INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA DE INTELIGÊNCIA ARTIFICIAL




Topology-Dependent Privacy Risks in Decentralized Federated Learning

JOSÉ INÁCIO ANTUNES DE GOUVEIA
Junho de 2026

Topology-Dependent Privacy Risks in Decentralized Federated Learning

José Inácio Antunes De Gouveia
Student No.: 1211089

A dissertation submitted in partial fulfillment of the requirements for the degree of Master of Science, Specialisation Area of Artificial Intelligence Engineering

Supervisor: Dra. Isabel Cecília Correia da Silva Praça Gomes Pereira
Co-Supervisor: Dra. Ivone De Fátima Da Cruz Amorim

Evaluation Committee:
President:
Luis Conceição, Assistant Professor, Institute of Engineering, Polytechnic of Porto

Members:
Cristiana Neto, Assistant Professor, University of Minho
Isabel Cecília Correia da Silva Praça Gomes Pereira, Associate Professor, Institute of Engineering, Polytechnic of Porto

Porto, June 23, 2026

Statement Of Integrity

I hereby declare that I have conducted this academic work with integrity.

I have not plagiarised or applied any form of undue use of information or falsification of results along the process leading to its elaboration.

Therefore, the work presented in this document is original and was authored by me, having not been previously used for any other purpose. The exceptions are explicitly acknowledged in the section that addresses ethical considerations. This section also states how Artificial Intelligence tools were used and for what purpose.

I further declare that I have fully acknowledged the Code of Ethical Conduct of P.PORTO.

ISEP, Porto, June 23, 2026

Abstract

Federated Learning (FL) has established itself as a leading paradigm for collaborative machine learning, allowing participants to train models collectively without sharing their private data. Despite its privacy-preserving design, the periodic exchange of model updates leaves these systems vulnerable to information leakage. Notable threats include Membership Inference Attacks (MIA), which exploit model overfitting to determine if specific data samples were used during training, and Gradient Inversion Attacks (GIA), which attempt to reconstruct the exact training images from shared gradients. While existing literature has proposed various active defenses and investigated privacy risks within star and mesh network topologies, a systematic evaluation of how attack effectiveness evolves across training rounds over a diverse spectrum of network topologies remains a critical research gap.

This dissertation presents a comprehensive empirical analysis of how different network topologies influence data privacy in centralized and decentralized FL systems over time. By isolating the network topology as the primary variable, we evaluate the vulnerability of six distinct topologies, star, tree, line, ring, full mesh, and partial mesh, against three MIA variants and a GIA. The experiments were conducted using the MNIST and CIFAR10 datasets under both Independent and Identically Distributed and Non-Independent and Identically Distributed (Non-IID) data distributions to capture the temporal evolution of these attacks across the training rounds.

Our findings show that network topologies fundamentally dictate the severity and localization of privacy leakage. For instance, intermediate aggregation in the tree topology acts as a native privacy shield, effectively hiding memorized features from a central root node, though it inadvertently shifts the primary risk to intermediate edge nodes. Furthermore, the analysis reveals that extreme data heterogeneity (Non-IID) significantly aggravates vulnerabilities across all topologies, heavily increasing MIA and GIA success rates on complex visual tasks. Moreover, the results establish that restricting an adversary's awareness of the broader network topology severely impedes their ability to accurately execute GIA, as successful data reconstruction depends heavily on precise knowledge of the aggregation phase in federated systems. Ultimately, this research highlights that network topology can be strategically leveraged as passive defense mechanisms.

Keywords: Collaborative Machine Learning, Federated Learning, Privacy, Topology, Membership Inference Attack, Gradient Inversion Attack

Resumo

O *Federated Learning* (FL) é um paradigma de referência para a aprendizagem colaborativa, permitindo que os participantes treinem modelos de *machine learning* em conjunto sem partilharem os seus dados privados. Apesar da sua premissa orientada para preservação de privacidade, a troca regular de atualizações dos modelos faz com que este tipo de sistemas sejam vulneráveis a fugas de dados. Ameaças importantes incluem os *Membership Inference Attacks* (MIA), que exploram o *overfitting* do modelo para determinar se amostras de dados específicas foram utilizadas durante o treino, e os *Gradient Inversion Attacks* (GIA), que tentam reconstruir as imagens exatas usadas no treino a partir dos gradientes partilhados. Embora a literatura existente tenha proposto vários mecanismos de defesa e investigado os riscos de privacidade em topologias de estrela e malha, uma avaliação sistemática de como a eficácia destes ataques evolui ao longo de todo o ciclo de treino, em diferentes topologias de rede, continua a ser uma lacuna na investigação.

Esta dissertação apresenta uma análise empírica abrangente sobre o modo como diferentes topologias de rede influenciam a privacidade dos dados em sistemas de FL ao longo do tempo. Ao isolar o grafo da rede como variável principal, avaliamos a vulnerabilidade de seis topologias de rede distintas, estrela, árvore, linha, anel, malha completa e malha parcial, contra três variantes de MIA e um GIA avançado. As experiências foram conduzidas utilizando os conjuntos de dados MNIST e CIFAR10, sob distribuições de dados *Independent and Identically Distributed* e *Non-Independent and Identically Distributed* (Non-IID), com o objetivo de capturar a evolução temporal destes ataques ao longo das rondas de treino.

Os resultados obtidos mostram que as propriedades estruturais fundamentalmente definem a gravidade e a localização da fuga de privacidade. Por exemplo, a agregação intermédia em topologias em árvore atua como um escudo de privacidade nativo, ocultando eficazmente as características memorizadas de um servidor central raiz, embora transfira inadvertidamente o risco principal para os nós intermédios. Além disso, a análise revela que a extrema heterogeneidade dos dados (*Non-IID*) agrava significativamente as vulnerabilidades em todas as topologias, aumentando fortemente as taxas de sucesso dos MIA e GIA em tarefas visuais complexas. Os resultados sugerem ainda que ao restringir o conhecimento do adversário sobre a topologia global da rede se impede a sua capacidade de executar GIA com precisão, uma vez que a reconstrução bem sucedida dos dados depende fortemente do conhecimento exato dos caminhos de agregação. Em última análise, esta investigação destaca que a topologia da rede pode ser estrategicamente usada como mecanismos de defesa passivos.

Palavras-chave: Collaborative Machine Learning, Federated Learning, Privacy, Topology, Membership Inference Attack, Gradient Inversion Attack

Acknowledgement

I would like to express my gratitude to everyone who, directly or indirectly, helped me complete this dissertation.

I am sincerely grateful to my supervisors, Dr. Ivone Amorim and Dr. Eva Maia, for their continuous support and guidance throughout the development of this dissertation.

To Dr. Isabel Praça, for proposing the topic of this dissertation and providing me with the opportunity to develop this research.

To Ivan Silva for his help and guidance when I started this dissertation.

I would also like to thank my team: Pedro, Diana, João, Paulo, Maria and Mariana.

This dissertation was done and funded in the scope of the PHRESH project (COMPETE2030-FEDER-01232900 - 17556).

Contents

List of Figures

List of Tables

List of Acronyms

ADMM	Alternating Direction Method of Multipliers.
AI	Artificial Intelligence.
CFL	Centralized Federated Learning.
CNN	Convolutional Neural Network.
DNN	Deep Neural Network.
DP	Differential Privacy.
D-PSGD	Decentralized Parallel Stochastic Gradient Descent.
DFL	Decentralized Federated Learning.
FedAVG	Federated Averaging.
FL	Federated Learning.
GDPR	General Data Protection Regulation.
GIA	Gradient Inversion Attack.
IID	Independent and Identically Distributed.
MIA	Membership Inference Attack.
ML	Machine Learning.
MLP	Multi Layer Perceptron.
MPC	Secure Multiparty Computation.
Non-IID	Non-Independent and Identically Distributed.
P2P	Peer-to-Peer.
PCA	Principal Component Analysis.
PDMM	Primal-Dual Method of Multipliers.
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses.
PSNR	Peak Signal-to-Noise Ratio.
SA	Secure Aggregation.
SGD	Stochastic Gradient Descent.
SLR	Systematic Literature Review.
SPoF	Single Point of Failure.
SSIM	Structural Similarity Index Measure.

Chapter 1

Introduction

This chapter establishes the foundation for the research presented in this thesis. It begins by providing the context and motivation, highlighting the evolution from centralized Machine Learning (ML) to Federated Learning (FL), and subsequently to Decentralized Federated Learning (DFL), driven by increasing data privacy concerns. Following this, a specific research problem regarding the influence of network topology on privacy risks is defined. The chapter then outlines the research questions and specific objectives of this work. Finally, the main contributions of this work are summarized, concluding with an outline of the document's structure.

1.1 Context and Motivation

With the rapid proliferation of Artificial Intelligence (AI) and ML, modern applications require vast amounts of data to train robust ML models [1]. As this demand for data grows, so do concerns regarding data integrity and privacy. A distinct manifestation of this concern is the implementation of the General Data Protection Regulation (GDPR) by the European Union [2], which aims to protect user data privacy. This regulation governs the collection, storage, and sharing of personal data, granting individuals rights over their own data, including the right to access, rectify, and request the deletion of their personal information. These kind of regulations present a significant challenge for traditional ML, which relies on centralized data aggregation, a practice that often violates GDPR principles. Furthermore, data is frequently siloed across multiple institutions (e.g., hospitals, banks, insurance companies), where cross-institutional sharing is prohibited by legal or competitive constraints.

To address these challenges, Google researchers proposed FL in 2016 [3]. This paradigm, now referred to as Centralized Federated Learning (CFL), allows participants to collaboratively train a ML model without sharing raw data. In this architecture, each participant acts as a client node, training a local model using their private data. After training, the nodes send their model updates (gradients or weights) to a central server that aggregates them into a new global model. This global model is then distributed back to all nodes for the next round of training. This process is repeated until convergence is reached or a certain stopping criteria is met. By keeping raw data local, CFL achieves a performance comparable to traditional centralized training while mitigating many of the privacy risks associated with data centralization [4]. However, while CFL addresses the primary challenges of raw data centralization, it introduces architectural vulnerabilities. The reliance on a central server creates a Single Point of Failure (SPoF) [5], because, if the server fails or is compromised, the entire system is jeopardized. Additionally, the server can become a communication bottleneck as it must manage traffic from thousands or millions of nodes simultaneously [6].

To mitigate the reliance on a central server, DFL was proposed. In this paradigm, the central server is eliminated, instead, nodes communicate directly via Peer-to-Peer (P2P) protocols. Each node is responsible for training its local model, exchanging updates with a specific set of neighbors, and performing local aggregation. This decentralized approach eliminates SPoF and alleviates the communication bottleneck inherent in CFL.

Unlike the star topology characteristic of CFL, DFL systems employ a diverse range of network topologies that dictate how nodes are interconnected. These decentralized topologies can be structured into line, ring, or tree topologies, as well as various configurations of mesh topologies [5], [7], [8]. In these setups, the chosen topology dictates the specific paths which model updates propagate. Consequently, the arrangement of the network not only influences training convergence and communication overhead, but also reshapes these pathways, thereby altering the exposure of shared updates to potential adversaries.

Despite the structural advantages of DFL, this paradigm remains susceptible to privacy attacks [9], [10]. Adversaries can exploit shared model updates, with Membership Inference Attack (MIA) and Gradient Inversion Attack (GIA) representing two critical vulnerabilities. On one hand, MIA attempts to determine whether a specific data sample was used to train an ML model. This poses a severe privacy risk by potentially exposing sensitive associations, for instance, an adversary could determine whether an individual's medical record was included in a hospital's training dataset, thereby revealing their status as a patient with a specific medical condition. On the other hand, GIA represents another severe threat, which an adversary attempts to reconstruct the original training data directly from the exchanged model updates. This exploit is highly critical, as it allows the adversary to recover the private training samples of the ML model.

A critical yet underexplored factor that influences these privacy risks is the network topology. Prior works have evaluated the effectiveness of MIA across training rounds in different FL systems [11]–[15]. However, most studies focus on the star network topology, with limited attention given to alternative topologies [16]. Similarly, the effectiveness of GIA across training rounds remains underexplored. Most existing literature on GIA focuses on the star topology [9], [17]–[19]. While a few studies have investigated different network topologies [16], [20]–[23], they primarily focus on the mesh topology, leaving other structures largely unexplored.

To address this research gap, this work systematically investigates the temporal evolution of both MIA and GIA efficacy across diverse FL systems following different network topologies. By evaluating the performance of these attacks over consecutive training rounds, we provide a comprehensive analysis of how the network topology impacts participant privacy and overall system resilience.

1.2 Problem Statement and Objectives

Privacy risk is a major concern in FL systems, as malicious nodes can exploit the network's topology to obtain private data held by other nodes. The choice of network topology in FL systems plays a significant role in determining the privacy risks inherent to these systems. Different network topologies possess distinct structural properties, that can mitigate or exacerbate these risks.

The primary purpose of this thesis is to investigate how different network topologies in FL systems influence these privacy risks. The main research question this thesis aims to answer

is: "How can the choice of network topology in Decentralized Federated Learning influence the privacy risks inherent to these systems?". To answer this central question, the following sub-research questions are proposed:

- **RQ1** - Which network topologies are commonly used in DFL systems and how do they differ from each other?
- **RQ2** - How does the choice of a network topology affect the system vulnerability to attacks?
- **RQ3** - How does the temporal progression of the federated learning process affect the efficacy of privacy attacks across different network topologies?

Considering the problem stated, the main objective of this thesis is to analyze the impact of different network topologies on privacy risks in FL systems and to identify structural configurations that inherently mitigate these risks. The specific objectives required to achieve this goal are:

- **O1** - Identify and categorize the structural properties of distinct network topologies commonly utilized in Decentralized Federated Learning systems.
- **O2** - Evaluate and compare the susceptibility of these network topologies to prominent privacy threats.
- **O3** - Analyze the temporal evolution of these privacy attacks across training rounds to determine the impact of network topologies on the efficacy of these attacks.

1.3 Thesis Contributions

This dissertation provides several key contributions to the field of FL, specifically focusing on the intersection of network topologies and data privacy in both centralized and decentralized environments. The primary contributions are outlined as follows.

First, it delivers a comprehensive systematization of the distinct network topologies utilized in FL systems. This contribution categorizes and details the structural characteristics, practical advantages, and inherent limitations of configurations such as the star, tree, line, ring, full mesh, and partial mesh topologies.

Second, it presents a review of the state-of-the-art regarding identified privacy threats and existing countermeasures in DFL. By synthesizing recent literature, this analysis exposes the complex dynamics between network topology, node connectivity, and privacy leakage. Furthermore, it highlights how the current defensive landscape heavily relies on active cryptographic or perturbation-based mechanisms. Doing so explicitly identifies a crucial research gap regarding the native, unadulterated privacy boundaries provided by the network structures themselves.

Third, this dissertation conducts a empirical comparative analysis of the aforementioned topologies against prominent privacy threats. By strictly isolating the network graph as the primary independent variable, the thesis evaluates the vulnerability of each system to three variants of MIA and one advanced GIA. This evaluation provides a detailed breakdown of attack effectiveness across different training stages, dataset complexities, data distributions, and adversary-victim placements within the network.

Finally, the work presented in this dissertation resulted in a research paper titled "Evaluating Membership Inference Attacks in Hierarchical Federated Learning". This paper has been submitted to the Workshop on Hot Topics in Distributed Security (HotDiSec 2026), co-located with the European Symposium on Research in Computer Security (ESORICS 2026), and is currently under review.

1.4 Thesis Outline

This thesis is structured as follows: Chapter 1 presents the context and motivation for this work, the problem in question, objectives to be achieved, and the contributions of this thesis. Chapter 2 provides the necessary background to understand the concepts and techniques used throughout this work. Chapter 3 presents a systematic literature review regarding the state of the art on the topic of this thesis. Chapter 4 details the experimental setup, describing the selected network topologies, the datasets and data distribution strategies utilized, and the specific model configurations adopted to evaluate privacy leakages under both centralized and decentralized environments. Chapter 5 presents and analyzes the experimental results concerning MIAs, discussing how different structural configurations and data heterogeneity influence model susceptibility across various training rounds. Chapter 6 follows a similar structure to evaluate GIAs, detailing the extent to which private training data can be reconstructed under specific topological bottlenecks and communication protocols. Chapter 7 explores the broader concepts of data privacy and security within FL systems. It examines the legal implications of decentralized roles under the GDPR, analyzes how network structures alter trust boundaries and attack surfaces for passive adversaries, and discusses the ethical dimensions of collaborative learning. Finally, Chapter 8 summarizes the main findings of this dissertation, draws concluding remarks regarding the impact of network topology on FL privacy, and outlines potential directions for future research.

Chapter 2

Preliminaries

This chapter presents the background necessary to understand the concepts and techniques used in this thesis. In Section 2.1, we describe the CFL paradigm, its architecture, and its main characteristics. Section 2.2 introduces the DFL paradigm, highlighting its differences from CFL. Section 2.3 defines the mathematical frameworks for network topologies. Section 2.4 outlines Secure Aggregation protocols used to protect model updates. Finally, Section 2.5 discusses the privacy threats relevant to these systems.

2.1 Centralized Federated Learning

CFL is a collaborative ML paradigm that allows a set of nodes $i = \{1, 2, ..., n\}$ to train a global ML model, without disclosing their sensitive local training dataset D_i [24]. This paradigm relies on a central server, that coordinates and manages the system. Each node i trains their local model w_i with their private dataset D_i. After the node finishes training, it sends the new model update, in the form of gradients or model weights, to the central server. The central server after receiving all model updates from every node, aggregates them into one global model w^t_{global}, where t denotes the current communication round, and distributes this global model back to all participants. This process is repeated until a fixed number of training rounds is completed or a consensus is reached [25]. Figure 2.1 depicts the network topology of a CFL system.

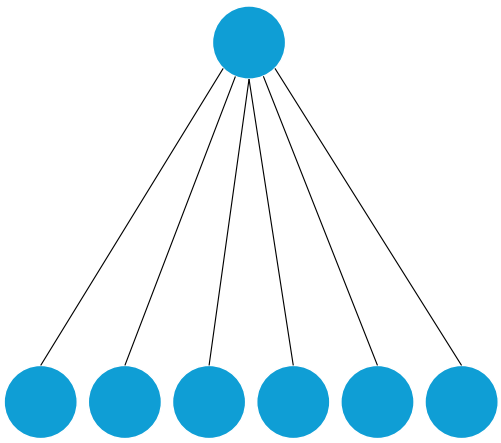


Figure 2.1: Centralized Federated Learning

This aggregation process is most commonly performed using the Federated Averaging (FedAVG) algorithm [25], [26], which computes the weighted average, proportional to the volume of data used in training, of all received model updates. The aggregation process can be mathematically represented by:

$$w_{global}^{t+1} = \sum_{i \in S_t} \frac{n_i}{n} w_i^{t+1} \tag{2.1}$$

where w_{global}^{t+1} is the updated global model after round $t+1$, S_t is the set of selected nodes that participated in round t, n_i is the number of data samples used by node i for training, n is the total number of data samples used by all selected nodes, and w_i^{t+1} is the local model update from node i after round $t+1$.

2.2 Decentralized Federated Learning

DFL is an emerging paradigm that eliminates the need for a central server, therefore eliminating the SPoF and bottleneck issues present in CFL [8]. Instead, nodes communicate directly via P2P, as shown in Figure 2.2. This new approach has gained significant attention in recent years, as it solves some major problems present in CFL [27]–[30].

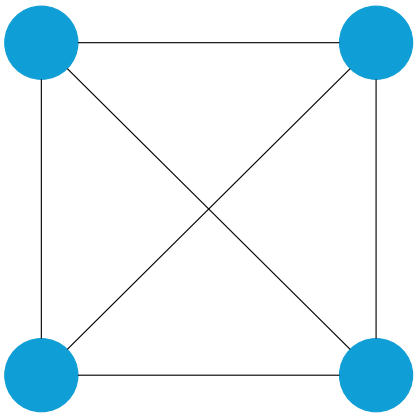


Figure 2.2: Decentralized Federated Learning (Full mesh topology)

In DFL each node $i \in \{1, 2, ..., n\}$ trains its local model w_i with their own private training dataset D_i, and after a training round is completed the node achieves an updated local model $w_i^{t+\frac{1}{2}}$. Instead of sending the model updates to a central server, the node will exchange its model updates with a select number of neighboring nodes, $\mathcal{N}_i$, determined by the defined network topology. After this exchange, the node aggregates the received model updates from its neighbors finally achieving a new state of their local model w_i^{t+1}, which will be used to continue training on their local data. Similar to in CFL, this process is repeated until a determined number of training rounds is completed or a consensus is reached [31].

Mathematically, this aggregation step is typically defined as a weighted sum of the neighbors' models following Decentralized Parallel Stochastic Gradient Descent (D-PSGD) protocol [32]:

$$w_i^{t+1} = \sum_{j \in \mathcal{N}_i \cup \{i\}} W_{ij} w_j^{t+\frac{1}{2}} \tag{2.2}$$

where $\mathcal{N}_i$ represents the set of neighbors of node i, and W_{ij} is the weight (or mixing value) assigned to the connection between node i and node j. This process repeats until convergence is reached [31].

2.3 Network Topology Definition

The network topology in a DFL system defines the structural interconnection and communication pathways between participating nodes. Formally, we represent the topology as an undirected graph $G = (V, E)$, where $V = \{1, \ldots, n\}$ denotes the set of n nodes in the system, and $E \subseteq V \times V$ represents the set of active communication links [5]. Within this structure, the *neighborhood* of a specific node i, denoted as $\mathcal{N}_i = \{j \in V \mid (i, j) \in E\}$, consists of all peers with whom node i can directly exchange information. The size of this neighborhood is defined as the node's degree, $d_i = |\mathcal{N}_i|$. This metric is particularly relevant in decentralized privacy analysis, as the degree determines the number of adversaries that may directly observe a specific user's updates.

Another way of describing network topologies is through matrix representations, which facilitate the algebraic analysis of the aggregation process. The structural connectivity is captured by the Adjacency Matrix $A = [a_{i,j}] \in M_n(\{0, 1\})$, where $a_{i,j} = 1$ if a connection exists between nodes i and j, and 0 otherwise. However, to perform the actual consensus algorithm, the system employs a *Mixing Matrix* (or Weight Matrix) $W \in^{n \times n}$. This matrix is fundamental to algorithms such as the D-PSGD. protocol [32], where each entry W_{ij} determines the weight node i assigns to the update received from node j. For the decentralized training to converge to the true global average, this matrix is typically designed to be *doubly stochastic*, meaning that both the rows and columns sum to 1 ($\sum_j W_{ij} = 1$ and $\sum_i W_{ij} = 1$).

2.4 Secure Aggregation

While FL allows for data to remain local within each participant's system, the transmission of model updates exposes nodes to privacy attacks, as sensitive information can be reconstructed from individual gradients. To mitigate this, *Secure Aggregation* (SA) protocols are employed to ensure that the central server computes only the aggregated result (e.g., the weighted average) of the user updates, without ever inspecting the individual contributions [33]. The state-of-the-art protocol for FL, proposed by Bonawitz, Ivanov, Kreuter, *et al.* [33], relies on Secure Multiparty Computation (MPC) designed specifically to handle high-dimensional vectors and the unreliability of mobile devices. The core mechanism is *pairwise additive masking*. For every pair of users (u, v), a common random seed $s_{u,v}$ is negotiated via Diffie-Hellman Key Exchange [34]. When masking their local model update w_u, user u adds the expanded mask of $s_{u,v}$ if $u < v$, and subtracts it if $u > v$. In a perfect execution where all users participate, these pairwise masks cancel out in the final summation, leaving only the sum of the private updates:

$$\sum_{u \in \mathcal{U}} y_u = \sum_{u \in \mathcal{U}} \left(w_u + \sum_{u<v} s_{u,v} - \sum_{u>v} s_{u,v} \right) = \sum_{u \in \mathcal{U}} w_u \tag{2.3}$$

However, a major challenge in FL is that devices frequently drop out during training. If a user drops out after exchanging keys but before uploading their update, the corresponding pairwise masks of surviving users would not be canceled, rendering the aggregation incorrect. To resolve this, the protocol integrates Shamir's t-out-of-n Secret Sharing [35]. During the setup phase, users distribute shares of their private secrets to each other. If a user drops out, the server can request these shares from the surviving users to reconstruct the necessary masks and subtract them from the aggregate.

To ensure privacy during this recovery phase, the protocol employs a "double masking" scheme. Users add a personal mask b_u in addition to the pairwise masks. The server is permitted to reconstruct the pairwise masks ($s_{u,v}$) only for users who dropped out, and the personal masks (b_u) only for users who survived. This guarantees that the server never possesses enough information to recover an individual honest user's update w_u, provided that at least a threshold t of users remain honest and online. Despite these cryptographic guarantees, the protocol remains efficient for deep learning, incurring a communication expansion factor of less than $2\times$ compared to sending plaintext updates [33].

2.5 Privacy Attacks

Despite the collaborative nature of FL systems, they remain vulnerable to various privacy attacks that aim to infer information or recover the original training data directly from the exchanged model updates [9], [36]–[38]. In this section, we address two critical privacy threats within FL systems, MIAs and GIAs.

2.5.1 Membership Inference Attacks

MIAs exploit the extent to which an ML model overfits to its training data. This phenomenon is reflected in the model's output, which typically exhibits higher prediction probabilities for data seen during training than for unseen data. Consequently, these attacks are formulated as a binary classification problem where an adversary aims to determine whether a given data sample x is a member or non-member of the training dataset D_i [39].

Depending on the adversary's access level to the victim model, these attacks are categorized into black-box and white-box approaches, and they rely either on training an auxiliary attack model or on evaluating prediction output against defined thresholds. In this dissertation, three distinct variants of these approaches are considered, the black-box attack introduced by Shokri et al. [40], the white-box approach developed by He et al. [41], and the metric-based modified prediction entropy method proposed by Song et al. [42].

Black-Box MIA

In a black-box scenario, the adversary can only observe the final output of the victim ML model, such as prediction logits, probabilities, or the predicted label, to infer membership [39]. This type of attack fundamentally exploits the extent to which an ML model overfits to its training data. Typically, an overfitted ML model produces systematically higher and more confident prediction probabilities for data seen during training (members) than for unseen data (non-members).

To capitalize on this behavior, the foundational black-box MIA proposed by Shokri et al. [40] relies on training class-specific binary classifiers, known as attack models, for each class

in the dataset. Each individual attack model is designed to perform membership inference exclusively on data samples belonging to the specific class it was trained to evaluate. These attack models are trained on a dataset containing the prediction vectors produced by an ML model, alongside their corresponding membership status (member or non-member). One way to generate this dataset is when the adversary has prior knowledge of some training samples from the victim. In this scenario, the adversary can train a local model with a similar architecture to the victim's model on these known data samples, subsequently querying it with both member and non-member data to extract the required prediction vectors. However, in more realistic cases where the adversary does not know any data samples from the victim, they must instead employ a shadow training technique. To do so, the adversary leverages a disjoint, private dataset, D_{adv}, to train N shadow models.

The data generation process operates as follows, for a given shadow model k (where $k \in \{1, \ldots, N\}$), a unique shadow dataset D^k_{shadow} is assembled. This dataset is evenly partitioned into a training set $D^{train,k}_{shadow}$ and a testing set $D^{test,k}_{shadow}$, both randomly sampled from D_{adv} with no overlap. Each shadow model is then trained relying solely on its respective $D^{train,k}_{shadow}$. Following this training phase, the full dataset D^k_{shadow} is passed through the model to collect softmax prediction vectors for member samples (originating from $D^{train,k}_{shadow}$) and non-member samples (from $D^{test,k}_{shadow}$). These resulting prediction vectors, combined with their true class labels and binary membership status, are aggregated into a comprehensive dataset. Because this specific attack architecture evaluates membership on a per-class basis, this global dataset is ultimately partitioned by class label. Subsequently, individual class-specific attack models are trained to accurately map the input prediction vectors to their corresponding binary membership status.

Once the class-specific attack models are fully trained, the adversary can evaluate any target data sample against the actual victim model. To perform the inference, the adversary queries the victim model with the target sample x, extracts its final softmax prediction vector, and identifies its corresponding class label. This prediction vector is then passed to the specific attack model trained for that particular class. Finally, the selected attack model processes the vector and outputs a prediction regarding the sample's membership status.

White-Box MIA

In contrast to black-box methods, white-box MIAs assume the adversary has privileged access to the victim model's internal parameters, such as weights and gradients. A prominent white-box attack, proposed by He et al. [41], leverages not only the victim model's final prediction outputs but also its internal state. By exposing these finer traces of data memorization embedded within the parameter updates during training, this approach enables significantly more powerful and accurate membership inference attacks.

Unlike the black-box approach, which requires multiple class-specific classifiers, this white-box attack utilizes a singular, global attack model to evaluate membership across all classes simultaneously. To train this comprehensive attack model, the authors assume that the adversary already possesses prior knowledge of a subset of the victim's actual training data (member samples). Under this assumption, the adversary instantiates a local model with the exact same architecture as the victim and trains it using these known data samples. By querying this local model with both the member samples and a disjoint set of non-member samples, the adversary can directly extract the necessary training features.

This attack extracts a richer set of features compared to black-box methods. Specifically, the adversary captures the parameter gradients originating from the layers which carries the most amount of information, typically the final convolutional layer and the last fully connected layer [41], alongside the final softmax prediction vector. Given the immense dimensionality of the raw gradient space, Principal Component Analysis (PCA) is applied to compress these gradient values into their two primary principal components [41]. These reduced, two-dimensional gradient summaries are subsequently concatenated with the corresponding softmax prediction outputs to form a definitive, unified feature set. These concatenated feature vectors function as the direct inputs for the global attack model, mapping the intricate correlations between the model's prediction confidence, its internal parameter updates, and the final membership status.

Once the global attack model is fully trained, the adversary can evaluate any target data sample against the actual victim model. To execute the attack, the adversary queries the victim model with the target sample x, extracts the resulting gradients from the specified key layers, and captures the final softmax output. Following the exact same preprocessing steps, the adversary applies PCA to compress the gradients and concatenates them with the prediction vector. Finally, this unified feature vector is fed into the attack model, which evaluates the extracted data to accurately infer the sample's membership status.

Modified Prediction Entropy MIA

While the previously discussed methods rely on training a dedicated attack model, alternative approaches utilize purely metric-based comparisons. A prominent example is the black-box attack proposed by Song et al. [42], which eliminates the need for a trained attack classifier. Instead, it evaluates a model's prediction uncertainty using a modified prediction entropy metric (M_{entr}), formulated as follows:

$$M_{entr}(x, y) = -(1 - F(x)_y)\log(F(x)_y) - \sum_{i \neq y}(1 - F(x)_i)\log(1 - F(x)_i) \tag{2.4}$$

Here, x is the target data sample, y represents its ground-truth label, and $F(x)$ is the prediction vector outputted by the victim model. Consequently, $F(x)_y$ denotes the confidence probability assigned to the correct class y, whereas $F(x)_i$ captures the probabilities distributed among the incorrect classes. While traditional entropy quantifies overall prediction uncertainty, M_{entr} explicitly factors in the true label, transforming it into a highly precise metric for detecting memorization. The underlying principle is that an overfitted ML model will exhibit significantly lower modified entropy for data it memorized during training compared to unseen data.

To execute this attack, membership is determined using class-dependent thresholds, if the computed M_{entr} for a target sample falls below the specific threshold established for its class, the adversary classifies the sample as a member. To derive these optimal thresholds without access to the victim's actual training data, the adversary relies once again on the shadow training technique [42]. First, a shadow model is trained on a dataset disjoint from the victim's. The adversary then queries this shadow model with known member and non-member data, calculating the M_{entr} for every resulting output. By analyzing the distribution of these values, the adversary can identify the best threshold that best separates members from non-members for each individual class.

Once these class-specific thresholds are established, the adversary can evaluate any target sample against the actual victim model. To perform the inference, the adversary queries the victim model with the target sample, extracts the prediction vector, and computes its M_{entr}. Finally, this value is compared against the corresponding shadow-derived threshold for the sample's class; if the computed modified entropy falls below this threshold, the target sample is classified as a member, and is otherwise classified as a non-member.

2.5.2 Gradient Inversion Attacks

A GIA represents a more severe form of privacy-disrupting attack specifically designed for collaborative ML paradigms, where it is common practice for participants to exchange model updates rather than raw data. Unlike MIAs, which merely attempt to infer whether a specific data sample was used during training, GIAs aim to completely reconstruct the original training data solely from these exchanged updates [9], [17], [38]. To demonstrate this vulnerability, we rely on the prominent attack proposed by Geiping et al. [9], which was explicitly developed to operate and compromise data privacy within FL systems.

The attack process begins when a victim node trains it's local ML model on its local training data and subsequently shares the resulting model update, w_i^t. However, to successfully execute the reconstruction, the adversary must first isolate the specific updates originating from the victim's local training data. Because clients in FL transmit updated model weights rather than raw gradients, the adversary deduces this individual contribution by computing the difference between the previous global model and the newly submitted model from the victim, $w_i^{t-1} - w_i^t$.

Once these individual updates are extracted, the adversary initiates the reconstruction process. First, the adversary generates a set of "dummy" inputs and labels, typically composed of random noise, which are then passed through a copy of the previously shared local model, $w_{i,dummy}^{t-1}$. By performing a forward and backward pass on this synthetic data, the adversary computes the corresponding dummy gradients. Next, the adversary calculates the cosine similarity to measure the distance between these generated dummy gradients and the victim's extracted updates [9]. Using this metric, the adversary iteratively updates the dummy data to minimize this distance. This optimization process is repeated over several iterations until the distance is minimized and the resulting dummy weights begin to get closer to w_i^t, indicating that the dummy inputs have successfully converged to match the victim's original private training data.

Chapter 3

Literature Review

This chapter presents a comprehensive literature review conducted to answer the research questions outlined in Chapter 1. This review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines for systematic reviews to ensure a rigorous and transparent approach [43].

3.1 Search Strategy

The search strategy for this literature review was designed to comprehensively capture relevant literature on the impact of network topologies in FL systems and their implications for privacy. The process involved several key steps: identification of databases, definition of search terms to address RQ1 and RQ2, establishment of inclusion and exclusion criteria, works selection, and a supplementary snowballing method to gather literature for RQ3.

3.1.1 Identification

The first step in the literature review process involved identifying and defining the search sources to be used while executing the literature review. For this work, four distinct academic databases were used to gather relevant articles, namely IEEE Xplore, Web of Science, the ACM Digital Library, and ScienceDirect (Table 3.1), considering works published in journals and conference proceedings.

Table 3.1: Databases Used for Article Search

Database	**URL**
IEEE Xplore	`https://ieeexplore.ieee.org`
Web of Science	`https://www.webofscience.com`
Association for Computing Machinery	`https://dl.acm.org`
ScienceDirect	`https://www.sciencedirect.com`

3.1.2 Definition of Search Terms

To investigate the impact of network topologies in FL systems and their implications for privacy, and specifically to gather literature to answer RQ1 and RQ2, a systematic search query was constructed. The search terms were derived to cover three distinct dimensions, the FL paradigm, the network topology, and privacy attacks.

To ensure a comprehensive retrieval of different network topologies, specific topologies such as "mesh", "ring", "tree", and "star" were explicitly included alongside general terms like "network topology" and "network structure". This approach was taken to ensure that studies focusing on individual configurations, rather than only the generic concept of network topology, were adequately captured.

Furthermore, to filter for results relevant to the security-centric research question, the query mandated the inclusion of terms related to privacy preservation and adversarial threats (e.g., "inference", "leakage", and "attack"). This constraint was applied to exclude research focused solely on the convergence speed or communication efficiency of topologies, thereby narrowing the results to those specifically discussing the intersection of network topology and security/privacy. Finally, the search was executed across all available metadata fields (including titles, abstracts, and keywords), requiring that the specified keywords appear in at least one of these metadata fields. The resulting query string is presented below:

("decentralized federated learning" OR "topology federated learning" OR ("topology" AND "federated learning"))*
AND
("topology" OR "network structure" OR "mesh" OR "ring" OR "tree" OR "star")
AND ("privacy" OR "attack" OR "inference attack" OR "leakage")

3.1.3 Study Selection

In order to filter and select the most relevant articles for this literature review, specific inclusion and exclusion criteria which are detailed in Tables 3.2 and 3.3 respectively. Articles which met at least one of the inclusion criteria were considered, while those that met any of the exclusion criteria were disregarded.

Table 3.2: Inclusion Criteria for Article Selection

ID	Inclusion Criteria
IC1	Papers published between 2019 and 2026.
IC2	Papers that explicitly analyze or compare specific network topologies (e.g., mesh vs. ring).
IC3	Papers that address privacy implications, attacks, or defenses related to the network structure.
IC4	Papers providing quantitative measurements of privacy or data leakage.

Table 3.3: Exclusion Criteria for Article Selection

ID	Exclusion Criteria
EC1	Studies focused solely on convergence speed or communication efficiency without a privacy analysis.
EC2	Studies focused exclusively on centralized topologies (standard FL) without decentralized comparison.
EC3	Papers where "topology" refers to the neural network architecture (Model Topology) rather than the communication graph.
EC4	Non-English papers, abstracts, or non-peer-reviewed pre-prints.

3.2 Results

Figure 3.1 illustrates the article selection process following the PRISMA guidelines. The methodology was split into two primary avenues to address the different research questions.

To address RQ1 and RQ2, database searches were conducted, resulting in an initial total of 1116 articles. The screening phase began with the removal of duplicate records, leaving 702 unique articles. Subsequently, a screening of titles and abstracts led to the exclusion of 652 articles, resulting in 50 articles being assessed for full-text eligibility. After applying the defined inclusion and exclusion criteria during the full-text review, 35 articles were excluded, yielding a set of 15 articles from the database search. To address RQ3, an alternative identification method was employed using snowballing, which successfully identified 4 additional relevant records. Combining the 15 records from the database searches and the 4 records from the snowballing process, a final set of 19 articles was included in this review.

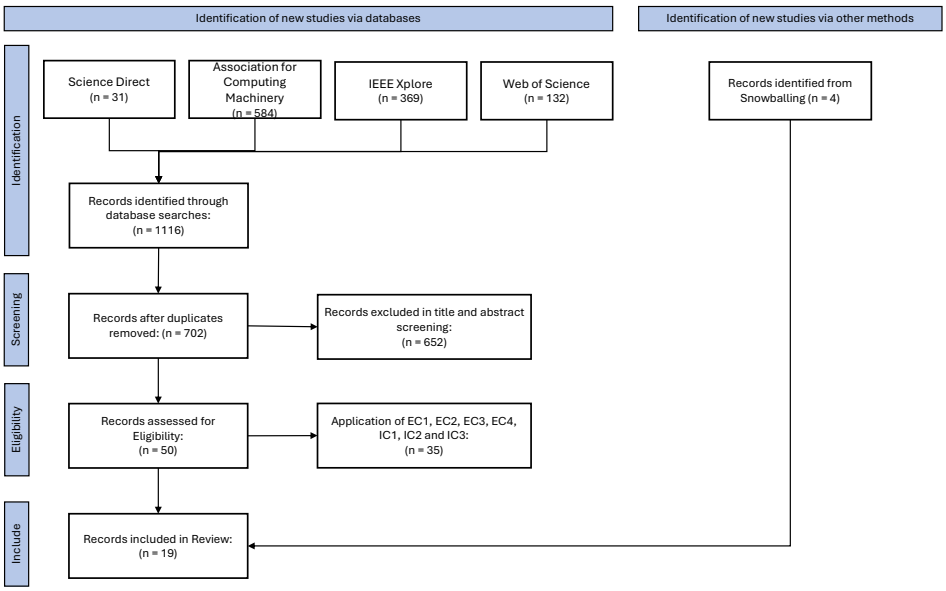


Figure 3.1: PRISMA Flow Diagram of the Literature Review Process

3.3 Network Topologies in Decentralized Federated Learning and Their Structural Properties

Within the context of DFL, the network topology defines the structural arrangement of nodes and the specific pathways used for communication, as explained in Chapter 2. The choice of a network topology is a critical factor, as it heavily influences the privacy, robustness, generalization capabilities, and overall performance of DFL systems [44]. Literature on the

subject generally categorizes these architectures into five primary types: star, line, ring, mesh, and tree.

It is important to note that while the systematic review process yielded a total of 15 primary studies, Table 3.4 focuses exclusively on the eight papers that address in detail these topological structures.

Table 3.4: Network topologies identified in the selected literature

Paper	Star	Tree	Line	Ring	Mesh
Yuan, Wang, Sun, *et al.* [8]	✓		✓	✓	✓
Martínez Beltrán, Pérez, Sánchez, *et al.* [44]	✓			✓	✓
Jin, Kannengießer, Rank, *et al.* [10]	✓	✓			✓
Khan, Khan, Rizwan, *et al.* [45]	✓	✓		✓	✓
Wu, Dong, Leung, *et al.* [5]	✓	✓			✓
Majeed and Hwang [7]	✓			✓	
Chen, Wang, Long, *et al.* [46]	✓	✓			✓

3.3.1 Star topology

The star network topology, the one used in CFL systems, is characterized by having a central node that connects to all participants and each participant only communicates with this central node. This topology is present and discussed by Wu, Dong, Leung, *et al.* [5], Majeed and Hwang [7], Yuan, Wang, Sun, *et al.* [8], Jin, Kannengießer, Rank, *et al.* [10], Martínez Beltrán, Pérez, Sánchez, *et al.* [44], Khan, Khan, Rizwan, *et al.* [45], and Chen, Wang, Long, *et al.* [46]. This topology, despite being a centralized architecture, can be implemented in a decentralized manner by rotating the role of the central node among all nodes in the network over different training rounds, as happens in the swarm learning framework [47]. In DFL this topology is characterized by the fact that only one node acts as an aggregator, while all other nodes train their models locally and send their updates to this aggregator node. Figure 3.2 presents an illustration of the star topology.

These previous works describe this topology as the most simple and straightforward to implement, since all nodes communicate exclusively with the central node, simplifying network management. However, having a central node introduces a SPoF in the system. Furthermore, this central node can become a significant communication bottleneck, creating scalability challenges as the network expands [45].

3.3.2 Tree topology

The tree network topology is characterized by a hierarchical structure where nodes are organized in distinct levels. This network topology is presented and discussed by Wu, Dong, Leung, *et al.* [5], Jin, Kannengießer, Rank, *et al.* [10], Khan, Khan, Rizwan, *et al.* [45], and Chen, Wang, Long, *et al.* [46]. In a DFL system, this network topology involves leaf nodes which perform local training and send their model updates to an intermediate edge node. This intermediate edge node aggregates the received updates from its leaf nodes and then forwards those aggregated updates to its own parent node; this process continues recursively until the root node is reached. The root performs the final aggregation and sends the new global model back down the tree. As noted by Wu, Dong, Leung, *et al.* [5], if a

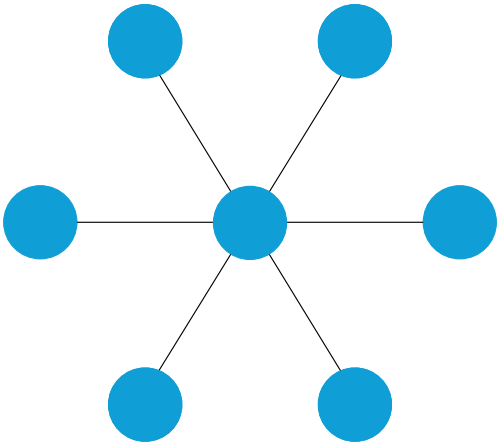


Figure 3.2: Star topology

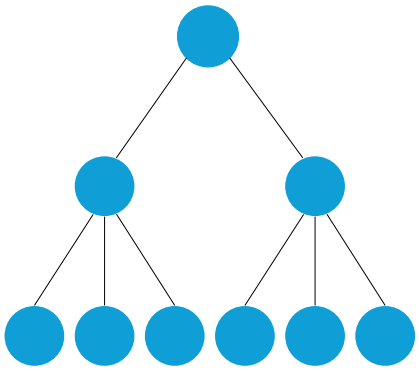


Figure 3.3: Tree topology

tree network topology is reduced to fewer than two hierarchical levels, it effectively becomes a star topology. Figure 3.3 presents an illustration of the tree topology.

These previous works describe this topology as a way to enhance scalability by distributing communication loads across multiple hierarchical layers, thereby reducing the burden on any single node. However, just as with the star topology, these studies emphasize that the tree network is susceptible to a SPoF, as the failure of the root node can disrupt the entire system. Furthermore, they point out that if an intermediate node fails, an entire branch of the network becomes isolated, preventing the downstream leaf nodes from participating in the training process.

3.3.3 Line topology

The line network topology arranges nodes in a sequential manner, forming a linear chain. This specific configuration was only discussed in detail by Yuan, Wang, Sun, *et al.* [8] among all works found in the literature. In a DFL context, this topology is characterized by a sequential communication pattern where each node performs local training and shares its model updates exclusively with its immediate successor in the chain. This propagation

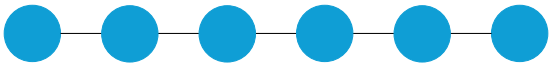


Figure 3.4: Line topology

continues node by node until the end of the line is reached. Figure 3.4 presents an illustration of the line topology.

This previous work describes this topology as simple to implement and maintain due to its linear structure. However, the authors point out that it suffers from significant drawbacks in terms of robustness and performance. They note that this topology is still vulnerable to a SPoF due to its sequential nature, since the failure of any single node will interrupt the communication chain for the entire system. Additionally, the study highlights that the sequential propagation of updates can lead to slow convergence rates, and updates from the nodes at the beginning of the line may become diluted by the time they reach the nodes at the end of the line.

3.3.4 Ring topology

The ring network topology addresses the connectivity limitations of the line network topology by establishing a cyclical connection among nodes. This topology is present and discussed by Majeed and Hwang [7], Yuan, Wang, Sun, *et al.* [8], Martínez Beltrán, Pérez, Sánchez, *et al.* [44], and Khan, Khan, Rizwan, *et al.* [45]. Similar to the line network topology, each node performs local training and shares its model updates with its immediate successor. However, in this configuration, the final node in the sequence connects back to the initial node, forming a closed loop. This cyclical exchange allows updates to circulate repeatedly until the system reaches convergence or completes a predefined number of full rotations. An illustration of the ring topology is presented in Figure 3.5.

Despite solving the convergence issue from the line network topology, the ring topology still suffers from the SPoF problem. If any single node fails, it breaks the communication cycle, effectively isolating parts of the network [8].

3.3.5 Mesh topology

The final primary network topology identified in the literature is the mesh network topology. This topology is distinguished by the absence of a rigid hierarchical or sequential structure, instead, nodes interconnect with multiple peers to form a complex, web-like arrangement. This topology is present and discussed by Wu, Dong, Leung, *et al.* [5], Yuan, Wang, Sun, *et al.* [8], Jin, Kannengießer, Rank, *et al.* [10], Martínez Beltrán, Pérez, Sánchez, *et al.* [44],

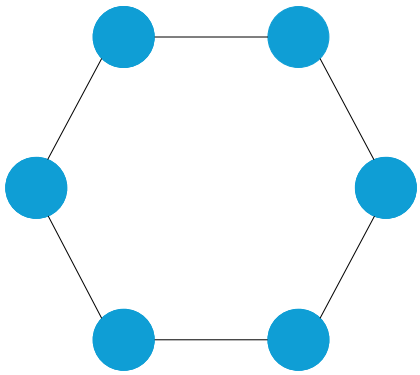


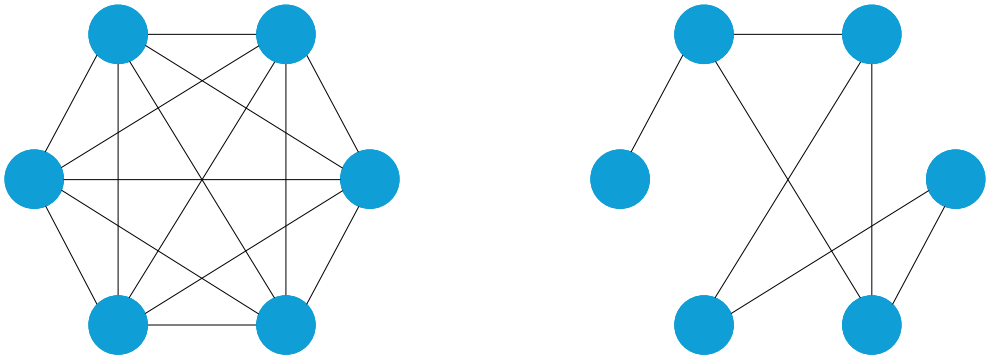
Figure 3.5: Ring topology

Khan, Khan, Rizwan, *et al.* [45], and Chen, Wang, Long, *et al.* [46]. Implementations of this topology spectrum from partially connected meshes, where nodes maintain links with a specific subset of peers, to fully connected meshes, where every node is directly connected to every other node in the network. Illustrations of both fully and partially connected mesh topologies are presented in Figures 3.6a and 3.6b, respectively.

Previous works note that this topology improves fault tolerance, as there is no SPoF in the system, the failure of an individual node rarely disrupts overall communication. However, it also introduces new challenges, such as more complex implementation and maintenance, especially as the network size increases. Furthermore, the communication overhead can increase significantly, as nodes are required to manage multiple simultaneous connections [8].

Beyond the general classification of fully or partially connected meshes, the specific degree distribution within the network topology fundamentally influences both system robustness and privacy guarantees. Li, Yu, Ji, *et al.* [21] differentiate between mesh topologies following a Poisson distribution and those following a Power Law distribution. Poisson network topologies resemble random structures with a relatively uniform degree distribution, where most nodes possess a connection count close to the average. Conversely, Power Law networks, which mirror many real-world social and internet architectures, are characterized by a few highly connected "hub" nodes.

In Table 3.5 is presented a summary of key aspects of every network topology discussed. This comparison highlights that no single architecture offers a perfect solution for DFL. While hierarchical and sequential structures such as star, tree, and line, often provide simpler implementation and lower initial overhead, they remain vulnerable to bottlenecks and system wide disruptions caused by a SPoF. Conversely, distributed topologies like the ring and mesh offer superior robustness and fault tolerance, which are essential for truly decentralized environments, but they introduce significant implementation complexity and higher communication costs. Consequently, the choice of topology represents a strategic trade-off that must align with the specific constraints of the deployment environment and the scenario of application, such as bandwidth availability, reliability requirements, and the desired level of decentralization.



(a) Mesh topology (fully connected)

(b) Mesh topology (partially connected)

Figure 3.6: Mesh Network Topologies

3.4 The Impact of Network Topology on System Privacy

Despite being designed as a privacy-preserving approach for ML, DFL systems remain susceptible to diverse privacy attacks due to the exchange of model updates among participants. Generally, these attacks aim to infer sensitive information regarding the local training datasets of honest nodes by analyzing shared model parameters or gradients [46]. These types of attacks can be influenced by the network topology adopted in the DFL system.

Li, Yu, Ji, *et al.* [21] studied the relationship between network topology and privacy leakage in DFL, under the assumption that the system can compute the average of gradients without the presence of any privacy-preserving mechanism, such as Differential Privacy (DP). The authors modeled a scenario featuring an "honest-but-curious" adversary who colludes with other nodes to intercept shared updates. The study's core contribution is that privacy loss in DFL is dependent on the "honest component" of the network topology, this being the subset of nodes that participate honestly in the training phase. The authors mathematically demonstrated that an adversary inevitably learns the partial sum of gradients within the honest component. By quantifying the size of this component as the number of nodes in the connected honest subgraph, they concluded that a larger honest component leads to lower privacy loss, as individual contributions are more effectively masked. To validate their theoretical findings, the authors compared two mesh topologies, one following a Poisson distribution and another following a Power Law distribution. Experiments were conducted using MIA and GIA on the CIFAR-10 dataset [48]. Their experimental results align with their theoretical analysis, as the honest component decreases, the effectiveness of both attacks increases. They also concluded that network topologies following a Poisson distribution provide better privacy guarantees.

Pasquini, Raynal, and Troncoso [16] present a critical re-evaluation of the privacy claims surrounding DFL, arguing that decentralization inherently expands the attack surface rather than reducing it. They refute the belief that DFL offers better privacy than CFL, showing that a decentralized adversary can achieve capabilities equivalent to a malicious central server. The authors identify two conflicting leakage sources, "local generalization" and "system knowledge". In sparse topologies (where nodes have few number of neighbors), updates propagate slowly, causing nodes to overfit to their local data (local generalization) in early

Table 3.5: Synthesis of Network Topologies in Decentralized Federated Learning

Topology	Structure & Communication	Key Advantages	Key Limitations
Star	Central Node connects to all nodes. Participants communicate exclusively with the central node.	Simple to implement. Straightforward network management.	Central node is a SPoF. Communication bottleneck. Scalability challenges.
Tree	Hierarchical levels (leaves $\rightarrow$ intermediates $\rightarrow$ root). Aggregation is performed recursively.	Enhances scalability. Distributes communication load across layers.	Root node is a SPoF. Intermediate node failure causes branch isolation.
Line	Sequential linear chain. Nodes share updates only with their immediate successor.	Simple implementation due to linear structure.	SPoF (any node failure breaks the chain). Slow convergence. Dilution of updates.
Ring	Cyclical connection (last node connects to first). Updates circulate repeatedly.	Solves the convergence/dilution issues of the Line topology.	SPoF (node failure breaks the communication cycle/isolates parts).
Mesh	Non-hierarchical, web-like structure. Nodes connect to multiple peers (partially or fully).	High fault tolerance. No SPoF. High robustness.	Complex implementation and maintenance. Significant communication overhead.

stages of training, which significantly increases vulnerability to MIA. Conversely, in dense topologies, adversaries gain "system knowledge" by observing multiple distinct neighbor updates. This increased visibility allows the adversary to approximate the baseline model of the system, by computing the difference between this baseline and the received updates, they can isolate and calculate the individual update from victim model. Furthermore, the study extends beyond the passive "honest-but-curious" model to analyze active adversaries. They introduce novel attacks such as the "echo attack", where an adversary reflects a victim's parameters to force overfitting, and the "state-override attack", which allows an attacker to gain full control over a victim's local model aggregation. They conclude that a fundamental trade-off exists: increasing network density to mitigate local generalization inevitably increases the adversary's system knowledge. Consequently, they argue that no network topology can provide better privacy than CFL without incurring communication costs that negate the efficiency benefits of decentralization.

In contrast to these findings, Ji, Heusdens, Maag, *et al.* [23] challenge the notion that decentralization does not bring privacy benefits. Instead, they argue that the privacy of DFL is

rigorously upper and lower bounded by the capabilities of CFL. Adopting the same "honest-but-curious" threat model, and considering a malicious central server in the case of CFL. Four distinct configurations were tested: CFL and DFL, both with and without Secure Aggregation (SA) ($CFL_{w/SA}$, $CFL_{w/oSA}$, $DFL_{w/SA}$, $DFL_{w/oSA}$). Through theoretical analysis, they concluded that $DFL_{w/oSA}$ consistently leaks less information than $CFL_{w/oSA}$. Furthermore, they found that when SA is applied, $CFL_{w/SA}$ provides the strongest theoretical privacy guarantees, whereas $DFL_{w/SA}$ offers a robust middle ground. These results were validated experimentally on the CIFAR-10, CIFAR-100 [48], and MNIST [49] datasets. By quantifying the quality of reconstructed images using Structural Similarity Index Measure (SSIM), they confirmed that image reconstruction is most successful in $CFL_{w/oSA}$, while $DFL_{w/oSA}$ and $DFL_{w/SA}$ maintained better privacy preservation. Finally, they highlighted the role of network topology density in DFL. They also studied the effect of graph density concluding that increasing the topology density leads to higher privacy leaks as each node exposes its gradients to more neighbors, the extreme case is a fully connected mesh, which approaches the leakage levels of $CFL_{w/oSA}$. However, if SA is applied, this trend is reversed, SA inputs from additional neighbors effectively mask individual contributions, thereby enhancing privacy.

Shifting the focus to optimization-based DFL protocols, approaches that formulate the collaborative learning task as a constrained optimization problem Yu, Li, Lopuhaä-Zwakenberg, *et al.* [20] offer a counter-narrative to the vulnerabilities exposed by Pasquini, Raynal, and Troncoso [16]. The authors investigate distributed solvers, such as Alternating Direction Method of Multipliers (ADMM) and Primal-Dual Method of Multipliers (PDMM). In these methods, an auxiliary variable is initially exchanged between neighbors through a secure channel. During training, each node uses this variable to update its local model. Subsequently, the variable is updated, and the nodes only exchange the difference between the old and new values across the network. This mechanism makes it unnecessary to directly exchange the model weights or raw gradients. Conducting a rigorous information-theoretic analysis, they demonstrate that the privacy loss in these DFL schemes is theoretically upper bounded by the loss in CFL. The authors argue that unlike standard approaches where direct local gradients are exposed, these optimization-based protocols only reveal gradient differences and aggregated sums. This effectively breaks the direct mapping between model updates and training data, obscuring the label information necessary for high-fidelity reconstruction and rendering analytical label reconstruction infeasible. This theoretical stance is supported by empirical validations using logistic regression and Deep Neural Networks (DNNs). While the authors observed no privacy advantage for simple logistic regression models, their experiments on DNNs showed that GIA and MIA were significantly less effective against DFL than CFL. In particular, SSIM measurements indicated a notable degradation in reconstruction quality within the decentralized setting. Finally, they highlight a topological advantage similar to SA; as the number of honest neighbors increases, the aggregation of updates acts effectively as an increased batch size. This dilutes the specific information available to adversaries, further widening the privacy gap between DFL and CFL.

Continuing the analysis of gradient-based attacks, Lu, Xiong, Li, *et al.* [22] investigated the resilience of DFL against reconstruction techniques like GIA, proposing a framework named DEFEAT. They argue that the structural absence of a central server in DFL creates a fundamental barrier to gradient inversion compared to centralized approaches. By analyzing the attack surface where malicious nodes attempt to recover data from their neighbors, the authors emphasize that privacy is heavily modulated by network topology. Notably, their findings regarding graph density contrast with the baseline results of Ji, Heusdens, Maag, *et*

al. [23]. While the latter found that density increases leakage in the absence of SA, Lu, Xiong, Li, *et al.* [22] observe that increasing graph density actually enhances privacy protection. They attribute this to the aggregation process itself, as a node aggregates parameters from a larger number of neighbors, individual contributions are obfuscated within the sum. This effectively creates a "hiding in the crowd" effect, a benefit that Ji, Heusdens, Maag, *et al.* [23] only observed when SA was explicitly enabled. Furthermore, the authors identify a critical trade off between convergence and privacy. They demonstrate that schemas prioritizing rapid global consensus (via multi-hop communication) degrade privacy by allowing attacker models to closely approximate victim models, whereas single-hop communication schemas maintain significantly higher resistance to data reconstruction.

In conclusion, the relationship between network topology and privacy in DFL remains a subject of active debate, as summarized in Table 3.6. While some studies argue that decentralized structures inherently expand the attack surface specifically by allowing adversaries to exploit high connectivity for "system knowledge", others demonstrate that dense topologies can actually mitigate leakage through a "hiding in the crowd" effect, when applying mechanisms like SA, ADMM or PDMM. Consequently, there is no single "optimal" topology for privacy, rather, the security of the system depends on a complex interplay between graph density, the presence of honest neighbors, and the specific aggregation protocol used.

3.5 The Evolution of Privacy Risk During Training

Analyzing how privacy threats like MIAs and GIAs evolve across training rounds is essential, as vulnerability in distributed systems changes dynamically with the model's state. MIA success is tightly coupled with generalization [16], [39], these attacks present higher accuracies on overfitted ML models. Conversely, reconstruction threats like GIAs depend on the magnitude of shared updates [38], which carry less invertible information as the model approaches convergence and updates shrink. Consequently, tracking these distinct risks over successive rounds is critical to identifying periods of peak vulnerability in FL systems.

Yin, Li, Bai, *et al.* [11] explicitly acknowledge that the accuracy of a MIA in FL, following a star topology, is highly dependent on the training round. Their research demonstrates that attack accuracy peaks at a specific intermediate "key round". The authors identify this key round as the exact stage of model overfitting, typically occurring in the middle to late stages of training, where the difference in prediction confidence between member and non-member data samples becomes most pronounced. At this level of overfitting, the model's accuracy on training data continues to rise while its performance on unseen data staggers. Because of this difference, evaluating the model at this intermediate stage allows an adversary to separate members from non-members far more effectively than if they only analyzed the final model.

The work of Pasquini, Raynal, and Troncoso [16] aligns with this previous idea. Evaluating their work on DFL systems following mesh topologies, they demonstrated that as local models are overfitted to their training datasets, MIA becomes more effective, even though they did not evaluate the attacks directly across sequential training rounds. This behavior dictates the accuracy of the attack. In the early-to-middle phases of training, when models across the system have not yet finished training, MIAs executed directly on the received updates are successful because the parameters are biased toward the local training sets.

While decentralized systems exhibit early-stage vulnerabilities due to local parameter biases, the risk of continuous training is equally evident in standard FL. By tracking the evolution

Table 3.6: Summary of Literature on Network Topology and Privacy in Decentralized Federated Learning

Paper	**Focus**	**Impact of Topology / Density**	**Key Conclusion**
Li, Yu, Ji, *et al.* [21]	Topology-dependent bounds (No DP)	Privacy improves with larger "honest components". Uniform degree distributions offer better protection than skewed ones.	Privacy loss is mathematically tied to the size of the honest subgraph, larger honest groups mask individual contributions better.
Pasquini, Raynal, and Troncoso [16]	Critical Re-evaluation & Active Attacks	Sparse topologies lead to local overfitting. Dense topologies increase adversary "system knowledge".	Decentralization expands the attack surface. No topology offers better privacy than CFL without prohibitive costs.
Ji, Heusdens, Maag, *et al.* [23]	Info. Theoretic Analysis (DFL vs CFL)	Without SA, higher density increases leakage. With SA, higher density improves privacy via neighbor masking.	$DFL_{w/oSA}$ leaks less than $CFL_{w/oSA}$. SA provides a strong middle ground for privacy in DFL.
Yu, Li, Lopuhaä-Zwakenberg, *et al.* [20]	Optimization-based (ADMM/PDMM)	More honest neighbors effectively increase batch size, diluting information available to adversaries.	Optimization protocols (revealing sums/diffs rather than gradients) break direct data mapping, making DFL safer than CFL for DNNs.
Lu, Xiong, Li, *et al.* [22]	DEFEAT Framework (Gradient Inversion)	Higher density enhances privacy ("hiding in the crowd") via aggregation. Multi-hop schemas degrade privacy.	Structural absence of a server hinders inversion. A trade-off exists between convergence speed (multi-hop) and privacy (single-hop).

of model updates across successive rounds, Gu, Bai, and Xu [50] and Zhu, Li, Gu, *et al.* [51] demonstrated that the effectiveness of MIAs grows rapidly during the early phases of training as the model starts to memorize training data. As the global model continues to train and reaches a high degree of overfitting, the attack accuracy eventually stabilizes at its peak in the later stages. However, despite providing valuable insights into how continuous observation over time enhances MIAs, these studies predominantly limit their scope to centralized star topologies.

Huang, Chen, and Roos [52] investigated GIAs in FL systems under a standard star topology. They found that these attacks are more accurate during the early rounds of training. In these initial phases, the gradients shared by a client are representative of their raw local data because the model parameters have not yet been heavily affected by the global model updates of other users. As the training progresses, local updates become increasingly entangled with the network's global state, which diminishes the reconstruction capabilities of the attacks in later rounds.

In conclusion, the relationship between the training process and privacy vulnerability remains highly dynamic, as summarized in Table 3.7. While some studies demonstrate that MIAs reach their peak effectiveness during intermediate key rounds or the later stages of training due to accumulated model overfitting, others show that privacy risks can maximize much earlier. For instance, in decentralized partial mesh topologies, early-to-middle rounds expose clients to severe leakage due to local parameter biases. Furthermore, across all topologies, reconstruction threats like GIAs, which remain under-explored in comparison to MIAs, are consistently most dangerous at the very beginning of training before gradients are diluted by model aggregations. Consequently, there is no single safe phase during training, the security of the system depends on a complex interplay between the current training round, the underlying network topology and the specific attack being executed.

Table 3.7: Summary of Literature on the Temporal Evolution of Privacy Risks During Training

Paper	**Stage of Highest Attack Accuracy**	**Topology Studied**
Yin, Li, Bai, *et al.* [11]	MIA accuracy peaks at a specific intermediate "key round" that marks the exact onset of model overfitting, where the confidence gap between members and non-members is the most pronounced.	Star
Pasquini, Raynal, and Troncoso [16]	MIAs are most effective during the early-to-middle phases of training, before the network has converged, because the model parameters are still heavily biased toward the local training sets.	Partial Mesh
Gu, Bai, and Xu [50]	The effectiveness of MIAs grows rapidly during the initial rounds, but the accuracy ultimately peaks and stabilizes in the later stages of training once the global model reaches a high degree of overfitting.	Star
Zhu, Li, Gu, *et al.* [51]	Corroborates that MIA effectiveness grows early on and peaks in the later stages of training, utilizing the temporal tracking of client updates across multiple sequential communication rounds.	Star
Huang, Chen, and Roos [52]	GIAs achieve higher accuracy during the early rounds of training, as the shared gradients are highly representative of the raw local data and have not yet become heavily entangled with the global state of the network.	Star

3.6 Summary

This chapter presented a comprehensive Systematic Literature Review (SLR) following the PRISMA guidelines, analyzing 15 selected articles to understand the intersection of network topology and privacy in DFL. The review was structured around three key dimensions, the typology of network topology, their inherent privacy implications, and the adaptation of privacy-preserving mechanisms.

Regarding the first research question, the literature classifies DFL topologies into five primary categories: star, tree, line, ring, and mesh. The analysis reveals a fundamental trade off between structural robustness and management complexity. While centralized topologies (star, tree) offer simplicity, they are plagued by SPoF and bottlenecks. Conversely, decentralized topologies (mesh) provide high fault tolerance but incur significant communication overheads. Linear and ring topologies serve as intermediate solutions but often suffer from latency and slower convergence rates.

In addressing the second research question, the review highlights that there is no consensus in the literature regarding whether centralized or decentralized topologies offer superior privacy. The findings present a polarized landscape, while some theoretical analyses argue that DFL inherently expands the attack surface, suggesting that its privacy is strictly upper-bounded by CFL capabilities, other studies demonstrate that specific DFL configurations can outperform centralized approaches by obscuring the direct mapping between gradients and raw data. Consequently, the literature indicates that neither topology is universally advantageous, instead, privacy guarantees are highly contextual, depending on the interplay between topology density, the presence of secure aggregation mechanisms, and the size of the "honest component" within the network.

Finally, concerning the third research question, the review demonstrates that the temporal progression of the training process affects system's vulnerability to privacy attacks. The literature reveals that there is no universally "safe" phase during training. For MIAs, which exploit data memorization, the risk is intrinsically tied to overfitting. In centralized star topologies, MIA accuracy typically peaks at intermediate training rounds or stabilizes at high levels during the later stages of training. However, in decentralized mesh topologies, MIAs are highly effective much earlier, during the early-to-middle rounds, because model parameters remain heavily biased toward local training sets before the network reaches consensus. Conversely, reconstruction threats like GIAs, which remain under-explored in the literature compared to MIAs, pose the greatest risk at the very beginning of the training process across all topologies. During these initial phases, shared updates are highly representative of raw local data and have not yet been entangled or diluted by global aggregations. Ultimately, the efficacy of privacy attacks evolves dynamically as the model trains, dictating that robust defense mechanisms must be both topologically and temporally aware to secure FL systems across their entire training.

Despite the valuable insights provided by these studies, this literature review reveals a critical gap in the current research landscape, there is no thorough, empirical evaluation of these privacy attacks across different network topologies. While theoretical frameworks identify various structures, such as ring, line, tree, and mesh, and debate their implementation complexity and characteristics, a dedicated privacy evaluation of these configurations is missing, with existing concrete attack implementations still defaulting to the standard centralized star topology. Furthermore, the few studies that do explore decentralized architectures often limit their scope to comparing a star topology against mesh topologies, largely ignoring other topologies found within the literature. Consequently, the complex, real-world interplay between the specific network structure, the temporal stage of the training process, and the practical execution of both MIAs and GIAs remains deeply under-explored. This fundamental gap underscores the need for comprehensive, cross-topology vulnerability assessments to accurately measure, understand, and mitigate privacy risks in practical DFL deployments.

Chapter 4

Assessing Privacy Risks Across Network Topologies

This chapter presents the experimental methodology designed to evaluate privacy risks across diverse network topologies. It first defines the problem statement and the proposed solution, establishing the scope of the empirical evaluation. Subsequently, it details the experimental setup, the selected topologies, datasets, and data distributions, before defining the threat models and evaluation metrics utilized to systematically quantify system vulnerabilities against both MIAs and GIAs.

4.1 Problem Statement

While FL inherently minimizes data exposure by keeping training data localized, the exchange of intermediate model updates introduces significant vulnerabilities to privacy threats, specifically MIAs and GIAs. To mitigate the bottlenecks and single points of failure inherent in standard CFL, the research community has increasingly shifted towards DFL, introducing a variety of network topologies such as rings, lines, trees, and meshes. These structures offer distinct trade-offs regarding communication overhead, fault tolerance, and convergence speed. However, as established in the previous chapter, the privacy implications of deploying these alternative network structures remain highly debated and empirically under-explored.

The fundamental problem is that current privacy evaluations in federated environments predominantly default to the star topology. The literature often evaluates the meshes topology, ignoring other topologies such as the line, ring and tree topology. Consequently, there is no consensus on whether decentralization inherently expands the attack surface or mitigates leakage through the aggregation of multiple honest neighbors.

Furthermore, this topological ambiguity is compounded by the temporal dynamics of the training process. Privacy leakage in FL is not a static metric, it fluctuates continuously as the global model converges. Existing studies indicate that GIAs pose the highest risk in the initial training rounds before gradients become entangled with the global state. Conversely, MIAs peak at different stages depending on the topology, exploiting local model biases during early to middle rounds in decentralized meshes, while relying on model overfitting during intermediate or late rounds in centralized stars.

Despite these isolated insights, there is currently no unified, empirical study that systematically evaluates how these variables intersect. The lack of comprehensive, cross topology vulnerability assessments leaves a critical blind spot in the deployment of DFL systems. Therefore, this dissertation aims to solve this problem by designing and executing a robust

empirical evaluation to quantify the efficacy of both MIAs and GIAs across different network topologies. By evaluating these threats at various stages of training, this dissertation seeks to uncover the precise interplay between network structure, temporal progression, and privacy leakage, ultimately guiding the design of more secure and resilient federated learning systems.

4.2 Proposed Solution

To address the limitations and ambiguities identified in the current literature, this dissertation proposes a comprehensive and systematic empirical study to evaluate the impact of network topology on privacy in FL. Unlike existing studies that predominantly focus on the standard star topology, our solution provides a comparison across a diverse network topologies. Specifically, we will evaluate six distinct network topologies: star, tree, line, ring, full mesh, and partial mesh.

To execute this evaluation, we implemented six different FL systems, each following a distinct network topology, under controlled and identical environmental conditions. This approach ensures that the structural configuration is the sole independent variable, allowing for an accurate isolation of its effects on privacy leakage.

Within each FL system, the training process will be conducted using two standard machine learning benchmarking datasets, MNIST [49] and CIFAR-10 [48]. These datasets provide a baseline to evaluate attacks on both simple and more complex data distributions. Furthermore, to accurately reflect the real-world heterogeneity of decentralized environments, the experiments will be executed under both Independent and Identically Distributed (IID) and Non-Independent and Identically Distributed (Non-IID) data distributions. Incorporating both distribution types is critical, as data heterogeneity directly influences convergence speed, the degree of local overfitting, and the distinctiveness of shared gradientsall of which are primary drivers for the success of both MIAs and GIAs. By systematically deploying these attacks across the six topologies, multiple datasets, and varying data distributions throughout different stages of the training lifecycle, this proposed solution will provide a definitive, empirical understanding of how network structures govern privacy risks in FL.

4.2.1 Federated Learning and ML model

To assess the data leakage associated with various network topologies in FL systems, as previously mentioned, six FL systems following distinct network topologies were implemented. In these implementations, the line and ring topologies are strictly unidirectional, model updates propagate sequentially from node 0 to node 1, then to node 2, and so on. Every configuration was tested using different datasets under both IID and Non-IID data distributions. Specifically, in an IID scenario, the data is uniformly shuffled and distributed, ensuring that each client's local dataset contains a balanced, representative sample of all classes from the global distribution. Conversely, the Non-IID scenario reflects a more realistic, heterogeneous environment where clients hold highly unbalanced subsets of data. By altering only the network topology across these systems while keeping other variables constant, we can effectively isolate and evaluate its specific influence on data privacy.

To ensure a fair comparison, the training protocol is kept strictly uniform across all implemented systems. The base number of training nodes is set to six for every configuration. However, the star and tree topologies require additional nodes. Specifically, the star topology

incorporates a seventh node that functions as a central server dedicated entirely to model aggregation. The tree topology requires three extra nodes, two edge servers designated for intermediate aggregations, and one central server that exclusively aggregates the models received from these edge servers. Every node holds a specific local partition of the global dataset, which is utilized to execute local training on its corresponding model.

To examine the impact of irregular connectivity, the partial mesh topology was designed using a non-uniform configuration. In this setup, node 0 connects to nodes 1, 3, and 5, while node 1 links with nodes 0, 3, and 4. Node 2 communicates with nodes 3 and 4, whereas node 3 connects to nodes 0, 1, and 2. Additionally, node 4 maintains connections with nodes 1 and 2, and finally, node 5 connects solely to node 0. Resulting in node degrees ranging from one to three, this configuration provides a rigorous foundation for evaluating how localized structural bottlenecks influence both model weight synchronization and the resulting privacy leakage. This partially connected structure is depicted in Figure 4.1.

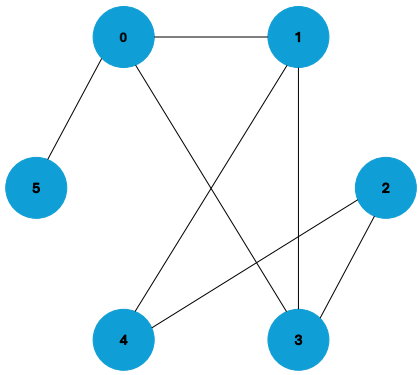


Figure 4.1: Partially connected mesh topology used in simulation

The ML model employed in our experiments is based on the architecture adopted by Shokri et al. [40]. It consists of a Convolutional Neural Network (CNN) comprising two convolutional blocks (each featuring a 3×3 kernel, a Tanh activation function, and 2×2 max pooling), followed by a fully connected layer with 128 units, and a final classification layer producing a logits vector.

To assess the vulnerability of these setups to MIA, we apply two different aggregation algorithms suited to the specific architecture of each topology. The FL systems using the star and tree topologies, which rely on a centralized aggregation, use the FedAVG aggregation method. In contrast, the decentralized topologies, specifically the line, ring, full mesh, and partial mesh topologies, utilize D-PSGD. For this assessment, model training is conducted with the Adam optimizer, utilizing a learning rate of 1×10^{-3}, a batch size of 64, and the cross-entropy loss function. Regarding the overall training duration, systems employing FedAVG are evaluated over 36 training rounds, with clients configured to execute 5 local epochs prior to each aggregation step, whereas systems utilizing D-PSGD are trained for 3500 rounds.

On the other hand, evaluating GIA necessitates specific modifications to the experimental configuration. Within the literature, the majority of GIA implementations depend on satisfying particular conditions to function effectively [9], [17]–[19], [38]. These prerequisites include reducing the batch size, restricting models to train on individual batches while

sharing gradients, possessing knowledge of the true labels for the target data samples, and most notably, adopting a simpler optimizer such as Stochastic Gradient Descent (SGD). Consequently, a primary adjustment is switching from the Adam optimizer to SGD, configured with a learning rate of 1×10^{-2}, a weight decay of 5×10^{-4}, and a momentum of 0.9. Another significant modification is lowering the batch size from 64 to 8, which helps mitigate the impact of duplicated labels [19] within the chosen datasets, both containing 10 classes. When multiple samples in a batch share the same label, their features become entangled, preventing standard gradient inversion methods from distinguishing between the individual samples, typically resulting in the reconstruction of blended, meaningless features rather than distinct images. Since a batch size of 64 across only 10 classes guarantees severe label collisions, reducing the batch size to 8 minimizes this probability and preserves the unique feature representations required for successful reconstruction. Ultimately, the final adjustment alters the core mechanics of the FedAVG aggregation method, training nodes are now restricted to training on a single batch prior to the aggregation phase, effectively replicating the experimental design of Pasquini, Raynal, and Troncoso [16]. Finally, to ensure a consistent and comprehensive evaluation, the total training duration for all topologies in this GIA setup is standardized to 27,000 rounds.

4.2.2 Datasets and Data Distribution

To train the ML models across the previously established FL systems and evaluate potential privacy leakages, two standard datasets were utilized. The first is the MNIST [49] dataset, which contains 70,000 images of handwritten digits ranging from 0 to 9. The second is the CIFAR-10 [48] dataset, comprising 60,000 32×32 color images categorized into 10 distinct classes.

Regarding data allocation, the global training dataset is partitioned into mutually exclusive subsets among the six training nodes, ensuring that each node operates entirely on its own private local data. In contrast, a common, fully balanced testing dataset is shared among all nodes. This global test set is held out strictly for validation purposes and does not influence the training process, providing an unbiased benchmark to assess and compare the generalization capabilities of the models as they train. To evaluate the impact of data heterogeneity, the local training datasets are allocated using two distinct distribution strategies. The first approach utilizes an IID setup, in which the data is evenly distributed across all participants. Although this ideal distribution is uncommon in practical FL deployments, it serves a precise experimental function: by removing data heterogeneity as a variable, this baseline enables the isolation and quantification of privacy vulnerabilities stemming solely from graph connectivity and topological structure.

Conversely, the second approach employs a Non-IID partitioning scheme to replicate the realistic, heterogeneous environments typical of practical FL deployments. To simulate this scenario where participants possess highly unbalanced data subsets, the label skew across local clients was generated using a Dirichlet distribution [53] configured with a concentration parameter of $\alpha = 0.5$. This methodology naturally yields varying data volumes and skewed class representations, creating a substantial discrepancy between local and global data distributions. By forcing the models to adapt to highly client-specific features, this challenging learning environment inherently induces local overfitting. Consequently, it provides a rigorous test to evaluate how data skewness aggravates privacy risks and impacts the performance of MIAs and GIAs against unique, non-representative data samples.

4.2.3 Threat Model

We consider an honest-but-curious adversary who faithfully executes the defined training protocol while actively attempting to infer private information from other participants. Strict data isolation is enforced across all systems, consequently, no node, including the adversary, has direct access to, or prior knowledge of, the local datasets held by other nodes. Instead, the adversary relies entirely on the model updates transmitted during the training rounds to execute the attacks. Because the adversary's capabilities are inherently linked to their position within the network, we evaluated these topologies through mutually exclusive test runs. In each iteration, exactly one node was designated as the adversary, strictly targeting the incoming updates from its direct topological neighbors.

Throughout the training phase, the designated adversary independently assesses the vulnerability of each adversary-victim pair by analyzing the intercepted model updates. The formation of these pairs is strictly defined by the network's topology. In decentralized setups, the adversary targets its connected nodes. In hierarchical topologies, such as the star and tree topologies, central, root or edge nodes act as the adversaries targeting their neighbor nodes.

To comprehensively evaluate privacy leakage across these diverse architectures, we subjected the intercepted updates to both MIA and GIA. The specific mechanics and implementation details of these attacks are outlined in the following subsections.

4.2.4 Membership Inference Attacks

To evaluate the different systems against MIA we implemented three different variants of the attack. The black-box attack introduced by Shokri et al. [40], the white-box approach developed by He et al. [41], and the modified prediction entropy method proposed by Song et al. [42]

Black-Box MIA

The black-box MIA follows the methodology outlined by Shokri et al. [40], which necessitates the creation of class-specific attack models. To generate the necessary training data for these classifiers, the adversary employs a shadow training strategy.

Specifically, the adversary leverages a disjoint, private dataset, D_{adv}, to train 20 shadow models. Although the original formulation utilized up to 100 shadow models, empirical testing demonstrated that training 20 shadow models achieves comparable efficacy for both the MNIST and CIFAR-10 datasets while significantly reducing computational overhead. For a given shadow model k (where $k \in \{1, \ldots, 20\}$), a unique shadow dataset D^k_{shadow} is assembled. This dataset is evenly split into a training set $D^{train,k}_{shadow}$ and a testing set $D^{test,k}_{shadow}$, both randomly sampled from D_{adv} with no overlap. Each shadow model is then trained for 50 epochs relying solely on its respective $D^{train,k}_{shadow}$. Following this training phase, the full D^k_{shadow} is passed through the model to collect softmax prediction vectors for member samples (originating from $D^{train,k}_{shadow}$) and non-member samples (from $D^{test,k}_{shadow}$). These resulting prediction vectors, combined with their true class labels and binary membership status, are joined into a comprehensive dataset. Because the attack architecture requires identifying membership on a per-class basis, this global dataset is ultimately partitioned by class label. Subsequently, the individual attack models are trained to map the input prediction vectors to their corresponding binary membership status.

Architecturally, each attack model consists of a Multi Layer Perceptron (MLP) featuring a single 64-neuron hidden layer equipped with a ReLU activation function, concluding with a two-logit linear output layer. The optimization is handled by the Adam optimizer, utilizing a cross-entropy loss function and a learning rate of 1×10^{-3}.

White-Box MIA

For the white-box evaluation, we adapted the MIA proposed by He et al. [41], which assumes the adversary can inspect the victim model's internal parameters. While the original attack assumes the adversary possesses a subset of the victim's actual training data, we constrain the adversary to a more rigorous threat model where only their private dataset is available. To compensate for this limitation, the adversary once again relies on the previously described shadow training technique, although the amount of shadow models is reduced to 10.

Different from the black-box method, this attack extracts both the model's internal parameter gradients and their softmax prediction vectors. Specifically, the adversary captures gradients originating from the final convolutional layer and the ultimate fully connected layer. Given the immense dimensionality of this gradient space, PCA is employed to compress these values into their two primary principal components [41]. These 2D gradient summaries are subsequently concatenated with the corresponding softmax prediction outputs.

Following the exact implementation described by He et al. [41], this white-box attack employs a singular, global attack model to evaluate membership across all classes, contrasting with the class-specific classifiers used in the black-box setting. The concatenated feature vectors function as the direct inputs to predict the binary membership target. To accurately map the intricate correlations between predictions and parameter updates, the attack model utilizes a deeper, unified MLP architecture. It incorporates two hidden layers containing 64 and 32 neurons, respectively, both utilizing ReLU activations. The network terminates in a single neuron with a sigmoid activation, outputting a continuous membership probability score.

Modified Prediction Entropy MIA

In contrast to model-based techniques, we also evaluated the metric-driven attack proposed by Song et al. [42]. This technique eliminates the need for a trained attack model, relying instead on a modified prediction entropy metric (M_{entr}), formulated as follows:

$$M_{entr}(x, y) = -(1 - F(x)_y) \log(F(x)_y) - \sum_{i \neq y} (1 - F(x)_i) \log(1 - F(x)_i) \tag{4.1}$$

Here, x is the target data sample, y represents its ground-truth label, and $F(x)$ is the prediction vector outputted by the victim model. Consequently, $F(x)_y$ denotes the confidence probability of the correct class y, whereas $F(x)_i$ captures the probabilities distributed among the incorrect classes. While traditional entropy quantifies overall prediction uncertainty, M_{entr} factors in the true label, transforming it into a highly precise metric for detecting memorization.

To conduct this attack, membership is determined using class-dependent thresholds. While the authors utilized shadow training to approximate these thresholds [42], we opted to derive the optimal thresholds directly from the victim model's actual predictions. By doing so, we bypass estimation errors and establish a theoretical upper bound on the adversarys attack accuracy. Therefore, the outcomes of this modified prediction entropy attack represent a

worst-case scenario for data privacy, rather than a strictly practical deployment of the attack. Assessing this rigorous upper bound is critical for security evaluations, as it guarantees that any identified privacy risks reflect the maximum potential information leakage of the system, rather than the limitations of the attacker's methodology.

4.2.5 Gradient Inversion Attacks

To evaluate the system's vulnerability to data reconstruction across different architectures, we implemented the GIA proposed by Geiping et al. [9]. As detailed in Section 2.5.2, successfully executing this attack requires the adversary to extract the victim's individual pseudo-gradients by computing the difference between a baseline model state and the victim's submitted model weights. In our evaluation, the specific method for establishing this baseline and extracting the difference depends heavily on the underlying network topology.

In hierarchical architectures, such as the star and tree topologies, this extraction process is straightforward. The adversarial server simply calculates the difference between the global model from the previous round and the victim's newly exchanged model. Specifically, the adversary computes $w^t_{global} - w^{t+1}_i$ in the star topology, or $w^t_{root} - w^{t+1}_i$ in the tree topology, to precisely isolate the local gradient contribution of node i.

For decentralized topologies, where model aggregation is performed by the D-PSGD aggregation method, the extraction strategy adapts to the adversary's visibility within the network. In a full mesh topology, every node is connected to all other participants. Consequently, at the end of a training round t, the models are effectively synchronized across the network. This convergence allows the adversary to treat its synchronized state as a baseline, successfully isolating the victim's individual contribution by calculating $w^t_{global} - w^{t+1}_i$.

In a partial mesh topology, the adversary lacks complete visibility over the entire network. To adapt, the attacker applies a similar differencing approach but calculates the global model approximation using only the subset of models received from its direct neighbors. The victim's contribution is then extracted relative to this partial baseline.

Finally, the line and ring topologies enforce strict connectivity constraints. In these configurations, we assume the adversary cannot access or intercept model updates from nodes beyond its explicitly defined immediate neighbors. Because establishing a global or robust localized baseline is impossible, the adversary must rely on temporal changes. To approximate the pseudo-gradients, the attacker computes the difference between two consecutive model updates transmitted by the same victim node i, expressed as $w^t_i - w^{t+1}_i$.

Chapter 5

Membership Inference Attack Results

This chapter presents the empirical results obtained from the experimental setup described in Chapter 4 regarding MIA. The primary objective is to quantify how the choice of a network topology affects the effectiveness of this attack in FL environments.

5.1 Star Topology

Figure 5.1 displays the average training and testing accuracies across all training nodes during the 36 communication rounds of the FL system following the star topology. It is evident that model performance is highly dependent on both datasets complexity and data distribution. On the simpler MNIST dataset, models achieve rapid convergence, reaching training and testing accuracies close to 1.0 with a negligible generalization gap under both IID and Non-IID configurations. In contrast, the more intricate CIFAR10 dataset presents a substantial generalization gap, pointing to severe overfitting. Although its training accuracy consistently near 1.0, testing accuracy plateaus around 0.65 in the IID environment and decreases to roughly 0.50 in the Non-IID scenario. This emphasizes that for complex visual tasks within a FL framework, data heterogeneity strongly impacts the model's capacity to generalize.

Regarding the attack outcomes, Figure 5.2 depicts the mean attack accuracy averaged across every adversary-victim pair for the three MIAs tested within the FL architecture. These findings reveal a clear correlation between the generalization gap shown in Figure 5.1 and the model's practical susceptibility to membership inference.

On the CIFAR10 dataset, attack accuracies increase rapidly over the first 10 communication rounds, closely reflecting the model's overfitting behavior. The modified prediction entropy MIA stands out as the most effective threat, recording the highest overall accuracy and reaching a peak of about 0.85 under the Non-IID configuration. Moreover, compared to the IID setup, the Non-IID distribution typically worsens privacy leakage for both the black-box MIA and the modified prediction entropy MIA. This suggests that highly imbalanced local datasets compel the model to intensely memorize client-specific characteristics, thereby rendering training samples more identifiable. As an interesting exception, the white-box MIA performs significantly worse in the Non-IID environment than in the IID case for CIFAR10.

Conversely, the evaluated MIAs prove mostly ineffective on the MNIST dataset, where the majority of attack accuracies remain close to random guessing (between 0.50 and 0.55). This observation perfectly matches the previously noted absence of overfitting for this dataset. A

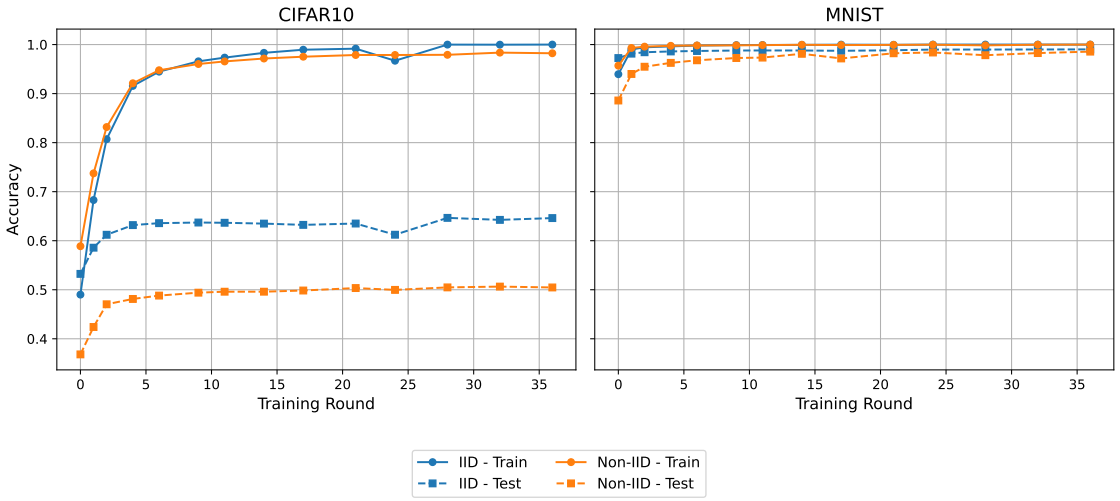


Figure 5.1: Training And Testing Accuracies on Star Topology

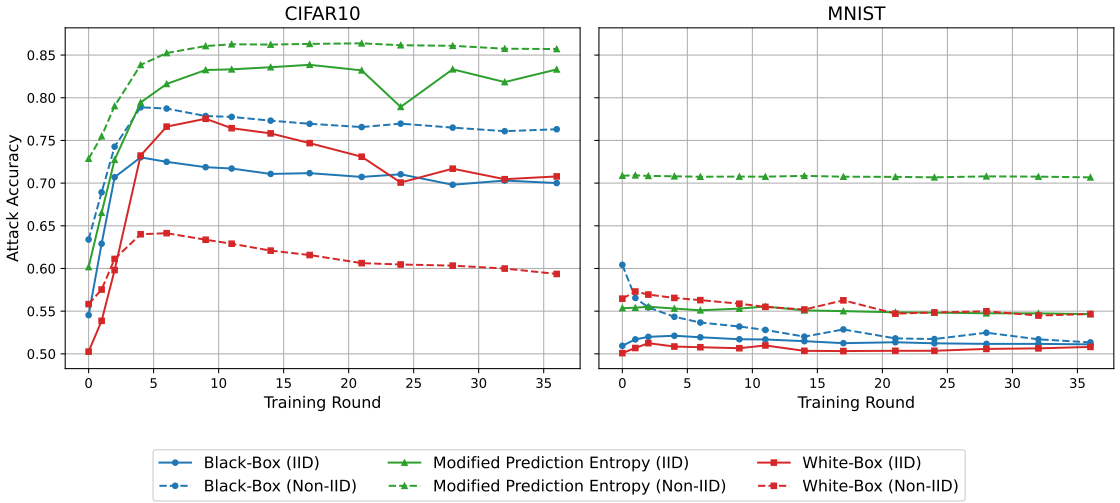


Figure 5.2: Mean Attacks Accuracy in Star Topology

significant exception, however, is the modified prediction entropy MIA in the Non-IID setup, which maintains a stable and relatively high accuracy of about 0.70 throughout all training rounds. This suggests that even if a model exhibits good generalization, the severe data heterogeneity of a Non-IID split produces unique confidence score patterns that a specialized modified prediction entropy MIA can effectively leverage for membership inference.

5.2 Tree Topology

Figure 5.3 illustrates the average training and testing accuracies for all training nodes throughout the 36 communication rounds within the tree topology. Much like in the star topology, the outcomes emphasize a performance difference based on dataset complexity. On the less complex MNIST task, models achieve rapid convergence with an insignificant generalization gap, reaching accuracies close to 1.0 under both IID and Non-IID conditions. Conversely, the more complicated CIFAR10 dataset demonstrates considerable overfitting, while training accuracies close in on 1.0, testing accuracies level off at a much lower point.

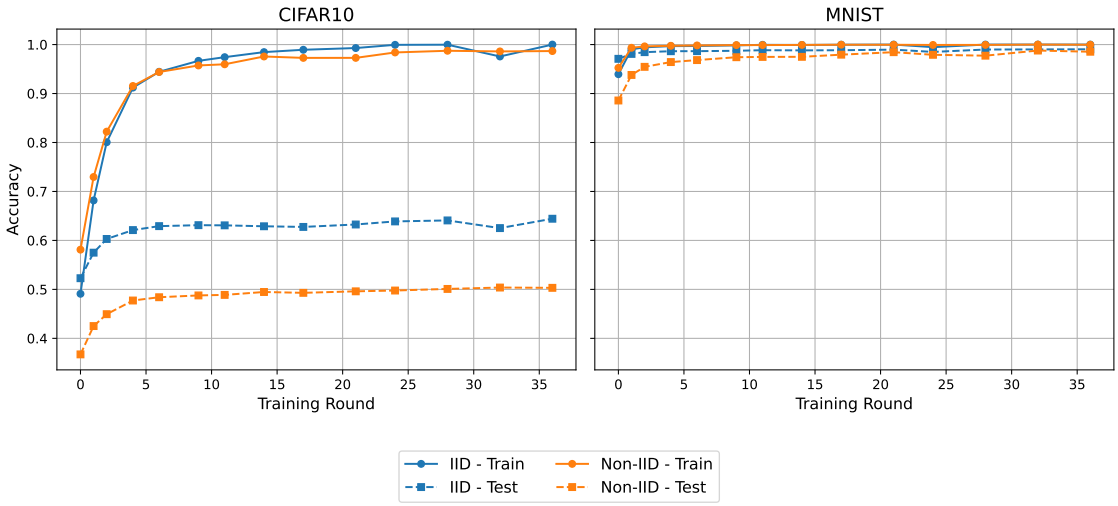


Figure 5.3: Training And Testing Accuracies on Tree Topology

Moreover, the effect of data heterogeneity is still very noticeable for CIFAR10, where testing accuracy hits roughly 0.65 in the IID setting but is drastically hindered in the Non-IID configuration, plateauing near 0.50. This shows that although the tree topology effectively facilitates collaborative learning, it continues to face the same core generalization issues when dealing with complex tasks and Non-IID data partitions.

Figure 5.4 displays the average attack accuracy, for each adversary-victim pair, when intermediate edge nodes function as adversaries targeting specific leaf nodes. Under these circumstances, the privacy leakages heavily resemble those seen in the star topology (Fig. 5.2). Since edge nodes obtain direct, non-aggregated model updates from their corresponding clients, the security flaws continue to be severe. On the CIFAR10 dataset, attack accuracies escalate quickly as overfitting begins, and the modified prediction entropy MIA once more emerges as the most potent threat, reaching a high of roughly 0.85 in the Non-IID configuration. Likewise, the MNIST dataset proves mostly resilient to the majority of attacks, except for the modified prediction entropy MIA in the Non-IID environment, which maintains an accuracy of about 0.70. This verifies that at the foundational tier of the hierarchy, clients experience the same direct vulnerability to MIAs as they would in a standard FL system.

On the other hand, Figure 5.5 depicts the threat environment when the central root server conducts MIAs on the semi-aggregated updates submitted by edge nodes. At this level, the architectural advantages of the tree topology are clear. For the CIFAR10 dataset, there is a visible reduction in general attack accuracy, especially within Non-IID setups. Most prominently, the highest accuracy for the modified prediction entropy MIA under Non-IID conditions falls drastically from 0.85 (at the edge tier) to about 0.70 (at the root tier). This decrease highlights that the intermediate aggregation executed by middle nodes functions as an inherent privacy defense, successfully masking the granular, client-specific memorization patterns before they arrive at the main server. Nevertheless, this defense is not foolproof; under the IID configuration, the modified prediction entropy MIA progressively increases to almost 0.80 across the 36 rounds. This indicates that if data features are evenly spread among clients, semi-aggregated models can still gradually expose membership details.

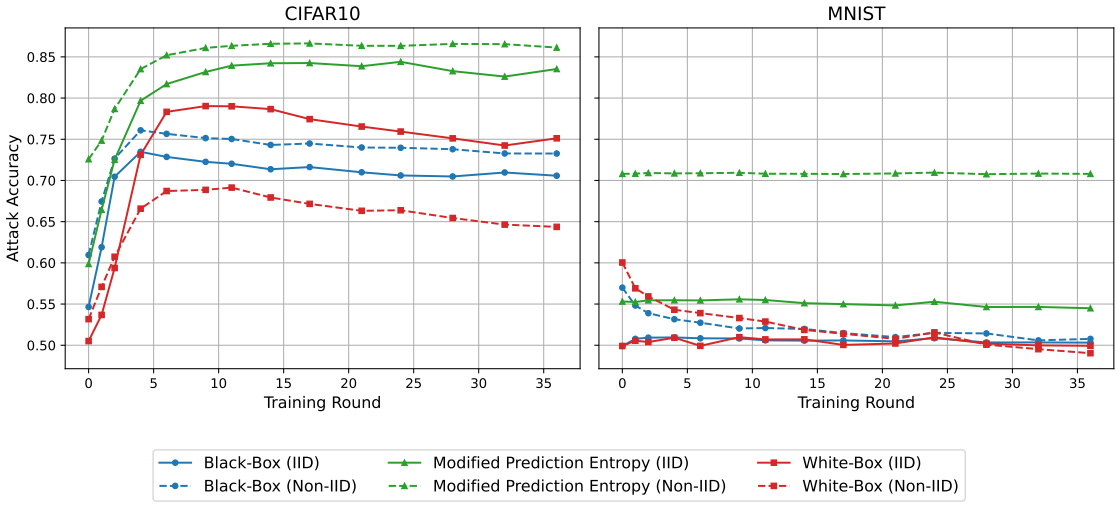


Figure 5.4: Mean Attacks Accuracy in Tree Topology (Edges attacks Leafs)

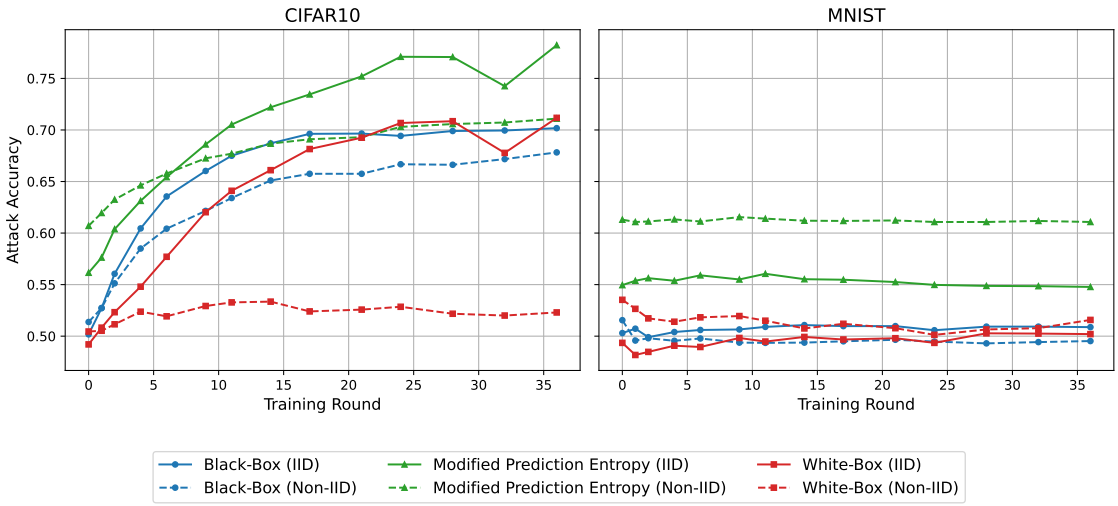


Figure 5.5: Mean Attacks Accuracy in Tree Topology (Root attacks Edges)

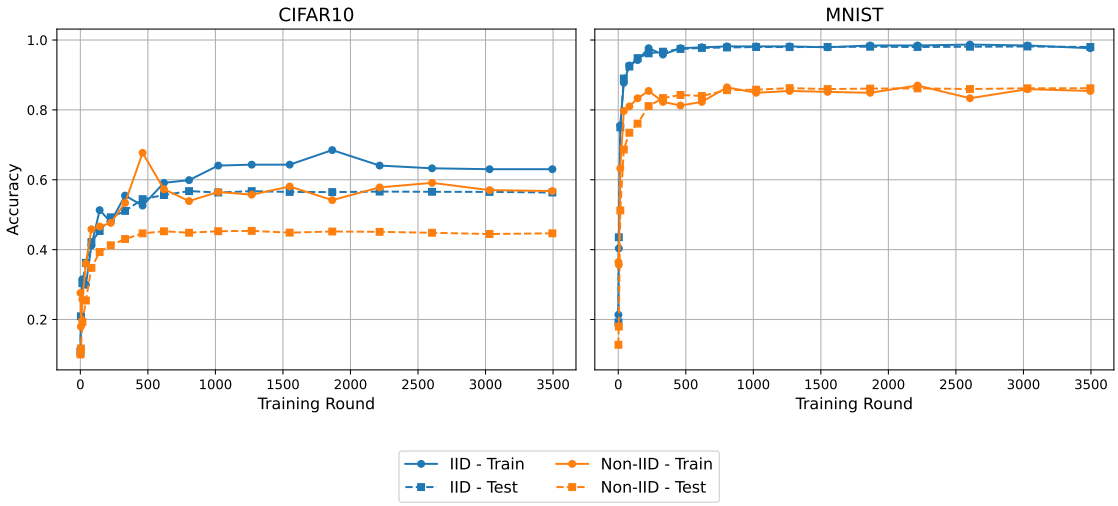


Figure 5.6: Training And Testing Accuracies on Line Topology

5.3 Line Topology

In Figure 5.6 is depicted the average training and testing accuracies of all nodes in the line topology through several training rounds. Like in previous topologies, the accuracy of the model depends heavily on the complexity of the dataset. On the CIFAR10 dataset, we can appreciate that the model stabilizes at around 0.60 accuracy on both data distribution types. On the MNIST dataset, once again accuracies peak at around 1.0 on the IID data distribution, whereas the Non-IID data distribution stabilizes at a slightly lower peak of approximately 0.85. Furthermore, on the MNIST dataset, the training and testing curves remain closely aligned throughout the entire process, indicating good generalization without significant overfitting.

Regarding the attack outcomes for the line topology, Figure 5.7 illustrates the average attack accuracy for the three MIAs evaluated within the line topology. On the CIFAR10 dataset, attack accuracies exhibit a steady increase over the communication rounds, closely mirroring the model's overfitting trend. The modified prediction entropy MIA emerges as the most significant threat, achieving the highest overall accuracy and peaking at approximately 0.75 under the Non-IID configuration. Furthermore, the Non-IID distribution consistently aggravates privacy leakage compared to the IID setup, particularly for the black-box and modified prediction entropy MIAs. The white-box MIA proves to be the least effective on this dataset, though it is worth noting that it performs marginally better in the Non-IID environment than in the IID case, remaining below 0.60 accuracy for both.

Conversely, on the MNIST dataset, the evaluated MIAs are largely ineffective, with the majority of attack accuracies hovering near random guessing (between 0.50 and 0.55). This aligns seamlessly with the previously noted lack of overfitting and high generalization for this simpler dataset. However, a prominent exception is the modified prediction entropy MIA under the Non-IID setup, which maintains a stable and elevated accuracy of nearly 0.70 from the very first communication rounds. This reaffirms that even in the absence of severe overfitting, the high data heterogeneity inherent to Non-IID partitions generates distinct confidence score patterns that a specialized modified prediction entropy MIA can effectively exploit to compromise privacy.

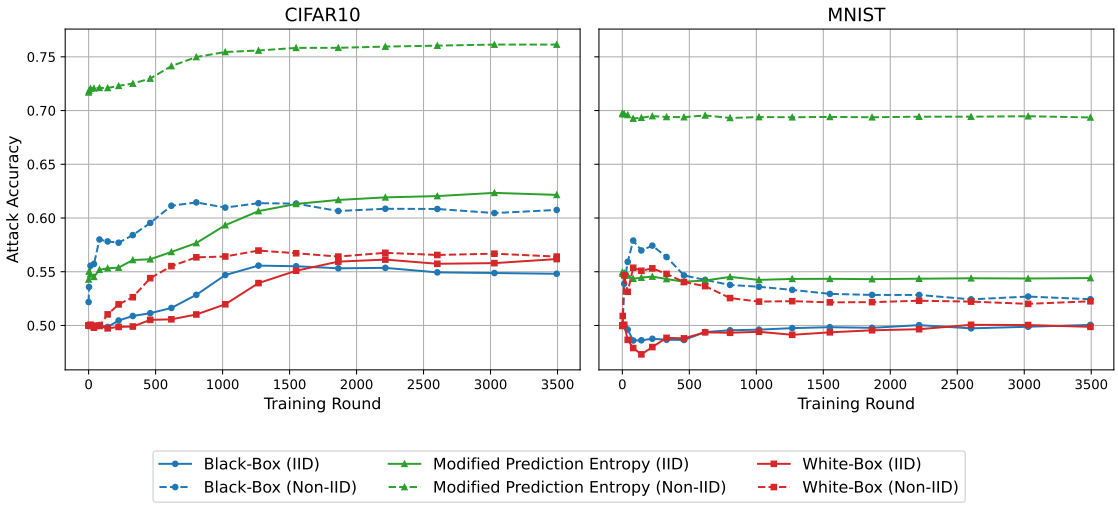


Figure 5.7: Mean Attacks Accuracy in Line Topology

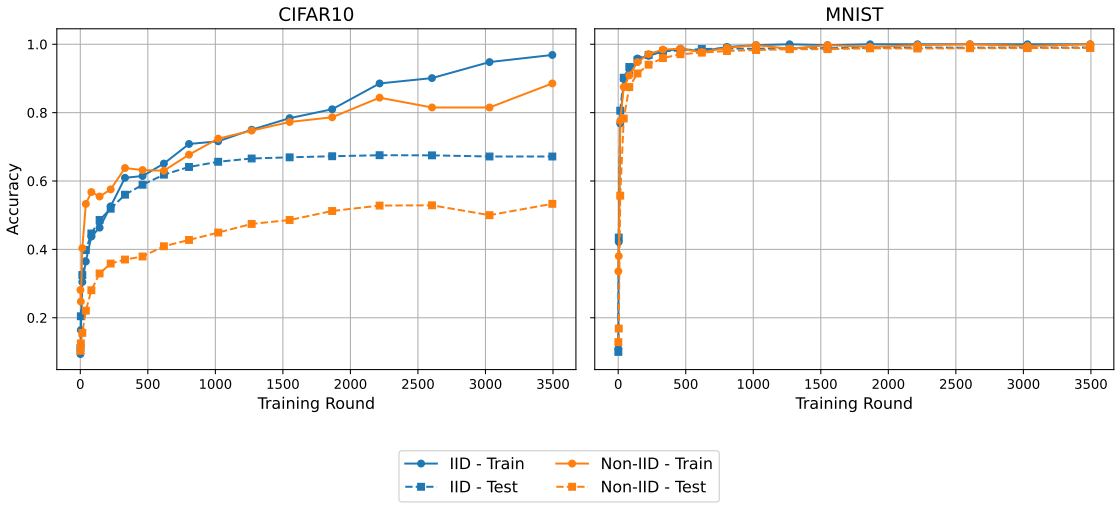


Figure 5.8: Training And Testing Accuracies on Ring Topology

5.4 Ring Topology

Figure 5.8 presents the average training and testing accuracies for all nodes in the ring topology across the communication rounds. On the CIFAR10 dataset, we can observe a pronounced overfitting. While the training accuracy steadily climbs, nearly reaching 1.0 for the IID setup and approaching 0.90 for the Non-IID setup, the testing accuracies plateau early on. The IID test accuracy stalls around 0.65, and the Non-IID test accuracy fails to surpass 0.55, creating a substantial generalization gap. Conversely, on the MNIST dataset, the model exhibits excellent generalization; both the training and testing curves remain tightly aligned and quickly converge to approximately 1.0 under both data distribution types, indicating no significant overfitting.

Regarding the membership inference threat, Figure 5.9 illustrates the average attack accuracies evaluated within this topology. On the CIFAR10 dataset, the continuous rise in attack accuracies clearly mirrors the growing generalization gap observed during training. The modified prediction entropy MIA operating in the Non-IID configuration stands out as

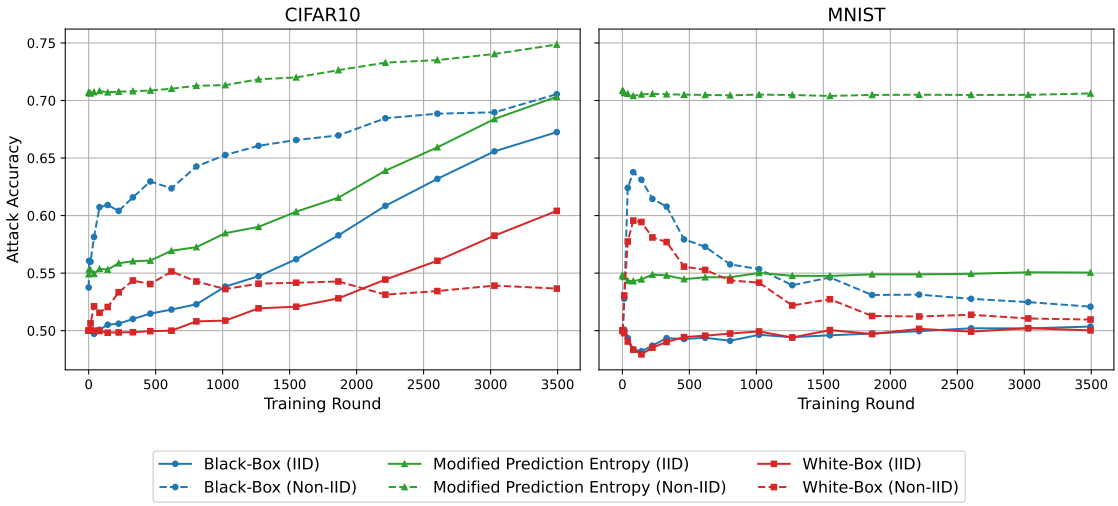


Figure 5.9: Mean Attacks Accuracy in Ring Topology

the most severe threat, steadily climbing to peak at nearly 0.75. Other approaches, such as the black-box and modified prediction entropy MIAs under the IID distribution, also show a pronounced upward trajectory, reaching approximately 0.70. Interestingly, the white-box MIA demonstrates a significant discrepancy, while it shows a moderate upward trend under IID conditions (reaching roughly 0.63), it remains mostly flat and ineffective in the Non-IID scenario, hovering below 0.55.

On the MNIST dataset, the robust generalization of the model successfully mitigates most MIAs, bringing their accuracies down to a random guessing baseline of approximately 0.50 after a brief spike in the earliest rounds. However, the prominent exception remains the modified prediction entropy MIA in the Non-IID setting, which persistently sustains an accuracy of roughly 0.70 from the beginning to the end of the training process. This further corroborates that the extreme data heterogeneity of a Non-IID split causes the model to produce distinct confidence score distributions that specialized attacks can reliably exploit, entirely independent of the model's overfitting status.

5.5 Full Mesh Topology

In Figure 5.10 we can observe the mean training and testing accuracies of nodes in the full mesh topology. Regarding the CIFAR10 dataset, under IID conditions, the mean node's accuracy almost reaches 1.0 at the end of training. However, the testing accuracy for this same setup stalls at approximately 0.70, indicating some level of overfitting. Similarly, under the Non-IID configuration, the model struggles with a noticeable generalization gap, the training accuracy approaches 0.70, while the testing accuracy rounds 0.55. On the MNIST dataset, by contrast, the training and testing curves overlap almost perfectly, rapidly converging to near 1.0 accuracy for both the IID and Non-IID setups. This demonstrates excellent generalization and a complete absence of overfitting.

Turning to the membership inference vulnerabilities, Figure 5.11 displays the mean attack accuracies within this topology. On the CIFAR10 dataset, the increasing attack accuracies over the communication rounds heavily reflect the model's overfitting trend. The modified prediction entropy MIA under the Non-IID distribution poses the greatest threat, steadily

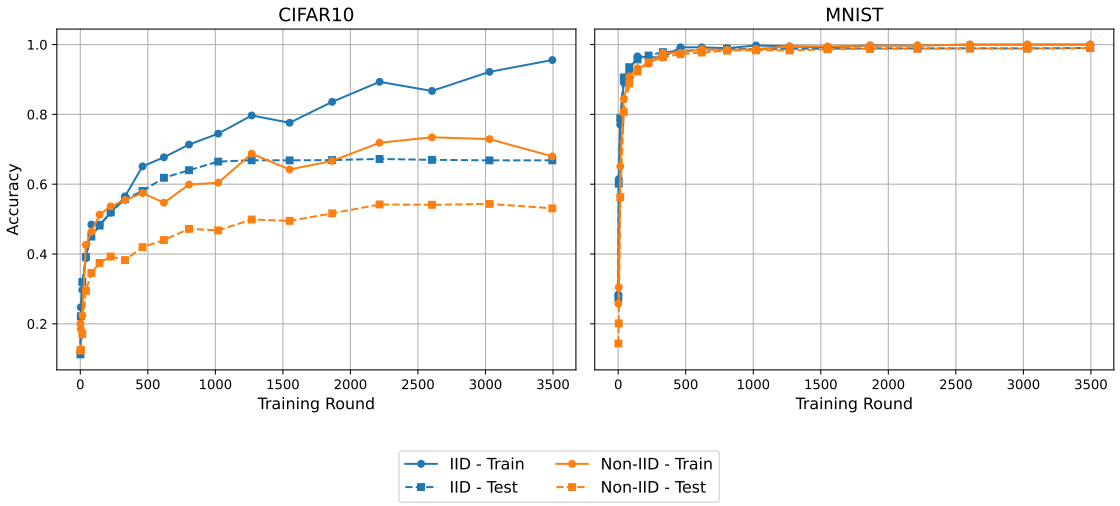


Figure 5.10: Training And Testing Accuracies on Full Mesh Topology

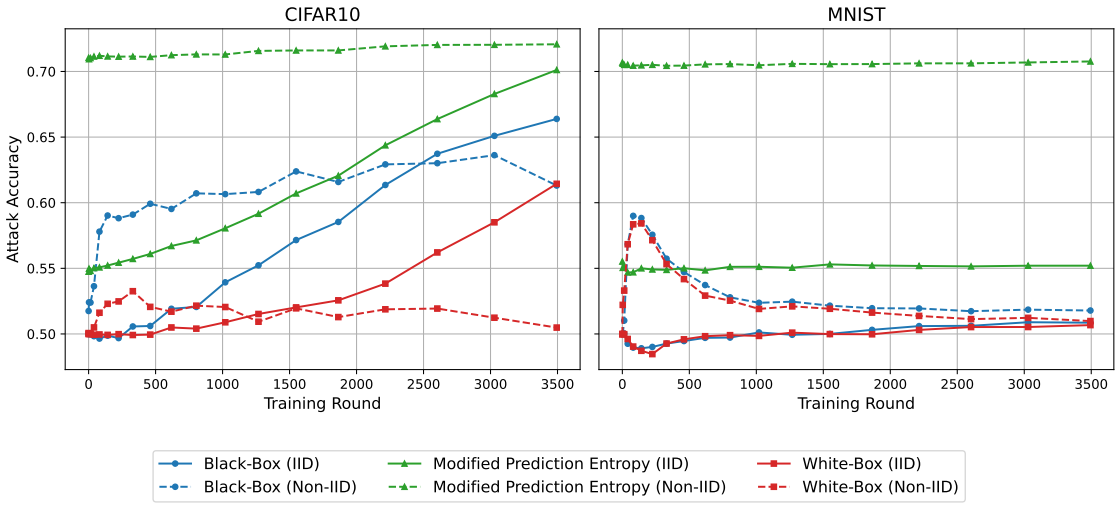


Figure 5.11: Mean Attacks Accuracy in Full Mesh Topology

climbing to peak at approximately 0.75. In the IID setting, both the modified prediction entropy and black-box MIAs exhibit a pronounced upward trajectory, reaching roughly 0.72 and 0.67, respectively. Notably, the white-box MIA diverges significantly in its effectiveness depending on the data distribution, while it rises to around 0.60 under IID conditions, it remains mostly flat and fails to surpass 0.50 in the Non-IID scenario.

Conversely, on the MNIST dataset, the strong generalization of the model renders most MIAs ineffective. After a brief initial perturbation during the early rounds, the accuracies for the black-box and white-box attacks quickly regress to the random guessing threshold of 0.50. However, the modified prediction entropy MIA in the Non-IID configuration consistently maintains a disproportionately high accuracy of over 0.70 throughout the entire training process. This persistent vulnerability confirms that even when a model generalizes perfectly without overfitting, the severe data heterogeneity intrinsic to Non-IID distributions produces distinct confidence scores that a specialized modified prediction entropy attack can easily exploit.

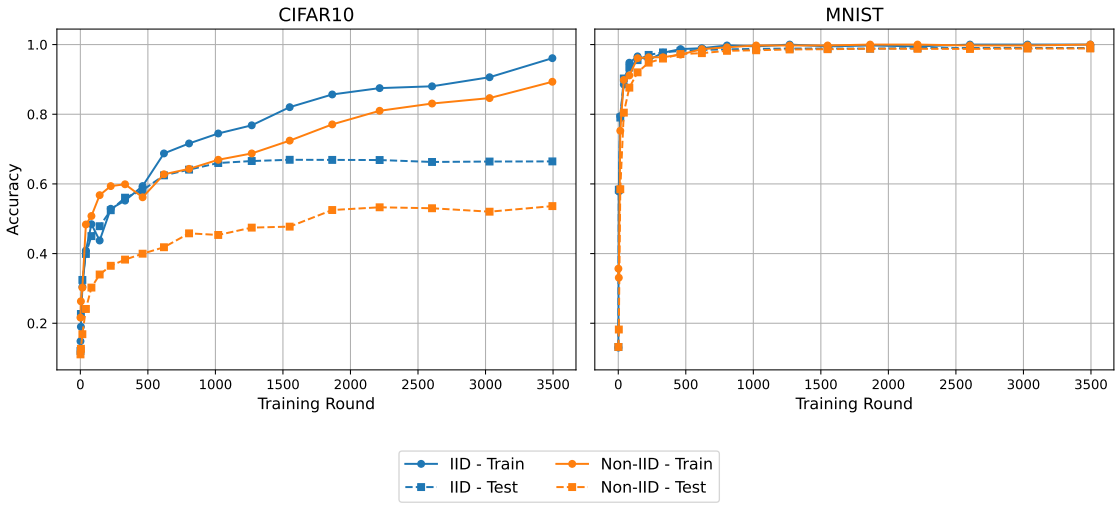


Figure 5.12: Training And Testing Accuracies on Partial Mesh Topology

5.6 Partial Mesh Topology

Figure 5.12 depicts the average training and testing accuracies across all nodes within the partial mesh topology. For the CIFAR10 dataset, a pronounced overfitting pattern is evident. Under IID conditions, the training accuracy increases to approximately 0.95, while the testing accuracy plateaus near 0.65. The generalization gap is similarly severe in the Non-IID setting, where the training accuracy reaches almost 0.90 but the testing accuracy struggles to surpass 0.55. In stark contrast, on the MNIST dataset, the model demonstrates robust generalization. Both the training and testing curves overlap closely and rapidly converge to near 1.0 under both data distribution scenarios, indicating a complete lack of overfitting.

Regarding the privacy evaluation, Figure 5.13 illustrates the mean attack accuracies for the partial mesh topology. On the CIFAR10 dataset, the increasing success of most attacks directly correlates with the widening generalization gap observed during training. The modified prediction entropy MIA under the Non-IID configuration represents the most significant vulnerability, starting high and ultimately peaking at roughly 0.75. Under IID conditions, both the modified prediction entropy and black-box MIAs also show a steady upward trend, reaching approximately 0.75 and 0.70, respectively. Interestingly, the white-box MIA again demonstrates a strong sensitivity to the data distribution; it manages to climb to about 0.60 in the IID setup, but remains relatively ineffective and flat, hovering below 0.55, in the Non-IID environment.

Conversely, the excellent generalization on the MNIST dataset effectively neutralizes most membership inference threats. For both the black-box and white-box MIAs, accuracies quickly settle near the 0.50 random guessing baseline after some minor initial fluctuations. Nevertheless, the modified prediction entropy MIA in the Non-IID configuration stands out as a persistent exception, maintaining a remarkably stable accuracy of 0.70 across all training rounds. This reinforces the conclusion that the extreme data heterogeneity inherent to a Non-IID partition inherently produces distinctive confidence score profiles, allowing specialized modified prediction entropy attacks to succeed even when the model does not overfit.

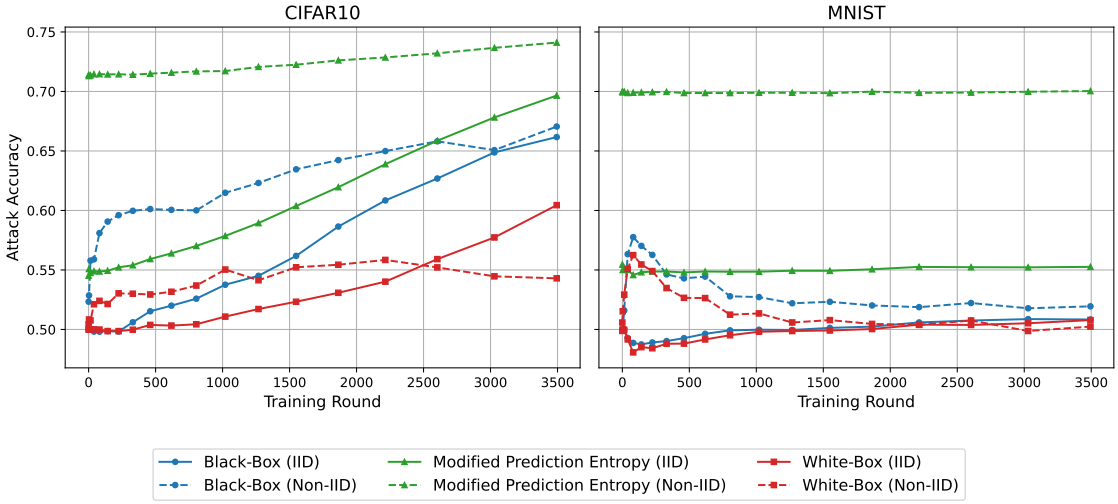


Figure 5.13: Mean Attacks Accuracy in Partial Mesh Topology

5.7 Summary from MIA

This section summarizes the maximum accuracy and training round for the black-box, white-box, and modified prediction entropy MIAs across all evaluated topologies. By consolidating these results, we highlight how the interplay between network topology and data distribution affects privacy. This analysis provides a comparative view of MIA across the CIFAR10 and MNIST datasets under both IID and Non-IID setups.

5.7.1 Black-Box Attack Analysis

Table 5.1 details the maximum attack accuracy and the corresponding training round for the black-box MIA. The results highlight a strong correlation between dataset complexity, data heterogeneity, and privacy leakage. Attack accuracies are substantially higher on the CIFAR10 dataset than on MNIST. Furthermore, the Non-IID distribution consistently exacerbates vulnerability across almost all setups, with the highest overall leakage occurring in the Star topology under Non-IID conditions, peaking at an accuracy of 0.817 in the star topology on Non-IID data distribution.

In the star and tree topologies, which employ the FedAVG aggregation method, the attack consistently achieves maximum accuracy during the early training stages. The sole exception occurs when the root node attacks the edge nodes when using the CIFAR10 dataset, where the peak accuracy is observed at the end of training for both data distributions. In contrast, for the remaining network topologies trained on the CIFAR10 dataset, the attack accuracy peaks during later training stages; for instance, in the full mesh topology, the peak occurs at round 3494 under a Non-IID data distribution. Notably, the attack in the line topology continues to peak at early training stages across both data distributions. Regarding the MNIST dataset, the attack accuracy peaks at later training stages under an IID data distribution, again with the line topology serving as the sole exception. Conversely, under the Non-IID data distribution, the attack accuracy peaks during the early training rounds in all cases, specifically before the 100th training cycle.

A broader comparison between the two datasets reveals that model memorization and subsequent privacy leakage are highly dependent on task complexity. For the simpler MNIST

dataset, IID configurations effectively neutralize the attack to near-random guessing, yielding accuracies between 0.514 and 0.532. However, the severe data imbalances in the MNIST Non-IID setup still allow the attack to reach an accuracy of approximately 0.70 across most topologies. Conversely, the more complex CIFAR10 dataset exhibits high vulnerability even under IID conditions, maintaining baseline accuracies above 0.70. Finally, the results from the tree topology on both datasets clearly demonstrate the protective nature of hierarchical aggregation. The attack is notably weaker when the root targets intermediate edges compared to when edge nodes target their leaf nodes, confirming that intermediate aggregations successfully dilute client-specific memorization to provide a natural layer of privacy protection.

Table 5.1: Maximum Attack Accuracy and Respective Round for Black-Box MIA.

Dataset	Setup	Dist.	Max Accuracy	Round
CIFAR10	Star	IID	0.735	6
		Non-IID	0.817	4
	Tree, Edge attacks Leafs	IID	0.743	4
		Non-IID	0.809	4
	Tree, Root attacks Edges	IID	0.706	36
		Non-IID	0.679	36
	Line	IID	0.771	1268
		Non-IID	0.791	804
	Ring	IID	0.712	3029
		Non-IID	0.781	3494
	Full Mesh	IID	0.713	2603
		Non-IID	0.693	2603
	Partial Mesh	IID	0.719	3494
		Non-IID	0.788	3029
MNIST	Star	IID	0.532	2
		Non-IID	0.677	0
	Tree, Edge attacks Leafs	IID	0.516	2
		Non-IID	0.624	0
	Tree, Root attacks Edges	IID	0.514	1
		Non-IID	0.550	0
	Line	IID	0.526	1549
		Non-IID	0.700	81
	Ring	IID	0.514	3494
		Non-IID	0.697	40
	Full Mesh	IID	0.522	3494
		Non-IID	0.726	40
	Partial Mesh	IID	0.517	3029
		Non-IID	0.705	40

5.7.2 White-Box Attack Analysis

Table 5.2 presents the maximum attack accuracy and the respective training round for the white-box MIA. In contrast to the black-box results, the white-box approach demonstrates higher effectiveness under IID conditions for the CIFAR10 dataset, suggesting that the gradient and parameter information utilized by this attack becomes less reliable when local

client datasets exhibit extreme heterogeneity. The highest recorded vulnerability for this method occurs in the line topology under IID conditions, peaking at an accuracy of 0.816.

Regarding the temporal distribution of these peak accuracies, systems employing the FedAVG aggregation method consistently register their highest attack accuracies during the early stages of training. The sole exception is observed in the tree topology when the root targets intermediate edges, where the peak accuracy is delayed to the final stages of the training process. For the remaining decentralized topologies on the CIFAR10 dataset, the white-box attack typically reaches its maximum accuracy during the late stages, with the notable exception of the full mesh topology under a Non-IID data distribution, where the peak is reached significantly earlier at round 224. On the MNIST dataset, the peak accuracy for IID configurations is consistently delayed to the late training stages, whereas under the Non-IID data distribution, the attack peaks during the early training rounds across nearly all topologies.

A comparative analysis of the two datasets underscores the impact of task complexity on privacy leakage. For the MNIST dataset, IID configurations largely mitigate the white-box attack, maintaining accuracies between 0.510 and 0.537. However, the Non-IID setup introduces meaningful vulnerabilities, pushing accuracy as high as 0.729 in the full mesh topology. The CIFAR10 dataset remains highly susceptible across both data distributions, consistently yielding higher attack accuracies than MNIST. Finally, the results from the tree topology reinforce the protective nature of hierarchical aggregation, the attack is notably less successful when the root targets intermediate edges compared to when edge nodes target their leaf nodes, confirming that intermediate aggregation effectively obscures internal model states and protects client-specific parameters.

5.7.3 Modified Prediction Entropy Attack Analysis

Table 5.3 details the maximum attack accuracy and the training round for the modified prediction entropy MIA. This method proves to be the most potent among all evaluated attacks, consistently achieving exceptionally high accuracy across the majority of configurations. However, it is crucial to reiterate that these results represent a theoretical upper bound on privacy leakage. Because the attack was executed under optimal conditions for the adversary, deriving thresholds directly from the victim model, these figures reflect a true worst-case scenario for data privacy rather than a standard practical deployment. Under this worst-case scenario, the CIFAR10 dataset exhibits extreme susceptibility to the attack, reaching a peak accuracy of 0.911 in the line topology under a Non-IID data distribution. Across all examined architectures, the Non-IID setup consistently results in higher attack accuracy than the IID counterpart, further confirming that severe data heterogeneity provides a rich source of information for this inference strategy.

Regarding the training rounds where the attack peaked in accuracy, systems utilizing the FedAVG aggregation method generally register peak effectiveness during the early to mid stages of training, typically between rounds 11 and 36. In stark contrast, all decentralized topologies trained on the CIFAR10 dataset exhibit a persistent and dominant trend, where the attack reaches its absolute maximum accuracy exclusively at the final training round. While the MNIST dataset displays more variabilitywith several Non-IID cases peaking almost immediately at the start of trainingthe decentralized architectures maintain a tendency to favor late-stage convergence for their maximum accuracy across IID distributions.

Table 5.2: Maximum Attack Accuracy and Respective Round for White-Box MIA.

Dataset	Setup	Dist.	Max Accuracy	Round
CIFAR10	Star	IID	0.787	9
		Non-IID	0.692	4
	Tree, Edge attacks Leafs	IID	0.801	11
		Non-IID	0.742	11
	Tree, Root attacks Edges	IID	0.720	36
		Non-IID	0.537	14
	Line	IID	0.816	2215
		Non-IID	0.724	1268
	Ring	IID	0.723	3494
		Non-IID	0.635	3494
	Full Mesh	IID	0.743	3494
		Non-IID	0.632	224
	Partial Mesh	IID	0.696	3494
		Non-IID	0.675	1864
MNIST	Star	IID	0.517	2
		Non-IID	0.635	11
	Tree, Edge attacks Leafs	IID	0.533	4
		Non-IID	0.665	0
	Tree, Root attacks Edges	IID	0.509	14
		Non-IID	0.547	0
	Line	IID	0.533	3029
		Non-IID	0.647	618
	Ring	IID	0.514	3029
		Non-IID	0.667	142
	Full Mesh	IID	0.537	2215
		Non-IID	0.729	3
	Partial Mesh	IID	0.523	3494
		Non-IID	0.633	81

Comparing the two datasets, it is evident that even simpler tasks like MNIST are not immune to privacy leakage when subjected to this specialized entropy-based attack. While the IID MNIST configurations are relatively robust with accuracies peaking around 0.56 to 0.57, the Non-IID setup triggers a dramatic increase in success, with most topologies reaching approximately 0.77. Finally, the tree topology results demonstrate that hierarchical aggregation remains a strong defensive measure. The attack accuracy is consistently higher when edge nodes target their immediate leaf nodes compared to when the root node attempts to infer membership from the intermediate edges. This confirms that the hierarchical structure successfully filters and aggregates confidence score patterns, effectively masking individual data contributions from top-level adversaries.

5.8 Final Remarks

By evaluating a worst-case scenario for data privacy with the modified prediction entropy MIA, we establish a theoretical upper bound on system vulnerability. Under these optimal adversarial conditions, the comprehensive evaluation demonstrates that model memorization and subsequent privacy leakage heavily depend on the complexity of the dataset and the data

Table 5.3: Maximum Attack Accuracy and Respective Round for Modified Prediction Entropy MIA.

Dataset	Setup	Dist.	Max Accuracy	Round
CIFAR10	Star	IID	0.847	21
		Non-IID	0.898	21
	Tree, Edge attacks Leafs	IID	0.851	24
		Non-IID	0.902	17
	Tree, Root attacks Edges	IID	0.782	36
		Non-IID	0.721	36
	Line	IID	0.889	3494
		Non-IID	0.911	3494
	Ring	IID	0.784	3494
		Non-IID	0.819	3494
	Full Mesh	IID	0.783	3494
		Non-IID	0.773	3494
	Partial Mesh	IID	0.767	3494
		Non-IID	0.868	3494
MNIST	Star	IID	0.566	2
		Non-IID	0.768	0
	Tree, Edge attacks Leafs	IID	0.565	11
		Non-IID	0.769	0
	Tree, Root attacks Edges	IID	0.562	11
		Non-IID	0.634	11
	Line	IID	0.561	3494
		Non-IID	0.733	0
	Ring	IID	0.560	1864
		Non-IID	0.766	3494
	Full Mesh	IID	0.572	0
		Non-IID	0.769	0
	Partial Mesh	IID	0.563	224
		Non-IID	0.768	3029

distribution. The more complex CIFAR10 dataset consistently aggravates attack accuracy across all tested architectures compared to the simpler MNIST dataset, ultimately achieving a peak overall accuracy of 0.911 in the line topology under a Non-IID data distribution. The pronounced success of this attack at the upper bound highlights its capability to exploit extreme local overfitting. In highly skewed data distributions, local models produce highly polarized output vectors with near-perfect prediction confidence for members and high entropy for non-members. By directly evaluating these confidence scores to measure maximum leakage, the modified prediction entropy MIA effectively distinguishes membership without requiring prior knowledge of the client's specific data distribution.

Conversely, the experimental data reveals a counter-intuitive behavior regarding the theoretically more powerful white-box MIA, which generally performs worse under data heterogeneity. This performance drop can be attributed to the fundamental mechanics of the attack and its reliance on shadow models. In an IID setting, uniform data distribution ensures that the internal gradients of the adversary's shadow models closely align with those of the victim's model. However, under a Non-IID setup, local client datasets become heavily skewed. Because the adversary lacks prior knowledge of this specific client-side data imbalance, the shadow models fail to accurately mimic the victim's internal gradient behavior, generating an

unrepresentative training set for the attack classifier. Furthermore, reducing these complex gradients to just two principal components via PCA likely discards the variance needed to identify membership in heterogeneous environments.

An examination of the training rounds at which these attacks reach their maximum accuracy reveals distinct temporal patterns dictated by the network topology and the aggregation method. In architectures utilizing the FedAVG aggregation method, such as the star and tree topologies, the attacks most commonly peak during the very early stages of training. This suggests that membership inference becomes effective almost immediately once local models begin to overfit. In contrast, decentralized topologies trained via D-PSGD, such as the line, ring, and mesh networks, significantly delay this peak to the later stages of training, frequently reaching maximum vulnerability only at the final communication round. However, severe data imbalances can disrupt this delay, as seen under Non-IID conditions where extreme data skew forces the model to rapidly expose identifiable gradients and confidence scores during the early training cycles.

Finally, the special case of the tree topology quantitatively demonstrates the defensive properties of hierarchical intermediate aggregation. The vulnerability footprint of the leaf nodes when an intermediate edge node acts as the adversary is nearly indistinguishable from that of clients facing a central node in a star topology. This confirms that without additional cryptographic measures, edge nodes represent a critical privacy bottleneck. However, when the root server attacks the intermediate edge nodes, attack accuracies experience a significant drop across all methods. Moreover, this intermediate aggregation not only weakens the attacks but also significantly delays privacy leakage, consistently pushing the peak vulnerability to the very end of the training process. This temporal delay provides a meaningful operational window during which a system administrator could detect anomalous behavior or deploy countermeasures before a top-level adversary can extract actionable membership information.

Chapter 6

Gradient Inversion Attack Results

This chapter presents the empirical results obtained form the experimental setup described in Chapter 4 regarding GIA. The primary objective is to quantify how the choice of a network topology affects the effectiveness of this attack in FL environments.

6.1 Mean PSNR Across Training Round

In this section, we evaluate different network topologies to the GIA. The effectiveness of the attack is measured by calculating the mean Peak Signal-to-Noise Ratio (PSNR) of the reconstructed images across various training rounds for each adversary-victim pair. The PSNR is a metric used to quantify the reconstruction quality of an image [9]. It measures the ratio between the maximum possible power of a signal, which represents the original image, and the power of corrupting noise, representing the reconstruction error, typically expressed in decibels (dB). In the context of a GIA, a higher PSNR value indicates that the reconstructed image is highly similar to the original private data of the victim, signifying a successful attack and a severe privacy leak. Conversely, a lower PSNR indicates a noisy and unrecognizable reconstruction.

The results of the GIA for the CIFAR10 dataset are presented in Figure 6.1 for the IID distribution and Figure 6.2 for the Non-IID distribution. Similarly, the results for the MNIST dataset are shown in Figure 6.3 and Figure 6.4.

When observing the progression of the training process across all datasets, data distributions, and topologies, it becomes evident that the attack is significantly more effective during the early stages of training. Specifically between rounds 0 and 5000, model weights undergo larger updates, resulting in gradients that carry more information about the local training data. As the training progresses and the models converge, the gradient updates become smaller and more generalized, causing the mean PSNR to drop significantly. By round 20000, the PSNR for almost all configurations drops below 10 dB, rendering the reconstructed images virtually indistinguishable from noise.

The choice of network topology also plays a major role in the inherent privacy of the system. The tree, star, and full mesh topologies consistently exhibit the highest vulnerability in the early rounds, reaching peak PSNR values of nearly 18 dB for the CIFAR10 dataset and over 22 dB for the MNIST dataset under IID conditions. This severe vulnerability is directly tied to the ability of the adversary to accurately calculate the previous model utilized by the victim prior to their local training step.

In highly centralized topologies or fully connected networks, an adversary can observe all the necessary incoming model updates either by being strategically positioned at a hub or by

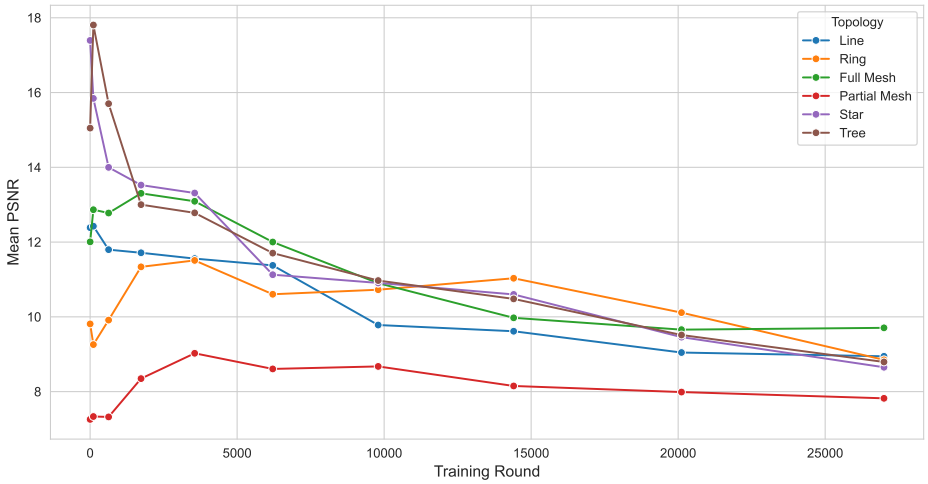


Figure 6.1: Mean PSNR of reconstructed images across training rounds for the CIFAR10 dataset under an IID data distribution.

directly receiving updates from all peers in the mesh. This visibility allows the adversary to perfectly reconstruct the exact starting weights that the victim uses for their local training. Possessing this precise reference model is critical for cleanly isolating the exact gradient update produced by the victim, which is a fundamental prerequisite for successful gradient inversion.

Conversely, the partial mesh topology demonstrates remarkable resilience across all scenarios. In this decentralized structure, a victim aggregates models from a subset of neighbors of which the adversary is not fully aware. Consequently, the adversary is unable to accurately determine the exact previous model used by the victim. This inevitable mismatch in the reference weights introduces severe calculation errors into the inversion process, inherently obscuring the gradient information and effectively mitigating the GIA. As a result, the partial mesh maintains a consistently low mean PSNR even during the earliest and most vulnerable training rounds.

Falling between these two extremes are the line and ring topologies. In these network topologies, each node connects to at most two neighbors. When an adversary acts as one of these neighbors, they can observe their own interactions with the victim but remain blind to the model updates arriving from the victim's other neighbor. Similar to the partial mesh, this missing information prevents the adversary from perfectly calculating the victim's exact starting weights. However, because the degree of connectivity is minimal, the adversary is only missing a single incoming update rather than multiple. The resulting calculation error during gradient isolation is therefore less severe than the multi-neighbor uncertainty present in a partial mesh. Consequently, the line and ring topologies offer a moderate degree of privacy, yielding initial PSNR trajectories that sit noticeably below those of the fully observable topologies but remain higher than the partial mesh.

Furthermore, the characteristics of the dataset and how the data is distributed influence the overall success of the attack. The GIA yields higher peak PSNR values on the MNIST dataset compared to CIFAR10. Because MNIST consists of simpler grayscale images, the gradients leak structural information much more easily than the complex color images found

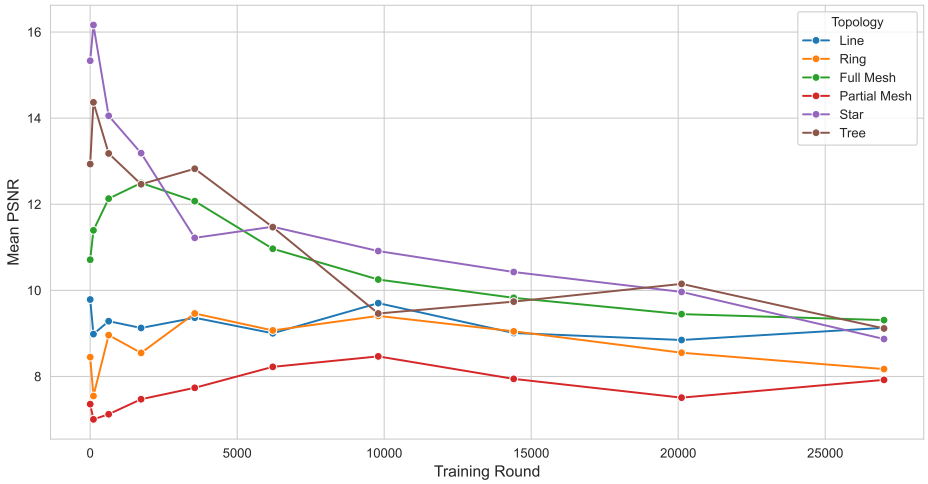


Figure 6.2: Mean PSNR of reconstructed images across training rounds for the CIFAR10 dataset under a Non-IID data distribution.

in CIFAR10. When comparing the data distributions, the IID configurations generally expose slightly higher initial PSNR peaks than their Non-IID counterparts. In Non-IID settings, the high variance in local data distributions may act as a natural noise mechanism, slightly diminishing the ability of the adversary to perfectly invert the gradients of specific target classes. Ultimately, utilizing highly connected but non-centralized topologies, such as a partial mesh, provides a significant advantage in minimizing the risk of gradient leakage in decentralized learning environments.

6.2 Effects of Isolating Individual Node Contributions

Beyond quantitative metrics, a visual inspection of the reconstructed images provides concrete evidence of how network topology and the aggregation process directly impact the privacy of individual data samples. To illustrate this, we examine the qualitative results of the GIA executed at an identical early stage of training, specifically at the 111th training round, across two different network topologies. Figure 6.5 displays the visual output of the attack within a line topology, whereas Figure 6.6 demonstrates the attack against the same victim within a ring topology. In both figures, the top row represents the original images belonging to the victim, and the bottom row displays the images reconstructed by the adversary.

The success of the reconstruction in the line topology can be attributed to the specific structural position of the victim. In a line topology, the nodes at the extreme edges of the network only possess a single neighbor. In the scenario depicted in Figure 6.5, the victim is located at the edge of the line and its only connection is the attacker. Consequently, the adversary has complete visibility over every model update the victim receives. This absolute knowledge allows the adversary to perfectly calculate the exact reference model used by the victim for local training. Because this occurs at an early stage of training where the model updates are large and the gradients carry significant structural information, the adversary can cleanly isolate the victim's gradient. This results in high quality reconstructions where the original features of the images, such as the shape and color of objects, are clearly visible.

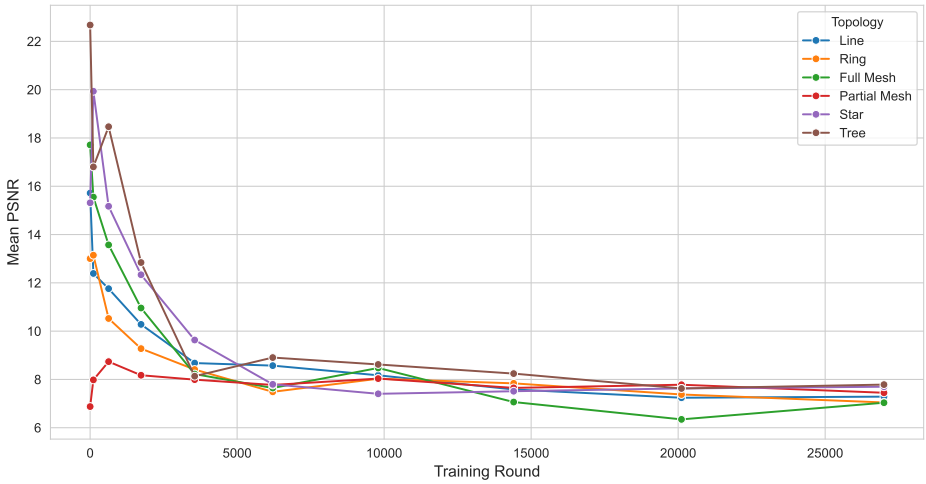


Figure 6.3: Mean PSNR of reconstructed images across training rounds for the MNIST dataset under an IID data distribution.

Conversely, the ring topology fundamentally alters the connectivity and, by extension, the knowledge available to the adversary. In a ring structure, every node is connected to exactly two neighbors. Therefore, the victim receives model updates from both the adversary and an additional, independent neighbor. Because the adversary cannot observe the model updates sent by this second neighbor, they are entirely unable to determine the final aggregated model that the victim uses as a starting point for their training round.

This information gap is critical to the defense of the system. The scenario in Figure 6.6 takes place at the exact same early stage of training as the line topology example, meaning the underlying gradients still inherently carry a vast amount of sensitive information. However, the direct consequence of the adversary failing to accurately calculate the previous model state used by the victim results in a completely failed reconstruction. The inevitable calculation error in the reference weights propagates through the inversion algorithm, preventing it from aligning the features correctly. As a result, the extracted data is reduced to a matrix of colorful noise, fully preserving the visual privacy of the victim despite the theoretical vulnerability of the gradients themselves.

This same fundamental principle extends to the MNIST dataset, as demonstrated in Figure 6.7 and Figure 6.8. Operating within the line topology at the identical training round, the adversary exploits their position as the victim's sole neighbor to perfectly calculate the reference model, achieving a near-perfect reconstruction of the simpler grayscale digits. However, when the topology is shifted to a ring, the adversary loses absolute visibility over the victim's aggregation phase.

To further highlight the resilience of decentralized structures, Figure 6.9 and Figure 6.10 illustrate the attack outcomes within a partial mesh topology. In this configuration, the adversary is only one of several nodes connected to the victim. Because the victim aggregates models from multiple other neighbors that remain completely hidden from the attacker, the adversary faces an even greater information deficit than in the ring topology. This multi-neighbor uncertainty makes it practically impossible to estimate the victim's starting weights. Furthermore, because this attack occurs at round 111, the models have largely converged.

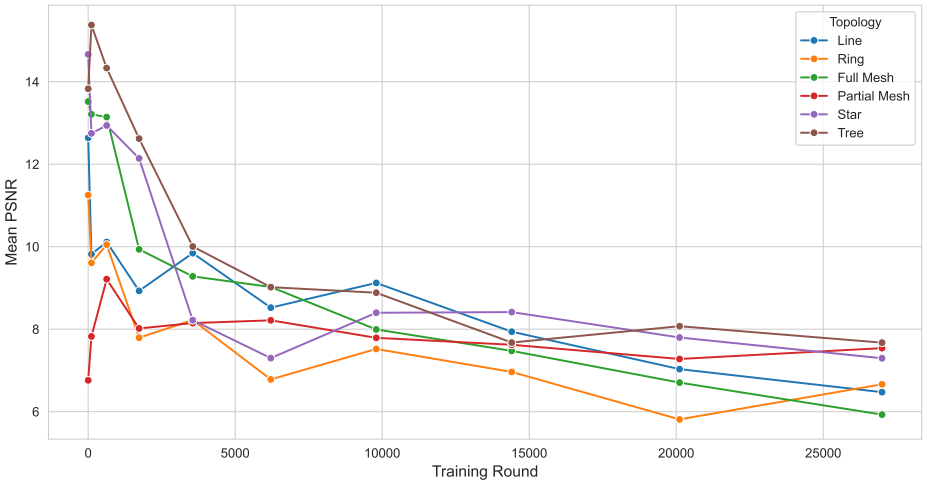


Figure 6.4: Mean PSNR of reconstructed images across training rounds for the MNIST dataset under a Non-IID data distribution.

Figure 6.5: Visual reconstruction results of the GIA on the CIFAR10 dataset under an IID distribution using a line topology.

The exchanged gradients are therefore significantly smaller and carry far less structural information about the local data compared to the early rounds. The combination of these two factors, the topological defense of the partial mesh and the temporal degradation of gradient information, results in a total failure of the attack. As seen in the figures, the reconstructions for both the complex CIFAR10 images and the simpler MNIST digits are reduced to completely unrecognizable patterns and visual noise, ensuring robust data privacy throughout the entirety of the training lifecycle.

6.3 Effects of the Inability to Isolate Individual Node Contributions

The structural design of a network topology directly dictates the flow of information and, consequently, the amount of knowledge an adversary can gather about a victim's state. When an adversary possesses complete knowledge of the exact model weights used by a victim prior to their local training step, the vulnerability to the GIA reaches its absolute maximum. To visually demonstrate the catastrophic impact of this complete knowledge, we examine the qualitative results of the attack executed at round 111 on the CIFAR10 dataset under an IID distribution. The results are shown across three highly vulnerable topologies:

Figure 6.6: Visual reconstruction results of the GIA on the CIFAR10 dataset under an IID distribution using a ring topology.

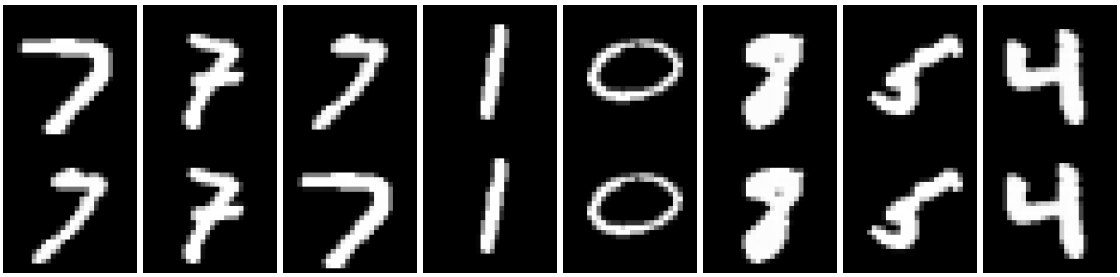


Figure 6.7: Visual reconstruction results of the GIA on the MNIST dataset under an IID distribution using a line topology.

the star topology in Figure 6.11, the tree topology in Figure 6.12, and the full mesh topology in Figure 6.13. In all three cases, the top row displays the original private images, and the bottom row shows the attacker's reconstructed images.

The exceptional quality of the reconstructed images across these three figures is a direct consequence of the adversary's ability to perfectly deduce the victim's starting reference model. In the star topology, the adversary is positioned as the central node. Because all communication flows through this node, the adversary knows the exact aggregated model that is sent out to the victim node. Similarly, in the tree topology, an adversary positioned as a intermediate edge node has total visibility over the model state that is passed down to its leaf victim. The full mesh topology achieves the same effect through a different mechanism. Because every node is connected to every other node in a full mesh, the adversary receives the exact same set of incoming model updates as the victim. This allows the adversary to perfectly simulate the victim's aggregation phase and independently arrive at the exact same starting weights.

Possessing this precise reference model is the critical prerequisite for a successful attack. When the adversary knows exactly what weights the victim started with, they can subtract this reference model from the updated model received from the victim, cleanly and perfectly isolating the exact gradient produced during the victim's local training step. Because this attack takes place at round 111, an early stage in the training process, these isolated gradients are large and carry rich, undisturbed structural information about the local data. The combination of early training gradients and zero calculation error in the reference model allows the inversion algorithm to perform exceptionally well. As evidenced by the visual results, the reconstructed images are nearly indistinguishable from the original private data, clearly exposing fine details, complex shapes, and accurate colors. This confirms that topological structures that grant an adversary total visibility over a victim's reference model essentially

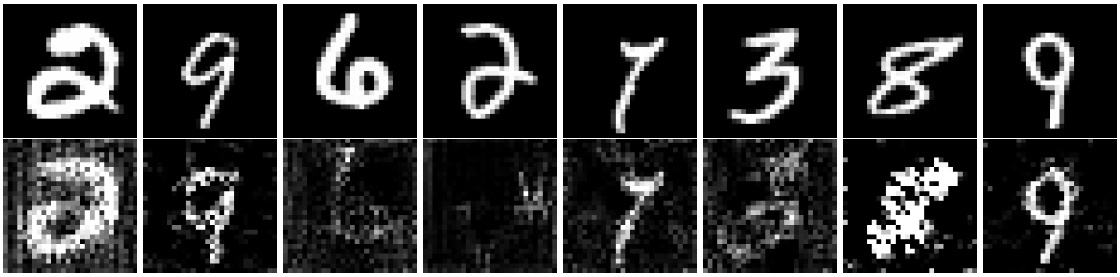


Figure 6.8: Visual reconstruction results of the GIA on the MNIST dataset under an IID distribution using a ring topology.

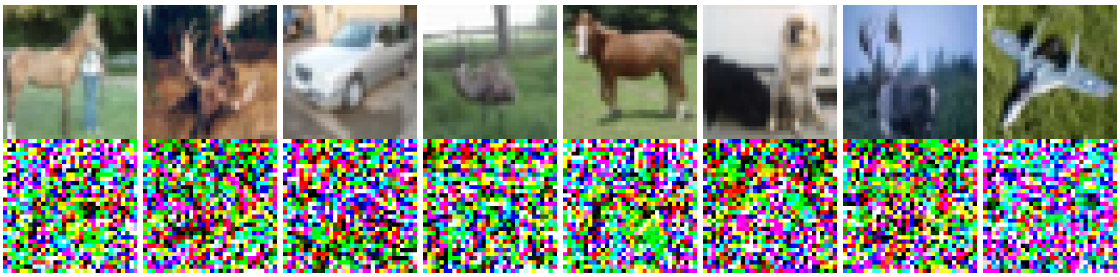


Figure 6.9: Visual reconstruction results of the GIA on the CIFAR10 dataset under an IID distribution using a partial mesh topology.

nullify any inherent privacy in decentralized learning, leaving the system completely defenseless against gradient inversion.

6.4 Final Remarks

The comprehensive evaluation of the GIA across different network topologies provides critical insights into the privacy vulnerabilities of FL systems. One of the most notable observations is that the effectiveness of the attack is not strictly dependent on the underlying data distribution. While a Non-IID setup introduces a natural variance that slightly dampens the attack compared to an IID setting, the overall threat level remains exceptionally high. The GIA demonstrates a robust capability to extract sensitive features regardless of how the training data is partitioned among the network participants, indicating that data heterogeneity alone cannot be relied upon as a primary defense mechanism against GIA.

The structural design of the network topology plays a far more decisive role in determining the system's vulnerability, particularly during the early stages of the learning process. Topologies such as the star, tree, and full mesh exhibit severe security flaws in these initial rounds. Because these structures centralize information flow or establish complete peer connectivity, they inadvertently expose the exact model updates circulating through the network. During the early training rounds, model weights undergo significant adjustments, resulting in large gradients that encapsulate rich structural information about the local training data. When these exposed structures are combined with the high information density of early gradients, the GIA achieves its highest PSNR values, allowing an adversary to reconstruct private images with near perfect accuracy.

However, the threat posed by the GIA is highly sensitive to the temporal progression of the training process. As the decentralized models begin to settle and converge towards

Figure 6.10: Visual reconstruction results of the GIA on the MNIST dataset under an IID distribution using a partial mesh topology.

Figure 6.11: Visual reconstruction results of the GIA on the CIFAR10 dataset under an IID distribution using a star topology.

a global solution, the magnitude of the gradient updates shrinks considerably. These late stage updates represent generalized, fine-tuning adjustments rather than the extraction of core structural features from the raw data. Consequently, the effectiveness of the attack plummets in the later rounds of training. Regardless of the underlying dataset or network topology, the gradients eventually lose the specific information required for image inversion, rendering the adversary's reconstructions completely indistinguishable from random noise.

A fundamental driver of the attack's success or failure is the specific position of the adversary within the network topologies. The ability to successfully reconstruct an original image hinges entirely on the attacker's capacity to accurately calculate the exact reference model utilized by the victim prior to their local training step. If the adversary is positioned strategically as a centralized node, they can monitor all incoming updates to the victim and perfectly isolate the victim's subsequent contributions. Conversely, if the attacker's position only affords them a partial view of the victim's connections, the resulting information deficit introduces severe calculation errors that are completely detrimental to GIA success.

Ultimately, this reliance on total network visibility highlights the partial mesh topology as a highly effective, natural defense mechanism against gradient inversion. By ensuring that a victim aggregates models from a subset of neighbors that remain hidden from the adversary, the partial mesh structure inherently obscures the victim's starting weights. This architectural uncertainty proves to be remarkably resilient across all tested datasets and distributions, successfully degrading the quality of the attack even during the most vulnerable early stages of training. Therefore, deploying sparsely connected, decentralized architectures stands out as a critical strategy for preserving data privacy and mitigating the severe risks associated with gradient leakage.

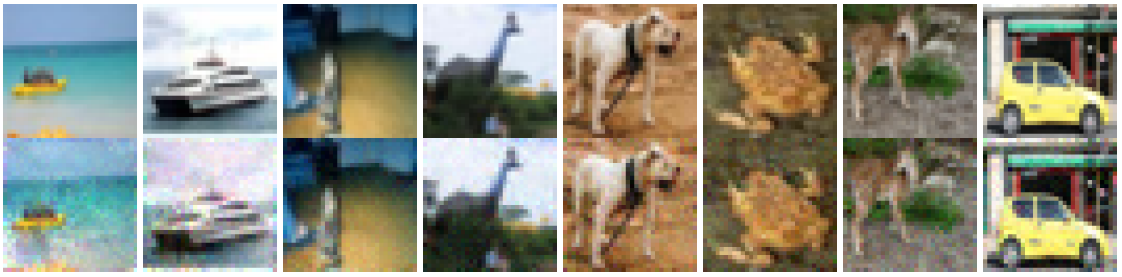


Figure 6.12: Visual reconstruction results of the GIA on the CIFAR10 dataset under an IID distribution using a tree topology.

Figure 6.13: Visual reconstruction results of the GIA on the CIFAR10 dataset under an IID distribution using a full mesh topology.

Chapter 7

Data Privacy and Security

While FL is frequently advocated as a privacy-preserving paradigm due to the retention of raw data on local devices, it does not inherently guarantee the confidentiality of sensitive information. The privacy of a FL system is a multi faceted concept encompassing regulatory compliance, technical security, and ethical considerations, all of which are heavily influenced by the used network topology that dictates information flow. This chapter analyzes these risks by examining the legal implications of decentralized roles under the GDPR, analyzing how network structure alters the trust boundary and attack surface for passive adversaries, and discussing the ethical dimensions of power dynamics and transparency in collaborative learning environments.

7.1 Data Protection in Federated Learning Environments

Data protection refers to the legal and regulatory frameworks designed to safeguard personal information and ensure the privacy rights of individuals. In the context of ML, and specifically FL, these regulations, most notably the GDPR in the European Union [2] dictate how data must be handled, processed, and stored. While FL is often proposed as a privacy-preserving paradigm because raw data remains local, the exchange of model updates still constitutes the processing of personal data, as these updates can be reverse engineered to reconstruct sensitive information [9].

7.1.1 Definition of Roles: Controller and Processor

Under the GDPR, it is essential to distinguish between the Data Controller and the Data Subject. A Data Controller is the legal entity that determines the "purposes and means" of data processing, essentially deciding why and how the data is used. A Data Subject is the identified or identifiable natural person to whom the data relates. In the context of CFL, typically organized in a star topology, these mappings are clearly delineated, the central server acts as the Data Controller, as it defines the global model architecture and orchestrates the training rounds. The participating nodes map to the Data Subjects, as they hold the local training data that pertains to them [54]. In this hierarchy, the central entity bears the sole legal liability for security and compliance.

In DFL, however, this clear distinction collapses, leading to a more complex legal structure. Since there is no central server to define the processing means, the participating nodes assume dual or even triple roles. First, they remain Data Subjects regarding their own local data. Second, because the group of nodes collectively determines the evolution of the global model, they effectively map to Joint Data Controllers [55]. Legally, this implies that every participant shares liability for the system's compliance [2, Art. 26]. Finally, the role of the

Data Processor, an entity that processes data on behalf of a controller, also falls upon the nodes. In topologies like the ring or mesh, when a node receives updates from neighbors, aggregates them, and forwards the result, it is technically acting as a Data Processor for those neighbors. This creates a significant friction point for DFL deployment, as a single user device is burdened with the legal responsibilities of a Controller, a Subject, and a Processor simultaneously.

7.1.2 Data Minimization and Network Topology

The principle of Data Minimization [2, Art. 5] mandates that data processing should be limited to the minimum necessary for the purposes for which it is being processed. Network topology plays a direct role in adherence to this principle. In a star topology, minimization is achieved by ensuring the central server only receives the minimum necessary updates. However, in DFL, the choice of topology dictates the redundancy of data sharing and the subsequent privacy risk [21].

High density topologies, such as the fully connected mesh, often challenge the concept of minimization. By broadcasting model updates to every other node in the network, a participant node exposes its model updates to all other nodes in the system. This level of exposure may be considered excessive if the algorithm could converge with fewer connections. Conversely, sparse topologies strictly limit data exposure. In a ring or line topology, a node shares its update with only one or two neighbors. From a data protection perspective, this architecture adheres more strictly to minimization principles [20], as the "need-to-know" circle is restricted to the immediate neighborhood, significantly reducing the surface area for potential data misuse.

7.1.3 Compliance Challenges

Ultimately, the alignment of DFL with data protection laws depends on the ability to control data flow [54]. If the network topology is static and known, such as a fixed ring of hospitals, legal contracts can establish trust and define liability. However, in dynamic, ad-hoc topologies, such as mobile devices in a mesh, ensuring that every peer adheres to data protection standards is nearly impossible [44]. Consequently, while DFL addresses the issue of data locality by keeping raw data on the device, it complicates the issue of accountability [56]. The more complex and dense the network topology, the harder it becomes to identify the controlling entity and to ensure that the rights of the data subject are preserved across the entire P2P network [54].

7.2 Security in Federated Learning Systems

While data protection focuses on the legal rights of the individual, security in DFL addresses the technical mechanisms required to enforce those rights and protect the confidentiality, integrity, and availability of the model. The transition from a centralized to a decentralized architecture fundamentally alters the security landscape. In CFL, security is perimeter based, focusing on hardening the central server against external intrusion. In contrast, DFL relies on a distributed security model where the network topology itself dictates the exposure of sensitive information. This section analyzes how the choice of network structure defines the trust boundary, the attack surface, and the system's resilience against passive adversaries.

7.2.1 The Shifting Trust Boundary

The trust boundary defines the perimeter within which data and operations are considered secure. In CFL, this boundary is explicit, participants effectively enter a trust contract with the central server, assuming it will not inspect or leak their individual updates. The network topology is a star, and the security of the entire system relies on the integrity of that single central node. DFL effectively dissolves this clear perimeter. By utilizing P2P communication, DFL pushes the trust boundary out to the edge devices. In this paradigm, a node must implicitly trust its neighbors not to collude or analyze the received model updates [16]. Consequently, the topology determines the extent of this required trust [32]. A sparse topology, such as a ring, minimizes the trust requirement to two immediate neighbors, whereas a dense mesh topology requires a node to extend trust to a significant portion of the network, thereby increasing the probability of encountering a malicious actor [16] within one's own trust boundary.

7.2.2 Topology as an Attack Surface

The attack surface represents the sum of all points where an unauthorized user can try to enter data to or extract data from a system. The choice of network topology creates a direct trade-off between the availability attack surface and the privacy attack surface. Centralized topologies exhibit a minimal privacy attack surface because updates are only exposed to one entity, but they suffer from a massive availability attack surface due to the SPoF at the central server. DFL topologies invert this risk profile. A mesh topology eliminates the SPoF, ensuring high availability even if multiple nodes fail. However, this resilience comes at the cost of a significantly expanded privacy attack surface. Since model updates are broadcast to multiple peers, the number of potential vantage points for an attacker increases linearly with the degree of the nodes. Therefore, in highly connected topologies, the vector for privacy leakage is not a single server, but a multitude of potentially compromised edge devices.

7.2.3 Passive Adversaries and Inference Risks

The most prevalent threat to data privacy in FL systems comes from passive adversaries, known as "honest-but-curious" adversaries. These adversaries adhere strictly to the training protocol, performing local training and aggregation correctly, but silently analyze the model updates they receive to infer private data. The effectiveness of such adversaries is intrinsically linked to the network topology. Attacks such GIA rely on isolating a victim's gradient to reverse engineer the training data [38]. In sparse topologies, an adversary may only observe a simplified or aggregated update, making the isolation of specific data points difficult. However, in dense topologies, a passive adversary can observe distinct gradient vectors from multiple neighbors. If the adversary colludes with other nodes in the mesh topology, they can calculate the specific contributions of a victim node, stripping away the anonymity usually provided by the aggregation process. Thus, while denser topologies offer faster convergence, they provide an ideal environment for passive adversaries to execute reconstruction attacks with higher fidelity [21].

7.3 Ethics Aspects of Federated Learning Systems

Ethical considerations in DFL extend beyond strict legal compliance with regulations such as GDPR. While data protection focuses on legality and security focuses on robustness, ethics

addresses the moral implications of system design [1], including fairness, transparency, and the power dynamics inherent in the chosen network structure. The shift from a centralized authority to a peer-based collaborative model introduces unique ethical challenges regarding how participants interact, how the burden of training is distributed, and whether users are truly informed about the destination of their data.

7.3.1 Power Dynamics and Topological Hierarchy

The selection of a network topology is not merely a technical decision but an ethical one that defines the power relations within the system [57]. In hierarchical structures, such as the star or tree topology, specific nodes are elevated to positions of authority, acting as aggregators or communication bridges for others. This creates a disparity in power: the central node in a star or the root node in a tree possesses a disproportionate ability to influence the global model and, potentially, to censor or isolate specific branches of the network [58]. This raises ethical questions regarding the selection of these key nodes. If a standard user device is arbitrarily designated as a root node, it assumes a burden of responsibility and power that may not be ethically justifiable. True decentralization implies an egalitarian structure, therefore, topologies that enforce rigid hierarchies may contradict the ethical spirit of DFL by recreating centralized control structures under the guise of decentralization [8].

7.3.2 Fairness and the Free Rider Problem

A core ethical tenet of collaborative systems is fairness: the idea that those who benefit from the collective resource should contribute to its creation. In DFL, this is challenged by "free riders" participants who utilize the global model without contributing meaningful local training [59], or who intentionally submit trivial updates to save their own computational resources. The impact of this unethical behavior is modulated by the network topology. In a mesh topology, the negative impact of a free rider is often diluted by the presence of numerous honest neighbors, distributing the "harm" across the network. However, in sequential topologies like the line or ring, the ethical violation becomes acute. A free rider in a ring topology directly degrades the model quality for their immediate successor, creating a localized unfairness where specific users are disproportionately penalized by the behavior of their neighbors. Thus, the choice of topology determines whether the burden of unethical behavior is shared collectively or imposed on specific victims.

7.3.3 Informed Consent and Transparency

The principle of informed consent requires that users understand how their data is being used and by whom. In traditional CFL, users consent to a trusted service provider processing their data. DFL fundamentally alters this contract, often without the user's explicit realization [60]. In a P2P setting, a user's model updates, which effectively encode their private data patterns, are transmitted not to a secure corporate server, but to the devices of other users. From an ethical standpoint, this demands a higher standard of transparency [61]. A user in a fully connected mesh topology should arguably be informed that their gradients are being broadcast to every other participant in the network, rather than being aggregated in a private enclave. The complexity of dynamic topologies further obfuscates this process, making it difficult for users to assess their own privacy exposure. Ethical system design therefore mandates that the network structure is not treated as a black box, but is transparently

communicated to the user, ensuring they consent to the specific peer-based transmission paths that the topology entails [60].

Chapter 8

Conclusions and Future Work

This concluding chapter synthesizes the empirical findings, structural insights, and overall contributions of this dissertation regarding privacy preservation in FL systems. It first provides an overview of the key conclusions regarding how network topology and data distribution influence vulnerability to MIAs and GIA. Next, it discusses the limitations of the current experimental framework before outlining future research to further secure distributed learning environments.

8.1 Dissertation Overview

This dissertation presented a comprehensive analysis of privacy vulnerabilities within FL systems, successfully fulfilling its three primary research objectives. The first objective was to identify and categorize the structural properties of distinct network topologies commonly utilized in DFL systems, which was thoroughly addressed within the state-of-the-art review. Building upon this foundation, the second objective sought to evaluate and compare the susceptibility of these network topologies to prominent privacy threats. Finally, the third objective aimed to analyze the temporal evolution of these privacy attacks across training rounds to determine the precise impact of network topologies on attack efficacy over time. Through extensive empirical evaluations, we demonstrated that network topologies, data distribution, and the temporal progression of the training process fundamentally dictate the privacy boundaries of decentralized environments.

Regarding MIAs, our findings showed that privacy leakage is strongly tied to the model's generalization gap. Complex tasks that induce overfitting, such as the CIFAR-10 dataset, rapidly escalate attack accuracy compared to simpler datasets like MNIST. This vulnerability is severely aggravated under Non-IID conditions, where extreme data heterogeneity forces local models to overfit on local data. Consequently, the modified prediction entropy MIA emerged as the most impactful threat, achieving an accuracy of 0.911 in the line topology. Although the adversary did not require prior knowledge of the client's specific data distribution, this evaluation was deliberately modeled as a worst-case scenario, therefore, these exceptionally high results represent a theoretical upper bound for privacy leakage rather than a typical operational baseline. Conversely, the theoretically powerful white-box MIA performs worse under data heterogeneity. Shadow models fail to accurately mimic the victim's skewed internal ML model state, and dimensionality reduction via PCA discards the variance crucial for identifying membership in heterogeneous environments.

The temporal peak of MIAs is heavily dictated by the network topology and aggregation method. In systems utilizing FedAVG, such as the star and tree topologies, attacks typically peak during the very early stages of training as models immediately begin to overfit. In

contrast, decentralized topologies trained via D-PSGD significantly delay this peak to the final communication rounds, though severe data imbalances can disrupt this delay. Notably, the tree topology quantitatively demonstrated the defensive properties of hierarchical intermediate aggregation. While edge nodes remain a critical privacy bottleneck for leaf nodes, intermediate aggregation significantly weakens attacks from the root server and pushes peak vulnerability to the very end of the training process, providing a crucial window to detect anomalous behavior.

In contrast, the GIA behaves fundamentally differently. Because the optimization process of gradient inversion operates directly on isolated gradient updates, the overall system data distribution, whether IID or Non-IID, does not significantly hinder the attack. Instead, the system's vulnerability to gradient inversion is primarily determined by the temporal stage of training. Topologies that centralize information flow or establish complete peer connectivity, such as the star, tree, and full mesh, exhibit severe security flaws in the initial rounds. During these early stages, large gradient adjustments encapsulate rich structural information, allowing the GIA to reconstruct private images with near-perfect accuracy. However, as models converge in late stages of training, updates represent generalized fine tuning, causing the effectiveness of the attack to plummet and rendering reconstructions indistinguishable from random noise.

Ultimately, GIA success hinges entirely on the adversary's position and network visibility. To reconstruct an original image, the attacker must perfectly isolate the victim's subsequent contributions by accurately calculating their exact reference model. While centralized positions allow this visibility, partial topologies introduce calculation errors. This reliance highlights the partial mesh topology as a highly effective, natural defense mechanism. By ensuring a victim aggregates models from a subset of hidden neighbors, the partial mesh structure inherently obscures the victim's starting weights. This architectural uncertainty proves resilient, successfully degrading the quality of the attack even during the most vulnerable early stages of training.

8.2 Current Limitations

Despite the insights obtained, this dissertation acknowledges some limitations in the current experimental framework. Primarily, the evaluated white-box MIA demonstrated unexpectedly low effectiveness, particularly within Non-IID environments where its performance frequently dropped below that of black-box alternatives. This suggests that the implemented white-box attack may not fully capitalize on the complete parameter transparency it is afforded, meaning the true upper bound of white-box privacy leakage remains partially unmapped.

Furthermore, the simulations were conducted using a restricted number of participating nodes to maintain computational feasibility. While this effectively demonstrates the foundational dynamics of P2P and hierarchical communications, it may not perfectly capture the complex aggregation behaviors found in distinct networks topologies deployed in the real world scenarios. Finally, the evaluations were limited to the MNIST and CIFAR10 datasets paired with relatively straightforward CNN architectures. While sufficient for benchmarking, these configurations do not encompass the intricacies of high resolution image datasets or the immense memorization capacities of deeper, state-of-the-art models.

8.3 Future Work

Considering the conclusions drawn from this dissertation, several promising avenues for future research emerge to further secure distributed learning systems. A critical first step should be the exploration and implementation of more sophisticated white-box MIAs tailored specifically for highly heterogeneous Non-IID environments. Establishing a rigorous and realistic upper bound for privacy leakage under full adversary knowledge is essential for understanding worst-case scenarios. Additionally, extending this evaluation to deeper network architectures, such as ResNets and Vision Transformers, and applying them to higher resolution datasets will clarify how massive parameter counts and increased memorization capacity influence both MIA and GIA vulnerabilities, as well as how they impact the privacy shield effect of hierarchical topologies.

While the tree topology provides passive mitigation, active defense mechanisms are necessary for comprehensive security. Future research could evaluate the integration of Secure Aggregation directly into the most vulnerable nodes. Specifically, investigating the deployment of these defenses at intermediate edge nodes is suggested to observe whether it can neutralize localized vulnerabilities while maintaining the inherent architectural protection against root level adversaries. Finally, scaling the experimental framework to support thousands of nodes would provide a much clearer picture of how extreme decentralization and massive partial mesh connectivity behave in production, allowing for the observation of scalability trade-offs between communication efficiency, model convergence, and privacy guarantees.

Bibliography

[1] A. Bhushan and P. Misra, "Unlocking the potential: Multimodal AI in biotechnology and digital medicineeconomic impact and ethical challenges," *npj Digital Medicine*, vol. 8, no. 1, p. 619, Oct. 20, 2025, issn: 2398-6352. doi: 10.1038/s41746-025-01992-6.

[2] P. Voigt and A. v. d. Bussche, *The EU General Data Protection Regulation (GDPR): A Practical Guide*, 1st. Springer Publishing Company, Incorporated, 2017, isbn: 3319579584.

[3] J. Konený, H. B. McMahan, D. Ramage, and P. Richtárik, *Federated optimization: Distributed machine learning for on-device intelligence*, 2016. eprint: 1610.02527 (cs.LG). [Online]. Available: https://arxiv.org/abs/1610.02527.

[4] S. Garst, J. Dekker, and M. Reinders, "A comprehensive experimental comparison between federated and centralized learning," *Database*, vol. 2025, baaf016, Mar. 2025, issn: 1758-0463. doi: 10.1093/database/baaf016.

[5] J. Wu, F. Dong, H. Leung, Z. Zhu, J. Zhou, and S. Drew, *Topology-aware federated learning in edge computing: A comprehensive survey*, New York, NY, USA, 2024. doi: 10.1145/3659205.

[6] Q. Yang, Y. Liu, T. Chen, and Y. Tong, *Federated machine learning: Concept and applications*, New York, NY, USA, Jan. 2019. doi: 10.1145/3298981.

[7] A. Majeed and S. O. Hwang, *A multifaceted survey on federated learning: Fundamentals, paradigm shifts, practical issues, recent developments, partnerships, trade-offs, trustworthiness, and ways forward*, 2024. doi: 10.1109/ACCESS.2024.3413069.

[8] L. Yuan, Z. Wang, L. Sun, P. S. Yu, and C. G. Brinton, *Decentralized federated learning: A survey and perspective*, 2024. doi: 10.1109/JIOT.2024.3407584.

[9] J. Geiping, H. Bauermeister, H. Dröge, and M. Moeller, "Inverting gradients - how easy is it to break privacy in federated learning?" In *34th International Conference on Neural Information Processing Systems*, ser. NIPS '20, Vancouver, BC, Canada: Curran Associates Inc., 2020, isbn: 9781713829546.

[10] D. Jin, N. Kannengießer, S. Rank, and A. Sunyaev, *Collaborative distributed machine learning*, New York, NY, USA, 2024. doi: 10.1145/3704807.

[11] X. Yin, Y. Li, C. Bai, Q. Han, and Y. Chen, "Enhancing membership inference attacks in federated learning based on overfitting property," in *Proceedings of the 20th International Conference on Mobility, Sensing and Networking (MSN)*, 2024, pp. 989–996. doi: 10.1109/MSN63567.2024.00135.

[12] L. Zhang, L. Li, X. Li, *et al.*, "Efficient membership inference attacks against federated learning via bias differences," in *Proceedings of the 26th International Symposium on Research in Attacks, Intrusions and Defenses*, ser. RAID '23, Hong Kong, China: Association for Computing Machinery, 2023, pp. 222–235, isbn: 9798400707650. doi: 10.1145/3607199.3607204.

[13] X. He, Y. Xu, S. Zhang, W. Xu, and J. Yan, "Enhance membership inference attacks in federated learning," *Computers Security*, vol. 136, p. 103 535, 2024, issn: 0167-4048. doi: 10.1016/j.cose.2023.103535.

[14] Y. Zhao, J. Chen, J. Zhang, *et al.*, "User-level membership inference for federated learning in wireless network environment," *Wireless Communications and Mobile Computing*, vol. 2021, no. 1, p. 5 534 270, 2021. doi: 10.1155/2021/5534270.

[15] A. Luqman, A. Chattopadhyay, and K.-Y. Lam, "Membership inference vulnerabilities in peer-to-peer federated learning," in *Proceedings of the 2023 Secure and Trustworthy Deep Learning Systems Workshop*, ser. SecTL '23, Melbourne, VIC, Australia: Association for Computing Machinery, 2023, isbn: 9798400701818. doi: 10.1145/3591197.3593638.

[16] D. Pasquini, M. Raynal, and C. Troncoso, "On the (in)security of peer-to-peer decentralized machine learning," in *44th IEEE Symposium on Security and Privacy*, 2023, pp. 418–436. doi: 10.1109/SP46215.2023.10179291.

[17] Y. Huang, S. Gupta, Z. Song, K. Li, and S. Arora, "Evaluating gradient inversion attacks and defenses in federated learning," in *Proceedings of the 35th International Conference on Neural Information Processing Systems*, ser. NIPS '21, Red Hook, NY, USA: Curran Associates Inc., 2021, isbn: 9781713845393.

[18] H. Yin, A. Mallya, A. Vahdat, J. M. Alvarez, J. Kautz, and P. Molchanov, "See through gradients: Image batch recovery via gradinversion," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 16 332–16 341. doi: 10.1109/CVPR46437.2021.01607.

[19] K. Xiang, H. Yang, M. Hao, *et al.*, "An advanced gradient leakage attack against duplicate labels via model outputs reconstruction," *IEEE Transactions on Dependable and Secure Computing*, vol. 23, no. 3, 2026. doi: 10.1109/TDSC.2026.3659196.

[20] W. Yu, Q. Li, M. Lopuhaä-Zwakenberg, M. Græsbøll Christensen, and R. Heusdens, *Provable privacy advantages of decentralized federated learning via distributed optimization*, 2025. doi: 10.1109/TIFS.2024.3516564.

[21] Q. Li, W. Yu, C. Ji, and R. Heusdens, "Topology-dependent privacy bound for decentralized federated learning," in *49th IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024, pp. 5940–5944. doi: 10.1109/ICASSP48485.2024.10446209.

[22] G. Lu, Z. Xiong, R. Li, N. Mohammad, Y. Li, and W. Li, *Defeat: A decentralized federated learning against gradient attacks*, 2023. doi: 10.1016/j.hcc.2023.100128.

[23] C. Ji, R. Heusdens, S. Maag, and Q. Li, "Re-evaluating privacy in centralized and decentralized learning: An information-theoretical and empirical study," in *50th IEEE International Conference on Acoustics, Speech and Signal Processing*, 2025, pp. 1–5. doi: 10.1109/ICASSP49660.2025.10888257.

[24] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, *Federated learning: Challenges, methods, and future directions*, 2020. doi: 10.1109/MSP.2020.2975749.

[25] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *20th International Conference on Artificial Intelligence and Statistics*, A. Singh and J. Zhu, Eds., ser. Proceedings of Machine Learning Research, vol. 54, PMLR, 2017, pp. 1273–1282.

[26] P. Qi, D. Chiaro, A. Guzzo, M. Ianni, G. Fortino, and F. Piccialli, *Model aggregation techniques in federated learning: A comprehensive survey*, 2024. doi: 10.1016/j.future.2023.09.008.

[27] X. Zhou, W. Liang, A. Kawai, K. Fueda, J. She, and K. I.-K. Wang, *Adaptive segmentation enhanced asynchronous federated learning for sustainable intelligent transportation systems*, 2024. doi: 10.1109/TITS.2024.3362058.

[28] Z. Zhai, Q. Wu, S. Yu, R. Li, F. Zhang, and X. Chen, *Fedleo: An offloading-assisted decentralized federated learning framework for low earth orbit satellite networks*, 2024. doi: 10.1109/TMC.2023.3304988.

[29] B. B. Gupta, A. Gaurav, and V. Arya, *Secure and privacy-preserving decentralized federated learning for personalized recommendations in consumer electronics using blockchain and homomorphic encryption*, 2024. doi: 10.1109/TCE.2023.3329480.

[30] S. Baghersalimi, T. Teijeiro, A. Aminifar, and D. Atienza, *Decentralized federated learning for epileptic seizures detection in low-power wearable systems*, 2024. doi: 10.1109/TMC.2023.3320862.

[31] A. Lalitha, S. Shekhar, T. Javidi, and F. Koushanfar, "Fully decentralized federated learning," in *3rd workshop on bayesian deep learning*, vol. 12, 2018.

[32] X. Lian, C. Zhang, H. Zhang, C.-J. Hsieh, W. Zhang, and J. Liu, "Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent," in *31st International Conference on Neural Information Processing Systems*, ser. NIPS'17, Long Beach, California, USA: Curran Associates Inc., 2017, pp. 5336–5346, isbn: 9781510860964.

[33] K. Bonawitz, V. Ivanov, B. Kreuter, *et al.*, "Practical secure aggregation for privacy-preserving machine learning," in *24th ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '17, Dallas, Texas, USA: Association for Computing Machinery, 2017, pp. 1175–1191, isbn: 9781450349468. doi: 10.1145/3133956.3133982.

[34] W. Diffie, *Diffie-hellman key exchange*, 1976.

[35] A. Shamir, "How to share a secret," *Communications of the ACM journal*, vol. 22, pp. 612–613, 1979. [Online]. Available: https://cir.nii.ac.jp/crid/1572543026020340992.

[36] H. Hu, Z. Salcic, L. Sun, G. Dobbie, P. S. Yu, and X. Zhang, *Membership inference attacks on machine learning: A survey*, New York, NY, USA, Sep. 2022. doi: 10.1145/3523273.

[37] F. Boenisch, A. Dziedzic, R. Schuster, A. S. Shamsabadi, I. Shumailov, and N. Papernot, "When the curious abandon honesty: Federated learning is not private," in *8th European Symposium on Security and Privacy*, 2023, pp. 175–199. doi: 10.1109/EuroSP57164.2023.00020.

[38] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," in *33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA: Curran Associates Inc., 2019.

[39] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE Symposium on Security and Privacy (SP)*, 2017, pp. 3–18. doi: 10.1109/SP.2017.41.

[40] L. Liu, Y. Wang, G. Liu, K. Peng, and C. Wang, *Membership inference attacks against machine learning models via prediction sensitivity*, 2023. doi: 10.1109/TDSC.2022.3180828.

[41] P. He, X. Wang, W. Zhang, *et al.*, "Evaluating differential privacy in federated learning based on membership inference attacks," in *Proceedings of the 10th International Conference on Big Data Computing and Communications (BigCom)*, 2024, pp. 196–203. doi: 10.1109/BIGCOM65357.2024.00035.

[42] L. Song and P. Mittal, "Systematic evaluation of privacy risks of machine learning models," in *Proceedings of the 30th USENIX Security Symposium (USENIX Security 21)*, USENIX Association, Aug. 2021, pp. 2615–2632, isbn: 978-1-939133-24-3.

[43] M. J. Page, J. E. McKenzie, P. M. Bossuyt, *et al.*, "The prisma 2020 statement: An updated guideline for reporting systematic reviews," *BMJ*, vol. 372, 2021. doi: 10.1136/bmj.n71.

[44] E. T. Martínez Beltrán, M. Q. Pérez, P. M. S. Sánchez, *et al.*, *Decentralized federated learning: Fundamentals, state of the art, frameworks, trends, and challenges*, 2023. doi: 10.1109/COMST.2023.3315746.

[45] Q. W. Khan, A. N. Khan, A. Rizwan, R. Ahmad, S. Khan, and D.-H. Kim, *Decentralized machine learning training: A survey on synchronization, consolidation, and topologies*, 2023. doi: 10.1109/ACCESS.2023.3284976.

[46] H. Chen, H. Wang, Q. Long, D. Jin, and Y. Li, *Advancements in federated learning: Models, methods, and privacy*, New York, NY, USA, 2024. doi: 10.1145/3664650.

[47] S. Warnat-Herresthal, H. Schultze, K. L. Shastry, *et al.*, *Swarm learning for decentralized and confidential clinical machine learning*, Jun. 1, 2021. doi: 10.1038/s41586-021-03583-3.

[48] A. Krizhevsky, G. Hinton, *et al.*, "Learning multiple layers of features from tiny images," 2009.

[49] L. Deng, *The mnist database of handwritten digit images for machine learning research [best of the web]*, 2012. doi: 10.1109/MSP.2012.2211477.

[50] Y. Gu, Y. Bai, and S. Xu, "Cs-mia: Membership inference attack based on prediction confidence series in federated learning," *Journal of Information Security and Applications*, vol. 67, p. 103 201, 2022, issn: 2214-2126. doi: 10.1016/j.jisa.2022.103201.

[51] G. Zhu, D. Li, H. Gu, Y. Yao, L. Fan, and Y. Han, "Fedmia: An effective membership inference attack exploiting "all for one" principle in federated learning," in *Proceedings of the 42th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 20 643–20 653. doi: 10.1109/CVPR52734.2025.01922.

[52] J. Huang, L. Y. Chen, and S. Roos, "On quantifying the gradient inversion risk of data reuse in federated learning systems," in *2024 43rd International Symposium on Reliable Distributed Systems (SRDS)*, 2024, pp. 235–247. doi: 10.1109/SRDS64841.2024.00031.

[53] T.-M. H. Hsu, H. Qi, and M. Brown, *Measuring the effects of non-identical data distribution for federated visual classification*, 2019. arXiv: 1909.06335 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1909.06335.

[54] A. Brauneck, L. Schmalhorst, M. M. Kazemi Majdabadi, *et al.*, *Federated machine learning, privacy-enhancing technologies, and data protection laws in medical research: Scoping review*, Mar. 2023. doi: 10.2196/41588.

[55] J. Casaletto, A. Bernier, R. McDougall, and M. Cline, *Federated analysis for privacy-preserving data sharing: A technical and legal primer*, May 2023. doi: 10.1146/annurev-genom-110122-084756.

[56] R. Xu, N. Baracaldo, Y. Zhou, A. Anwar, S. Kadhe, and H. Ludwig, "Detrust-fl: Privacy-preserving federated learning in decentralized trust setting," in *15th International Conference on Cloud Computing*, 2022, pp. 417–426. doi: 10.1109/CLOUD55607.2022.00065.

[57] L. Yuan, Z. Wang, and C. G. Brinton, *Digital ethics in federated learning*, 2024. doi: 10.1109/MIC.2024.3370408.

[58] L. Palmieri, C. Boldrini, L. Valerio, A. Passarella, and M. Conti, "Impact of network topology on the performance of decentralized federated learning," *Computer Networks*, vol. 253, p. 110 681, 2024, issn: 1389-1286. doi: https://doi.org/10.1016/j.comnet.2024.110681.

[59] J. Kang, Z. Xiong, D. Niyato, Y. Zou, Y. Zhang, and M. Guizani, *Reliable federated learning for mobile networks*, 2020. doi: 10.1109/MWC.001.1900119.

[60] K. M. Kostick-Quenet, M. C. Compagnucci, M. Aboy, and T. Minssen, *Patient-centric federated learning: Automating meaningful consent to health data sharing with smart contracts*, Apr. 2025. doi: 10.1093/jlb/lsaf003.

[61] A. Tariq, M. A. Serhani, F. M. Sallabi, *et al.*, *Trustworthy federated learning: A comprehensive review, architecture, key challenges, and future research prospects*, 2024. doi: 10.1109/OJCOMS.2024.3438264.