the final pairs of tokens for watermarking but frequency appearances are yet to be modified to create the watermarked histogram. Therefore, *FreqyWM* modifies the frequencies of the selected tokens where the frequencies of a pair of tokens would be equal to 0 (as a watermark embedding rule) in some modulo that is calculated based on secrets and tokens in the pair. To make it more comprehensible and show how the modifications occur, let us assume that the frequencies of a chosen pair, e.g., `youtube.com` and `instagram.com`, are 1098 and 537, respectively. Assume also that a modulo value, say 129, is computed based on the secrets and the tokens (e.g., `youtube.com` and `instagram.com`). The difference between the two frequencies in modulo 129 is 45. To set the difference to 0, we need to change the appearance frequencies for Youtube and Instagram in the dataset. 45 is divided (by 2) as 23 (by ceiling) and 22 (by flooring). The new frequencies of `youtube.com` and `instagram.com` need to become $1098 - 23 = 1075$ and $537 + 22 = 559$

[2]Although in our experiments we use cosine similarity, any similarity metrics can be deployed without any loss of security and change in *FreqyWM*.

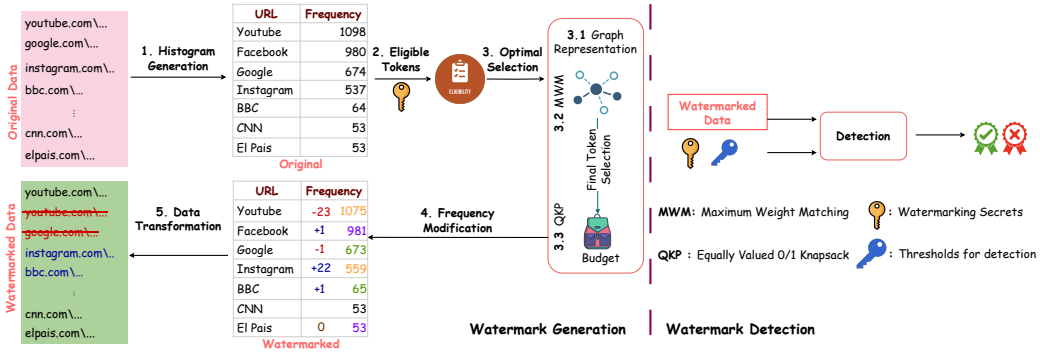


Fig. 1: *FreqyWM* illustrated based on a (Top Level Domain, TLD) URL dataset. URLs chosen as a pair for watermarking are represented with the same colored frequencies (e.g., `Youtube` and `Instagram`) while the ones not selected are colored black (e.g., `CNN`).

such that $(1075 - 559) \mod 129 \equiv 0$. We can do that by removing 23 instances of Youtube from the dataset, and adding 22 more instances of Instagram. However, when the remainder (i.e., $(1098 - 537) \mod 21 \equiv 16$) is greater than half of the modulo, we add the modulo result calculated as $(\lceil (1098 - 537) \div 21 \rceil) \times (1098 - 537)$ to the difference. This way, we never have to eliminate remainders that exceed half of the modulo. As it will be evident in the next section, this observation enables us to determine eligible tokens.

• *Data Transformation. FreqyWM* adds/removes tokens based on the frequencies and produces a watermarked dataset D_w.

Watermark Detection. An owner wishes to verify if a (watermarked) dataset D'_w (a modified version of D_w) is watermarked by using the secrets stored from the watermark generation. To determine the confidence level in the detection (e.g., the minimum number of detected watermarked tokens), the owner provides some threshold values (k and t). With the watermarking secret and the thresholds, the detection returns *accept/verified* or *reject*.

Algorithm I: Watermark Generation

Input: D_o, b
Output: D_w, L_{sc}
$D_o^{hist} = \textbf{Preprocess}(D_o)$
$R \leftarrow \{0,1\}^\lambda$, $z \leftarrow Z^+$
foreach $\{tk_i, tk_j\}_{i \neq j} \in D_o$ **do**
  $s_{ij} = H(tk_i || H(R || tk_j)) \mod z$
end
$L_e \leftarrow \textbf{Eligible}(D_o^{hist}, \{s_{ij}\})$
$L_{wm} \leftarrow \textbf{OptMatch}(D_o^{hist}, L_e, \{s_{ij}\}, b)$
foreach $\{tk_i, tk_j\} \in L_{wm}$ **do**
  $D_o^{hist}.\textbf{Update}(f_i^w, f_j^w, s_{ij})$
end
$D_w \leftarrow \textbf{Create}(D_o^{hist}, D_o)$
$L_{sc} = \{L_{wm}, R, z\}$
Result: $D_w, L_{sc} = \{L_{wm}, R, z\}$

B. Detailed Description of FreqyWM

1) Watermark Generation: The data owner holds the original data D_o and defines a budget b that decides how much distortion a watermark can introduce. For comprehensibility, assume that D_o is a single-dimensional dataset, e.g., a dataset with one attribute (see Section IV-C for how to apply *FreqyWM* to multi-dimensional datasets). D_o consists of repeating values called *tokens* that can be of *any* data type, which enables *FreqyWM* to be data-type agnostic. The goal of watermarking is to generate the optimal watermark, i.e., *with the largest number of watermarked pairs* within the given budget b.

The generation algorithm (Algorithm I) follows these steps:

Histogram Generation: It pre-processes D_o to generate a histogram $D_o^{hist} = \textbf{Preprocess}(D_o)$. D_o^{hist} consists of a set of tokens, $\{tk_0, \ldots, tk_{|D_o^{hist}|}\}$ (e.g., tk_0 =`youtube.com`) where each tk_i has an (original) appearance frequency f_i^o (e.g., there are 1098 YouTube visits). The histogram D_o^{hist} is sorted in a descending order of frequency. To keep the distortion introduced by the watermark at minimum (e.g., after watermarking, YouTube is still the most visited, followed by Facebook, although their frequencies may have changed), we calculate two boundaries for each token tk_i: an upper boundary u_i and a lower boundary l_i. The boundaries allow us to determine how much change we can introduce to the token and whether a token pair is eligible as explained later. Naturally, for the token with the highest frequency in the histogram, it is $u_0 = \infty$ because we can increase the frequency of tk_0 as much as we want, while the lower boundary of the last token, $tk_{|D_o^{hist}|-1}$, is set to its frequency as $l_{|D_o^{hist}|-1} = f_{|D_o^{hist}|-1}$ because we can remove at so many appearances. For the rest of the boundary calculations of each token tk_i, u_i is defined as the difference between $f_{(i-1)}^o$ and f_i^o, while l_i is assigned as $f_i^o - f_{(i+1)}^o$. Note that once the boundaries are set, they remain same until frequency modification.[3]

Generation of Eligible Tokens: In cryptography, $\lambda \in \mathbb{N}$ is a security parameter, i.e., a variable measuring the probability with which an adversary can break a cryptographic scheme [60]. In other words, λ provides a way of measuring how difficult it is for an adversary to break a cryptographic scheme. *FreqyWM* requires randomization to be secure by ensuring that an attacker has only negligible advantage to recover the watermark and create collision for false claim (e.g., coming up with another watermarking secret which returns accept on

[3]The frequencies of some tokens may have high importance. An owner can filter the dataset and exclude them from watermarking.

data not watermarked by it). Thus, we choose a hash function to overcome the collision. In detail, a hash function H (chosen from a family of such functions) is a deterministic function from an arbitrary size input to a fixed size output, denoted $H : \{0,1\}^* \rightarrow \{0,1\}^\lambda$. The hash function [60] is *collision resistant* if it is hard to find two different inputs $m_0 \neq m_1$ that hash to the same output $H(m_0) = H(m_1)$.

Based on the above, to determine token pairs for watermarking, *FreqyWM* first generates a high entropy random number, i.e., secret, $R \leftarrow \{0,1\}^\lambda$ and an integer $z \in \mathbb{Z}^+$. Then, it uses R and z to compute s_{ij} values for modulo operation as: $s_{ij} = H(tk_i||H(R||tk_j)) \mod z$, where $||$ denotes concatenation. A set L_e of all eligible pairs is generated by an algorithm $Eligible$ based on given pairs $\{tk_i, tk_j\}$ and corresponding s_{ij} values as $L_e \leftarrow Eligible(D_o^{hist}, \{s_{ij}\})$. A pair is accepted as eligible if it satisfies that the boundaries of each token in the pair are at least $\lceil s_{ij}/2 \rceil$ where $s_{ij} \geq 2$. s_{ij} cannot be 0 or 1 because of modulo operation since modulo 0 is undefined and modulo 1 is 0. Note that the size of L_e is bounded by $[0, \binom{|D_o^{hist}|}{2}]$ where 0 means that there is *no eligible pair* while $\binom{|D_o^{hist}|}{2}$ means that *all the possible pairs* of tokens are eligible. After the eligible pairs are constituted, the boundary check is not necessary anymore since whichever set of pairs (that does not have a common token among) is chosen, the ranking will be preserved.

Optimal Selection: The eligible pairs are defined by ensuring the ranking constraint. However, to determine which subset of eligible pairs shall be selected such that chosen optimal number of pairs of tokens, denoted by a set L_{wm}, respect the budget constraint, it runs optimal matching algorithm from the eligible pairs L_e using the frequencies and s_{ij} values as $L_{wm} \leftarrow OptMatch(D_o^{hist}, L_e, \{s_{ij}\}, b)$. In Section III-B2, we show that for our optimal selection solution, we acutely reduce our problem to *Maximum Weight Matching (MWM)* and *Equally Valued 0/1 Knapsack problem (QKP)* problems to solve. We also devise two heuristics: *greedy* and *random*.

Frequency Modification: Based on L_{wm}, the algorithm creates new frequencies of tokens chosen from the optimal matching algorithm $(f_i^w - f_j^w) \mod s_{ij} \equiv 0$. This, of course, changes the boundaries of tokens; however, we do not need to update the boundaries as they are not needed anymore.

Data Transformation: It generates or removes tokens based on new frequencies. Note that the position of where to add tokens is important for security of *FreqyWM* against guess attack. Therefore, new tokens should be added in random positions (see Section IV-C for more discussion). As a final step, it returns the list of tokens D_w and stores L_{wm}, z value, and the random value R as a list L_{sc}.

2) Optimal and Heuristic Approaches : Given that all watermarked pairs have equal value in terms of proving ownership of the data, an optimal watermark is just a watermark of maximum size in terms of watermarked pairs, within the defined constraints (*similarity* and *ranking*). *Optimal Matching*. Let us now define our optimal watermarking. Let $G = \{V, E\}$ be a connected undirected graph which is the representation of frequencies driven from eligible pairs L_e. $V = \{v_1, v_2, \ldots, v_{|V|}\}$ where v_i represents tk_i and $E = \{e_1, e_2, \ldots, e_{|E|}\}$ where $e(v_i, v_j)$ is the edge between v_i and v_j. The weight of an edge $e(v_i, v_j)$, $w(e_i)$, is equal to $T - ((f_i^o - f_j^o) \mod s_{ij})$ where T is a big value (e.g., $T > C$ where C is the highest difference between two frequencies in the eligible pairs). Then, our optimal watermarking problem reduces to finding the maximum number of edges (pairs) such that *no edge* has a common vertex and b is not exceeded.

Definition 1 (Optimal Watermarking)**.** *Let $OptWM(G(V,E), b)$ be the optimal watermarking with a budget of b among an eligible set of items L_e represented as a connected undirected graph $G(V,E)$. The optimal watermarking produces the maximum number of edges (pairs) while not exceeding the budget b defined below:*

$$MAX\ |M^w|,\ \ M^w = \{e_1, ..., e_{|M|}\}\ s.t.\ sim(D_o^{hist}, D_w^{hist}) \geq (100 - b)$$

where M^w denotes the chosen pairs for watermarking.

This problem is reduced to two well-known problems with polynomial time solutions: *Maximum Weight Matching* (MWM) and *Equally Valued 0/1 Knapsack problem* (QKP) (which we have a special case where all values are equal). While the general $0/1$ Knapsack problem is known to be NP-Hard [42], this special equally valued $0/1$ Knapsack problem would have a polynomial time (greedy) solution. Hence, our optimal pairing problem is reduced and solved as follows:

• Find the maximum weight matching $M = e_1, e_2, \ldots, e_{|M|}$ as $M = MWM(G(V,E))$. Notice that M includes the edges such that the sum gives the maximum weight. It refers to minimum weight for us since the weights are defined as $T - (f_i - f_j \mod s_{ij})$ which makes the highest frequency difference have the smallest weight and the smallest one have the highest weight. With this conversion, we identify the edges distorting the histogram minimally.

• After finding the edges via MWM, we have one more constraint which is the budget b. The matching algorithm has to return the maximum number of matchings for which the distortion (e.g., based on cosine similarity) does not exceed b which can be solved via QKP where the value of each item is 1, and the weight is recomputed as $T - w(e_i)$. Recomputation is necessary because for the QKP we want to add as many items as possible that will be bounded by b. Therefore, it finds the set of edges L_{wm} in M such that the selected edges do not exceed the budget b by employing the QKP as $L_{wm} = QKP(M, b)$ where $L_{wm} = e_1, e_2, \ldots, e_{|L_{wm}|}$ and value of each e_i is 1 ($val(e_i) = 1$). Showing the optimality of the resulting watermark according to Definition 1 can be proven via proof-by-contradiction. In a nutshell, if our solution is not optimal, it means that one of the solutions produced by MWM and QKP cannot be optimal. However, since MWM and QKP are both assumed optimal, this contradicts our statement and thus our solution is optimal.

Heuristic Matching Algorithms. We define two heuristic algorithms: 1) *greedy*; and 2) *random*. In the *greedy* algorithm, all the eligible token pairs are sorted in an ascending order by their remainders as $rm_{ij} \equiv (f_i^o - f_j^o) \mod s_{ij}$. The algorithm

starts selecting a pair respectively for watermarking where b would not be exceeded when it is chosen (i.e., comparing the similarities of original and watermark histograms). This continues until b is exhausted or there is no more item to visit. The *random* matching algorithm follows the same steps as the greedy algorithm except it does not sort the eligible pairs but rather selects a pair randomly from L_e.

3) Watermark Detection: In detection, the data owner wishes to know if there is a watermark of its in a token dataset D'_w to claim ownership. The owner holds its secret list $L_{sc} = \{L_{wm}, R, z\}$ where L_{wm} is the list of watermarked token pairs, R is the high entropy value, and z is the (modulo) integer, all generated by the watermark generation, along with two thresholds: (1) t, a *threshold* to decide if a certain pair is watermarked; and (2) k, the *minimum number of watermarked pairs* required to conclude whether D'_w is a watermarked dataset. How to set t and k depends on the robustness an owner wants (see Sections III-B4 and IV-A2). If the owner wants to prove ownership to a third party, it has to reveal its secrets to that party. This causes to prove the ownership once in public (see Section V-D). Our watermark detection algorithm (Algorithm II) proceeds as follows:

(1) It builds the histogram list D_w^{hist} of the suspected dataset D'_w as in the watermark generation algorithm. The algorithm does not calculate the boundaries, just the token frequencies.

(2) For each token pair $\{tk_i, tk_j\}$ in L_{wm}, if the pair exists in D_w^{hist}, the algorithm generates s_{ij} values as $s_{ij} = H(tk_i||H(R||tk_j)) \mod z$.

(3) Then, it decides whether it will accept a given token pair (tk_i, tk_j), as watermarked or not by checking if the following statement holds: $(f_i - f_j) \mod s_{ij} \leq t$.

(4) After finding which pairs are watermarked, it checks whether their number is over the *minimum number of pairs*, k, needed to conclude that D'_w is watermarked by the owner, and returns $accept$ (verified) or $reject$, accordingly.

4) Probabilistic Analysis of False Positives: We develop a statistical bound in the form of the closed form expression derived from Markov's inequality theorem, to demonstrate that the false positive probability (i.e., accepting a dataset as watermarked when it is not) goes to zero if we increase the minimum number of pairs k that has to be accepted, or if we decrease the threshold t to accept a pair as watermarked. Recall that the m-th token pair $\{tk_i, tk_j\} \in L_{wm}$ is accepted as watermarked, if $(f_i - f_j) \mod s_{ij} \leq t$. We represent the probability that this "watermarking statement" holds as $P(X_m = 1) = p_m$, for $m = 1, ..., n$. Let us assume that p_m's follow a `Uniform[0,1]` distribution. The probability of having at least k successes in n trials can be written as $P(S_n \geq k) = \sum_{i=k}^{n} P(S_n = m)$. We now study the behavior of $P(S_n \geq k)$ depending on the behavior of t and k by using the *Sandwich Rule* and Markov's upper bound obtained by its inequality theorem $P(S_n \geq k) \leq \frac{\mu}{k}$, where $\mu = \sum_{m=1}^{n} p_m$ is the mean of S_n. Our analysis shows that if we decrease t, the probability of accepting a dataset as watermarked goes to zero and if we increase k, it will be hard to "accept" a dataset as watermarked. For further details, see the full version [61].

Algorithm II: Watermark Detection

Input: $D'_w, L_{sc} = \{L_{wm}, R, z\}, k, t$
Output: $accept/reject$
D_w^{hist} = **Preprocess**(D'_w) $count = 0$, $result = reject$
foreach $\{tk_i, tk_j\} \in L_{wm}$ **do**
  if ***Found***(tk_i, tk_j, D_w^{hist}) **then**
    $s_{ij} = H(tk_i||H(R||tk_j)) \mod z$
    if $(f_i - f_j) \mod s_{ij} \leq t$ **then**
      $count++$
    end
  end
end
if $count \geq k$ **then**
  $result = accept$
end
Result: result

IV. Experimental Evaluation

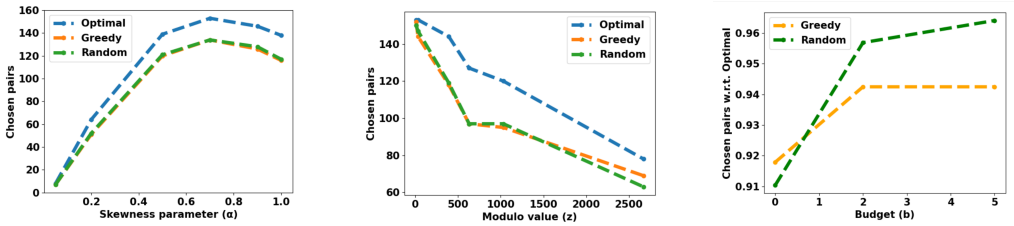
All of our experimental results are produced on a standard laptop machine with dual-core Intel Core(TM) i7 − 5600U CPU 2.5GHz, 16.00 GB RAM, 64-bit OS, and implemented in Python language. We deployed SHA256 as a hash function.

A. Synthetic Experiments

For our synthetic experiments, we generated synthetic datasets using a $power - law$ distribution [64] with different skewness values α as $[0.05, 0.2, 0.5, 0.7, 0.9, 1]$. The sample size is $1M$ and the number of tokens is $1K$ for each different α value. When α is 0, it is a uniform distribution in which there are no eligible tokens to watermark. When α is 1, the original dataset D_o is skewed with a very long tail with almost equal values. In this setting, we evaluate how the parameters (a modulo value z, a budget b, and skewness parameter α) are affecting the number of chosen pairs for watermarking and the performance of *optimal*, *greedy*, and *random* approached in terms of number of chosen pairs.

Figure 2a shows the correlation between skewness of a dataset α and the size of chosen pairs when budget $b = 2$ and modulo value $z = 1031$. When a dataset is almost uniform (i.e., $\alpha = 0.05$), the solutions can select very few pairs since there are not many eligible items (i.e., the upper and lower boundaries are not enough, in fact many of them are 0). When α starts increasing, the differences between the frequencies of tokens increase. Thus, the number of eligible items increases which also affects the number of chosen pairs under a given budget. However, at some point (i.e., $\alpha = 0.7$), the number of chosen pairs decreases due to the tail of (histogram) frequencies becoming uniform. The same figure shows the superior performance of the optimal solution. The gap between optimal and the heuristics is around 20% for most α values while the two heuristics perform similar the one with the other (0.02% in average).

Figure 2b illustrates how the modulo value z affects the size of chosen pairs. When we pick smaller modulo value z, we would have a higher number of chosen pairs. The reason is that a smaller z leads to smaller remainders s_{ij} that need to be eliminated, thereby yielding more selected pairs within a given budget b. When z is very small (i.e., 10), the three



(a) Effect of different skewness parameters (α) on chosen pairs by *Optimal*, *Greedy*, and *Random*.

(b) Effect of different modulo values (z) on chosen pairs by *Optimal*, *Greedy*, and *Random*.

(c) Chosen pairs by *Greedy* and *Random* with respect to *Optimal*.

Fig. 2: Effects of parameters on the size of chosen pairs for watermarking.

Dataset	Size	Token	Distinct Tokens	$\lvert L_e \rvert$	Optimal	Greedy	Random	Gen (sec)	Detect (sec)
Chicago Taxi [62]	9.68GB	`Taxi ID`	6573	33308	805	770	773	182.51	0.609
eyeWnder [58]	247MB	`URL`	11479	257	38	33	31	420.81	0.053
Adult [63]	4MB	`Age`	73	72	21	20	17	0.03	0.001

TABLE II: Validation results on real world datasets. **Dataset**: Dataset used **Size**: The size of original dataset. **Token**: Definition of the token (e.g., the name(s) of the attributes). **Distinct Token**: The number of distinct tokens. $|L_e|$: The number of eligible pairs. **Optimal**: The number of chosen pairs by the optimal matching. **Greedy**: The number of chosen pairs by the greedy matching. **Random**: The number of chosen pairs by the random matching. **Gen**: Running time of watermark generation. **Detect**: Running time of watermark detection.

approaches are very close (also see Section V-A for the effect of z in terms of security). However, when z increases, our optimal approach selects many more pairs than greedy and random. Figure 2c shows how the budget selection affects the performance comparison between the heuristics and the optimal. We set modulo value $z = 1031$ and use the dataset with the skewness $\alpha = 0.7$. When we increase the budget b, the heuristics get closer to the optimal performance. This is expected since even the optimal algorithm cannot select more than all the eligible pairs and with a large budget even the heuristics can approach that.

1) Limit of z: We stated that z is selected from Z^+; however, by analyzing the frequency histogram we can derive the upper and lower boundaries. Since the minimum value (lower bound) z can take is 2, we delve into investigation of the upper bound of z. Note that since the token with the highest frequency has the upper bound of infinity, there will be *at least* one pair that could be used for watermarking. Assume a watermarking pair candidate (tk_i, tk_j). Their frequencies, f_i and f_j, are changed such that $(f_i^w - f_j^w) \mod s_{ij} \equiv 0$. To have an upper bound for z, let us investigate which pair of tokens results in the highest difference. If we can determine the highest difference, say r_{max}, then r_{max} can be assumed as the upper bound for z since it is the highest remainder. Now, considering D_o^{hist}, the highest difference is between tk_0 (the token having the highest frequency) and $tk_{|D_o^{hist}|-1}$ (the token having the lowest frequency). That means that the largest remainder for any pair is $r_{max} = (f_0^0 - f_{|D_o^{hist}|-1}^0)$. Thus it is natural to accept that the upper bound of z is r_{max}. To conclude, z can be chosen from $(2, r_{max})$. Overall, r_{max} can be calculated as $\forall f_i, f_j \in D_o^{hist}$ s.t. $f_i \geq f_j$; $r_{max} = \max(\{f_i - f_j\})$. Hence, the upper bound for z is calculated. However, note that this value can be small and can be an advantage to an attacker. As discussed in Section IV-A, z affects the number of chosen pairs; thus, it correlates with the mix of possible attacks and is use-case scenario dependent. We plan to investigate this observation theoretically and experimentally in terms of security, robustness, and utility in the future.

2) Limit of t: Another critical parameter is $t \in [0, s_{ij} - 1]$. Note that since s_{ij} has an upper bound as $z - 1$, the highest value assigned to t is $z - 1$. While in our experimental study we chose t as a constant value, t could be also a percentage. Assume that an owner wishes to state that it wants 50% of error tolerance. Now, setting $t = s_{ij} \times 0.5$ states that a pair, say tk_i and tk_j, will be accepted as a watermarked if $(f_i - f_j) \mod s_{ij} \leq s_{ij}/2$. Thus, t represents the robustness level an owner desires. For instance, if $t = 0$ then the watermark becomes fragile since it cannot tolerate any changes in D_w, thus missing watermarked pairs (i.e., high false negatives). On the other hand, when $t = 100$, it is more robust and can tolerate modifications in D_w; however, it also means that it accepts more false positives (see also Section III-B4).

B. Validation Using Real World Datasets

Next we apply *FreqyWM* to three real world datasets from different domains: (1) `Chicago Taxi` dataset [62]; (2) A real click-stream dataset logging the URLs visited by a group of users of the `eyeWnder` advertisement detection add-on [58]; (3) `Adult` dataset [63]. Our intention is to validate our main conclusion using real data from the previous evaluation with synthetic data, i.e., that the heuristic approaches perform well enough compared to the optimal. Furthermore, we aim to report the real processing time on an ordinary machine for generating and detecting the watermark using these datasets.

For our watermark generation, we set the modulo value $z = 131$ and the budget $b = 2$. We run our algorithm 30 times and take the mean of total computations. Table II presents our validation results. `Taxi ID`, `URL`, and `Age` were chosen as tokens for `Chicago Taxi`, `eyeWnder`, and `Adult`, respectively. After generation, for `Chicago Taxi`, `eyeWnder`, and `Adult` datasets, our optimal solution chose 805, 38, and 21 pairs, respectively. Considering the heuristic approaches, greedy chose 770 pairs for `Chicago Taxi`, 33 pairs for `eyeWnder`, and 20 pairs `Adult` while random chose 773, 31, and 17 pairs, for `Chicago Taxi`, `eyeWnder`, and `Adult`, respectively. Running times of computations for watermark generation on the `Chicago Taxi`, `eyeWnder`, and `Adult` datasets were 182.51 secs, 420.81 secs, and 0.03 secs, respectively (where we exclude histogram and watermarked data generations). For watermark detection, the total detection time for each watermarked datasets was less than 1 sec. As it can be interpreted from Table II, the number of chosen pairs increases when the number of eligible pairs increases. For instance, while `eyeWnder` has more distinct tokens (11479) than `Chicago Taxi` has (6573), `eyeWnder` has fewer eligible pairs (257) than `Chicago Taxi` has (33308). Thus, *FreqyWM* has selected more pairs for `Chicago Taxi` (805) than it selected for `eyeWnder` (38).

C. Watermarking Multi-Dimensional Data

During our discussion so far, we set the token as a single attribute. However, as we previously stated, a token does not necessarily need to be restricted to a single attribute of a multi-dimensional dataset. Therefore, a token can be also defined as combination of more than one attributes (e.g., `[Age, WorkClass]`) in the `Adult` dataset. We ran *FreqyWM* on such token represented as `[Age, WorkClass]` with the same parameter setting in Section IV-B and the number of tokens (i.e., distinct `[Age, WorkClass]` attributes in the real dataset) were 481. The size of pairs chosen for watermarking was 20. With multi-dimensional data removing a token appearance is as simple as with single-dimensional data. Increasing, however, a token's frequency is more involved. The reason is that just repeating the value of the token would leave other fields not being part of the token with a value to be set, e.g., all the other fields beyond Age and WorkClass in the `Adult` dataset. A naive solution would be select a random appearance of the token and copy its other fields every time that an additional instance of the token needs to be added to the watermarked dataset. This, however, could create semantic inconsistencies if there are constraints to be respected for individual attributes or combinations of them. Making sure that added appearances of a token do not lead to semantic inconsistencies that, in addition to degrading the quality of the data, could also give away the existence of a watermarked pair to an attacker. This analysis requires domain knowledge about what each dataset represents. Such knowledge, however, is orthogonal to all previous steps of our algorithms and, thus, can be appended as a last step based on one's domain knowledge of the data whenever a token's frequency needs to be increased. We are currently investigating them and the effect of *FreqyWM* on data utility of such dataset with unique attributes as it is difficult to determine as addressed by [23].

D. Comparison to Related Works

As stated previously, we cannot directly apply (numerical) database watermarking to datasets similar to the ones we used for validation. However, one naive approach would be to convert a dataset to a numerical representation (e.g., a histogram) and watermark this numerical representation. In a nutshell, the histogram of a given dataset based on a predefined token is generated and then, the histogram is treated as a two dimensional database consisting of primary keys which are the tokens and an attribute consisting of integer values which are the frequencies. Later, a database watermarking is employed on this histogram. Then as in *FreqyWM* data transformation (e.g., removing/adding tokens) occurs according to the (new) watermarked histogram computed by the database watermarking. However, applying a numerical database watermarking is not really straightforward since it will distort the underlying dataset unexpectedly as a result of change in histogram data (e.g., cosine similarity) and would require modification in their watermarking techniques (e.g., how to create a watermarked dataset from the watermarked numerical representation). However, since this is the closest and simplest approach, we compare against it.

To actualize the above approach, we considered two numerical database watermarkings: 1) Shehab et al. [35] (referred as WM-OBT) due to partitioning approach (i.e., grouping tokens before watermarking) similar to *FreqyWM*; 2) Li et al. [30] (referred as WM-RVS) due to being one of the most recent reversible watermarking schemes introducing very small distortion compared to other same family of watermarkings.

More specifically, WM-OBT follows a data partition approach in which a watermark, defined as a bit sequence, is inserted on a group of partitions. Each data partition is filled by tokens and the frequencies of the tokens in each partition are modified/distorted by solving a minimization (if a watermark bit is 0) or maximization (if a watermark bit is 1) problem via a genetic algorithm [65], in which the objective function is in the form of a sum of sigmoid functions. WM-RVS treats each numeric value individually and changes its decimal part by selecting the random least significant position based on the watermarking key/bit and attributes. Furthermore, to apply WM-OBT and WM-RVS on a histogram generated from a dataset, we adjusted them such that their solutions produced are integers since a frequency count cannot be a decimal value.

For comparison, we investigate them based on two constraints: 1) change in the original histogram after watermarking (i.e., cosine similarity with watermarked histogram), and 2) the ranking of the tokens after watermarking.

We ran *FreqyWM*, WM-OBT, and WM-RVS on our synthetic data with skewness parameter 0.5 (with $1K$ distinct tokens and $1M$ sample size) where we set $b = 2$, and $z = 131$ for *FreqyWM*. We set parameters for WM-OBT and WM-RVS such that the parameters are proportional to the

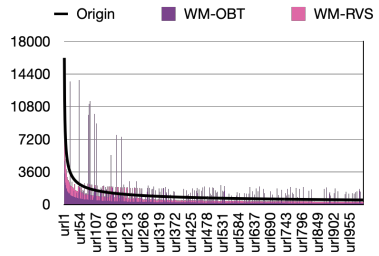


Fig. 3: Comparisons of the watermarked histograms generated from WM-OBT (purple) and WM-RVS (fuchsia) w.r.t. the original data histogram (black) for the synthetic dataset with dummy token names.

experimental settings of Shehab et al. [35] and Li et al. [30]. For WM-OBT, we use genetic algorithm (GA) technique for optimization [65] where we fix the number of partitions as 20 (where each partition has around 50 tokens), watermark bit sequence as $[1, 1, 0, 1, 0]$, condition as 0.75, and we allow the change (constraint) between $[-0.5, 10]$. The decoding threshold minimizing the probability of decoding error is calculated as 0.0966. For WM-RVS, we use the same bit sequence as in WM-OBT without creating it from the chaotic encryption. Also, let us note that WM-OBT took more than 30 minutes to run for such a small size dataset due to its optimization while WM-RVS was in the order of seconds. Figure 3 visualizes how the watermarked data histograms look like with respect to the original data histogram after applying WM-OBT and WM-RVS based on the experiments.

Similarity. In *FreqyWM*, even with 2% budget, the similarity between the original histogram and the watermarked histogram is 99.9998%, indicating that not all the budget was exhausted. On the other hand, for WM-OBT and WM-RVS, the similarities are 54.28% and 96%, respectively. The mean and standard deviation of the changes introduced to the histogram by WM-OBT are 444 and 855.91, respectively while they are -69.43 and 414.10 for WM-RVS, respectively.

Ranking. Preserving the ranking is important because it allows us not to sacrifice the utility of a dataset, e.g., preserving the popularity of URLs. *FreqyWM* by definition maintains the ranking of tokens. However, our analysis showed that WM-OBT and WM-RVS changed the ranking of 998 and 987 out of the total 1000 tokens, respectively!

The results on similarity and ranking support our claim that applying a numerical database technique on histogram data would result in unexpected and uncontrolled distortion that seriously undermines the utility of the original data.

V. Security and Robustness Analysis

This section discusses the security and robustness of our *FreqyWM* method against four attacks: **guess**, **sampling**, **destroy**, and **re-watermarking (false-claim)** attacks which are well-known attacks in watermarking as studied by [23].In order to measure the robustness against sampling and destroy attacks, we run our optimal solution on a dataset where the skewness parameter $\alpha = 0.5$ (with $1K$ distinct tokens and $1M$ sample size), unless stated otherwise, the modulo value $z = 131$, and the budget $b = 2$ and it selected 139 pairs for watermark. We run the experiments for 100 times and compute the average accepted pairs over all repetitions.

A. Guess (Brute-Force) Attack

In the guess attack, the probabilistic polynomial time adversary tries to guess the watermark, i.e., the secret embedded in the data. This is possible only if it can figure out a subset of token pairs $\{tk_i, tk_j\}_l$ (where $\binom{|D_w^{hist}|}{2} \geq l \geq k$) based on the watermarked data D_w, the random value R, and the modulo value z where the watermark detection algorithm based on these inputs (for some fixed k and t) returns `accept`. Assuming that the hash function is collision resistant, R is random, and z is an integer, the probability of the attacker being successful can be formally defined as:

$$Pr[R \leftarrow \{0,1\}^{\lambda}; (D_w, L_{sc} = \{\{tk_i, tk_j\}_{|L_{wm}|}, R, z\}) \leftarrow WmGenerate(D_o, b) : \mathcal{A}(D_w) \rightarrow L'_{sc} = \{\{tk_i, tk_j\}_l, R^*, z^*\}| WmDetect(D_w, L'_{sc}, k, t) = 1] \leq \mathsf{negl}(\lambda)$$

Considering the typical parameter values, the probability of success becomes negligible.

B. Sampling Attack

In this attack, $\mathcal{A}$ copies a random subsample from the watermarked dataset D_w in an attempt to exploit (pirate/steal) it while hoping that the watermark won't be detectable within the extracted sample. The attack is run for different sample sizes from 1% to 90%, extracted from the original watermarked dataset D_w. For each percentage and subsample we apply the detection algorithm and compute the percentage of accepted pairs. Also, for each subsample detection experiment, we deploy different values of the threshold t for accepting a pair as watermarked as $t = \{0, 1, 2, 4, 10\}$. The attack scenario is as follows: $\mathcal{A}$ randomly selects $x\%$ of D_w where x defines the percentage for the sampling attack (e.g., 1) as a subsample size of $1M \times \frac{x}{100}$. When the owner suspects the dataset (possible subsampled), it scales it up to the size of D_w by multiplying the frequency counts by $\frac{100}{x}$ by using its info from the (original) watermarked dataset (e.g., via info added to its metadata). For instance, for 1% sampling attack, a subsample would have total of $1M \times 0.01 = 10K$ where each f_i is multiplied by approximately 0.01. Note that if the sample size is greater than the number of distinct tokens, which is the number of items in D_w^{hist}, the sample will have all the distinct tokens with a high chance. This also means that all the chosen watermarked pairs are in the subsample. Our results show that the size of the extracted subsample does not greatly affect the number of accepted pairs if it is greater than the number of unique tokens ($1K$). Since the frequencies of the tokens vary, the value of t does affect the result of the detection. For example, with $t = 0$ the detection algorithm can detect around 36% (in average) of the watermarked pairs. When t increases from 1 to 10, the performance of the detection increases (in average) from 72% to 99.5%.

Let us now see the results when the size of the extracted subsample is very low, so that it might not contain at least

1 token from the total $1K$ of unique tokens that the original watermarked dataset has. Figure 4 shows the results for sample size proportions between 0.0007% and 0.5%. Observe that if the sample size is greater than $5\times$ the number of unique tokens ($1K$), the detection algorithm stabilizes its performance for detecting the watermark. Below $2\times$ ($2K$), the performance starts to decrease with higher velocity. In these extreme cases, the detection algorithm will have more difficulties to detect the data as watermarked. However, the utility of the data is highly degraded since the subsample sizes are very small compared to the original size of $1M$ tokens. This causes a small number of distinct tokens to be found in the subsample.

Effect of modulo bases. As seen previously, t is crucial for detecting whether a pair is watermarked. For small values of t to be sufficient to fend off sampling attacks, the remainders need to be small numbers that are covered by t. One way to achieve this is by ensuring that the modulo bases used (i.e., the s_{ij}'s) are relatively small numbers when compared to the actual appearance frequencies of watermarked pairs. When this does not apply, the method will of course fail. For instance, assume a watermarked pair involving frequencies $f_i = 540, f_j = 440$ which under base $s_{ij} = 100$ leave a remainder of 0. W.l.o.g, lets assume that a 50% frequency attack leads to a dataset with $f_i = 270, f_j = 220$ which leads to a remainder of $(270 - 220) \mod 100 \equiv 50$. Now if t is chosen smaller than 50 then the watermarked pair will not be detected. The reason is that mod 100 leaves large remainders when applied to f_i and f_j that are in the same order with 100. In our experimental results f_i's were always much larger numbers than the employed s_{ij}, thereby, even small t's would detect a pair under a sampling attack. To determine the optimal t and how robust it is against attacks, a further investigation is needed as it depends on various parameters such as z, s_{ij} as well on the mix of expected attacks as discussed later. Note that our experimental results show that s_{ij} values are ~ 2 order of magnitude smaller than z. Furthermore, we also tested the sampling attack in other watermarked datasets with different values of the skewness parameter and obtained similar results.

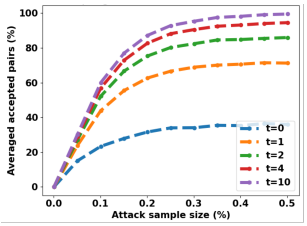


Fig. 4: Sampling attack results with very small sample and $\alpha = 0.5$.

C. Destroy Attack

In this case the attacker $\mathcal{A}$ tries to damage the watermark. The no-security-by-obscurity principle [66] allows $\mathcal{A}$ to know that it can destroy the watermark in a way that it cannot be detected by the owner. $\mathcal{A}$ computes the histogram of watermarked data D_w. $\mathcal{A}$ modifies the frequencies of tokens as it pleases by allowing re-ordering (changing the popularity/rank of the tokens) or without allowing re-ordering. We define these two attacks and discuss *FreqyWM*'s robustness against them.

1) Destroy Attack without re-ordering: In this attack type, $\mathcal{A}$ can modify the frequencies without changing the order of frequencies. We introduce two types: (1) attacker changes the frequencies randomly by the given boundaries and (2) attacker changes the frequencies by (at most) some percentage.

Changing the frequencies randomly within the boundaries. $\mathcal{A}$ calculates the boundaries for each token. Then, $\mathcal{A}$ chooses a random value r_i for each tk_i as $r_i \leftarrow (-l_i, u_i)$. $\mathcal{A}$ changes the frequency of tk_i and updates u_{i+1} of tk_{i+1} by r_i.

Changing the frequencies by (at most) some percentage. $\mathcal{A}$ changes the frequencies of tokens up to some percentage (e.g., 1%). To illustrate, $\mathcal{A}$ calculates the boundaries as u_i and l_i for each tk_i where it sets the percentage to 1%. It calculates $u'_i = floor(u_i \times 0.01)$ and $l'_i = floor(l_i \times 0.01)$. Then it gets a random value r_i between $(-l'_i, u'_i)$. It hereby changes tk_i by at most $\pm 1\%$. After every change ($f'_i = f_i + r_i$), the boundary of the next element is updated. Thus, the attack never changes the ranking/ordering since l'_i and u'_i are already in the boundaries.

Figure 5 shows how robust *FreqyWM* is against these two destroy attacks. We compare the success rate (the percentage of accepted token pairs given threshold for accepting a pair t) of detection algorithm with respect to modified watermarked data after the attacks. We also include in the figure a second dataset of skewness $\alpha = 0.7$ *that does not carry the watermark*, and report on how many of its pairs would be falsely verified for different values of t. For an attack in which the frequencies are changed by (at most) some percentage (represented by the red line in the figure), *FreqyWM* can detect around 90% of the pairs when $t = 0$. When t is increased, after a point where $t \geq f_i - f_j \mod s_{ij}$, the success rate converges at around 90%. For an attack where the frequencies are changed randomly within the upper and lower frequency boundaries (green line in Figure 5), *FreqyWM* can detect more than 35% of the pairs when $t = 0$. Note that the latter is more powerful than the former attack. There is a direct proportion between t and the success rate. As shown, the success rate reaches to 90% when t goes to 10.

From Figure 5, we can interpret in what parameter setting false negative (rejecting a watermarked pair) and false positive (accepting a pair as watermarked while it is not) can be avoided. Thus, the watermarking detection algorithm can successfully detect a watermarked dataset attacked and reject a dataset that was not watermarked. For instance, the rate of false positive increases when the threshold for accepting a pair t increases while the minimum number of accepted pairs for detection k decreases which is the area under the results of the dataset (not watermarked) with a different skewness parameter (the area under the orange line). On the other hand, the rate of false negative increases when the threshold accepting a pair t decreases while threshold for detecting a watermark k increases which is the area above the results of the attack without re-ordering (the area above the green line) if we consider a very strong attack. To avoid false negatives/positives, convenient parameter settings (i.e., t and k) for detecting a

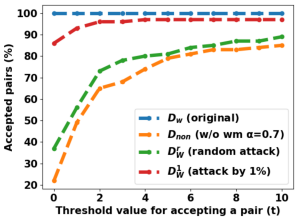


Fig. 5: Percentage of verified pairs for the following datasets: (1) D_w : the original watermarked dataset $\alpha = 0.5$ without any attack/modification, (2) D_{non} : a non-watermarked dataset defined over the same token space but with $\alpha = 0.7$, (3) D^r_w : D_w after attacked by random attack without reordering, (4) D^1_w : D_w after attacked by changing frequencies at most 1%.

watermark lie between these two areas (between the orange and the green lines in Figure 5). However, if a weaker attack (changing the frequencies by some percentage) is considered, the range of these parameters increases (the area between the red and orange line). Hence, the detection algorithm can detect a watermarked dataset and reject a dataset not watermarked by the owner with a careful parameter setting. For instance, adjusting t (and k) based on the nature of the data and the specific application context can enable us to reduce the false positives/negatives. This is an interesting future work.

2) Destroy Attack with re-ordering: In this attack type, an attacker $\mathcal{A}$ can modify the frequencies as it pleases without observing any ordering restrictions. Note that this attack introduces more noise than the attack without re-ordering which reduces the usability of watermarked data D_w. $\mathcal{A}$ modifies the frequencies with various percentages $[10\%, 30\%, 50\%, 60\%, 80\%, 90\%]$ where the success rates are $[94\%, 88\%, 82\%, 79\%, 78\%, 76\%]$ respectively. *FreqyWM* can detect the watermark with 76% chance up to modifications of 90% in frequencies approximately (where $t = 4$).

D. Re-watermarking/ False-Claim Attack

This attack is mounted by an attacker $\mathcal{A}$ creating a new watermark on the watermarked data D_w, generated by an honest owner. $\mathcal{A}$ generates its own watermarked data by simply inserting D_w into the watermark generation algorithm as data to produce $D^{\mathcal{A}}_w$. Then $\mathcal{A}$ can present $D^{\mathcal{A}}_w$ and claim the ownership of $D^{\mathcal{A}}_w$ (since $\mathcal{A}$ can prove its ownership claim by introducing its watermarking secret list $L^{\mathcal{A}}_{sc}$). This attack creates a dispute since both the real owner, who created D_w, and $\mathcal{A}$ have proofs of their ownerships. The dispute can be arbitrated by introducing a judge (a trusted third party as suggested by [67]) to the watermarking scheme. Both parties, $\mathcal{A}$ and the real owner, introduce their secrets and their watermarked data. $\mathcal{A}$ sends its secrets $L^{\mathcal{A}}_{sc}$ and its watermarked data $D^{\mathcal{A}}_w$. The real owner sends its secrets L_{sc} and its watermarked data D_w. The judge computes watermark detection algorithm on each received data for each secret which creates four outputs. The judge compares these results and identifies the real owner since only the secret of the real owner can produce accept on both data. To show practicality of our defense against the re-watermarking attack, we implemented the attack above. Our results show that the first watermark is detected with 92% on $D^{\mathcal{A}}_w$ under $t = 0$. The attacker's only way to succeed is to perform successful guess or destroy attack which it cannot perform as shown previously.

VI. Discussion

We propose possible adjustments to *FreqyWM* for more sophisticated properties and discuss some challenges as follows:

• *Incremental FreqyWM.* In the literature, there exist watermarking techniques that allow to update a watermark on a dataset without computing insertion from scratch [57]. We believe that an incremental FreqyWM can be built on top of dynamic maximum weighted matching [68, 69] works but we leave such investigations to future work.

• *Multi-watermarks.* One might watermark a dataset multiple times with different intentions: 1) to have a chronological order in the versions (e.g., data provenance); and 2) to falsely claim ownership (see Section V-D). Non-withstanding the motivation, we run an experiment to calculate the discrepancy between the original (histogram) one and the final one after 10 insertions assuming a budget $b = 0.002$ for each iteration on a sample dataset with $\alpha = 0.5$. The resulting similarity between the original and the latest watermarked version is 0.003%. As it is evident, *FreqyWM* did not introduce 0.02% but rather very tiny distortion (see also the full version [61] for further analysis, i.e., ML analysis). Hence, successive re-watermaking can be practical with *FreqyWM*, but we plan to extensively investigate it in the future.

•*Challenging datasets.* Apart from datasets with close to uniform frequencies, *FreqyWM* can also be challenged when the range of token values is too wide, e.g., sales' datasets with many decimal values, resulting to very few (if any) repetition of the same value. One natural solution to this is to first bucketize (cluster) the widely ranged data and then apply *FreqyWM* at the level of the bucket as opposed to the exact token value.

VII. Conclusions and Future Work

We proposed *FreqyWM*, a novel frequency watermarking technique for protecting the ownership of data in the emerging new data economy. We analysed the performance of *FreqyWM* and showed how *FreqyWM* can encode watermarks with minimal distortion on the original data, provided that the data has sufficient variability in terms of token frequencies. We analysed *FreqyWM*'s robustness to generic attacks. *FreqyWM* is applicable to large numbers of tuples sold in wholesale manner in modern DMs. An interesting, yet challenging, research direction is to consider how to watermark small sets or even individual tuples used in distributed data operations such as replication and remote hosting and/or query execution. We are currently looking at more attack scenarios and at devising systematic procedures for optimizing the parameters and also how to apply *FreqyWM* to multidimensional datasets by overcoming the challenges mentioned in Section IV-C. We also investigate integrating data privacy (e.g., differentially-private fingerprinting [70]).

Acknowledgements

Devriş İşler was supported by the European Union's HORIZON project DataBri-X (101070069). Nikolaos Laoutaris was supported by the MLEDGE project (REGAGE22e00052829516), funded by the Ministry of Economic Affairs and Digital Transformation and the European Union-NextGenerationEU/PRTR.

References

[1] S. A. Azcoitia and N. Laoutaris, "A survey of data marketplaces and their business models," *SIGMOD Rec.*, vol. 51, no. 3, pp. 18–29, 2022. [Online]. Available: https://doi.org/10.1145/3572751.3572755

[2] A. Lutu, D. Perino, M. Bagnulo, E. Frias-Martinez, and J. Khangosstar, "A characterization of the covid-19 pandemic impact on a mobile network operator traffic," in *Proceedings of the ACM Internet Measurement Conference*, ser. IMC '20. New York, NY, USA: Association for Computing Machinery, 2020. [Online]. Available: https://doi.org/10.1145/3419394.3423655

[3] D. Evans, V. Kolesnikov, and M. Rosulek, "A pragmatic introduction to secure multi-party computation," *Found. Trends Priv. Secur.*, 2018. [Online]. Available: https://doi.org/10.1561/3300000019

[4] C. Gentry, "A fully homomorphic encryption scheme," Ph.D. dissertation, Stanford University, USA, 2009. [Online]. Available: https://searchworks.stanford.edu/view/8493082

[5] D. Boneh, A. Sahai, and B. Waters, "Functional encryption: Definitions and challenges," in *Theory of Cryptography Conference, TCC*. Springer, 2011. [Online]. Available: https://doi.org/10.1007/978-3-642-19571-6_16

[6] M. Sabt, M. Achemlal, and A. Bouabdallah, "Trusted execution environment: What it is, and what it is not," in *TrustCom/BigDataSE/ISPA*. IEEE, 2015. [Online]. Available: https://doi.org/10.1109/Trustcom.2015.357

[7] Y. Li, D. Ghosh, P. Gupta, S. Mehrotra, N. Panwar, and S. Sharma, "PRISM: private verifiable set computation over multi-owner outsourced databases," in *SIGMOD: International Conference on Management of Data, Virtual*. ACM, 2021. [Online]. Available: https://doi.org/10.1145/3448016.3452839

[8] R. Poddar, T. Boelter, and R. A. Popa, "Arx: An encrypted database using semantically secure encryption," *Proc. VLDB Endow.*, 2019. [Online]. Available: http://www.vldb.org/pvldb/vol12/p1664-poddar.pdf

[9] N. Anciaux, L. Bouganim, P. Pucheral, I. S. Popa, and G. Scerri, "Personal database security and trusted execution environments: A tutorial at the crossroads," *Proc. VLDB Endow.*, 2019. [Online]. Available: http://www.vldb.org/pvldb/vol12/p1994-anciaux.pdf

[10] X. Ren, L. Su, Z. Gu, S. Wang, F. Li, Y. Xie, S. Bian, C. Li, and F. Zhang, "HEDA: multi-attribute unbounded aggregation over homomorphically encrypted database," *Proc. VLDB Endow.*, 2022. [Online]. Available: https://www.vldb.org/pvldb/vol16/p601-gu.pdf

[11] W. Zhou, Y. Cai, Y. Peng, S. Wang, K. Ma, and F. Li, "Veridb: An sgx-based verifiable database," in *SIGMOD: International Conference on Management of Data*. ACM, 2021. [Online]. Available: https://doi.org/10.1145/3448016.3457308

[12] P. Jougleux, "Data ownership (and succession law)," in *Facebook and the (EU) Law: How the Social Network Reshaped the Legal Framework*. Springer, 2022, pp. 129–143.

[13] J. Kennedy, P. Subramaniam, S. Galhotra, and R. C. Fernandez, "Revisiting online data markets in 2022: A seller and buyer perspective," *SIGMOD Rec.*, vol. 51, no. 3, pp. 30–37, 2022. [Online]. Available: https://doi.org/10.1145/3572751.3572757

[14] R. C. Fernandez, P. Subramaniam, and M. J. Franklin, "Data market platforms: Trading data assets to solve data problems," *Proc. VLDB Endow.*, vol. 13, no. 11, pp. 1933–1947, 2020. [Online]. Available: http://www.vldb.org/pvldb/vol13/p1933-fernandez.pdf

[15] F. Banterle, "Data ownership in the data economy: a european dilemma," *EU Internet Law in the Digital Era: Regulation and Enforcement*, pp. 199–225, 2020. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3277330

[16] M. Asikuzzaman and M. R. Pickering, "An overview of digital video watermarking," *IEEE Trans. Circuits Syst. Video Technol.*, 2018. [Online]. Available: https://doi.org/10.1109/TCSVT.2017.2712162

[17] M. Begum and M. S. Uddin, "Digital image watermarking techniques: A review," *Inf.*, 2020. [Online]. Available: https://doi.org/10.3390/info11020110

[18] H. Ma, C. Jia, S. Li, W. Zheng, and D. Wu, "Xmark: Dynamic software watermarking using collatz conjecture," *IEEE Trans. Inf. Forensics Secur.*, 2019. [Online]. Available: https://doi.org/10.1109/TIFS.2019.2908071

[19] X. Zhou, H. Pang, K. Tan, and D. Mangla, "Wmxml: A system for watermarking XML data," in *International Conference on Very Large Data Bases (VLDB)*. ACM, 2005. [Online]. Available: http://www.vldb.org/conf/2005/papers/p1318-zhou.pdf

[20] R. Agrawal and J. Kiernan, "Watermarking relational databases," in *Proceedings ofInternational Conference on Very Large Data Bases, VLDB*, 2002. [Online]. Available: http://www.vldb.org/conf/2002/S05P03.pdf

[21] R. Agrawal, P. J. Haas, and J. Kiernan, "A system for watermarking relational databases," in *ACM SIGMOD International Conference*, 2003. [Online]. Available: https://doi.org/10.1145/872757.872865

[22] T. Wang and F. Kerschbaum, "RIGA: covert and robust white-box watermarking of deep neural networks," in *WWW: The Web Conference*, 2021. [Online]. Available: https://doi.org/10.1145/3442381.3450000

[23] S. Rani and R. Halder, "Comparative analysis

of relational database watermarking techniques: An empirical study," *IEEE Access*, vol. 10, pp. 27 970–27 989, 2022. [Online]. Available: https://doi.org/10.1109/ACCESS.2022.3157866

[24] N. Agarwal, A. K. Singh, and P. K. Singh, "Survey of robust and imperceptible watermarking," *Multim. Tools Appl.*, 2019. [Online]. Available: https://doi.org/10.1007/s11042-018-7128-5

[25] R. Agrawal, P. J. Haas, and J. Kiernan, "Watermarking relational data: framework, algorithms and analysis," *VLDB J.*, 2003. [Online]. Available: https://doi.org/10.1007/s00778-003-0097-x

[26] T. Ji, E. Yilmaz, E. Ayday, and P. Li, "The curse of correlations for robust fingerprinting of relational databases," in *RAID : International Symposium on Research in Attacks, Intrusions and Defenses*. ACM, 2021. [Online]. Available: https://doi.org/10.1145/3471621.3471853

[27] E. Quiring, D. Arp, and K. Rieck, "Forgotten siblings: Unifying attacks on machine learning and digital watermarking," in *IEEE European Symposium on Security and Privacy, EuroS&P*. IEEE, 2018. [Online]. Available: https://doi.org/10.1109/EuroSP.2018.00041

[28] A. Cohen, J. Holmgren, R. Nishimaki, V. Vaikuntanathan, and D. Wichs, "Watermarking cryptographic capabilities," *SIAM J. Comput.*, 2018. [Online]. Available: https://doi.org/10.1137/18M1164834

[29] X. Tang, Z. Cao, X. Dong, and J. Shen, "Pkmark: A robust zero-distortion blind reversible scheme for watermarking relational databases," in *IEEE International Conference on Big Data Science and Engineering*, 2021. [Online]. Available: https://doi.org/10.1109/BigDataSE53435.2021.00020

[30] W. Li, N. Li, J. Yan, Z. Zhang, P. Yu, and G. Long, "Secure and high-quality watermarking algorithms for relational database based on semantic," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–14, 2022.

[31] M. L. P. Gort, M. Olliaro, A. Cortesi, and C. F. Uribe, "Semantic-driven watermarking of relational textual databases," *Expert Syst. Appl.*, 2021. [Online]. Available: https://doi.org/10.1016/j.eswa.2020.114013

[32] C. Lin, T. Nguyen, and C. Chang, "LRW-CRDB: lossless robust watermarking scheme for categorical relational databases," *Symmetry*, 2021. [Online]. Available: https://doi.org/10.3390/sym13112191

[33] S. Kumar, B. K. Singh, and M. Yadav, "A recent survey on multimedia and database watermarking," *Multim. Tools Appl.*, vol. 79, no. 27-28, pp. 20 149–20 197, 2020. [Online]. Available: https://doi.org/10.1007/s11042-020-08881-y

[34] M. H. Jony, F. T. Johora, and J. F. Katha, "A robust and efficient numeric approach for relational database watermarking," in *IEEE International Conference on Sustainable Technologies for Industry 4.0 (STI)*, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9732582

[35] M. Shehab, E. Bertino, and A. Ghafoor, "Watermarking relational databases using optimization-based techniques," *IEEE Trans. Knowl. Data Eng.*, 2008. [Online]. Available: https://doi.org/10.1109/TKDE.2007.190668

[36] D. Ibosiola, B. A. Steer, Á. García-Recuero, G. Stringhini, S. Uhlig, and G. Tyson, "Movie pirates of the caribbean: Exploring illegal streaming cyberlockers," in *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM*. AAAI Press, 2018. [Online]. Available: https://aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17835

[37] W. Zhou, J. Hu, and S. Wang, "Enhanced locality-sensitive hashing for fingerprint forensics over large multi-sensor databases," *IEEE Trans. Big Data*, 2021. [Online]. Available: https://doi.org/10.1109/TBDATA.2017.2736547

[38] Y. Lei, Q. Huang, M. S. Kankanhalli, and A. K. H. Tung, "Locality-sensitive hashing scheme based on longest circular co-substring," in *Proceedings of the 2020 International Conference on Management of Data, SIGMOD*. ACM, 2020. [Online]. Available: https://doi.org/10.1145/3318464.3389778

[39] D. Chang, M. Ghosh, S. K. Sanadhya, M. Singh, and D. R. White, "Fbhash: A new similarity hashing scheme for digital forensics," *Digit. Investig.*, 2019. [Online]. Available: https://doi.org/10.1016/j.diin.2019.04.006

[40] C. N. K. Osiakwan and S. G. Akl, "The maximum weight perfect matching problem for complete weighted graphs is in pc*," *Parallel Algorithms Appl.*, 1995. [Online]. Available: https://doi.org/10.1080/10637199508915506

[41] Z. Galil, "Efficient algorithms for finding maximum matching in graphs," in *ACM CSUR*, 1986.

[42] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to algorithms*. MIT press, 2009.

[43] E. Ayday, E. Yilmaz, and A. Yilmaz, "Robust optimization-based watermarking scheme for sequential data," in *International Symposium on Research in Attacks, Intrusions and Defenses, RAID*, 2019. [Online]. Available: https://www.usenix.org/conference/raid2019/presentation/ayday

[44] T. Ji, E. Ayday, E. Yilmaz, and P. Li, "Robust fingerprinting of genomic databases," *CoRR*, vol. abs/2204.01801, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2204.01801

[45] M. Kamran and M. Farooq, "A comprehensive survey of watermarking relational databases research," in *arXiv preprint arXiv:1801.08271*, 2018.

[46] A. S. Panah, R. G. van Schyndel, T. K. Sellis, and E. Bertino, "On the properties of non-media digital watermarking: A review of state of the art techniques," *IEEE Access*, 2016. [Online]. Available: https://doi.org/10.1109/ACCESS.2016.2570812

[47] M. E. Farfoura, S. Horng, J. Lai, R. Run, R. Chen, and M. K. Khan, "A blind reversible method for watermarking relational databases based on a time-stamping protocol," *Expert Syst. Appl.*, 2012. [Online].

Available: https://doi.org/10.1016/j.eswa.2011.09.005

[48] Y. Li and R. H. Deng, "Publicly verifiable ownership protection for relational databases," in *Proceedings of th ACM Symposium on Information, Computer and Communications Security, ASIACCS*. ACM, 2006. [Online]. Available: https://doi.org/10.1145/1128817.1128832

[49] D. Hu, D. Zhao, and S. Zheng, "A new robust approach for reversible database watermarking with distortion control," *IEEE Trans. Knowl. Data Eng.*, 2019. [Online]. Available: https://doi.org/10.1109/TKDE.2018.2851517

[50] H. M. El-Bakry and M. Hamada, "A novel watermark technique for relational databases," in *Artificial Intelligence and Computational Intelligence - International Conference, AICI 2010, Sanya, China, October 23-24, 2010, Proceedings, Part II*, ser. Lecture Notes in Computer Science. Springer, 2010. [Online]. Available: https://doi.org/10.1007/978-3-642-16527-6_29

[51] S. M. Darwish, H. A. Selim, and M. M. El-Sherbiny, "Distortion free database watermarking system based on intelligent mechanism for content integrity and ownership control," *J. Comput.*, 2018. [Online]. Available: https://doi.org/10.17706/jcp.13.9.1053-1066

[52] Y. Zhang, B. Yang, and X.-M. Niu, "Reversible watermarking for relational database authentication," 2008.

[53] W. Wang, C. Liu, Z. Wang, and T. Liang, "FBIPT: A new robust reversible database watermarking technique based on position tuples," in *International Conference on Data Intelligence and Security, ICDIS*. IEEE, 2022, pp. 67–74. [Online]. Available: https://doi.org/10.1109/ICDIS55630.2022.00018

[54] G. Gupta and J. Pieprzyk, "Reversible and blind database watermarking using difference expansion," *Int. J. Digit. Crime Forensics*, 2009. [Online]. Available: https://doi.org/10.4018/jdcf.2009040104

[55] K. Jawad and A. Khan, "Genetic algorithm and difference expansion based reversible watermarking for relational databases," *J. Syst. Softw.*, 2013. [Online]. Available: https://doi.org/10.1016/j.jss.2013.06.023

[56] M. B. Imamoglu, M. Ulutas, and G. Ulutas, "A new reversible database watermarking approach with firefly optimization algorithm," *Mathematical Problems in Engineering*, 2017. [Online]. Available: https://doi.org/10.1155/2017/1387375

[57] C. Chang, T. Nguyen, and C. Lin, "A reversible database watermark scheme for textual and numerical datasets," in *IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, SNPD*. IEEE, 2021. [Online]. Available: https://doi.org/10.1109/SNPD51163.2021.9704991

[58] C. Iordanou, N. Kourtellis, J. M. Carrascosa, C. Soriente, R. Cuevas, and N. Laoutaris, "Beyond content analysis: detecting targeted ads via distributed counting," in *Proceedings of the 15th International Conference on Emerging Networking Experiments And Technologies, CoNEXT*. ACM, 2019. [Online]. Available: https://doi.org/10.1145/3359989.3365428

[59] G. Cormode, S. Maddock, and C. Maple, "Frequency estimation under local differential privacy," *Proc. VLDB Endow.*, 2021. [Online]. Available: http://www.vldb.org/pvldb/vol14/p2046-cormode.pdf

[60] J. Katz and Y. Lindell, *Introduction to Modern Cryptography, Second Edition*. CRC Press, 2014. [Online]. Available: https://www.crcpress.com/Introduction-to-Modern-Cryptography-Second-Edition/Katz-Lindell/p/book/9781466570269

[61] D. İşler, E. Cabana, A. Garcia-Recuero, G. Koutrika, and N. Laoutaris, "Freqywm: Frequency watermarking for the new data economy," IMDEA Networks Technical Report, Tech. Rep., 2022.

[62] "Chicago Data Portal," 2022, https://data.cityofchicago.org/Transportation/Taxi-Trips/wrvz-psew.

[63] "Adult Dataset," 1996, https://archive.ics.uci.edu/ml/datasets/Adult.

[64] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *SIAM Rev.*, 2009. [Online]. Available: https://doi.org/10.1137/070710111

[65] D. Goldberg and K. Sastry, *Genetic algorithms: the design of innovation*. Springer, 2007.

[66] A. Kerckhoffs, "A. kerckhoffs, la cryptographie militaire, journal des sciences militaires ix, 38 (1883)," in *Journal des sciences militaires*, 1883.

[67] A. Adelsbach, S. Katzenbeisser, and H. Veith, "Watermarking schemes provably secure against copy and ambiguity attacks," in *ACM workshop on Digital rights management*, 2003. [Online]. Available: https://doi.org/10.1145/947380.947395

[68] S. Behnezhad, "Dynamic algorithms for maximum matching size," in *ACM-SIAM Symposium on Discrete Algorithms, SODA*. SIAM, 2023. [Online]. Available: https://doi.org/10.1137/1.9781611977554.ch6

[69] S. Solomon, "Fully dynamic maximal matching in constant update time," in *IEEE Annual Symposium on Foundations of Computer Science, FOCS*. IEEE Computer Society, 2016. [Online]. Available: https://doi.org/10.1109/FOCS.2016.43

[70] T. Ji, E. Ayday, E. Yilmaz, and P. Li, "Differentially-private fingerprinting of relational databases," *CoRR*, vol. abs/2109.02768, 2021. [Online]. Available: https://arxiv.org/abs/2109.02768

Anexo 7: Understanding the price of data in commercial Data Marketplaces

Understanding the Price of Data in Commercial Data Marketplaces

1st Santiago Andrés Azcoitia
IMDEA Networks Institute
Universidad Carlos III de Madrid
Leganés, Spain
santiago.azcoitia@imdea.org

2nd Costas Iordanou
Cyprus University of Technology
Limassol, Cyprus
kostas.iordanou@cut.ac.cy

3rd Nikolaos Laoutaris
IMDEA Networks Institute
Leganés, Spain
nikolaos.laoutaris@imdea.org

***Abstract*—A large number of Data Marketplaces (DMs) have appeared in the last few years to help owners monetize their data, and data buyers optimize their marketing campaigns, train their ML models, and facilitate other data-driven decision processes. In this paper, we present a first of its kind measurement study of the growing DM ecosystem, focused on understanding which features of data are actually driving their prices in the market. We show that data products listed in commercial DMs may cost from few to hundreds of thousands of US dollars. We analyze the prices of different categories of data and show that products about telecommunications, manufacturing, automotive, and gaming command the highest prices. We also develop classifiers for comparing data products across different DMs, as well as a regression analysis for revealing features that correlate with data product prices of specific categories, such as update rate or history for financial data, and volume and geographical scope for marketing data.**

***Index Terms*—Data economy, data marketplaces, measurement, data pricing**

I. Introduction

Data-driven decision making powered by Machine Learning (ML) algorithms is changing how the society and the economy work and is having a profound positive impact on our daily life. A McKinsey report predicted that data-driven decision-making could reach US$2.5 trillion globally by 2025 [30], whereas a recent market study within the scope of the European Data Strategy estimates a size of 827 billion euro for the EU27 [14]. ML is driving up the demand for data in what has been called the fourth industrial revolution.

To satisfy this demand, several data marketplaces (DMs) have appeared in the last few years. DMs are mediation platforms that aim to connect data providers (acting as sellers) to data consumers (acting as potential buyers), and to manage data transactions between them. This ecosystem includes open data repositories [28], [33], general-purpose [2], [7], [18], [19], [21], and specialized or niche DMs targeting specific industries, such as automotive [13], [50], financial [8], [55], marketing [41], [42], and logistics [65], to name a few.

An issue of paramount importance is that of *data pricing*. Some marketplaces leave it to sellers to set a price for their data products. Many of them do not list prices of their products, but leave it to buyers and sellers to agree on a price after a negotiation. Due to the elusive nature of the traded "commodity", pricing is a very complex matter, even more than in the case of material goods [53]. Unlike oil, to which it is often compared [17], data can be copied / transmitted / processed with close to zero cost. Even the use of the term commodity is a gross oversimplification of what data is. Notice that whereas two liters of gasoline yield a similar mileage on two similar cars under similar driving styles, nothing of this sort applies to data since 1) two datasets of equal volume may carry vastly different amounts of usable information, 2) the same information may have tremendously different value for Service A than for Service B, and 3) even if the per usage value of two services is the same, Service A may use the data 1,000 times more intensely than Service B leading to extremely different produced benefits. Some authors compared data to labor, too [6]. However, unlike labor, data is non-rivalrous meaning that its supply is not affected by its consumption, and thus selling data for a Service A does not prevent a provider from selling (a copy of) the same data for a Service B.

The research community at the intersection between computer science and economics has studied several aspects of data pricing. Still its elusive nature, and the complex business models under which it is made available makes it very hard to prescribe a price for data. Ultimately it is the market that decides and sets prices via complex mechanisms and feedback loops that are hard to capture. Despite some other works trying to measure the price of personal data of individuals [12], [43], [51], there is no systematic measurement study about the price of data products traded in commercial data marketplaces.

Our Contributions: In this paper we present what is, to the best of our knowledge, the first systematic measurement study of marketplaces for B2B data products. This ecosystem, despite being quite vibrant commercially, remains completely unknown to the scientific community. Very basic questions such as "*What is the range of prices of data traded in modern DMs?*", "*Which categories and types of data products command the highest prices?*", "*Which are the features, if any, that correlate with the most expensive data products?*" appear to have no answer and evade most meaningful speculations.

Our research has been supported by MLEDGE project (REGAGE22e00052829516), funded by the Ministry of Economic Affairs and Digital Transformation and the European Union-NextGenerationEU/PRTR, and by the European Union's HORIZON project DataBri-X (101070069).

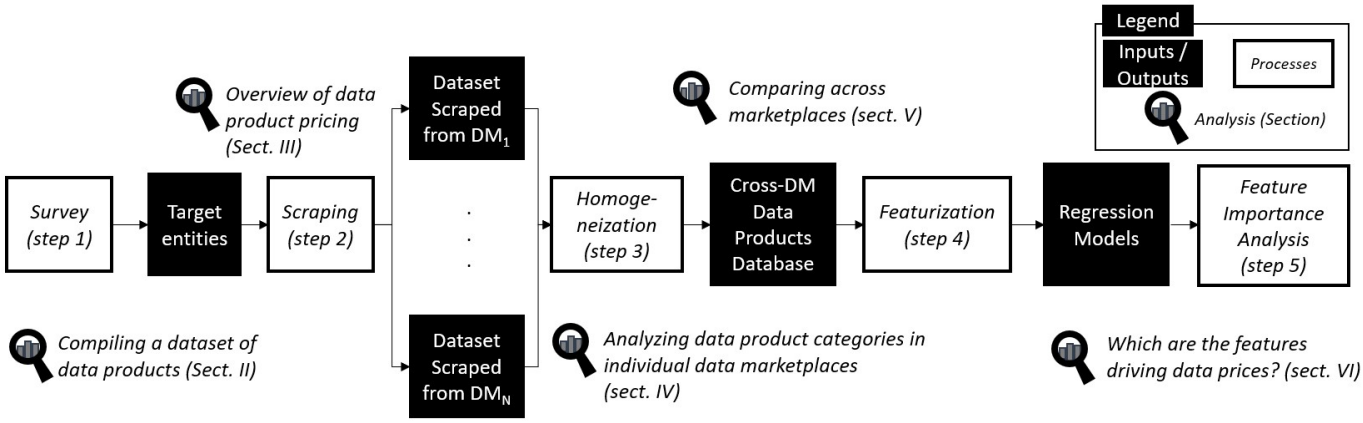


Fig. 1: Summary of our methodology

To answer such questions we followed the methodology summarized in Fig. 1. First, we checked existing surveys for compiling a list of data marketplaces [4], [57], [60]–[62]. We then selected 10 of them that fulfill necessary criteria for a measurement study. For these ones we developed custom crawlers for retrieving information about the products they trade. Using these crawlers, and adding the portfolio of another 30 data providers, we obtained information for more than 210,000 data products and a catalog of more than 2,100 distinct sellers[1]. We also developed data product category classifiers, meaning ML models for identifying products of similar categories across marketplaces, and executed 9 different regression models to understand which features are actually driving their prices.

Our Findings: Analyzing the collected data we observed that the majority of data products were either given for free, or did not carry a fixed price, but rather were up for direct negotiation between the seller and interested buyers. Focusing on the ones that carried a price, some 4,200 of them, we observed that:

• Prices vary in a wide range from few, to several hundreds of thousands of US dollars. The median price for data products sold under a *subscription* model is US$1,400 per month, and US$2,200 for those sold as an *one-off* purchase.

• Using classifiers, we enriched our sample by consistently labeling products according to AWS's categories.

• We found that those related to *telecoms*, *manufacturing*, *automotive* and *gaming* command the highest median prices, and that the most expensive ones relate to *retail and marketing*.

• Using regression models, we managed to fit the prices of commercial products from their features with R^2 above 0.84.

• Due to the heterogeneity of the sample there is no single feature that drives the prices, but instead we spotted meaningful features that drive the prices of specific categories of data. For example, data update rate is a key price driver for *financial* and *healthcare*-related products, whereas geo-spatial localization and the possibility of connecting data points from the same owner are for *marketing* data.

[1]Please, find datasets generated during our research, and code to reproduce our experiments at https://gitlab.com/sandresazcoitia1/data-pricing-tool.

• Overall our models use features related to the category and description of the different data products (i.e., 'Financial', 'Retail', 'stock', 'contact', 'list', etc.), features related to the data products volume and units, as well as singular characteristics extracted from the data products description (i.e., words like 'custom', 'accuracy', 'quality', etc.) to forecast the data product price. Features related to '*what*' and '*how much*' data a product contains are driving 66% of its price.

Like in all measurement studies of Internet-scale phenomena, we will refrain from claiming that any of our findings are "typical" or "representative". What we do claim, however, is that to the best of our knowledge, our measurement study is the first one that attempts to characterize the DM sector, and our above mentioned quantitative results were previously totally unknown. Also, as it will become evident from our methodology later, and to the best of our knowledge, we collected all publicly available pricing information that was accessible during the time of our study.

The remainder of the paper is structured as Fig. 1 shows. First, we frame the scope of our analysis and show some initial outcomes of our measurement study in Sect. II. In Sect. III, we present an analysis on data product pricing in commercial marketplaces. Furthermore, Sect. IV dives deeper into analyzing AWS' DM and DataRade, which account for the largest number of price references in our sample. We then develop tools for enriching our sample and we compare across DMs in Sect. V. Finally, in Sect. VI, we apply several methodologies for analyzing the importance of different metadata features in determining the price of commercial data products.

II. Compiling a dataset of data products

Existing works and surveys on commercial data marketplaces [4], [57], [60]–[62], an extensive web search and a consultation with experts in the area allowed us to compile a list of data marketplaces and understand the different business models they use to compete in this ecosystem. From our analysis, we identified a subset of DMs that fulfilled the criteria for using them as sources of data for a reproducible measurement study. Such criteria include that they grant access to their product catalog without requiring an account, or

through an account but without a vetting process or upfront paid registration, that they have a reasonably large catalog that includes sufficient descriptions of their data products, and that they include a clear description of their pricing policy. Out of the 180 initial DMs, only 10 companies fulfilled all of the above criteria. Most of them did not make it to the list simply because they do not allow non-paying users to browse their catalogs. For example, marketing-related private marketplaces such as *Liveramp*, *LOTAME* or *TheTradeDesk* neither provide public per-product information nor any price references. However, they do provide information about their data partners. By analyzing this information, we did find that 45% of providers in those private marketplaces sell through general-purpose public ones, such as *AWS* or *DataRade*, as well, and hence we have included their products in this study. We also discarded several otherwise *scrapable* general-purpose DMs such as *Data Intelligence Hub* (DIH), *Google Cloud DM* because they included only free data products. We chose to scrape the largest of these free open data marketplaces, *Advaneo*, to help in training our data product category classifiers.

TABLE I: Summary of scraped DMs

Marketplace	#Products	#Paid prod.	#Sellers
Advaneo	198,743	1	N/A
AWS	4,263	2,674	262
DataRade	1,592	1,592	1,262
Snowflake	889	889	200
Knoema	158	158	142
DAWEX	160	160	79
Carto	8,182	5,283	42
Crunchbase	9	9	15
Veracity	115	95	38
Refinitiv	187	187	76
Other providers	777	775	30

Table I lists the 10 DMs that we use as data sources in our study. Overall, we include 6 general-purpose and 4 niche DMs, as well as 30 data providers[2] that, in addition to commercializing their own 777 data products through DMs, provide valuable pricing information on their own websites.

We developed our own web crawler to render and download web pages, and specialized parsers for extracting metadata. We followed common crawling good practices [31]. For example, we avoided visiting several times the same product page in each scraping round and we set up a random wait time from 1 to 2 minutes after requesting a web page in order to avoid flooding the target servers with requests.

We collected information related to 215,075 products from 2,115 distinct sellers in total. We noticed the huge market fragmentation with lots of data providers working with a large number of marketplace platforms. This is natural in a cross-industry nascent market, though hard for data providers to manage. In fact, most data providers (81%) work with only one DM in addition to selling their products through their own web site. 45% of providers in niche financial and marketing-related marketplaces sell through general-purpose DMs, such as *AWS* or *DataRade*, as well. We also spotted DMs advertising and offering their products in other DMs (e.g., *Battlefin* or *CARTO* through *AWS*). Finally, small and niche providers (58% of them) are focusing on one product only.

We scraped all available metadata for data products such as the product id, title, description, source, seller and, when available, its geographic scope, volume, category, use cases, update rate, historic time span, format, etc. We searched for and eliminated duplicates from a single seller within the same DM. We paid special attention to information related to pricing and actual prices of data products.

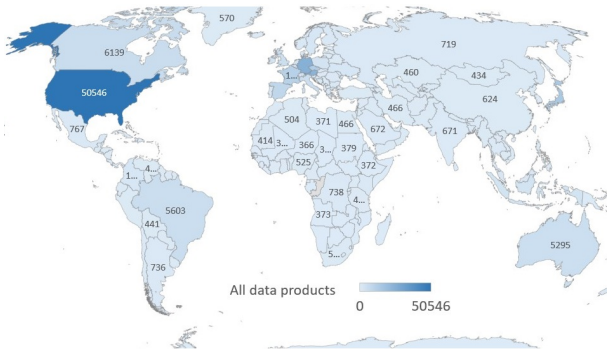


Fig. 2: Data products by country

Regarding the geographical scope of data products, we found that DMs aggregate information from different countries. 14,472 (7%) of the products did not inform about their scope, and 1,177 (around 10% out of the 11,823 paid products) claimed to be global. Figure 2 shows the number of data products covering each country. Regarding the number of *paid* data products, US leads this ranking: around 30% of paid products cover this country. Canada (9.3%), UK (9.2%), Germany (7.6%), France (7.4%), and Spain (7.1%) follow the US in the ranking of countries by number of *paid* products.

III. Overview of Data Product Pricing

It may appear initially surprising that, despite being commercial entities in the B2B space, most of the surveyed and some of the scraped DMs offer predominately free (most of the time open) data. Again we point to the fact that these are privately held companies [2], [21] and not open data NGOs or government initiatives. Our conjecture is that since DMs are two-sided platforms, pre-populating them with free data is a very reasonable bootstrapping strategy, since it can attract the initial "buyers", which in turn will attract commercial sellers and thus help the marketplace grow its revenue.

Next, we focus on the 11,823 paid data products, for which we managed to extract information about their pricing, and whose price is higher than zero. Despite being few compared to the free ones, this sample provides valuable insights about the current status of commercial DMs, as well as to where this segment of the economy is heading to, and how.

[2]42matters, Airbtics, Apptopia, Benzinga, Bizprospex, BoldData, BookYourData, bronID, BuiltWith, DataScouts, Demografy, ebCard, Enigma, ESGAnalytics, HGXN, IFDAQ, ipinfo.io, MultimediaLists, MyDex, OikoLab Weather, Onclusive. Open Corporate, PanXchange, Pipecandy, Shutterstock, Storm Glass, TelephoneListsBiz, Unwrangle, USASalesLeads, and Walklists.

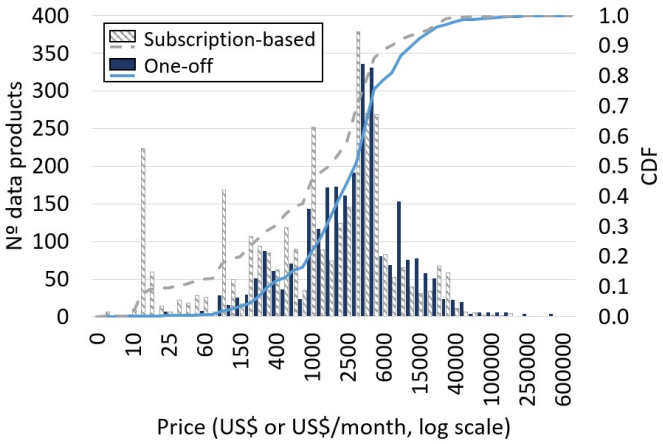


Fig. 3: Histogram and CDF of data products

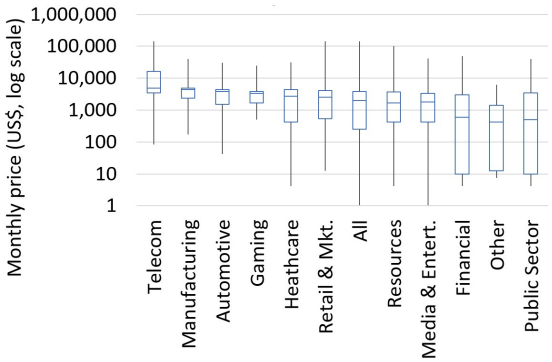
There is a great magnitude of pricing schemes for data products, such as seller-led, buyer-led (bidding), revenue-sharing, tiered-pricing, subject to negotiation, usage-based, etc [4], [44], [53]. Predominant among the 11,823 non-free data products are the *subscription-based* model (i.e., buyer paying for a subscription to get access to data for a period of time), and the *one-off* model (i.e., lump sum payment for data), seller-led in both cases. The first one is used mostly for "live" data usually accessed via an API (e.g., IoT sensor data), whereas the second is used for more static data, which are usually downloaded as one or more files.

4,162 products from 443 distinct providers provided clear information about their prices. Figure 3 shows a histogram and the corresponding CDF of monthly prices for data products. Regarding those offered under a *subscription model*, we see prices across a wide range up to US$150,000 per month. Cheap products below US$100 per month are often curated and cleaner versions of open data. For example, a seller offers a historical compilation of quarterly reports submitted to the US Securities and Exchange Commission (SEC), also downloadable from their websites. They also include low-cost "promotion samples" of more expensive products from well-known sellers, such as GIS data and supporting metadata for a small area of some US cities. The median price is US$1,417 per month. Almost one-third of all products, including targeted market data for example, are sold for US$2-5k monthly.

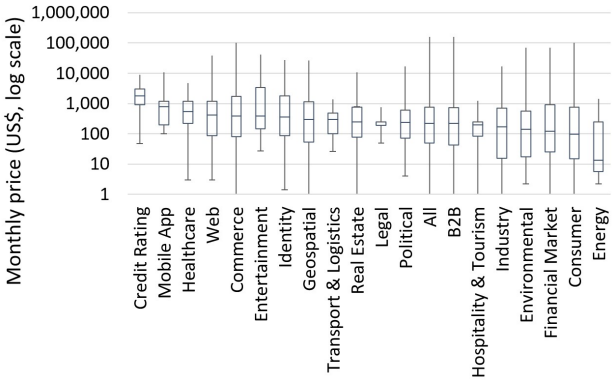
Comparing to products sold under a *one-off model*, (1) the latter tend to be more expensive: median price US$2,176 vs. US$1,417 per month for *subscription-based* products; maximum price US$500,000, more than 3 times higher than the maximum in *subscription-based* access, and (2) *one-off* products have a price histogram more normally distributed around its median at US$2,176. Within the heterogeneous set of products within the US$1,000-4,000 interval, we found a large group of voluminous targeted contact data. Interestingly, we observe a long tail of valuable products in Fig. 3.

IV. Analyzing Data Product Categories in Individual Data Marketplaces

To get a more in-depth understanding of data pricing, we analyzed the catalog of AWS' DM and DataRade, the ones with the largest base of paid products with prices. The former tags data products in 10 different categories, whereas the latter allows data products to be positioned in a hierarchy with more than 300 categories and more than one (out of 150) use cases. Specifically, a product can belong to none, one, or several categories. For instance, credit card transaction data products are classified both as '*Financial*' and '*Retail, Location and Marketing*', whereas those related to weather are not labeled in AWS. We have marked such unclassified products as '*Other*' in our sample.



(a) Subscription prices by industry in AWS.



(b) Subscription prices by category in DataRade.

Fig. 4: Monthly price by data category in AWS and DataRade.

Figure 4 shows box plots of products by first level category in AWS and DataRade. '*Telecom*', '*Manufacturing*' and '*Automotive*' categories exhibit a median price significantly above the global ($\times 2.6$, $\times 2.3$ and $\times 2$, respectively) in AWS, whereas '*Credit Rating*', '*Mobile App*' and '*Healthcare*' data show a higher median price ($\times 8.3$, $\times 3.7$ and $\times 2.5$ above the overall median, respectively) in DataRade. In both cases, the most expensive products are related to marketing ('*B2B*', '*Consumer*' and '*Commerce*' in DataRade).

V. Comparing across marketplaces

Comparing information about data products from different marketplaces is not a straightforward task since i) they provide metadata of different granularity and level of detail, and ii) they use different categorization to describe their products. To overcome these challenges, we developed a methodology to homogenize the categorization of data collected in order to be able to compare similar products across marketplaces.

A. Dealing with different levels of detail

Some marketplaces provide more information than others about their offers. To sort this out, we built a common cross-DM database utilizing a superset of all the different description fields found in different data marketplaces. Apart from their category and text descriptive fields, data product records include the time scope, the volume and units, any potential limitations (e.g., maximum number of users), add-ons, granularity of the information, geo-scope at country level, data delivery methods, update frequency and data format.

We normalized and stored in this cross-DM database all the information from the scraped datasets. We managed to fully automate the extraction of most of the fields (18 out of 27), which were directly scraped from the web pages of the different DMs. This extraction was semi-automated for 5 fields, meaning that they were automatically extracted for certain marketplaces, or retrieved from product descriptions for others, in a process that required a manual check afterwards. For example, *update rate* of data is usually included in the general description of a data product, but the presence of the word '*monthly*' may not necessarily point to a monthly update rate. Information about data volume or data subject units was automatically extracted only for DataRade and BookYourData, and required computer-aided manual typing in the rest of the DMs (we highlight and extract numbers and their context from data descriptions). Manual checks were performed by three different experts. Any ambiguities and disagreements were resolved by majority voting.

B. Dealing with different categorization systems

In Sect. IV we showed that every marketplace has its own way to classify data. Furthermore, boundaries between tags are often blurry, and the criteria followed by different DMs to label a data product with a certain category tag are not necessarily coherent. For example, only certain marketplaces mark '*credit card transaction*' data products as '*financial*', whereas all DMs label them as related to '*marketing*'. Thus, even if we find apparently comparable categories across different marketplaces, we may miss relevant data products due to inconsistencies in their categorization processes.

We addressed this issue by developing a series of natural language processing naïve Bayes (NB) classifiers [20], [22], [39]. In our first attempt, we wanted to identify similar data products – those that belong to the same category – between two different (source and destination) DMs. As a result, we trained both multinomial and complement versions of NB classifiers to detect data products from the source DM that belong in a certain category by using feature vectors based on the information provided by the data product description from the source DM. We used bag of words [36] and data preprocessing steps such as removing stop words and words with numbers, using stemming and TF-IDF transformation [47], [56]. Then we validated the resulting classifier against a manually labeled sample from the destination DM. Manual labeling was performed by three different experts. Any ambiguities and disagreements were resolved again by majority voting.

We utilized the above methodology to build different classifiers to help us compare data products between the two DMs including more price references, namely DataRade (destination DM) and AWS (source DM). We generated our feature vectors based on AWS data product descriptions (source DM) and applied the resulting classifiers to DataRade data products (destination DM). We were interested in finding out: (1) what percentage of products from those categories could we identify in DataRade, (2) whether categorization and pricing were coherent between them, and (3) whether we could enrich our metadata by adding AWS's inferred categories to all products.

We utilized our cross-DM database to generate the train/test datasets at 80/20 split in order to train and test the corresponding classifiers. We observed that multinomial classifiers outperformed the complement NB for this task so we proceeded with the former ones. The resulting classifiers yield an acceptable F_1 score above 0.85 (average for 50 executions with different random 80/20 train/test splits). In fact, they identified meaningful and reasonable stems when tagging products related to each category. For example, for the two categories including more data products:

Financial: 'system', 'sec', 'exchang', 'type', 'file', 'form', 'edgar', 'secur', 'act', and 'compani'.

Retail, Location and Marketing: 'locat', 'topic', 'b2b', 'score', 'echo', 'trial', 'compani', 'visit', 'intent', 'consum'.

We then validated the models against a manually labeled sample from DataRade. Manual labeling was performed by three different experts. Any ambiguities and disagreements were resolved again by majority voting. The validation set included 745 manually pre-labeled with both '*Financial*' and '*Retail, Location and Marketing*' tags. The models trained only with data from AWS did not perform so well on the validation set (F_1 scores of 0.73 and 0.43 for '*Financial*' and '*Retail, Location and Marketing*' data). To generalize further our methodology and improve its accuracy, we enriched the train data with information from other DMs. In particular:
(1) The ***Financial*** classifier was trained with 95,208 labeled descriptions of products from 4 different entities (Advaneo, Carto, AWS, and Refinitiv), and 45,298 financial products.
(2) The ***Retail, Location and Marketing*** classifier was trained with 3,828 descriptions from 3 entities (AWS, BookYourData and TelephoneLists), including 1,614 marketing products.

By adding products belonging to the same category from other DMs we observed better balance between precision and recall and an overall improvement of model generalization. We also observed an increase of the F_1 score in the test set. Particularly, adding information from Refinitiv improves the

F_1 score from 0.73 to 0.79. In the case of '*Retail, Location and Marketing*', adding information from specialized marketing DMs (e.g., BookYourData), drastically improves the F_1 score from 0.43 to 0.74. We tested multiple classifiers, with and without stemming, and we found that using word-based instead of stem-based features led in general to more accurate results in both cases (+5% F_1 score). Table II shows the accuracy obtained by both classifiers.

TABLE II: Score of data product classifiers

	Accuracy	**Precision**	**Recall**	**F_1 Score**
Test - Financial	0.93	0.97	0.81	0.88
Test - Retail	0.95	0.96	0.88	0.91
Val. - Financial	0.89	0.72	0.88	0.79
Val. - Retail	0.78	0.81	0.68	0.74

We used them to label data products in DataRade, and we located 619 and 701 '*Financial*' and '*Retail, Location and Marketing*' data products, which represent 39% and 44% of the total sample, respectively. As happened in AWS, not only do those categories contain the largest number of products in DataRade, but the most expensive ones are tagged as '*Retail, Location and Marketing*', as well.

We repeated the process for the rest of the 11 AWS data categories, and this way we managed to enrich our sample by homogeneously labeling products based on their descriptions. The four categories with highest median prices are the same as in AWS, but in a different order. Again, most products belong to '*Financial*' and '*Retail, Location and Marketing*', and the most expensive ones belong to the latter category.

Does this methodology work if we switch source and destination DMs? In order to answer this question, we trained NB classifiers to detect products in AWS related to relevant use cases and categories in DataRade. In this case, DataRade acted as the source DM, i.e., it provided descriptions and tagging information to train the classifiers, whereas AWS' role was the destination DM, whose products we labeled with some of DataRade's tags and driven by the criteria we learned from the source DM. In particular, we focused on products belonging to the '*B2B Marketing*', '*Audience Targeting*' and '*Risk Management*' use cases in DataRade, some 46, 48 and 30 products out of 745 respectively. Since the training set is imbalanced and the number of samples is low, complement NB outperformed multinomial NB in this case. We trained the classifiers and obtained the log-probability of belonging in each category for all the data products in AWS. As a result, at least 16 out of the top 20 data products showing the highest log-probability turned out to be useful for those specific use cases, according to the assessment of three different experts.

VI. Which are the features driving data prices?

So far we have seen an overview of data pricing, looked at the prices of particular categories, developed and applied a methodology to homogeneously label products across marketplaces in our sample. Our final goal is to understand the prices of data in commercial data marketplaces.

For that purpose, we first extract features to train regression models for predicting the prices of real commercial data products. We do not intend to build state-of-the-art price predictors, but rather to understand which features are driving the price of data. Therefore, we conduct feature importance analysis on the resulting regression models and we find out which features have the highest impact on the observed prices for the different data products in our corpus.

A. Building a feature matrix to feed regression models

An additional preprocessing step is needed in order to transform the fields of our cross-DM database into a set of valuable features that can be ingested by ML regression algorithms. This process uses the NLTK [9] and Scikit-learn [52] Python libraries and includes mainly the following steps:

1) Extraction of '*word*' features from the title and the textual description of each data product. We use bag of words [36] and data preprocessing steps such as removing stop words and words with numbers, TF-IDF transformation [56], and stemming [47]. In addition, we have sellers' names removed from the vocabulary, so as to avoid bias introduced by knowing their identity. Finally, we prepare matrices for different vocabulary lengths and optimize each algorithm for this parameter.
2) Breakdown of volume-related fields in 13 different groups depending on their nature. For example, we separate data products targeting 'entities' or 'companies', from those whose subjects are 'individuals' in different features. The resulting comparable units are in turn normalized, and a new overarching feature ('units') measuring the percentage of units covered is added to compare products across groups of units.
3) Calculation of country-level binary features to indicate whether a certain country is covered by a data product.
4) Homogenization of the units of time when measuring the time scope of the products, what we will call *history*.

Before feeding the models, we reduce the number of input features by discarding those that have a unique value, which may appear when filtering the complete dataset by *category*. Next, we unify groups of features showing a high cross-correlation among them, i.e., $R^2 \geq 0.9$.

As a result of this *featurization* process, we reduce each sample product to a feature vector and produce a feature matrix to train our regression models. Table III lists feature groups and some examples of their individual features. We organize features in 10 disjoint sets according to their nature and the basic questions they answer about data products.

We evaluated the linear correlation of individual features with respect to data product prices. Not surprisingly, it turns out that none of them is linearly correlated to price, as opposed to what we found for specific sellers. Our challenge now is measuring which features and groups of features are more significant in determining the price of data products in commercial marketplaces.

TABLE III: List of feature groups

Question	Group	Definition	Nº features	Example of features
What?	**Category**	Labels attached to the product that define the type of data it contains	11	'Weather', 'Gaming', 'Financial'
	Description	Stem-like features obtained from data product descriptions	up to 2000	'wordmarket', 'wordidentifi', 'wordlist'
	Identifiability	Tells whether the product allows the buyer to recognize the activity of individuals or to identify specific companies	2	'idSessions', 'IdCompanies'
How much?	**Volume**	Normalized nº units covered broken down by the nature of such units	14	'units', 'people', 'entities'
	Update rate	Defines the frequency between data updates as announced by the seller	11	'real time', 'monthly', 'hourly'
How?	**Delivery method**	Defines how the buyer can have access to data	8	'S3Bucket', 'Download', 'FeedAPI'
	Format	Defines the way in which data is arranged	17	'txt', 'shapefile', 'xls'
	Add-ons	Tells whether the product attaches any add-on or has any limitations	2	'ProfServices', 'Limitations'
When?	**History**	Time scope included	1	'History'
Where?	**Geo scope**	Metrics about countries included in the data product	up to 249	'Nº Countries', 'USA', 'Canada'

TABLE IV: Accuracy achieved by regression models

Model	Financial			Marketing			Healthcare			All		
	R^2	MAE	MSE	R^2	MAE	MSE	R^2	MAE	MSE	R^2	MAE	MSE
RF	0.85	0.2	0.14	0.86	0.21	0.13	0.78	0.25	0.15	0.84	0.23	0.16
kN	0.78	0.31	0.26	0.74	0.33	0.24	0.77	0.26	0.17	0.69	0.37	0.31
GB	0.82	0.23	0.16	0.8	0.28	0.19	0.73	0.27	0.19	0.79	0.3	0.22
DNN	0.73	0.33	0.35	0.77	0.30	0.22	0.68	0.26	0.18	0.72	0.33	0.28

TABLE V: Top 10 most relevant features not related to volume by category and regression model

Financial			Marketing			Healthcare		
RF	**kNN**	**GBR**	**RF**	**kNN**	**GBR**	**RF**	**kNN**	**GBR**
S3Bucket	Email	S3Bucket	IdSessions	History	csv	wordhealth	csv	wordlist
wordsubmit	Download	wordmonthli	Download	USA	yearly	wordtrend	daily	Del. Methods
Download	daily	wordstock	REST API	IdSessions	REST API	wordmedic	wordmarket	wordhospit
txt	IdCompanies	worddeliv	wordcustom	Nº Countries	wordqualiti	wordglobal	wordgo	wordidentifi
wordedgar	USA	Del. Methods	USA	Financial	wordaccur	csv	Limitations	wordamerica
wordcustom	wordmarket	txt	yearly	Others	wordidentifi	Del. Methods	location data	wordhealth
wordlist	Retail	wordneed	monthly	wordcontact	wordwebsit	wordinsight	wordpopul	wordreport
wordcontact	wordcontact	wordsubmit	IdCompanies	Email	UI Export	wordreport	wordprofil	wordstudi
wordsystem	real time	wordreport	wordname	UI Export	wordcover	wordregion	wordinsight	wordupdat
wordcompar	wordprice	wordcontact	location data	Download	wordfield	wordlist	Download	wordcontact

B. Analyzing feature importance

Regression models can be used for feature importance analysis. Next we use a range of such techniques to understand which features have the higher impact on data product prices.

1) Optimizing Regression models: Owing to their stochastic nature, training several regression algorithms and comparing their outcomes is key to obtaining robust conclusions. Consequently, we have tested variations of 9 different regressors with different values for their main parameters (e.g., num. of estimators, depth, etc.) as included in the Scikit-learn [52] Python library, and inputs of different vocabulary lengths. Such models work with the log instead of the absolute value of product prices as the dependent variable so as to normalize the distribution of prices and avoid negative price predictions. We were hoping to find at least 3 models that produce sufficiently accurate price predictions, measured as the R^2score of their output w.r.t. actual prices.

To reduce the complexity of each model, we removed low-value features, i.e., those that had a negative leave-one-out (LOO) value, provided the accuracy of the model was not negatively affected. A feature having negative LOO value means that the model improved its average accuracy in 10 random executions for different train and test data splits when such feature was removed from the input matrix. Finally, we performed a cross-validation to check the variance of the accuracy of the model when training and testing in 5-folds, and 20-random training-test splits of the input data.

We found that three target models worked reasonably well (i.e., they yield an R^2 score greater or equal to 0.70), namely Random Forest [10], k-Nearest Neighbours [38], and Gradient Boosting [23], [46] regression models. On the contrary, we discarded linear, Elastic-Net [68], Ridge [32], Bayesian Ridge [45], and Lasso [64] regressions even though they worked well in specific simulations.

In addition, we also tested a Deep Neural Network regressor using the TensorFlow [1] and Keras [34] libraries. We followed all common good practices recommended for such activity by first standardizing the input data. We tested RELU/Leaky RELU activation functions for all hidden layers, and a linear activation function for the output layer. As loss function we used the mean absolute error (MAE). To avoid overfitting we randomly applied Drop-out between training epochs and to avoid dying/exploding neurons we also applied Batch normalization between all layers. We used the Adam optimizer [35] with a tuned learning rate decay to train the

model faster at the beginning and then decrease the learning rate with further epochs to make training more precise. Finally, we used Callbacks to stop the training at the optimal epoch.

Table IV presents a summary of the accuracy obtained by regressor and category of data products, including the R^2 score, the MAE and the mean squared error (MSE) with regards to the actual log prices. For the sake of robustness, our results were consistent across subsequent 5-fold and 20 random train/test splits: R^2 score showed a standard deviation below 4% of the average in each round. Note that due to the total (low) number of observations that we have in our datasets, DNN models are not recommended, nevertheless, we wanted to explore them since we believe that they will further improve our results as soon as we manage to increase the overall size of our datasets. Consequently, we avoided using any DNN model in the feature importance analysis.

2) Analyzing the importance of individual features: We carried out this process for financial, marketing, healthcare and all data products in our sample. Financial and marketing data were the most popular data categories, whereas healthcare data was chosen as a relevant disjoint category of less though increasingly popular products showing a different behavior in terms of prices. As a result, we obtained at least one model that achieves a R^2 score of 0.78 by category and accurately fits the prices of data products (see Tab. IV). We ran two different individual feature importance analysis:

1) measuring the accuracy lost by randomly shuffling the values of a certain feature among samples (permutation importance analysis [63]), and
2) measuring the prediction accuracy lost when one individual feature is removed from the inputs (leave-one-out or LOO value).

We have found that 50% of the positive LOO and 67% of the ΔR^2 score by shuffling values owe to the top 10 most relevant features on average for specific categories of data. Note that we would need more than 25 features to achieve equivalent scores if we include all the products. Whereas features related to units and the volume of data clearly lead the ranking for financial and marketing data products, they are less important for healthcare-related ones.

We cross-validated our results in 5-fold executions of both methods and took averages in order to disregard features that showed to be important only in specific tests. As regards robustness, we compared the top-20 ranking of every individual test to the top-20 average ranking of that algorithm and category. It turns out that both rankings have at least 5 features in common in 95% of the cases, and a median of 13 common individual features.

Table V lists other features not related to data volume in descending order of importance. Next we provide some details about the most important features of each specific category:

Financial: Not only do volume-related features such as '*units*' and '*entities*' rank number one, but they are on average four times more important than the second feature in the ranking. Other features relate to specific characteristics of financial data products and help models identify data products either by their category (e.g., '*Retail*') or their description. For instance, RF relies on the word '*edgar*', which stands for SEC's Electronic Data Gathering, Analysis, and Retrieval System, all algorithms identify business '*contact*' lists, a family of financial products, and they also use '*stock*' and '*market*'. The word '*custom*' helps identify information about *customers*, but also refers to the valuable possibility of personalizing data products (e.g., select which companies we want financial data from). Features related to delivery methods (e.g., '*S3bucket*' or '*Download*') and update rate (e.g., '*real time*' or '*daily*') stand out in terms of relevance, as well.

Marketing: With regards to marketing data products, features related to volume, such as '*units*' and '*entities*' lead the ranking, as well. Again categories (e.g., '*Financial*', '*Others*') and specific words pointing to relevant characteristics of data play a relevant role, too. For example, words like '*contact*' are used to locate contact lists, a family of marketing products, the stems '*qualiti*' and '*accur*' refer to the high-quality and accuracy of data, as advertised by sellers. A number of features, such as the stem '*identifi*', emphasize the value of identification for marketing data. In addition, the presence of '*IdSessions*' and '*IdCompanies*' features indicates that being able to reconstruct sessions of anonymized individuals and being able to identify merchants are price drivers for marketing products. Unlike financial data, the fact that a dataset includes *'location data'* is also used to set prices of marketing data. Finally, the scope of data is important, as suggested by features like '*USA*' and '*Nº Countries*' ranking high in the results of RF and kNN models.

Healthcare: The '*what*' is more important than the '*how much*' when fitting the observed prices of healthcare products. This is due to the heterogeneity of data products belonging in this category, ranging from contact lists of healthcare practitioners and hospitals to data about clinical trials or specific medications. Therefore, stems like '*trial*', '*hospit*' or '*studies*' help in identifying what a dataset is about. The stem '*go*' refers to an official check-in and rating system that was used to limit the spread of COVID in the US. Features related to the update rate, data format ('*csv*'), the number of available delivery options ('*Del. Methods*') and the presence of '*Limitations*' (e.g., limited number of reports, or limited data exports included) determine product prices, too.

3) Analyzing the importance of groups of features: Since LOO is often negligible for individual features, we have repeated this analysis for groups of features answering to the same question regarding the data product (see Tab. III). In this case, we have used the following two methods:

1) Measuring the prediction accuracy lost when a group of features is removed from the input dataset (LOO).
2) Measuring the average (in 20 random train/test split executions) Shapley value of each group of features.

The Shapley value is defined as the average R^2 score added by combining the information of a certain group of features with every possible mix of the rest of groups. This is a well-known and widely-used concept in game theory, economics and ML [24], [59], and it is considered a 'fair' method to

TABLE VI: LOO values by feature group

Group	Financial			Marketing			Healthcare			All		
	RF	kNN	GBR	RF	kNN	GBR	RF	kNN	GBR	RF	kNN	GBR
Description	0.027	0.025	0.066	0.021	0.034	0.098	0.054	0.425	0.052	0.023	-0.020	0.079
Volume	0.092	0.182	0.167	0.171	0.138	0.199	0.048	0.014	0.052	0.138	0.123	0.142
Geo Scope	-0.005	-0.007	-0.001	-0.003	-0.006	0.000	0.015	0.000	-0.011	-0.003	-0.002	0.000
Del. Method	0.005	0.032	0.011	0.000	0.018	0.008	0.019	0.017	0.003	0.002	0.010	0.008
Format	0.002	0.004	0.010	0.007	0.001	0.023	0.007	0.030	0.000	0.002	0.007	0.006
Category	-0.002	0.001	0.001	-0.001	-0.003	0.001	0.013	-0.033	-0.006	0.001	0.000	0.003
Add-ons	-0.001	0.007	-0.001	-0.001	0.000	0.001	0.000	0.022	0.000	0.001	0.001	0.000
Identifiability	-0.002	0.016	0.002	-0.001	0.006	0.004	0.010	0.000	-0.009	0.000	0.008	0.000
History	-0.001	0.000	0.000	-0.003	0.004	0.000	0.009	0.000	0.000	0.002	0.000	-0.001
Update Rate	0.001	0.023	0.001	0.036	0.000	0.016	0.010	0.021	0.000	0.021	-0.002	0.014

TABLE VII: Shapley values by feature group

Group	Financial			Marketing			Healthcare			All		
	RF	kNN	GBR	RF	kNN	GBR	RF	kNN	GBR	RF	kNN	GBR
Description	0.155	0.266	0.222	0.247	0.153	0.152	0.232	0.290	0.236	0.113	0.176	0.187
Volume	0.211	0.216	0.184	0.290	0.241	0.241	0.168	0.125	0.131	0.211	0.210	0.174
Format	0.087	0.006	0.086	0.027	0.046	0.094	0.090	0.077	0.082	0.072	0.087	0.071
History	0.072	0.000	0.059	0.009	0.037	0.036	0.063	0.001	0.046	0.058	0.010	0.037
Update Rate	0.088	0.056	0.084	0.060	0.032	0.050	0.046	0.145	0.041	0.067	0.034	0.067
Del. Method	0.036	0.054	0.044	0.093	0.075	0.049	0.030	0.040	0.035	0.062	0.062	0.074
Identifiability	0.034	0.038	0.028	0.052	0.027	0.048	0.040	0.001	0.031	0.056	0.022	0.039
Geo Scope	0.056	0.046	0.050	0.032	0.044	0.036	0.030	0.001	0.040	0.061	0.015	0.024
Category	0.071	0.021	0.044	0.018	0.043	0.037	0.017	0.031	0.039	0.070	0.063	0.055
Add-ons	0.021	0.003	0.021	0.012	0.028	0.038	0.048	0.053	0.041	0.055	0.026	0.045

distribute the gains obtained by cooperation. In our case, we applied the Shapley value to distribute the gains in accuracy of our regression models among the groups of features that contributed to achieving such an accuracy. Furthermore, we ran 5-fold feature importance analysis in the case of LOO, in a similar way as we did for individual features, and 20 calculations of the Shapley values for random 80/20 train/test splits of our input data.

Whereas LOO measures gains or loses in accuracy of a model when features belonging in a group are removed from the input matrix, Shapley values better capture the complementarity among groups and take into consideration their individual predictive power, as well. Table VI and Table VII list the LOO and the Shapley values by group of features in descending order of importance. The standard deviation of Shapley values across executions is acceptable (average below 0.029 for financial and marketing datasets, 0.057 for healthcare-related data, and below 0.017 for all the data), and the ranking of relevant feature groups remains stable.

Figure 5 plots the percentage of the sum of Shapley and LOO values that each feature group represents, what we call their *predictive power*, and illustrates how important each group is for determining the prices of each category of products. We have piled together and colored in gradients groups responding to the same question about data products.

Note that the algorithms, in the absence of certain features, try to replace or infer them through other features in order to come up with the best estimation possible. We have observed that this happens with '**category**' labels or '**add-ons**', and it is also the reason why LOO values are generally smaller than the corresponding Shapley values.

By looking at Fig. 5, we can confirm that features related to '**volume**' and '**descriptions**' are the most relevant groups driving data prices: at least half of the predictive power owes to those two groups of features according to their Shapley

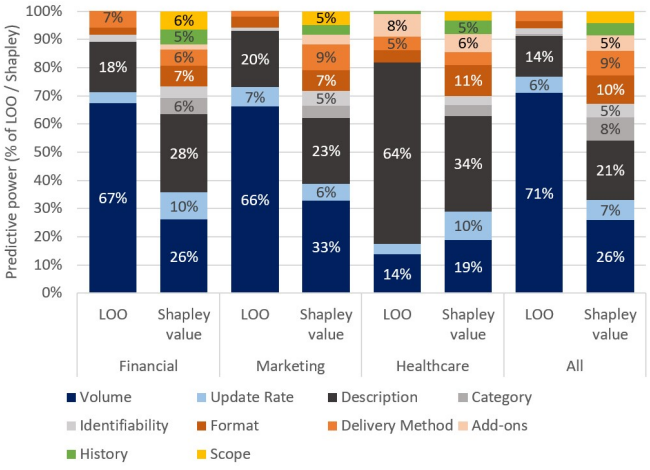


Fig. 5: Predicting power of feature groups

values. While '*volume*' is clearly the most relevant group for marketing data products, it is not so relevant for healthcare-related data due to the heterogeneity of products belonging in this category and due to their lower price-sensitivity to volume.

Data '**update rate**' and its '**format**' are consistently relevant across all data categories, but to a lesser extent (6-11% of the prediction score), whereas the Shapley values of the other groups differ across categories: '**history**' (meaning the time span of data delivered) is more relevant for financial and healthcare-related data, '**delivery methods**' are more relevant for marketing data, and ´**identifiability**' is important in general, but especially for marketing products. These results are in line with our discussion based on the relevance of individual features in the previous section.

In summary, it is mostly '*what*', as captured in product description and categories, and '*how much*' data is being traded that determine the price of a product. Since relevant descriptive features are diverse and strongly differ across

data categories, we failed to find a single feature other than '*units*' that, with some aforementioned exceptions, consistently shows a significant *predictive power*. However, we did find interesting features driving the prices of specific categories of data, such as update rate for financial products, and the ability to provide exact locations and those related to identifiability for marketing data. '*How*' data is delivered to buyers proved to be important too, and accounts for 15-24% of *predictive power* according to Shapley. Finally, historical time span ('*when*') and geographical scope ('*where*') of data products, whose score oscillates around 5% for every data category, are less relevant in driving their prices.

VII. Related works

Even though several surveys related to data marketplaces have been recently published [4], [57], [60]–[62], our work is, to the best of our knowledge, the first empirical measurement study that deals with the prices of data products sold in commercial data marketplaces.

In fact, the lack of empirical data around dataset prices is considered as a key challenge in data pricing research [53]. According to some authors, some techniques to set the prices of digital products [58] or cloud services [66] are applicable to data products, as well. Some authors proposed auction designs to set the prices of digital goods and data products [26], [27]. Novel AI/ML data marketplace architectures have been proposed under the concept of value-based pricing [3], [16], [49] and the value of privacy [48]. Moreover, some authors defined pricing strategies and marketplaces based on differential privacy [25], [40] or queries to a database [15], [37]. All of them work on analyzing the theoretical properties for fair, arbitrage-free pricing, but leave the responsibility of actually defining absolute prices to both buyers and sellers. Quality-based pricing [29] is the one closest to our approach. According to it, the value of data must be assessed by evaluating and assigning weights to certain quality features. Even though some additional works have provided data pricing strategies for sellers based on this idea [67], we are not aware of any measurement study that has been able to derive weights for such features from real market data.

The pricing of personal data of individuals has received attention from the privacy and measurement communities. There are measurement studies based on prices carried over the Real Time Bidding protocol [43], [51] as well as more traditional survey-based studies [12]. These works report prices for the data and the attention of individuals and, therefore, have nothing to do with B2B datasets traded in modern DMs.

Cross-marketplace analysis and discoverability of data has been pointed out as a significant challenge by data marketplace vision papers [54]. Google Dataset Search has proposed a standard for providing metadata for their crawlers [11]. Discoverability is the *leit motiv* of DMs and data aggregators, such as DataRade, but do not touch upon pricing questions.

Finally, part of this work explaining the challenges in scraping and comparing across data marketplaces and outlining the design of a data quotation tool was published as a workshop paper [5]. This paper is adding substantially new material, such as the procedures we used to populate a cross-DM database, the development and test of classifiers to compare across DMs (see Sect. V), and the training of regression models to fit data prices and carry out feature importance analysis (see Sect. VI).

VIII. Conclusions and future work

Our work has provided a first glimpse into the growing market for B2B data. Despite having worked in a range of pricing topics in the past, prior to conducting this study, we did not have the slightest idea even for fundamental questions such as "What are typical prices for data products sold online?", or "What types of data command higher prices?". Our work has produced answers to those and many other questions. We have seen that while the median price for data is few thousands, there exist data products that sell for hundreds of thousands of dollars. We have also looked at the categories of data and the specific per-category features that have the highest impact on prices. Having scraped metadata for hundreds of thousands of data products listed by 10 real-world data marketplaces and other 30 data providers we found fewer than ten thousand that were non-free and included prices. We believe that this is due to prices being often left to direct negotiation between buyers and sellers, and also because most marketplaces use free data to bootstrap their marketplace and attract the first "buyers" and then commercial sellers.

Moreover, the paper represents a first step towards developing a price recommendation tool for new data products [5], and has even provided a first implementation of some of its key components, namely i) the metadata and taxonomy required to describe data products, ii) crawlers and parsers to automate the collection of such information from key leading marketplaces, iii) classifiers to compare across them, and iv) regression models to understand which are the most relevant features driving product prices. The significant monthly growth rate we have seen at AWS and other marketplaces makes us believe that in the future the paid catalog of data marketplaces is bound to grow and therefore, we will continue monitoring them to see how they evolve.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Tal- war, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale Machine Learning on Heterogeneous Systems, 2015. Software available from tensorflow.org.

[2] Advaneo. Access to the world of data. https://www.advaneo-datamarketplace.de/. Last accessed: Oct'22

[3] A. Agarwal, M. Dahleh, and T. Sarkar. A Marketplace for Data: An Algorithmic Solution. In Proc. of ACM EC, 2019.

[4] S. Andrés Azcoitia and N. Laoutaris. A Survey of Data Marketplaces and their Business Models. SIGMOD Record, 2022.

[5] S. Andrés Azcoitia, C. Iordanou, N. Laoutaris, "Measuring the Price of Data in Commercial Data Marketplaces," ACM Data Economy Workshop, 2022.

[6] I. Arrieta-Ibarra, L. Goff, D. Jiménez-Hernández, J. Lanier, and E. G. Weyl. Should we Treat Data as Labor? Moving Beyond "Free". AEA Papers and Proceedings, 108:38–42, 2018

[7] AWS. Amazon Web Services Marketplace. https://aws.amazon.com/marketplace. Last accessed: Oct'22

[8] Battlefin. Better your investments using alternative data. https://www.battlefin.com/. Last accessed: Oct'22.

[9] E. L. Bird, Steven and E. Klein. Natural Language Processing with Python. O'Reilly Media Inc, 2009.

[10] Breiman. Random forests. Mach. Learn., 45(1):5–32, Oct. 2001.

[11] D. Brickley, M. Burgess, and N. Noy. Google Dataset Search: Building a Search Engine for Datasets in an Open Web Ecosystem. In Proc. of ACM WWW conf., 2019.

[12] J. P. Carrascal, C. Riederer, V. Erramilli, M. Cherubini, and R. de Oliveira. Your Browsing Behavior for a Big Mac: Economics of Personal Information Online. In Proc. of ACM WWW Conf., 2013

[13] Caruso. Your solution. one platform. multibrand in-vehicle data. https://www.caruso-dataplace.com/. Last accessed: Oct'22.

[14] G. Cattaneo, G. Micheletti, and al. The European Data Market Monitoring Tool. Key Facts and Figures, First Policy Conclusions, Data Landscape and Quantified Stories. Final Study Report. European Commission, 2020.

[15] S. Chawla, S. Deep, P. Koutris, and Y. Teng. Revenue Maximization for Query Pricing. Proc. of the VLDB Endow., 13, 2019.

[16] L. Chen, P. Koutris, and A. Kumar. Towards Model-Based Pricing for Machine Learning in a Data Marketplace. In Proceeding of ACM SIGMOD, 2019.

[17] Clive Humby. Data is the New Oil! Keynote at ANA Senior Marketer's Summit, Kellogg School, 2006.

[18] DataRade. Datarade. choose the right data with confidence. https://datarade.ai/. Last accessed: Oct'22.

[19] Dawex. DAWEX Data Exchange, unleash the value of your data. https://www.dawex.com/. Last accessed: Oct'22.

[20] L. Denoyer and P. Gallinari. Bayesian Network Model for Semi-structured Document Classification. Inf. Process. Manage., 40(5), 2004.

[21] DIH. Data intelligence hub. extract value from data securely. https://dih.telekom.net/. Last accessed: Oct'22.

[22] P. Domingos and M. Pazzani. On the optimality of the Simple Bayesian Classifier under Zero-one Loss. Mach. Learn., 1997

[23] J. Friedman. Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics, 29, 2000

[24] A. Ghorbani and J. Zou. Data shapley: Equitable Valuation of Data for Machine Learning. Proc. of the ICML, 2019.

[25] A. Ghosh and A. Roth. Selling privacy at auction. In Proc. of the ACM EC '11, 2011.

[26] A. V. Goldberg and J. D. Hartline. Competitiveness via Consensus. In Proc. of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '03, page 215–222, USA, 2003. Society for Industrial and Applied Mathematics.

[27] A. V. Goldberg, J. D. Hartline, and A. Wright. Competitive Auctions and Digital Goods. In Proc. of the ACM-SIAM Symposium on Discrete Algorithms. Society for Industrial and Applied Mathematics, 2001.

[28] Harvard. Dataverse. https://dataverse.harvard.edu/. Accessed: Oct'22.

[29] J. R. Heckman, E. Boehmer, E. H. Peters, M. Davaloo, and N. G. Kurup. A Pricing Model for Data Markets. In Proc. iConference 2015.

[30] N. Henke, J. Bughin, and al. The age of analytics: Competing in a Data-driven World. McKinsey Global Institute, 2016.

[31] M. Hils, D. W. Woods, and R. Böhme. Measuring the Emergence of Consent Management on the Web. In Proc. of the ACM IMC'20, page 317–332. 2020.

[32] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased Estimation for Nonorthogonal Problems. Technometrics, 42(1):80–86, Feb. 2000.

[33] Kaggle. Datasets. https://www.kaggle.com/datasets. Accessed: Oct'22.

[34] Keras. Simple. flexible. powerful. https://keras.io/. Accessed: Jun '22.

[35] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In Y. Bengio and Y. LeCun, editors, Proc of ICLR '15, 2015.

[36] Y. Ko. A Study of Term Weighting Schemes using Class Information for Text Classification. In Proc. of ACM SIGIR 2012.

[37] P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, and D. Suciu. Querymarket Demonstration: Pricing for Online Data Markets. Proc. of the VLDB Endow., 5, 2012.

[38] O. Kramer. Unsupervised k-Nearest Neighbor regression. 2011.

[39] G. Krishnaveni and T. Sudha. Naïve Bayes Text Classification - a Comparison of Event Models. Imperial Journal of Interdisciplinary Research, 3, 2016.

[40] C. Li, D. Y. Li, G. Miklau, and D. Suciu. A theory of pricing private data. ACM Transactions on Database Systems 39(4), 2015.

[41] LiveRamp. Data marketplace. https://liveramp.com/our-platform/data-marketplace/. Last accessed: Oct'22.

[42] LOTAME. Private data exchange (pdx). trusted data relationships made easy. https://www.lotame.com/pdx/. Last accessed: Oct'22.

[43] C. C. Lukasz Olejnik, Minh-Dung Tran. Selling off privacy at auction. In Proc. of the NDSS Symposium, 2014.

[44] A. Löser, F. Stahl, A. Muschalle, and G. Vossen. Pricing Approaches for Data Markets. In Proc. of the International Workshop on Business Intelligence for the Real-Time Enterprise, 2012.

[45] D. J. C. MacKay. Bayesian interpolation. Neural Comput., 4(3), 1992.

[46] L. Mason, J. Baxter, P. Bartlett, and M. Frean. Boosting Algorithms as Gradient Descent. In Proc. of the International Conference on Neural Information Processing Systems, 1999.

[47] S. Matic, C. Iordanou, G. Smaragdakis, and N. Laoutaris. Identifying Sensitive Urls at Web-scale. In Proceedings of the ACM IMC, 2020.

[48] C. Niu, Z. Zheng, F. Wu, S. Tang, X. Gao, and G. Chen. Unlocking the Value of Privacy: Trading Aggregate Statistics over Private Correlated Data. In Proc. of ACM SIGKDD, 2018.

[49] O. Ohrimenko, S. Tople, and S. Tschiatschek. Collaborative Machine Learning Markets with Data-replication-robust Payments. CoRR, 2019.

[50] Otonomo. One-stop shop for vehicle data. https://otonomo.io/. Last accessed: Oct'22.

[51] P. Papadopoulos, N. Kourtellis, P. R. Rodriguez, and N. Laoutaris. If you are not paying for it, you are the product: How much do advertisers pay to reach you? In Proc. of the ACM IMC, 2017.

[52] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. Journal of Machine Learning Research.

[53] J. Pei. Data Pricing – from Economics to Data Science. In Proc. of the ACM SIGKDD, page 3553–3554, 2020.

[54] M. F. Raul Castro Fernandez, Pranav Subramaniam. Data Market Platforms: Trading Data Assets to Solve Data Problems. In Proc. of the VLDB Endow., 2020.

[55] Refinitiv. Data catalog. our data, your way. https://www.refinitiv.com/en/financial-data. Last accessed: Oct'22.

[56] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. Info. Processing and Management, 1988.

[57] F. Schomm, F. Stahl, and G. Vossen. Marketplaces for data: An initial survey. ACM SIGMOD Record, 2013.

[58] C. Shapiro and H. R. Varian. Information Rules: A Strategic Guide to the Network Economy. Harvard Business School Press, 2000.

[59] L. S. Shapley. A Value for n-Person Games. RAND Corporation, 1952.

[60] M. Spiekermann. Data marketplaces: Trends and monetisation of data goods. Intereconomics, 2019.

[61] F. Stahl, F. Schomm, L. Vomfell, and G. Vossen. Marketplaces for digital data: Quo vadis? Computer and Information Science, 10, 2017.

[62] F. Stahl, F. Schomm, and G. Vossen. The Data Marketplace Survey Revisited. Westf. Wilhelms-Univ., ERCIS, 2014.

[63] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis. Conditional Variable Importance for Random Forests. BMC Bioinformatics, 2008.

[64] R. Tibshirani. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society (Series B), 1996.

[65] Veracity. Veracity by DNV GL. Find the Right Tools for your Industry Needs. https://store.veracity.com/. Last accessed: Oct'22.

[66] C. Wu, R. Buyya, and K. Ramamohanarao. Cloud Pricing Models: Taxonomy, Survey, and Interdisciplinary Challenges. ACM Computing Surveys, 2019.

[67] H. Yu and M. Zhang. Data Pricing Strategy based on Data Quality. Computers and Industrial Engineering, 112:1–10, 2017.

[68] H. Zou and T. Hastie. Regularization and Variable Selection via the Elastic Net. Journal of the Royal Statistical Society. Series B. (Statistical Methodology), 2005.

Anexo 8: Try Before You Buy

MLEDGE Report: Try-Before-You-Buy reloaded and federated

Alexandr Goultiaev Tolstokorov
alexandr.goultiaev@imdea.org
IMDEA Networks Institute
Leganés, Madrid, Spain

Santiago Andrés Azcoitia
santiago.azcoitia@imdea.org
IMDEA Networks Institute
Leganés, Madrid, Spain

Nikolaos Laoutaris
nikolaous.laoutaris@imdea.org
IMDEA Networks Institute
Leganés, Madrid, Spain

1 Motivation

As data is increasingly used to drive decision-making processes, companies need obtaining suitable data and insights from third parties to improve the accuracy and efficiency of their models. A number of commercial data marketplaces (DMs) have appeared in the market to mediate between providers and consumers and to manage data sharing and transactions between them [1].

The operation of a data marketplace entails complex technical and economic challenges due to the elusive nature of data as an economic good. In particular, the value of data is strongly dependent on the use case, and two similar datasets could be very valuable for a ML task A and not so valuable for a ML task B, and data from different providers may yield very different values in training a single ML model, irrespective of the amount of information they carry [2]. In summary, buyers are not able to realise the value of a piece of data for them until they are able to test the data on their own model.

A number of solutions have been proposed to circumvent this problem, known as Arrow's information paradox. On the one hand, commercial data marketplaces are addressing this by sharing outdated data samples, by providing metadata to potential buyers, and most interestingly by allowing potential data buyers try versions or samples of a dataset in "sandboxed" environments that do not allow buyers to download data. On the other hand, most DM proposals from the research community have responded to this challenge by allowing buyers to share their ML models with DMs and letting the platform provide data according to the price paid by users [3, 4], or find the combination of datasets [5, 6] that best suits the ML task. These solutions, which have not been implemented yet, share a common drawback: buyers are supposed to trust the platform and share a most likely sensitive and confidential ML model. Such ML models are often complex, which makes training them with different combinations of eligible datasets to select the most suitable ones computationally expensive. Finally, none of these solutions evaluate the cost of processing a data transaction, which can be prohibitive and hence can threaten the viability of the DM.

Our objective is to provide more scalable and explainable alternatives to these proposals that avoid buyers sharing confidential intellectual property with DMs, and make the data purchasing process more scalable. This solution will also be useful in federated environments in the cloud edge, where data will be distributed in different edge nodes, and federated clients in those nodes will be in charge of evaluating data assets and returning the results to a centralized control node that will be accessed by data buyers.

***Our planned contributions*:** This paper will introduce a practical architecture of a data marketplace that evaluates and charges for data and for processing transactions. Moreover, it introduces the idea of valuation functions (VFs) and puppet valuation models (PVMs) to reduce the processing costs, and shows that data buyers are able to find suitable data tailored to their use case without necessarily sharing information about their (often) complex ML tasks. Then we plan to test different PVMs and VFs for different use cases related to image classification problems and a forecasting taxi-ride demand in different districts of Chicago [7] use case, and we plan to investigate the efficiency vs. accuracy trade-off and evaluate the cost of processing and selling such data.

2 Problem Setting

Let us assume that the buyer is seeking to buy data in order to optimise a certain AI/ML model ($\mathcal{M}$), i.e., to maximise the value of a metric $a(\cdot)$ which we will call *accuracy* and can accommodate any model performance metrics such as F1 score, precision, mean absolute error, etc. We will assume the buyer has also a limited budget B to pay for the cost of such transaction.

Let us denote by $\mathcal{S}$ the set of sellers able to provide suitable data for the task $(\mathcal{M}, a)$ at a given DM. We will denote by $\mathcal{C}$ the catalogue (set of possible data inputs) for $(\mathcal{M}, a)$ based on data from those sellers. We will assume the pricing function $p : \mathcal{C} \longrightarrow \mathbb{R}^+$ is known by the DM and subadditive. Based on the way real data marketplaces work, we will assume that the price the buyer pays for data is set according to different criteria, namely i) the volume of samples purchased, ii) a realistic distribution obtained from existing prices in commercial DM catalogues, iii) a random uniform price distribution that has nothing to do with the value of data [1], or iv) according to the accuracy it brings to the task $(\mathcal{M}, a)$.

Apart from presenting the buyers different eligible datasets $d \subseteq \mathcal{C}$ and their prices $p(d)$, we will assume that the data marketplace can evaluate the performance of an AI/ML model trained with data it has access to, and buyers can therefore ask the DM for $a(d)$ to make the decision of which dataset(s) to buy. We will assume that the DM will charge the buyer for the corresponding processing cost a quantity that will depend on the task and the data to evaluate, denoted as $\gamma(\mathcal{M}, a, d)$. We will assume that this cost is proportional to the processing time invested in the transaction, and we will take real costs of IaaS in public cloud platforms to evaluate the processing charges issued by the DM.

For now, we will assume that the DM is also willing to perform some free valuation for the buyer, by providing him with the revealed accuracies for a small subset of $\mathcal{C}$ which we denote the known set $\mathcal{K} \subset \mathcal{C}$. The rest of the datasets for which accuracy is unknown belong to the unknown set $\mathcal{U} \subseteq \mathcal{C}$, therefore the datasets on sale in the catalogue are $\mathcal{C} = \mathcal{K} \cup \mathcal{U}$. This assumption is relaxed later as the purchasing strategies developed can work in the setting where nothing is provided for free $|\mathcal{K}| = 0$.

Therefore, we will assume that the cost of a data transaction for the buyers, denoted as $c(d)$, includes a fixed cost component (compensating, for example, for the registration of a transaction in a blockchain ledger, or for the resulting costs or transferring the data) denoted as c_f, the cost of acquiring the data ($p(d)$), and a variable cost component proportional to the amount of processing required to select d and calculate the corresponding payoffs for sellers, which we will denote as c_p and that will in turn depend on the datasets $d' \in K$ the buyer has asked the DM to evaluate to make this decision.

$$c(d) = p(d) + c_f + c_p = p(d) + c_f + \sum_{d' \in K} \gamma(\mathcal{M}, a, d') \tag{1}$$

2.1 Purchasing Strategies

In this setting, the problem of selecting suitable data products using the capabilities of data marketplaces is not straightforward and requires careful purchasing strategies to optimise $argmax_{d \in \mathcal{C}} a(d)$ subject to $c(d) < B$. Optimally, buyers would like to get the $d^\star, \forall d \subseteq \mathcal{C}, a(d^\star) \geq a(d)$. But for finding $d^\star$ a buyer needs to ask the DM to evaluate any possible combination of data products, which usually leads to prohibitive processing costs. Since it is not feasible to reveal the value of any possible combinations of elements in $\mathcal{C}$, buyers need purchasing strategies to select promising data to be evaluated by the DM and, eventually, be bought to increase the accuracy of their model.

***Impossibility of maximising absolute quality with a deterministic algorithm*:** Given that qualities of the datasets in $\mathcal{U}$ are unknown, it follows that no deterministic algorithm is guaranteed to be able to buy the affordable dataset of highest (*absolute*) quality.

Lemma 1 *Given prices and qualities for datasets in $\mathcal{K}$, prices for datasets in $\mathcal{U}$, and a budget B for buying datasets and revealing unknown qualities at cost R per individual dataset, it is impossible to guarantee that the dataset with highest quality and price up to B will be identified in $\mathcal{K} \cup \mathcal{U}$.*

This can be easily established via the following counter-example. Assume that o is the optimal dataset that has $p(o) \leq B$ and $a(o) \geq v$ for any $v \in \mathcal{C}$ and that $o \in \mathcal{U}$. Assume also that $|\mathcal{U}| > B/R$. This means that independently of the actual price $p(o)$, the buyer may not be able to identify o even if he uses all his budget to reveal the price of $\lfloor B/R \rfloor$ datasets in $\mathcal{U}$.

2.1.1 Greedy Algorithm

Granted the impossibility of maximising *absolute* quality, let us define the alternative notion of *guaranteed* quality for the buyer as the maximum quality dataset that he can afford at a certain point in time with the available information and budget at that time. Knowing the prices and qualities in $\mathcal{K}$, the prices in $\mathcal{U}$, and having a budget B, if we take the set $\mathcal{K}$ to be sorted in order of decreasing accuracy and the set $\mathcal{U}$ to be sorted in order of increasing price. Then the initial guaranteed quality of the buyer is the first and therefore highest accuracy dataset in the known set $a(k(1))$ where $k \in \mathcal{K}$. If, he starts exploring $\mathcal{U}$ then his guaranteed quality can become *at least*:

$$g^* = \max\left(a(k(1)), \max_{u \in \mathcal{U}}(a(u))\right) \tag{2}$$

Algorithm 1 Greedy Algorithm for Maximum Guaranteed Quality

Require: $\mathcal{K}$ (known datasets), $\mathcal{U}$ (unknown datasets), B (budget), R (cost to reveal)

Initialize:
    $remaining\_budget \leftarrow B - B - p(k(1))$
    $best\_dataset \leftarrow k(1)$
    $i \leftarrow 0$
    Sort $\mathcal{K}$ in order of descending accuracy, and $\mathcal{U}$ in order of increasing price
while $remaining\_budget > R$ **do**
    Pay R to learn the quality of $\mathcal{U}(i)$
    if $a(\mathcal{U}(i)) > best\_dataset$ and $p(\mathcal{U}(i)) \leq B - R * i$ **then**
        Set $best\_dataset = a(\mathcal{U}(i))$ and `leftover` $= B - p(\mathcal{U}(i)) - R * i$
    end if
    $i \leftarrow i + 1$
end while
return $best\_dataset$

2.1.2 Bandit Algorithm

Depending on the distributions of qualities and prices in the unveiled set $\mathcal{K}$ and the unknown set $\mathcal{U}$, heavier bias towards exploration might lead to better quality found and perhaps even the global maximum dataset being found. For this case, we can postulate that a buyer is willing to risk a certain drop in guaranteed quality g^* for an increase in the expected accuracy e^* in the hope of finding a dataset $u(k)$ whose quality is higher than the guaranteed maximum $a(u(k)) > g^*$. This can occur if the dataset $u(k)$ is outside the exploration depth achieved by the greedy algorithm, $k > \frac{B - p(g^*)}{R}$. We denote this *risk* factor s, which dictates how much sacrifice in quality the buyer is willing to incur. It is provable that reserving the price of a cheaper yet slightly lower accuracy dataset in $\mathcal{K}$ as the exploration starting point can lead to deeper exploration and perhaps better results. This alternate starting point in $\mathcal{K}$ we denote as $k(j) \in \mathcal{K}$.

$$e^* = \max\left(a(k(j)), \max_{u \in \mathcal{U}}(a(u))\right) \tag{3}$$

To choose our starting point, we compute the probability of choosing a dataset $k(j)$ in $\mathcal{K}$ and reserving its price $p(k(j))$, denoted as ε_j, as the ratio of sacrifice in quality $\Delta a(j) = a(k(1)) - a(k(j))$ to increase in leftover budget $\Delta l(j) = l(k(j)) - l(k(1))$, $\varepsilon_j = \frac{\Delta l(j)}{\Delta a(j)}$. The weights are normalized so that $\sum_{j=1}^{|\mathcal{K}|} \varepsilon_j = 1$.

Algorithm 2 Bandit Algorithm for Maximum Expected Quality

Require: $\mathcal{K}$ (known datasets), $\mathcal{U}$ (unknown datasets), B (budget), R (cost to reveal), s (risk)

Initialize:
    $datasets\_explored \leftarrow$ boolean array for datasets in $\mathcal{U}$
    $linear\_regression\_performance \leftarrow$ array for performance tracking
    $remaining\_budget \leftarrow B$
    $best\_dataset \leftarrow k(j) \in \mathcal{K}$
    $t_p \leftarrow 0$
while $remaining\_budget > R$ **do**
    Calculate $average\_error$ as average of $linear\_regression\_performance$
    Update t_p using GET_ADJUSTED_TP($average\_error, s, t_p$)
    Calculate ρ between known datasets in $\mathcal{K}$ and unveiled datasets in $\mathcal{U}$
    if t_p suggests using linear regression ($\rho > t_p$) **then**
        Apply linear regression on known datasets
        Predict accuracies in $\mathcal{U}$
        Select a dataset in $\mathcal{U}$ based on predictions and affordability
        Update $linear\_regression\_performance$
    else
        Select and reveal a dataset from $\mathcal{U}$ with heavy bias on cheaper datasets
    end if
    Update $datasets\_explored$ and $remaining\_budget$
    if revealed dataset has higher accuracy and is affordable **then**
        Update $best\_dataset$
    end if
end while
return accuracy of $best\_dataset$

3 Scenarios

For the Benchmark testing we have used the task of image classification on the MNIST dataset. To simulate different realistic DM scenarios where our purchasing strategies could be tested and compared we have introduced non-iid distributions of labels, dataset sizes, image quality, Moreover, as part of prior research and surveys in commercial data market [1], we have identified four pricing schemes commonly present in DMs, namely volume-based pricing, accuracy-correlated pricing, random pricing, and pricing according to a market-based distribution. Next we provide additional details on how we carried out non-iid distributions, and on the different price schemes considered in the tests.

3.1 Non-iid: Label and Size Distribution

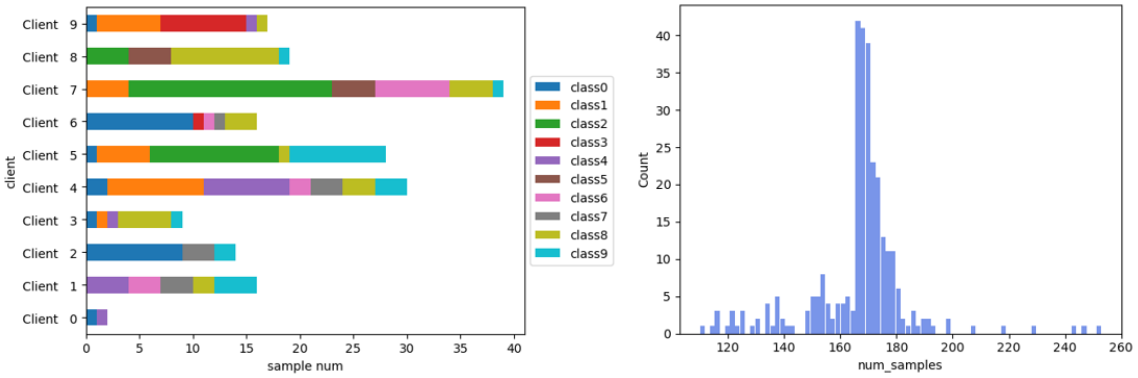
Using a Dirichlet distribution, we are able to split the MNIST dataset into partitions simulating the datasets on sale on a DM using a α parameter to decide the non-iid degree as shown in Figure 1.

3.2 Non-iid: Quality Distribution

To simulate the differing quality of images offered by the sellers in a DM we introduce different types of noise to the images, namely: Gaussian noise, speckle noise and salt and pepper noise as show in Figure 2

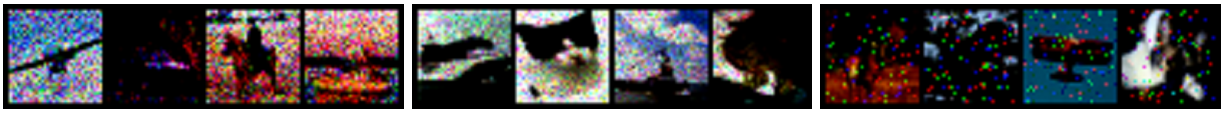
3.3 Pricing Schemes

Volume-based pricing: This is the case in which the price per image is the same for all sellers.



(a) Label Distribution of 10 sellers (b) Size distribution of 100 partitions

Figure 1: Representation of the resulting simulated DM catalogue with MNIST Non-iid split.



(a) Gaussian Noise representation (b) Speckle Noise representation (c) S&P Noise representation

Figure 2: Representation of the resulting images after applying different noises.

Random pricing: In this case the price per dataset is set independently and following a uniform random distribution.

Pricing correlated with accuracy: Some literature works propose that DMs charge prices relevant to the utility brought to the buyer. We will assume that in our case the utility of a dataset is equivalent to the accuracy of that dataset on the buyers model.

Market-based pricing: From surveying existing commercial DMs we have obtained a distribution for image dataset prices following two distinct log-normal distributions. The two classes represent "general purpose" datasets which are generally larger and less task dependent and "custom" datasets which are generally smaller and more tailored to a specific task. We simulate this case by adding noise to the larger partitions in our split and assigning those partitions the price distribution of the "general purpose" datasets, whereas the smaller cleaner datasets are assigned the price distribution of "custom" datasets.

4 Preliminary Results

To compare our algorithms, we have implemented five other algorithms present in the existing exploration vs exploitation research literature [8, 9, 10].

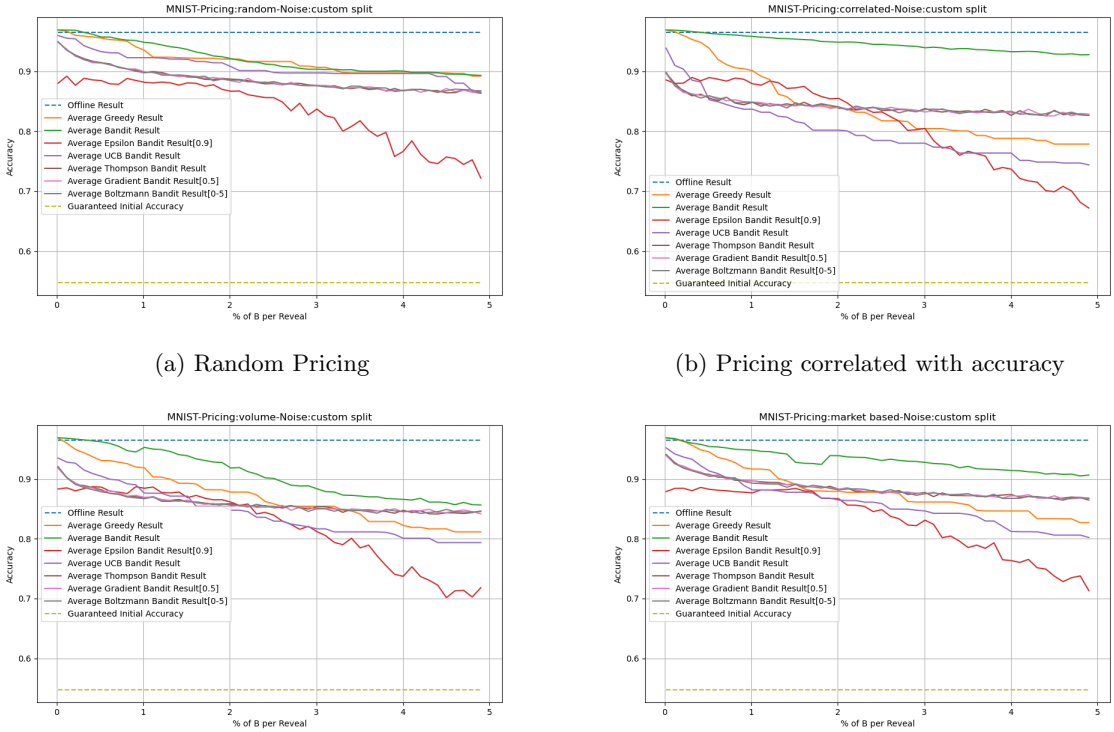
Epsilon Greedy Bandit: In the Epsilon-Greedy strategy, with probability epsilon, you explore by revealing the accuracy of a dataset in $\mathcal{U}$, and with probability 1 - epsilon, you exploit by choosing the best dataset found so far that is affordable with the remaining budget. The value of epsilon determines the balance between exploration and exploitation – a higher epsilon meaning more exploration.

Upper Confidence Bound (UCB) Bandit: UCB involves selecting the dataset that has the highest upper confidence bound, balancing between the mean accuracy and the uncertainty associated with it.

Thompson Sampling Bandit: Thompson sampling in our context would involve maintaining a probability distribution for the accuracy of each dataset in $\mathcal{U}$. As you reveal more datasets, you update these distributions.

Softmax (Boltzmann) Exploration: Softmax - aka Boltzmann - Exploration is a strategy where the probability of selecting an option is based on the relative estimated values of the options, using a softmax function. This approach is softer than the hard max decision rule used in UCB and can be more explorative.

Gradient bandit Algorithm: Gradient Bandit Algorithms use a numerical preference for each option. The preferences are updated based on received rewards, and the probability of selecting each option is determined using a softmax function over these preferences.



(a) Random Pricing

(b) Pricing correlated with accuracy

(c) Volume-based Pricing

(d) Market-based Pricing

Figure 3: Performance Comparison of our algorithms on MNIST.

5 Conclusion and Next Steps

As shown in Figure 3, our Bandit algorithm seems to outperform the greedy and the other algorithms from exploration vs exploitation literature in all four of our case scenarios. It proves to be more robust as it deals with even unfavourable pricing schemes such as random pricing while still maintaining its strength in the correlated and volume-based pricing scenarios.

Next Steps: We are progressing with this task in the following directions:

1. Scanning for optimal parameters both for the bandit and the existing literature solutions to properly show that our algorithm outperforms them,

2. Introducing VFs and PVMs and testing them to find out about the accuracy vs cost trade-off that is of interest to us in the scope of this work, and finally

3. Moving on to a second scenario to validate our purchasing strategies performance in that case and explore the effect and trade-off introduced by the use of VFs and PVMs. In this case, we will evaluate and select data of individual taxis for the task of forecasting taxi-ride demand per district in Chicago.

References

[1] S. Andrés Azcoitia and N. Laoutaris. A survey of data marketplaces and their business models. 2022.

[2] Santiago Andrés Azcoitia, Marius Paraschiv, and Nikolaos Laoutaris. Computing the relative value of spatio-temporal data in wholesale and retail data marketplaces, 2020.

[3] Anish Agarwal, Munther Dahleh, and Tuhin Sarkar. A marketplace for data: An algorithmic solution. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 701–726, 06 2019.

[4] Lingjiao Chen, Paraschos Koutris, and Arun Kumar. Towards model-based pricing for machine learning in a data marketplace. In *Proceedings of the 2019 International Conference on Management of Data*, SIGMOD '19, page 1535–1552, New York, NY, USA, 2019. Association for Computing Machinery.

[5] Santiago Andrés Azcoitia and Nikolaos Laoutaris. Try before you buy: A practical data purchasing algorithm for real-world data marketplaces, 2020.

[6] Xinlei Xu, Awni Hannun, and Laurens Van Der Maaten. Data appraisal without data sharing. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 11422–11437. PMLR, 28–30 Mar 2022.

[7] City of Chicago. Taxi trips: City of chicago: Data portal, Dec 2023.

[8] WILLIAM R THOMPSON. ON THE LIKELIHOOD THAT ONE UNKNOWN PROBABILITY EXCEEDS ANOTHER IN VIEW OF THE EVIDENCE OF TWO SAMPLES. *Biometrika*, 25(3-4):285–294, 12 1933.

[9] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256, 2002.

[10] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Anexo 9: RAGs to Riches: Decentralized Retrieval and Measuring Document Influence in Retrieval-Augmented Generation (RAG) Systems

Retrieval-Augmented Generation (RAG) systems enhance traditional large language models (LLMs) by retrieving external, relevant data during the generation process. While this approach improves accuracy and relevance—particularly in specialized domains such as finance, law, or medicine—it introduces challenges related to privacy, scalability, and incentivizing third-party data providers. This work proposes a novel framework for decentralized retrieval (DR) and document influence measurement to enable a sustainable business model for RAG systems.

A key issue lies in incentivizing third-party data contributors while protecting their proprietary information. LLM owners want to pay only for valuable, utilized data without exposing sensitive user queries, while data providers seek fair compensation without premature data disclosure. To address these conflicting interests, we propose decentralized retrieval mechanisms, ensuring that queries and third-party data remain private while still enabling retrieval.

Our approach adapts indirect Wasserstein distances—previously used in federated learning—to serve as a retrieval score, replacing traditional cosine similarity. This method not only preserves privacy but also offers competitive performance, as shown in initial evaluations. We address scalability challenges in DR through a filtering step such using randomized projection vectors, ensuring the method's efficiency for large-scale retrieval tasks.

In parallel, we tackle the novel problem of measuring the influence of retrieved documents on RAG outputs. Existing literature lacks methods to evaluate document contribution in our context. We explore multiple influence metrics, including:

- **Semantic and Transform Approaches:** Cosine similarity and Optimal Transport (OT) distances.
- **Model Dynamics:** Attention-based methods, such as L2 norm of attention weights and OT Attention-weighted distances.
- **Information Theory:** Jensen-Shannon Divergence to compare probability distributions.
- **LLM-based Scoring:** Prompting the LLM itself to evaluate relevancy and influence.

Our results demonstrate that the Attention-weighted OT Distance performs best, effectively capturing the relationship between retrieved documents and generated outputs. This metric, combined with decentralized retrieval, allows for precise document evaluation without compromising privacy.

The proposed system establishes a realistic and innovative RAG business model where LLM owners can securely retrieve and evaluate third-party data while compensating contributors fairly. Our contributions include a novel, scalable method for decentralized retrieval and a rigorous evaluation of document influence metrics, advancing the state of RAG systems.