Novelty Assessment Report

Paper: Protection against Source Inference Attacks in Federated Learning

PDF URL: https://openreview.net/pdf?id=1GMw3IwEHW

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2026-01-05

Abstract

Federated Learning (FL) was initially proposed as a privacy-preserving machine learning paradigm. However, FL has been shown to be susceptible to a series of privacy attacks. Recently, there has been concern about the Source Inference Attack (SIA), where an honest-but-curious central server attempts to identify exactly which client owns a given data point which was used in the training phase. Alarmingly, standard gradient obfuscation techniques with Differential Privacy have been shown to be ineffective against SIAs, at least without severely diminishing the accuracy.

In this work, we propose a defense against SIAs within the widely studied shuffle model of FL, where an honest shuffler acts as an intermediary between the clients and the server. First, we demonstrate that standard naive shuffling alone is insufficient to prevent SIAs. To effectively defend against SIAs, shuffling needs to be applied at a more granular level; we propose a novel combination of parameter-level shuffling with the residue number system (RNS). Our approach provides robust protection against SIAs without affecting the accuracy of the joint model and can be seamlessly integrated into other privacy protection mechanisms.

We conduct experiments on a series of models and datasets, confirming that standard shuffling approaches fail to prevent SIAs and that, in contrast, our proposed method reduce the attack's accuracy to the level of random guessing.

Core Task Landscape

This paper addresses: **defending against source inference attacks in federated learning**

A total of **50 papers** were analyzed and organized into a taxonomy with **9 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Privacy Attack Characterization and Threat Modeling**
- **Defense Mechanisms and Mitigation Strategies**
- **Evaluation Frameworks and Empirical Analysis**
- **Domain-Specific Applications and Implementations**

Complete Taxonomy Tree

- defending against source inference attacks in federated learning Survey Taxonomy
- Privacy Attack Characterization and Threat Modeling
  - Source and Membership Inference Attacks ★ (5 papers)
  - [0] Protection against Source Inference Attacks in Federated Learning (Anon et al., 2026) View paper
  - [1] Subject data auditing via source inference attack in cross-silo federated learning (Jiaxin Li, 2025) View paper
  - [6] Cmi: Client-targeted membership inference in federated learning (Tianhang Zheng, 2023) View paper
  - [7] A privacy preserving framework for federated learning in smart healthcare systems (Wenshuo Wang, 2023) View paper
  - [13] Interaction-level Membership Inference Attack Against Federated Recommender Systems (Wei Yuan, 2023) View paper
  - Data Leakage and Reconstruction Attacks (9 papers)
  - [2] Defending Batch-Level Label Inference and Replacement Attacks in Vertical Federated Learning (Tianyuan Zou, 2024) View paper
  - [4] Exploiting Defenses against GAN-Based Feature Inference Attacks in Federated Learning (Xinjian Luo, 2022) View paper
  - [5] FLSG: A Novel Defense Strategy Against Inference Attacks in Vertical Federated Learning (Kai Fan, 2024) View paper
  - [10] Evaluating Gradient Inversion Attacks and Defenses in Federated Learning (Yangsibo Huang, 2022) View paper
  - [12] Defending Against Inference and Backdoor Attacks in Vertical Federated Learning via Mutual Information Regularization (Tianyuan Zou, 2024) View paper
  - [14] From Risk to Resilience: Towards Assessing and Mitigating the Risk of Data Reconstruction Attacks in Federated Learning (X., 2025) View paper
  - [17] Defending Label Inference and Backdoor Attacks in Vertical Federated Learning (Yang Liu, 2021) View paper
  - [25] Link Inference Attacks in Vertical Federated Graph Learning (Oualid Zari, 2024) View paper
  - [30] Defending Label Inference Attacks in Split Learning under Regression Setting (Haoze Qiu, 2023) View paper
  - Comprehensive Privacy Attack Surveys and Taxonomies (7 papers)
  - [3] Membership inference attacks and defenses in federated learning: A survey (Li Bai, 2024) View paper
  - [15] Securing the collective intelligence: a comprehensive review of federated learning security attacks and defensive strategies (Vishal Kaushal, 2025) View paper
  - [22] Privacy inference attack and defense in centralized and federated learning: A comprehensive survey (Bosen Rao, 2024) View paper
  - [23] Cyber Threat Intelligence and Security for Federated Learning in Digital Forensics (S.S. Iyengar, 2025) View paper
  - [24] Threats, attacks and defenses to federated learning: issues, taxonomy and perspectives (Peng Liu, 2022) View paper

- [35] Privacy and Robustness in Federated Learning: Attacks and Defenses (Lyu, 2022) View paper
- [43] A Comprehensive Analysis of Inference Attacks in Federated Learning (Dhairya Ameria, 2023) View paper

- Defense Mechanisms and Mitigation Strategies
  - Cryptographic and Shuffling-Based Defenses (5 papers)
  - [11] FedMLU: Mitigating Source Inference Attacks in Federated Learning Without Losing Utility for Secure IoT Services (Mengmeng Cui, 2025) View paper
  - [27] Quality inference in federated learning with secure aggregation (Pejo Balazs, 2023) View paper
  - [33] Protection against Source Inference Attacks in Federated Learning using Unary Encoding and Shuffling (Jung, 2024) View paper
  - [36] MixNN: Protection of Federated Learning Against Inference Attacks by Mixing Neural Network Layers (Thomas Lebrun, 2022) View paper
  - [39] Poster: Protection against Source Inference Attacks in Federated Learning using Unary Encoding and Shuffling (Andreas Athanasiou, 2024) View paper
  - Differential Privacy and Noise-Based Defenses (6 papers)
  - [8] Accuracy-privacy trade-off in the mitigation of membership inference attack in federated learning (Sayyed Farid Ahamed, 2025) View paper
  - [18] On Defensive Neural Networks Against Inference Attack in Federated Learning (Hong-Kyu Lee, 2021) View paper
  - [29] Asynchronous Federated Learning With Local Differential Privacy for Privacy-Enhanced Recommender Systems (Xiaopeng Zhao, 2025) View paper
  - [34] Metric privacy in federated learning for medical imaging: Improving convergence and preventing client inference attacks (DÃ az, 2025) View paper
  - [37] Secure Intrusion Detection by Differentially Private Federated Learning for Inter-Vehicle Networks (Qian-Kun Xu, 2023) View paper
  - [47] MemDefense: Defending Against Membership Inference Attacks in IoT-Based Federated Learning via Pruning Perturbations (Meng Shen, 2024) View paper
  - Model-Centric and Representation Learning Defenses (6 papers)
  - [9] Digestive neural networks: A novel defense strategy against inference attacks in federated learning (Hongkyu Lee, 2021) View paper
  - [16] Ressfl: A resistance transfer framework for defending model inversion attack in split federated learning (Jingtao Li, 2022) View paper
  - [21] Soteria: Provable defense against privacy leakage in federated learning from representation perspective (Jingwei Sun, 2021) View paper
  - [28] Rec-Def: A recommendation-based defence mechanism for privacy preservation in federated learning systems (Chamara Sandeepa, 2023) View paper
  - [32] KDk: A Defense Mechanism Against Label Inference Attacks in Vertical Federated Learning (Marco Arazzi, 2024) View paper
  - [38] Information Bottleneck-Based Subgraphs Defending Against Inference Attacks in Federated Graph Learning Systems (Chenhan Zhang, 2026) View paper
  - Attack Detection and Adversarial Training (4 papers)
  - [19] Advanced Privacy and Security Techniques in Federated Learning Against Sophisticated Attacks (Hariprasad Holla, 2025) View paper
  - [31] Mia-bad: An approach for enhancing membership inference attack and its mitigation with federated learning (Soumya Banerjee, 2023) View paper
  - [48] Defending against Membership Inference Attacks in Federated learning via Adversarial Example (Yuanyuan Xie, 2021) View paper
  - [50] Towards Resilient Federated Learning in CyberEdge Networks: Recent Advances and Future Trends (Li Kai, 2025) View paper
- Evaluation Frameworks and Empirical Analysis (4 papers)
  - [20] Defending Against Membership Inference Attack for Counterfactual Federated Recommendation With Differentially Private Representation Learning (Xiu-wen Liu, 2024) View paper
  - [45] Free-riders in Federated Learning: Attacks and Defenses (Lin, 2022) View paper
  - [46] Mitigating Membership Inference Vulnerability in Personalized Federated Learning (Jung, 2025) View paper
  - [49] Find a Scapegoat: Poisoning Membership Inference Attack and Defense to Federated Learning (Li, 2025) View paper
- Domain-Specific Applications and Implementations (5 papers)
  - [26] ISPPFL: An incentive scheme based privacy-preserving federated learning for avatar in metaverse (Yang Bai, 2024) View paper
  - [40] Federated Learning: An Overview of Attacks and Defense Methods (K. M. Sameera, 2025) View paper
  - [41] Cyber Security Concerns and Mitigation Strategies in Federated Learning: A Comprehensive Review (Param Ahir, 2023) View paper
  - [42] A survey on securing federated learning: Analysis of applications, attacks, challenges, and trends (Helio N. Cunha Neto, 2023) View paper
  - [44] Privacy-Preserving Federated Learning in Healthcare, E-Commerce, and Finance: A Taxonomy of Security Threats and Mitigation Strategies (Rahul Kumar, 2025) View paper

Narrative

Core task: defending against source inference attacks in federated learning. The field organizes itself around four main branches that together capture the lifecycle of privacy research in federated settings. Privacy Attack Characterization and Threat Modeling establishes the adversarial landscape, examining how attackers can infer sensitive information about data sources, memberships, and client identities from model updates or aggregated parameters. Defense Mechanisms and Mitigation Strategies develops countermeasures ranging from differential privacy and gradient perturbation techniques to architectural modifications that limit information leakage. Evaluation Frameworks and Empirical Analysis provides the methodological backbone, offering benchmarks and metrics to quantify privacy risks and measure defense effectiveness across diverse scenarios. Domain-Specific Applications and Implementations translates these insights into practical deployments in healthcare, finance, and other sensitive domains where federated learning promises collaboration without direct data sharing.

Within the attack characterization branch, a dense cluster of works explores source and membership inference threats, revealing how adversaries can exploit gradient information or model behavior to identify training participants or reconstruct sensitive attributes. Source Inference Protection[0] sits squarely in this area, addressing the specific challenge of preventing attackers from determining which client contributed particular data samples. Its emphasis contrasts with broader membership inference studies like Membership Inference Survey[3], which surveys a wider range of inference attacks, and with works such as Client Targeted Membership[6] or Interaction Level Membership[13], which focus on finer-grained membership detection at the client or interaction level. Meanwhile,

defenses like FLSG Defense[5] and techniques surveyed in Privacy Inference Survey[22] illustrate the ongoing tension between utility preservation and privacy guarantees, highlighting open questions about scalability, robustness under adaptive attacks, and the trade-offs inherent in deploying privacy-preserving federated systems at scale.

Related Works in Same Category

The following **4 sibling papers** share the same taxonomy leaf node with the original paper:

1. Subject data auditing via source inference attack in cross-silo federated learning

Authors: Jiaxin Li, Marco Arazzi, Antonino Nocera, Mauro Conti | **Year/Venue**: 2025 | **URL**: View paper

Abstract

â¦ Source Inference Attack (SIA) in Federated Learning (FL) â¦ Therefore, we propose a Subject-Level Source Inference Attack (â¦ Finally, we propose to defend our SLSIA using item-level â¦

Relationship Analysis

Both papers belong to the Source and Membership Inference Attacks category, focusing on identifying which clients contributed specific data in federated learning. They overlap in addressing source inference attacks (SIA) in cross-silo FL settings and analyzing how data heterogeneity affects attack success. The original paper proposes a defense mechanism using parameter-level shuffling with RNS encoding to protect against SIAs, while the candidate paper proposes a novel attack method (SLSIA) that detects all clients using data from a target subject by analyzing model embeddings, representing opposite perspectives on the same privacy threat.

2. Cmi: Client-targeted membership inference in federated learning

Authors: Tianhang Zheng, Bao-Chun Li, Baochun Li | **Year/Venue**: 2023 | **URL**: View paper

Abstract

Membership inference is a popular benchmark attack to evaluate the privacy risk of a machine learning model or a learning scheme. However, in federated learning, membership inference is still under-explored due to several issues. For instance, some assumptions in prior works may not be practical in federated learning. Most existing membership inference methods stand on those impractical assumptions or lack generalization ability, which may misestimate the privacy risk. To address these issues, w...

Relationship Analysis

Both papers belong to the Source and Membership Inference Attacks category, addressing privacy attacks in federated learning that aim to infer information about training data. While the original paper focuses on defending against source inference attacks (identifying which client owns a specific data point) using shuffle-based mechanisms with RNS encoding, the candidate paper addresses membership inference attacks (determining if a data point was in the training set) through a client-targeted poisoning framework called CMI and proposes RR-Label as a defense mechanism. The key difference is that they target different types of inference attacks: source inference versus membership inference, with distinct attack methodologies and defense strategies.

3. A privacy preserving framework for federated learning in smart healthcare systems

Authors: Wenshuo Wang, Xu Li, Xiuqin Qiu, Xiang Zhang, Vladimir Brusic, et al. (7 authors total) | **Year/Venue**: 2023 | **URL**: View paper

Abstract

â¦ Federated Learning (FL) is a platform for smart healthcare systems that use wearables and other Internet of Things enabled devices. However, source inference attacks (SIAs) can infer â¦

Relationship Analysis

Both papers belong to the Source and Membership Inference Attacks category, focusing on privacy vulnerabilities in federated learning where adversaries attempt to infer which client contributed specific training data. The candidate paper addresses source inference attacks (SIAs) in smart healthcare systems using wearables and IoT devices, sharing the original paper's concern about SIA threats in FL. However, the original paper proposes a novel defense mechanism combining parameter-level shuffling with the residue number system (RNS) in the shuffle model, while the candidate paper appears to focus on a privacy-preserving framework specifically tailored for healthcare IoT environments, likely with different architectural assumptions and defense strategies.

4. Interaction-level Membership Inference Attack Against Federated Recommender Systems

Authors: Wei Yuan, Chaoqun Yang, Quoc Viet Hung Nguyen, Chao-Peng Yang, Li-Zhen Cui, et al. (9 authors total) | **Year/Venue**: 2023 • The Web Conference | **URL**: View paper

Abstract

The marriage of federated learning and recommender system (FedRec) has been widely used to address the growing data privacy concerns in personalized recommendation services. In FedRecs, usersâ attribute information and behavior data (i.e., user-item interaction data) are kept locally on their personal devices, therefore, it is considered a fairly secure approach to protect user privacy. As a result, the privacy issue of FedRecs is rarely explored. Unfortunately, several recent studies reveal t...

Relationship Analysis

Both papers belong to the Source and Membership Inference Attacks category, addressing privacy vulnerabilities in federated learning through inference attacks on training data. While the original paper focuses on defending against source inference attacks (identifying which client owns specific data points) in general federated learning using shuffle-based mechanisms with RNS encoding, the candidate paper addresses interaction-level membership inference attacks specifically in federated recommender systems, proposing a regularization-based defense that constrains public parameter updates. The key difference is the application domain (general FL vs. federated recommenders) and defense approach (shuffle model with encoding vs. regularization constraints).

Contributions Analysis

Overall novelty summary. The paper proposes a defense against source inference attacks in federated learning using parameter-level shuffling combined with the residue number system. It resides in the 'Source and Membership Inference Attacks' leaf under 'Privacy Attack Characterization and Threat Modeling', which contains five papers total. This leaf represents a moderately populated research direction within the broader taxonomy of 50 papers across approximately 36 topics. The sibling papers in this leaf focus on characterizing inference threats rather than proposing defenses, suggesting the paper bridges attack analysis with mitigation strategies.

The taxonomy reveals that defense mechanisms occupy a separate major branch with four distinct leaves covering cryptographic approaches, differential privacy, model-centric defenses, and attack detection. The paper's shuffle-based defense naturally connects to the 'Cryptographic and Shuffling-Based Defenses' leaf, which contains five papers exploring secure aggregation and encoding schemes. The scope notes clarify that while the paper sits taxonomically among attack characterization works, its defensive contribution positions it at the boundary between threat modeling and mitigation strategies, potentially explaining why it appears somewhat isolated from its immediate siblings.

Among the three contributions analyzed, the first (reconstruction attacks against standard shuffling) examined 10 candidates with zero refutations, suggesting relative novelty in demonstrating shuffling vulnerabilities. The second contribution (robust defense in shuffle model) examined 9 candidates and found 2 refutable matches, indicating more substantial prior work in shuffle-based defenses. The third contribution (experimental validation) examined 10 candidates with no refutations. These statistics reflect a limited search scope of 29 total candidates examined, not an exhaustive literature review, meaning the analysis captures top semantic matches rather than comprehensive field coverage.

Based on the limited search scope, the work appears to occupy a niche intersection between attack demonstration and defense design within shuffle-based federated learning. The taxonomy structure suggests this specific combination of parameter-level shuffling with residue number systems may be relatively unexplored, though the broader shuffle defense paradigm has established precedents. The analysis acknowledges uncertainty inherent in examining only 29 candidates from a field of 50 surveyed papers, leaving open questions about related work in adjacent cryptographic or encoding-based defense approaches.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Novel reconstruction attacks against standard shuffling in federated learning

Description: The authors introduce reconstruction algorithms for three shuffling granularities (model-level, layer-level, and parameter-level) that enable source inference attacks within the shuffle model of FL. These attacks demonstrate that standard shuffling alone is insufficient to protect against SIAs.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Scale-mia: A scalable model inversion attack against secure federated learning via latent space reconstruction

URL: View paper

Brief Assessment

Scale-mia[70] focuses on model inversion attacks that reconstruct training data from aggregated model updates in federated learning with secure aggregation. The original paper addresses source inference attacks (identifying which client owns specific data) and proposes reconstruction attacks against shuffling mechanisms at different granularities. These are fundamentally different attack types with different objectives and technical approaches.

---

2. {SoK}: Gradient Inversion Attacks in Federated Learning

URL: View paper

Brief Assessment

Gradient Inversion SoK[66] systematizes gradient inversion attacks that reconstruct training data from model updates, but does not address reconstruction attacks against shuffling mechanisms specifically. The original paper's contribution focuses on defeating shuffle-based privacy protections through remapping algorithms, which is a distinct problem from gradient inversion attacks that operate on standard FL without shuffling defenses.

---

3. Privacy in federated learning

URL: View paper

Brief Assessment

Privacy Federated Learning[67] is a survey paper that reviews existing privacy-preserving techniques in federated learning, including shuffling mechanisms. It does not present novel reconstruction attacks against shuffling protocols, which is the specific contribution of the original paper.

---

4. Revamping Federated Learning Security from a Defender's Perspective: A Unified Defense with Homomorphic Encrypted Data Space

URL: View paper

Brief Assessment

Homomorphic Encrypted Defense[73] focuses on defending against evasion data poisoning and model inversion attacks using homomorphic encryption in the data space, not on reconstruction attacks against shuffling mechanisms in federated learning.

---

5. Tazza: Shuffling Neural Network Parameters for Secure and Private Federated Learning

URL: View paper

Brief Assessment

Tazza Shuffling Parameters[72] focuses on neural network parameter shuffling for privacy preservation and integrity defense, not on reconstruction attacks against shuffling mechanisms. The candidate paper does not demonstrate prior work on attacking shuffle-based defenses in federated learning.

---

6. Separation of Powers in Federated Learning (Poster Paper)

URL: View paper

Brief Assessment

Separation of Powers[75] focuses on parameter-granularity disassembly and re-stitching for decentralized aggregation to mitigate data reconstruction attacks, not on developing reconstruction attacks against shuffling mechanisms themselves. The candidate paper is a defense mechanism, not an attack methodology.

---

7. Inverting Gradients -- How easy is it to break privacy in federated learning?

URL: View paper

Brief Assessment

Inverting Gradients[68] focuses on reconstructing original training data from gradient information in federated learning, not on defeating shuffling mechanisms. The paper does not address shuffle-based privacy protections or source inference attacks.

---

8. Towards accurate and stronger local differential privacy for federated learning with staircase randomized response

URL: View paper

Brief Assessment

Staircase Randomized Response[74] focuses on local differential privacy mechanisms with parameter shuffling for privacy amplification, not on developing reconstruction attacks against shuffling mechanisms. The paper addresses privacy through LDP perturbation rather than analyzing vulnerabilities of shuffling alone.

9. Agic: Approximate gradient inversion attack on federated learning

URL: View paper

Brief Assessment

Agic Gradient Inversion[71] focuses on gradient inversion attacks to reconstruct training data from model/gradient updates in federated learning, not on attacking shuffling mechanisms as a privacy defense. The candidate paper does not address shuffling-based defenses or source inference attacks.

10. GTV: Generating Tabular Data via Vertical Federated Learning

URL: View paper

Brief Assessment

GTV Vertical Learning[69] focuses on generating synthetic tabular data via vertical federated learning using GANs, not on reconstruction attacks against shuffling mechanisms in federated learning. The paper addresses privacy-preserving synthetic data generation across organizations with disjoint features, which is a fundamentally different problem domain from the original paper's focus on source inference attacks and shuffling defenses.

Contribution 2: First robust defense against source inference attacks in the shuffle model

Description: The authors present a defense mechanism that combines parameter-level shuffling with the residue number system (RNS) and unary encoding. This approach reduces SIA accuracy to random guessing without affecting joint model accuracy and can be seamlessly integrated into existing shuffle mechanisms.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Secure Federated Matrix Factorization via Shuffling Encrypted Parameters Between Devices

URL: View paper

Brief Assessment

Shuffling Encrypted Parameters[65] focuses on federated matrix factorization with parameter shuffling for privacy, but does not specifically address source inference attacks (SIA) or demonstrate defense mechanisms against them. The candidate's focus is on secure computation rather than SIA defense.

2. Poster: Protection against Source Inference Attacks in Federated Learning using Unary Encoding and Shuffling

URL: View paper

Prior Art Analysis

Unary Shuffling Poster[39] demonstrates that the ORIGINAL paper's claim to be the first robust defense against SIAs in the shuffle model is refuted. The candidate paper explicitly states it proposes 'a defense against sias by using a trusted shuffler' and employs 'a combination of unary encoding with shuffling' to prevent the central server from inferring information about individual client model updates. This directly overlaps with the ORIGINAL paper's core contribution of combining parameter-level shuffling with unary encoding. The candidate paper was published as a poster, indicating it predates or is contemporaneous with the ORIGINAL submission, thereby challenging the novelty claim of being 'first' to propose this defense mechanism.

Evidence

Evidence 1 - **Rationale**: Both papers propose defenses against SIAs using shuffling combined with encoding techniques (RNS+unary in ORIGINAL, unary+quantization in candidate). The candidate's use of unary encoding with shuffling directly overlaps with the ORIGINAL's claimed novel approach, refuting the claim of being 'first'. - **Original**: we propose the first robust defense against sias in the shuffle model of fl. our approach introduces a novel dimension-based shuffling method with higher granularity, using the residue number system (rns). the defense reduces attack accuracy to random guessing without affecting joint model accuracy ... - **Candidate**: in this work, we propose a defense against sias by using a trusted shuffler, without compromising the accuracy of the joint model. we employ a combination of unary encoding with shuffling, which can effectively blend all clients' model updates, preventing the central server from inferring informatio...

3. Secure Federated Matrix Factorization via Device-to-Device Model Shuffling

URL: View paper

Brief Assessment

Secure Matrix Factorization[63] focuses on location-based recommendation systems with device-to-device parameter shuffling to prevent server inference of location data. This is a different application domain (recommendation systems vs. general federated learning) and does not address source inference attacks as defined in the original paper.

4. When federated learning meets medical image analysis: A systematic review with challenges and solutions

URL: View paper

Brief Assessment

Medical Image Review[56] focuses on federated learning challenges in medical image analysis, not on source inference attacks or shuffle model defenses in general federated learning frameworks.

5. Protection against Source Inference Attacks in Federated Learning using Unary Encoding and Shuffling

URL: View paper

Prior Art Analysis

Unary Encoding Shuffling[33] demonstrates that a defense mechanism combining unary encoding with shuffling for source inference attacks in federated learning's shuffle model was proposed prior to the original paper. Both papers address the same problem (defending against SIAs in the shuffle model), employ similar technical approaches (shuffling with encoding), and target the same threat model (honest-but-curious central server with trusted shuffler). The candidate paper explicitly states it proposes 'a defense against sias by using a trusted shuffler' and employs 'a combination of unary encoding with shuffling,' which directly overlaps with the original paper's claimed novel contribution of combining 'parameter-level shuffling with the residue number system (rns) and unary encoding.'

Evidence

Evidence 1 - **Rationale**: Both papers explicitly claim to propose a defense against SIAs using a shuffle model architecture, indicating the candidate addresses the same problem space before the original paper. - **Original**: in this work, we propose a defense against sias within the widely studied shuffle model of fl, where an honest shuffler acts as an intermediary between the clients and the server. - **Candidate**: n this work, we propose a defense against sias by using a trusted shuffler, without compromising the accuracy of the joint model.

Evidence 2 - **Rationale**: The candidate paper describes using unary encoding with shuffling to blend model updates, which is a core component of the original paper's claimed novel approach. This shows prior work combining encoding schemes with shuffling for SIA defense. - **Original**: to effectively defend against sias, shuffling needs to be applied at a more granular level; we propose a novel combination of parameter-level shuffling with the residue number system (rns). - **Candidate**: we employ a combination of unary encoding with shuffling, which can effectively blend all clients' model updates, preventing the central server from inferring information about each client's model update separately.

Evidence 3 - **Rationale**: Both papers claim their defense mechanisms protect against SIAs while maintaining model accuracy, demonstrating that the candidate achieved similar goals prior to the original paper's submission. - **Original**: our approach provides robust protection against sias without affecting the accuracy of the joint model and can be seamlessly integrated into other privacy protection mechanisms. - **Candidate**: n this work, we propose a defense against sias by using a trusted shuffler, without compromising the accuracy of the joint model.

---

6. Model Fragmentation, Shuffle and Aggregation to Mitigate Model Inversion in Federated Learning

URL: View paper

Brief Assessment

Model Fragmentation Shuffle[62] addresses model inversion attacks (data reconstruction) rather than source inference attacks (identifying which client owns a data point). These are fundamentally different privacy threats with different attack mechanisms and defense requirements.

---

7. MixNN: Protection of Federated Learning Against Inference Attacks by Mixing Neural Network Layers

URL: View paper

Brief Assessment

MixNN[36] focuses on protecting federated learning against inference attacks (membership, property, and attribute inference) by mixing neural network layers through a proxy-based system. The original paper addresses source inference attacks (SIAs) in the shuffle model using parameter-level shuffling with RNS encoding. These are distinct attack types and defense mechanisms operating in different architectural contexts.

---

8. RAFLS: RDP-based adaptive federated learning with shuffle model

URL: View paper

Brief Assessment

RAFLS[61] focuses on differential privacy with adaptive noise injection and shuffle model for general data leakage protection, not specifically on defending against source inference attacks (SIAs) where an adversary identifies which client owns a particular data point.

---

9. Advanced Probabilistic Methods for Privacy Amplification: Cooperative and Non-Cooperative Approaches

URL: View paper

Brief Assessment

Probabilistic Privacy Amplification[64] appears to focus on privacy amplification methods in general, not specifically on source inference attacks in federated learning shuffle models. The minimal context provided does not demonstrate prior work on SIA defenses.

Contribution 3: Experimental validation across multiple models and datasets

Description: The authors provide empirical evaluation demonstrating that standard shuffling approaches fail to prevent SIAs, while their proposed method successfully reduces attack accuracy to the level of random guessing across various datasets and model architectures.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Membership Inference Attacks: Evaluation and Defenses

URL: View paper

Brief Assessment

Membership Evaluation Defenses[60] appears to be a master's thesis focused on membership inference attacks, not source inference attacks. The provided context is insufficient to assess overlap with the original paper's experimental validation of SIA defenses.

---

2. Evaluate Inference Attacks: Attack and Defense against 2D Semantic Segmentation Models

URL: View paper

Brief Assessment

Semantic Segmentation Attacks[57] focuses on inference attacks (membership, attribute, model inversion) against 2D semantic segmentation models in autonomous driving, not on evaluating source inference attack defenses in federated learning across image classification datasets.

---

3. privGAN: Protecting GANs from membership inference attacks at low cost to utility

URL: View paper

Brief Assessment

privGAN[53] focuses on protecting GANs from membership inference attacks using synthetic data generation, not on evaluating source inference attack defenses in federated learning across classification datasets.

---

4. When federated learning meets medical image analysis: A systematic review with challenges and solutions

URL: View paper

Brief Assessment

Medical Image Review[56] is a systematic review of federated learning in medical imaging and does not present experimental validation of source inference attack defenses on image classification datasets.

---

5. Synthetic image learning: Preserving performance and preventing membership inference attacks

URL: View paper

Brief Assessment

Synthetic Image Learning[54] focuses on training classifiers using synthetic data and evaluating membership inference attacks, not on evaluating source inference attack (SIA) defenses. The candidate does not address SIAs or shuffling-based defenses in federated learning.

---

6. Handling sensitive medical dataâ□□a differential privacy enabled federated learning approach

URL: View paper

Brief Assessment

Differential Privacy Medical[59] focuses on medical data privacy using differential privacy in federated learning, not on evaluating source inference attack defenses across image classification datasets.

---

7. MI: Multi-modal Models Membership Inference

URL: View paper

Brief Assessment

Multi-modal Membership[58] focuses on membership inference attacks against multi-modal models (image captioning), not on evaluating source inference attack defenses in federated learning across image classification datasets.

---

8. FinP: Fairness-in-Privacy in Federated Learning by Addressing Disparities in Privacy Risk

URL: View paper

Brief Assessment

FinP Fairness Privacy[55] focuses on fairness in privacy risk distribution across FL clients rather than evaluating defense mechanisms against SIAs. While both papers address SIAs, the candidate's primary contribution is a fairness framework, not defense evaluation methodology.

---

9. AugMixCloak: A Defense against Membership Inference Attacks via Image Transformation

URL: View paper

Brief Assessment

AugMixCloak[52] focuses on membership inference attacks (MIA) in federated learning, not source inference attacks (SIA). The experimental validation targets are fundamentally different attack types with distinct objectives.

---

10. Membership Inference Attacks and Defenses in Classification Models

URL: View paper

Brief Assessment

Classification Membership Attacks[51] focuses on membership inference attacks in classification models, not source inference attacks in federated learning. The experimental validation contexts are fundamentally different.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Protection against Source Inference Attacks in Federated Learning View paper
- [1] Subject data auditing via source inference attack in cross-silo federated learning View paper
- [2] Defending Batch-Level Label Inference and Replacement Attacks in Vertical Federated Learning View paper
- [3] Membership inference attacks and defenses in federated learning: A survey View paper
- [4] Exploiting Defenses against GAN-Based Feature Inference Attacks in Federated Learning View paper
- [5] FLSG: A Novel Defense Strategy Against Inference Attacks in Vertical Federated Learning View paper
- [6] Cmi: Client-targeted membership inference in federated learning View paper
- [7] A privacy preserving framework for federated learning in smart healthcare systems View paper
- [8] Accuracy-privacy trade-off in the mitigation of membership inference attack in federated learning View paper
- [9] Digestive neural networks: A novel defense strategy against inference attacks in federated learning View paper
- [10] Evaluating Gradient Inversion Attacks and Defenses in Federated Learning View paper
- [11] FedMLU: Mitigating Source Inference Attacks in Federated Learning Without Losing Utility for Secure IoT Services View paper
- [12] Defending Against Inference and Backdoor Attacks in Vertical Federated Learning via Mutual Information Regularization View paper
- [13] Interaction-level Membership Inference Attack Against Federated Recommender Systems View paper
- [14] From Risk to Resilience: Towards Assessing and Mitigating the Risk of Data Reconstruction Attacks in Federated Learning View paper
- [15] Securing the collective intelligence: a comprehensive review of federated learning security attacks and defensive strategies View paper
- [16] Ressfl: A resistance transfer framework for defending model inversion attack in split federated learning View paper
- [17] Defending Label Inference and Backdoor Attacks in Vertical Federated Learning View paper
- [18] On Defensive Neural Networks Against Inference Attack in Federated Learning View paper
- [19] Advanced Privacy and Security Techniques in Federated Learning Against Sophisticated Attacks View paper
- [20] Defending Against Membership Inference Attack for Counterfactual Federated Recommendation With Differentially Private Representation Learning View paper
- [21] Soteria: Provable defense against privacy leakage in federated learning from representation perspective View paper
- [22] Privacy inference attack and defense in centralized and federated learning: A comprehensive survey View paper
- [23] Cyber Threat Intelligence and Security for Federated Learning in Digital Forensics View paper
- [24] Threats, attacks and defenses to federated learning: issues, taxonomy and perspectives View paper
- [25] Link Inference Attacks in Vertical Federated Graph Learning View paper
- [26] ISPPFL: An incentive scheme based privacy-preserving federated learning for avatar in metaverse View paper
- [27] Quality inference in federated learning with secure aggregation View paper
- [28] Rec-Def: A recommendation-based defence mechanism for privacy preservation in federated learning systems View paper
- [29] Asynchronous Federated Learning With Local Differential Privacy for Privacy-Enhanced Recommender Systems View paper
- [30] Defending Label Inference Attacks in Split Learning under Regression Setting View paper
- [31] Mia-bad: An approach for enhancing membership inference attack and its mitigation with federated learning View paper
- [32] KDk: A Defense Mechanism Against Label Inference Attacks in Vertical Federated Learning View paper
- [33] Protection against Source Inference Attacks in Federated Learning using Unary Encoding and Shuffling View paper

- [34] Metric privacy in federated learning for medical imaging: Improving convergence and preventing client inference attacks View paper
- [35] Privacy and Robustness in Federated Learning: Attacks and Defenses View paper
- [36] MixNN: Protection of Federated Learning Against Inference Attacks by Mixing Neural Network Layers View paper
- [37] Secure Intrusion Detection by Differentially Private Federated Learning for Inter-Vehicle Networks View paper
- [38] Information Bottleneck-Based Subgraphs Defending Against Inference Attacks in Federated Graph Learning Systems View paper
- [39] Poster: Protection against Source Inference Attacks in Federated Learning using Unary Encoding and Shuffling View paper
- [40] Federated Learning: An Overview of Attacks and Defense Methods View paper
- [41] Cyber Security Concerns and Mitigation Strategies in Federated Learning: A Comprehensive Review View paper
- [42] A survey on securing federated learning: Analysis of applications, attacks, challenges, and trends View paper
- [43] A Comprehensive Analysis of Inference Attacks in Federated Learning View paper
- [44] Privacy-Preserving Federated Learning in Healthcare, E-Commerce, and Finance: A Taxonomy of Security Threats and Mitigation Strategies View paper
- [45] Free-riders in Federated Learning: Attacks and Defenses View paper
- [46] Mitigating Membership Inference Vulnerability in Personalized Federated Learning View paper
- [47] MemDefense: Defending Against Membership Inference Attacks in IoT-Based Federated Learning via Pruning Perturbations View paper
- [48] Defending against Membership Inference Attacks in Federated learning via Adversarial Example View paper
- [49] Find a Scapegoat: Poisoning Membership Inference Attack and Defense to Federated Learning View paper
- [50] Towards Resilient Federated Learning in CyberEdge Networks: Recent Advances and Future Trends View paper
- [51] Membership Inference Attacks and Defenses in Classification Models View paper
- [52] AugMixCloak: A Defense against Membership Inference Attacks via Image Transformation View paper
- [53] privGAN: Protecting GANs from membership inference attacks at low cost to utility View paper
- [54] Synthetic image learning: Preserving performance and preventing membership inference attacks View paper
- [55] FinP: Fairness-in-Privacy in Federated Learning by Addressing Disparities in Privacy Risk View paper
- [56] When federated learning meets medical image analysis: A systematic review with challenges and solutions View paper
- [57] Evaluate Inference Attacks: Attack and Defense against 2D Semantic Segmentation Models View paper
- [58] MI: Multi-modal Models Membership Inference View paper
- [59] Handling sensitive medical dataâ□□a differential privacy enabled federated learning approach View paper
- [60] Membership Inference Attacks: Evaluation and Defenses View paper
- [61] RAFLS: RDP-based adaptive federated learning with shuffle model View paper
- [62] Model Fragmentation, Shuffle and Aggregation to Mitigate Model Inversion in Federated Learning View paper
- [63] Secure Federated Matrix Factorization via Device-to-Device Model Shuffling View paper
- [64] Advanced Probabilistic Methods for Privacy Amplification: Cooperative and Non-Cooperative Approaches View paper
- [65] Secure Federated Matrix Factorization via Shuffling Encrypted Parameters Between Devices View paper
- [66] {SoK}: Gradient Inversion Attacks in Federated Learning View paper
- [67] Privacy in federated learning View paper
- [68] Inverting Gradients -- How easy is it to break privacy in federated learning? View paper
- [69] GTV: Generating Tabular Data via Vertical Federated Learning View paper
- [70] Scale-mia: A scalable model inversion attack against secure federated learning via latent space reconstruction View paper
- [71] Agic: Approximate gradient inversion attack on federated learning View paper
- [72] Tazza: Shuffling Neural Network Parameters for Secure and Private Federated Learning View paper
- [73] Revamping Federated Learning Security from a Defender's Perspective: A Unified Defense with Homomorphic Encrypted Data Space View paper
- [74] Towards accurate and stronger local differential privacy for federated learning with staircase randomized response View paper
- [75] Separation of Powers in Federated Learning (Poster Paper) View paper