Sybil-Based Data Poisoning Attack in Federated Learning

期末專題提案

F74124723資訊系周昀軒

Motivation

Sybil-based Virtual Data Poisoning Attacks in Federated Learning*

Changxun Zhu[1], Qilong Wu[1], Lingjuan Lyu[2] and Shibei Xue[1]

***Abstract*— Federated learning is vulnerable to poisoning attacks by malicious adversaries. Existing methods often involve high costs to achieve effective attacks. To address this challenge, we propose a sybil-based virtual data poisoning attack, where a malicious client generates sybil nodes to amplify the poisoning model's impact. To reduce neural network computational complexity, we develop a virtual data generation method based on gradient matching. We also design three schemes for target model acquisition, applicable to online local, online global, and offline scenarios. In simulation, our method outperforms other attack algorithms since our method can obtain a global target model under non-independent uniformly distributed data.**

***Keywords*—Federated learning, Sybil poisoning attack, Virtual data.**

I. INTRODUCTION

The revolution in sensing technology has enabled high-quality data acquisition and processing across diverse real-world applications. This technological progress has catalyzed significant advancements in artificial intelligence (AI), achieving state-of-the-art performance in specialized domains including natural language processing [1], recommender systems [2], [3], pose estimation [4], [5], intelligent transportation [6], [7], energy-related prediction [8]–[10].

However, with the growing emphasis on data privacy and the introduction of data protection regulations, traditional centralized machine learning approaches face significant obstacles [11]. To address this, federated learning (FL) [12] emerges as a privacy-preserving paradigm. FL establishes a shared model on a central server, distributes the model to clients for training on local data, and subsequently aggregates the locally trained models on the server. This framework avoids direct data transmission, thereby preserving client privacy [13].

While federated learning preserves data locality on client devices, it also introduces new challenges. The inability to filter user data results in non-independent and identically distributed (Non-IID) data, leading to model drift and prolonged convergence times for optimal performance [14]. Additionally, the lack of data filtering makes federated learning susceptible to attacks by malicious adversaries.

To address the aforementioned challenges, designing federated learning defense algorithms to enhance the stability of the federated learning process is a practical approach [15]. Another perspective is to deepen the study of federated learning attack algorithms to understand potential security risks, thereby improving the security and privacy protection of federated learning. The latter can better understand the attack process from the attacker's perspective, which is more conducive to our formulation of proactive defense strategies.

The earliest poisoning attack was introduced against Support Vector Machines (SVM) by flipping the labels of training data [16]. Although originally designed for centralized settings, this attack is found to be effective in federated learning scenarios [17]. Based on the structural characteristics of federated learning, the following poisoning attacks can be categorized into three types: data poisoning, model poisoning, and sybil-based poisoning attacks.

In data poisoning attacks, adversaries cannot directly manipulate users' models but can access and tamper with client training data to execute attacks. Ref. [17] first introduced label-flipping attacks to federated learning, where malicious actors flip sample labels, causing the trained model to deviate from the intended prediction boundary. However, the effectiveness of this approach is limited by the influence of non-malicious clients. To address this, Ref. [18] proposed a dynamic label-flipping strategy that selects the target label with the smallest loss, improving on static label-flipping methods. Beyond label-flipping attacks, clean-label poisoning is another common data poisoning approach. This technique retains original labels but injects malicious patterns into model parameters through image pixel optimization [16]. However, this approach is computationally expensive for deep neural networks. To overcome this limitation, heuristic methods have been proposed, as demonstrated in Refs. [19], [20], to achieve clean-label poisoning more efficiently.

However, the success rate of data poisoning attacks is directly proportional to the number of malicious clients controlled by the attacker, making such attacks costly in large-scale federated learning systems. To address this limitation, model poisoning attacks were introduced, enabling adversaries to manipulate the local training process. Ref. [21] demonstrated an attack executed when the global model nears convergence, modifying the local training process by adding an anomaly detection term to the loss function. In contrast, Ref. [22] proposed an attack targeting the early stages of global model training, before convergence is achieved. Additionally, Ref. [23] leveraged a regularization term in the objective function to embed malicious neurons into the redundant spaces of neural networks. This approach minimizes the impact of benign clients during model aggregation, allowing the attacker to execute poisoning attacks effectively.

Sybil-based attacks are another common method of dis-

*This work was supported in part by the National Natural Science Foundation of China under Grant 62273226 and Grant 61873162. (Corresponding author: Shibei Xue)

[1]Department of Automation, Shanghai Jiao Tong University, Shanghai 200240, P.R. China (e-mail: shbxue@sjtu.edu.cn).

[2]Sony AI

2025/05/15 Sybil-based Virtual Data Poisoning Attacks in Federated Learning

SPoiL: Sybil-Based Untargeted Data Poisoning Attacks in Federated Learning

Zhuotao Lian, Chen Zhang, Kaixi Nan, and Chunhua Su(✉)

Department of Computer Science and Engineering, The University of Aizu, Aizuwakamatsu, Japan
chsu@u-aizu.ac.jp

Abstract. Federated learning is widely used in mobile computing, the Internet of Things, and other scenarios due to its distributed and privacy-preserving nature. It allows mobile devices to train machine learning models collaboratively without sharing their local private data. However, during the model aggregation phase, federated learning is vulnerable to poisoning attacks carried out by malicious users. Furthermore, due to the heterogeneity of network status, communication conditions, hardware, and other factors, users are at high risk of offline, which allows attackers to fake virtual participants and increase the damage of poisoning. Unlike existing work, we focus on the more general case of untargeted poisoning attacks. In this paper, we propose novel sybil-based untargeted data poisoning attacks in federated learning (SPoiL), in which malicious users corrupt the performance of the global model by modifying the training data and increasing the probability of poisoning by virtualizing several sybil nodes. Finally, we validate the superiority of our attack approach through experiments across the commonly used datasets.

Keywords: Federated learning · Poisoning attacks · Sybil · Distributed learning

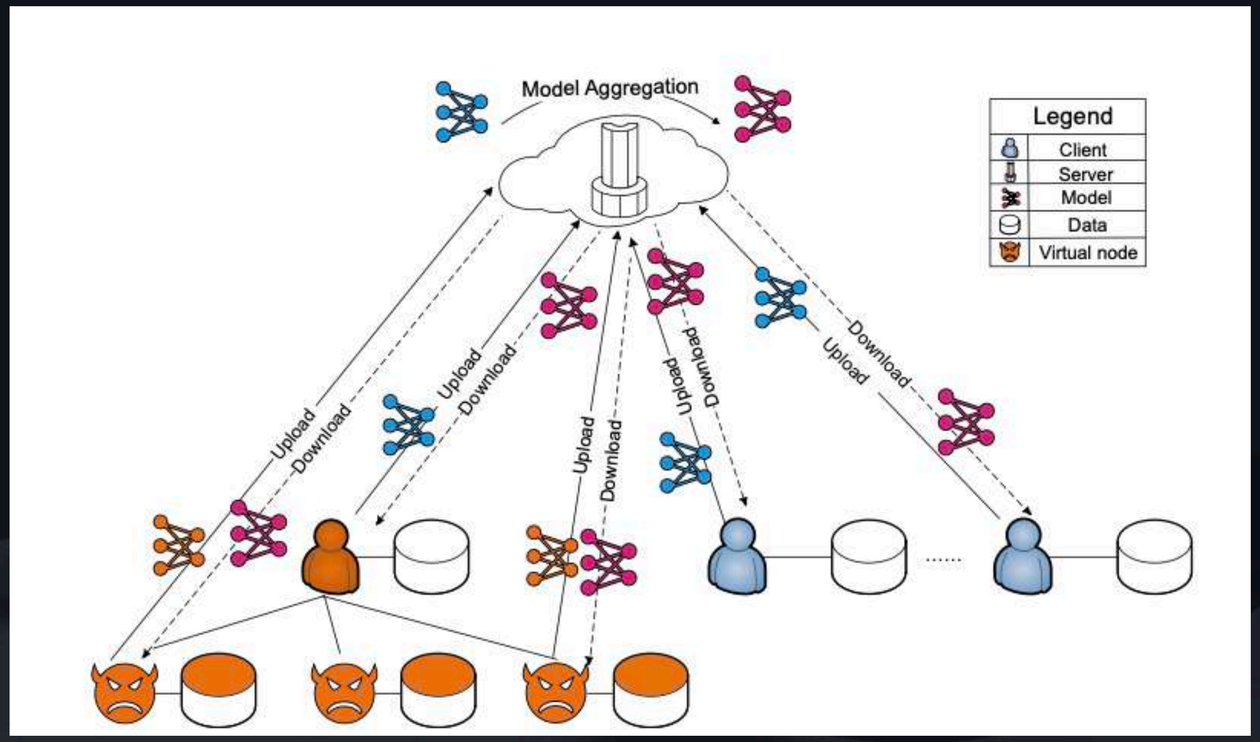
1 Introduction

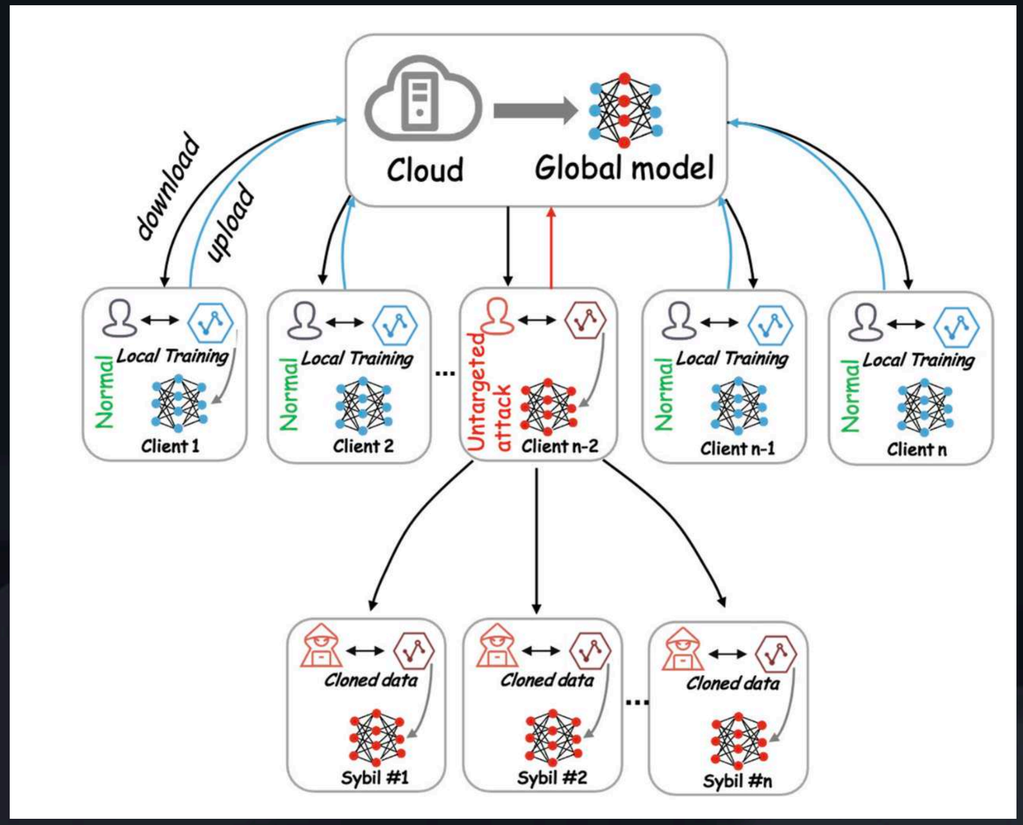
Federated learning has emerged as a prominent distributed machine learning paradigm that enables collaboration among data owners without the need to share sensitive data. It allows each participant to train a local model using their private data and then aggregate the models' parameters to create a global model. This approach has found applications in various domains, including finance, recommendation systems, and healthcare, due to its privacy-preserving nature and compliance with data privacy regulations [13]. One of the key advantages of federated learning is its ability to facilitate collaboration between distrustful clients, such as competing banks or mobile phone users [14]. By enabling collaboration without compromising data privacy, federated learning allows competitors to benefit from shared insights and advancements while preserving their business interests and privacy. For example, banks can collectively train credit

S. Li et al. (Eds.): NSS 2023, LNCS 13983, pp. 235–248, 2023.
https://doi.org/10.1007/978-3-031-39828-5_13

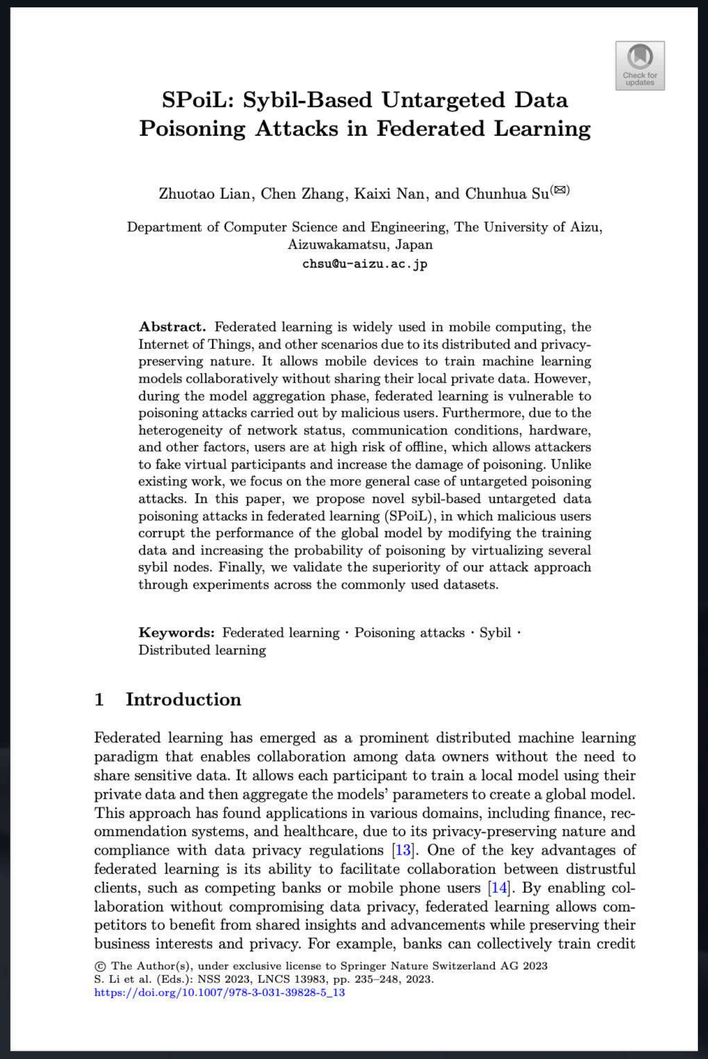
2023 SPoiL: Sybil-Based Untargeted Data Poisoning Attacks in Federated Learning

Motivation


Model Aggregation
Legend
Client
Server
Model
Data
Virtual node
Upload
Download


Cloud
Global model
download
upload
Normal
Local Training
Client 1
Client 2
Untargeted attack
Client n-2
Client n-1
Client n
Cloned data
Sybil #1
Sybil #2
Sybil #n

Motivation



SPoiL: Sybil-Based Untargeted Data Poisoning Attacks in Federated Learning

Zhuotao Lian, Chen Zhang, Kaixi Nan, and Chunhua Su(✉)

Department of Computer Science and Engineering, The University of Aizu, Aizuwakamatsu, Japan
chsu@u-aizu.ac.jp

Abstract. Federated learning is widely used in mobile computing, the Internet of Things, and other scenarios due to its distributed and privacy-preserving nature. It allows mobile devices to train machine learning models collaboratively without sharing their local private data. However, during the model aggregation phase, federated learning is vulnerable to poisoning attacks carried out by malicious users. Furthermore, due to the heterogeneity of network status, communication conditions, hardware, and other factors, users are at high risk of offline, which allows attackers to fake virtual participants and increase the damage of poisoning. Unlike existing work, we focus on the more general case of untargeted poisoning attacks. In this paper, we propose novel sybil-based untargeted data poisoning attacks in federated learning (SPoiL), in which malicious users corrupt the performance of the global model by modifying the training data and increasing the probability of poisoning by virtualizing several sybil nodes. Finally, we validate the superiority of our attack approach through experiments across the commonly used datasets.

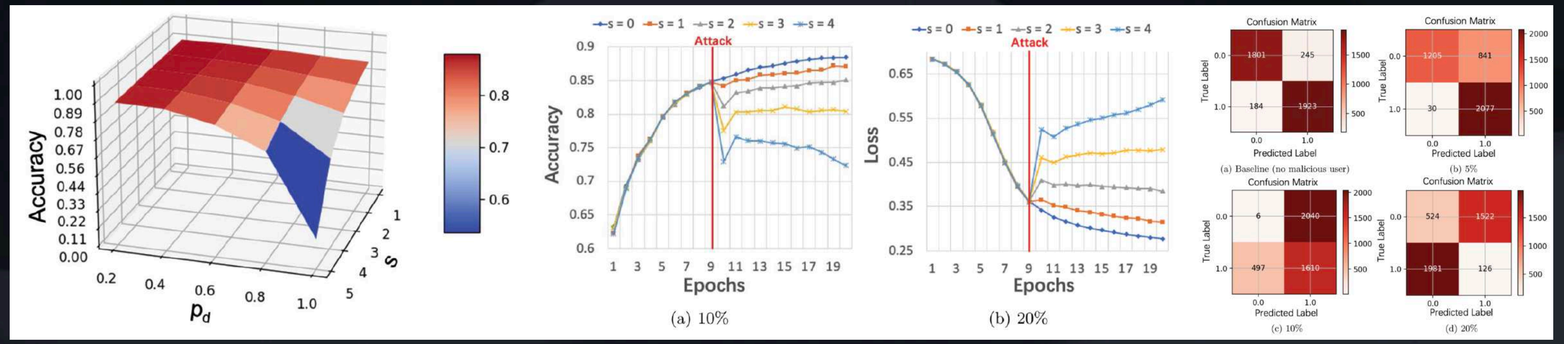
Keywords: Federated learning · Poisoning attacks · Sybil · Distributed learning

1 Introduction

Federated learning has emerged as a prominent distributed machine learning paradigm that enables collaboration among data owners without the need to share sensitive data. It allows each participant to train a local model using their private data and then aggregate the models' parameters to create a global model. This approach has found applications in various domains, including finance, recommendation systems, and healthcare, due to its privacy-preserving nature and compliance with data privacy regulations [13]. One of the key advantages of federated learning is its ability to facilitate collaboration between distrustful clients, such as competing banks or mobile phone users [14]. By enabling collaboration without compromising data privacy, federated learning allows competitors to benefit from shared insights and advancements while preserving their business interests and privacy. For example, banks can collectively train credit

S. Li et al. (Eds.): NSS 2023, LNCS 13983, pp. 235–248, 2023.
https://doi.org/10.1007/978-3-031-39828-5_13

* untargeted attack
  * 對data做隨機label flipping
* Sybil attack
  * 對受控制的參與者新增Sybil 節點
  * 將label flipping過後的資料複製到Sybil 節點並進行訓練
  * 放大中毒效果
* Result



2023 SPoiL: Sybil-Based Untargeted Data Poisoning Attacks in Federated Learning

Kaggle fake new , NN model

Motivation

arXiv:2505.09983v1 [cs.CR] 15 May 2025

Sybil-based Virtual Data Poisoning Attacks in Federated Learning*

Changxun Zhu[1], Qilong Wu[1], Lingjuan Lyu[2] and Shibei Xue[1]

***Abstract*— Federated learning is vulnerable to poisoning attacks by malicious adversaries. Existing methods often involve high costs to achieve effective attacks. To address this challenge, we propose a sybil-based virtual data poisoning attack, where a malicious client generates sybil nodes to amplify the poisoning model's impact. To reduce neural network computational complexity, we develop a virtual data generation method based on gradient matching. We also design three schemes for target model acquisition, applicable to online local, online global, and offline scenarios. In simulation, our method outperforms other attack algorithms since our method can obtain a global target model under non-independent uniformly distributed data.**

***Keywords*—Federated learning, Sybil poisoning attack, Virtual data.**

I. INTRODUCTION

The revolution in sensing technology has enabled high-quality data acquisition and processing across diverse real-world applications. This technological progress has catalyzed significant advancements in artificial intelligence (AI), achieving state-of-the-art performance in specialized domains including natural language processing [1], recommender systems [2], [3], pose estimation [4], [5], intelligent transportation [6], [7], energy-related prediction [8]–[10].

However, with the growing emphasis on data privacy and the introduction of data protection regulations, traditional centralized machine learning approaches face significant obstacles [11]. To address this, federated learning (FL) [12] emerges as a privacy-preserving paradigm. FL establishes a shared model on a central server, distributes the model to clients for training on local data, and subsequently aggregates the locally trained models on the server. This framework avoids direct data transmission, thereby preserving client privacy [13].

While federated learning preserves data locality on client devices, it also introduces new challenges. The inability to filter user data results in non-independent and identically distributed (Non-IID) data, leading to model drift and prolonged convergence times for optimal performance [14]. Additionally, the lack of data filtering makes federated learning susceptible to attacks by malicious adversaries.

To address the aforementioned challenges, designing federated learning defense algorithms to enhance the stability of the federated learning process is a practical approach [15]. Another perspective is to deepen the study of federated learning attack algorithms to understand potential security risks, thereby improving the security and privacy protection of federated learning. The latter can better understand the attack process from the attacker's perspective, which is more conducive to our formulation of proactive defense strategies.

The earliest poisoning attack was introduced against Support Vector Machines (SVM) by flipping the labels of training data [16]. Although originally designed for centralized settings, this attack is found to be effective in federated learning scenarios [17]. Based on the structural characteristics of federated learning, the following poisoning attacks can be categorized into three types: data poisoning, model poisoning, and sybil-based poisoning attacks.

In data poisoning attacks, adversaries cannot directly manipulate users' models but can access and tamper with client training data to execute attacks. Ref. [17] first introduced label-flipping attacks to federated learning, where malicious actors flip sample labels, causing the trained model to deviate from the intended prediction boundary. However, the effectiveness of this approach is limited by the influence of non-malicious clients. To address this, Ref. [18] proposed a dynamic label-flipping strategy that selects the target label with the smallest loss, improving on static label-flipping methods. Beyond label-flipping attacks, clean-label poisoning is another common data poisoning approach. This technique retains original labels but injects malicious patterns into model parameters through image pixel optimization [16]. However, this approach is computationally expensive for deep neural networks. To overcome this limitation, heuristic methods have been proposed, as demonstrated in Refs. [19], [20], to achieve clean-label poisoning more efficiently.

However, the success rate of data poisoning attacks is directly proportional to the number of malicious clients controlled by the attacker, making such attacks costly in large-scale federated learning systems. To address this limitation, model poisoning attacks were introduced, enabling adversaries to manipulate the local training process. Ref. [21] demonstrated an attack executed when the global model nears convergence, modifying the local training process by adding an anomaly detection term to the loss function. In contrast, Ref. [22] proposed an attack targeting the early stages of global model training, before convergence is achieved. Additionally, Ref. [23] leveraged a regularization term in the objective function to embed malicious neurons into the redundant spaces of neural networks. This approach minimizes the impact of benign clients during model aggregation, allowing the attacker to execute poisoning attacks effectively.

Sybil-based attacks are another common method of dis-

*This work was supported in part by the National Natural Science Foundation of China under Grant 62273226 and Grant 61873162. (Corresponding author: Shibei Xue)

[1]Department of Automation, Shanghai Jiao Tong University, Shanghai 200240, P.R. China (e-mail: shbxue@sjtu.edu.cn).

[2]Sony AI

* Virtual Data
  * 受控制的受害者先訓練一次(label flipping)產生出目標模型，並基於 Gradient Matching 的方式產生出中毒樣本傳給 Sybil 節點
  * 解決以前的攻擊成本昂貴
  * 解決non-IID引發的模型漂移、收斂時長延長等問題
  * Target attack
* Result
  * 在non-IID的分布下仍保持高效攻擊
  * 提供了高質量的投毒數據

Motivation

Sybil-based Virtual Data Poisoning Attacks in Federated Learning*

Changxun Zhu[1], Qilong Wu[1], Lingjuan Lyu[2] and Shibei Xue[1]

***Abstract*— Federated learning is vulnerable to poisoning attacks by malicious adversaries. Existing methods often involve high costs to achieve effective attacks. To address this challenge, we propose a sybil-based virtual data poisoning attack, where a malicious client generates sybil nodes to amplify the poisoning model's impact. To reduce neural network computational complexity, we develop a virtual data generation method based on gradient matching. We also design three schemes for target model acquisition, applicable to online local, online global, and offline scenarios. In simulation, our method outperforms other attack algorithms since our method can obtain a global target model under non-independent uniformly distributed data.**

***Keywords*—Federated learning, Sybil poisoning attack, Virtual data.**

I. INTRODUCTION

The revolution in sensing technology has enabled high-quality data acquisition and processing across diverse real-world applications. This technological progress has catalyzed significant advancements in artificial intelligence (AI), achieving state-of-the-art performance in specialized domains including natural language processing [1], recommender systems [2], [3], pose estimation [4], [5], intelligent transportation [6], [7], energy-related prediction [8]–[10].

However, with the growing emphasis on data privacy and the introduction of data protection regulations, traditional centralized machine learning approaches face significant obstacles [11]. To address this, federated learning (FL) [12] emerges as a privacy-preserving paradigm. FL establishes a shared model on a central server, distributes the model to clients for training on local data, and subsequently aggregates the locally trained models on the server. This framework avoids direct data transmission, thereby preserving client privacy [13].

While federated learning preserves data locality on client devices, it also introduces new challenges. The inability to filter user data results in non-independent and identically distributed (Non-IID) data, leading to model drift and prolonged convergence times for optimal performance [14]. Additionally, the lack of data filtering makes federated learning susceptible to attacks by malicious adversaries.

To address the aforementioned challenges, designing federated learning defense algorithms to enhance the stability of the federated learning process is a practical approach [15]. Another perspective is to deepen the study of federated learning attack algorithms to understand potential security risks, thereby improving the security and privacy protection of federated learning. The latter can better understand the attack process from the attacker's perspective, which is more conducive to our formulation of proactive defense strategies.

The earliest poisoning attack was introduced against Support Vector Machines (SVM) by flipping the labels of training data [16]. Although originally designed for centralized settings, this attack is found to be effective in federated learning scenarios [17]. Based on the structural characteristics of federated learning, the following poisoning attacks can be categorized into three types: data poisoning, model poisoning, and sybil-based poisoning attacks.

In data poisoning attacks, adversaries cannot directly manipulate users' models but can access and tamper with client training data to execute attacks. Ref. [17] first introduced label-flipping attacks to federated learning, where malicious actors flip sample labels, causing the trained model to deviate from the intended prediction boundary. However, the effectiveness of this approach is limited by the influence of non-malicious clients. To address this, Ref. [18] proposed a dynamic label-flipping strategy that selects the target label with the smallest loss, improving on static label-flipping methods. Beyond label-flipping attacks, clean-label poisoning is another common data poisoning approach. This technique retains original labels but injects malicious patterns into model parameters through image pixel optimization [16]. However, this approach is computationally expensive for deep neural networks. To overcome this limitation, heuristic methods have been proposed, as demonstrated in Refs. [19], [20], to achieve clean-label poisoning more efficiently.

However, the success rate of data poisoning attacks is directly proportional to the number of malicious clients controlled by the attacker, making such attacks costly in large-scale federated learning systems. To address this limitation, model poisoning attacks were introduced, enabling adversaries to manipulate the local training process. Ref. [21] demonstrated an attack executed when the global model nears convergence, modifying the local training process by adding an anomaly detection term to the loss function. In contrast, Ref. [22] proposed an attack targeting the early stages of global model training, before convergence is achieved. Additionally, Ref. [23] leveraged a regularization term in the objective function to embed malicious neurons into the redundant spaces of neural networks. This approach minimizes the impact of benign clients during model aggregation, allowing the attacker to execute poisoning attacks effectively.

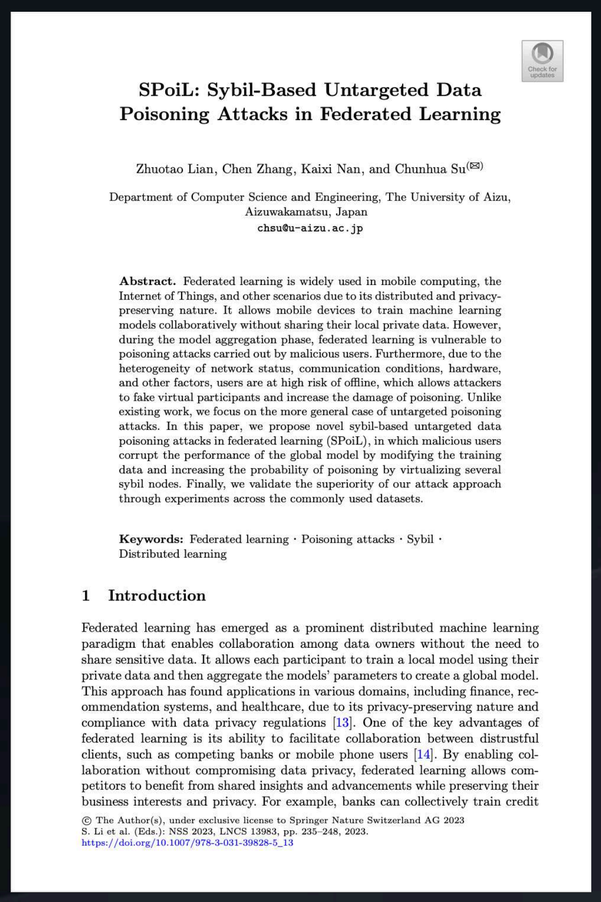
Sybil-based attacks are another common method of dis-

*This work was supported in part by the National Natural Science Foundation of China under Grant 62273226 and Grant 61873162. (Corresponding author: Shibei Xue)

[1]Department of Automation, Shanghai Jiao Tong University, Shanghai 200240, P.R. China (e-mail: shbxue@sjtu.edu.cn).

[2]Sony AI

+



SPoiL: Sybil-Based Untargeted Data Poisoning Attacks in Federated Learning

Zhuotao Lian, Chen Zhang, Kaixi Nan, and Chunhua Su(✉)

Department of Computer Science and Engineering, The University of Aizu, Aizuwakamatsu, Japan

chsu@u-aizu.ac.jp

Abstract. Federated learning is widely used in mobile computing, the Internet of Things, and other scenarios due to its distributed and privacy-preserving nature. It allows mobile devices to train machine learning models collaboratively without sharing their local private data. However, during the model aggregation phase, federated learning is vulnerable to poisoning attacks carried out by malicious users. Furthermore, due to the heterogeneity of network status, communication conditions, hardware, and other factors, users are at high risk of offline, which allows attackers to fake virtual participants and increase the damage of poisoning. Unlike existing work, we focus on the more general case of untargeted poisoning attacks. In this paper, we propose novel sybil-based untargeted data poisoning attacks in federated learning (SPoiL), in which malicious users corrupt the performance of the global model by modifying the training data and increasing the probability of poisoning by virtualizing several sybil nodes. Finally, we validate the superiority of our attack approach through experiments across the commonly used datasets.

Keywords: Federated learning · Poisoning attacks · Sybil · Distributed learning

1 Introduction

Federated learning has emerged as a prominent distributed machine learning paradigm that enables collaboration among data owners without the need to share sensitive data. It allows each participant to train a local model using their private data and then aggregate the models' parameters to create a global model. This approach has found applications in various domains, including finance, recommendation systems, and healthcare, due to its privacy-preserving nature and compliance with data privacy regulations [13]. One of the key advantages of federated learning is its ability to facilitate collaboration between distrustful clients, such as competing banks or mobile phone users [14]. By enabling collaboration without compromising data privacy, federated learning allows competitors to benefit from shared insights and advancements while preserving their business interests and privacy. For example, banks can collectively train credit

S. Li et al. (Eds.): NSS 2023, LNCS 13983, pp. 235–248, 2023.
https://doi.org/10.1007/978-3-031-39828-5_13

=

Sybil-based Virtual Untargeted Data Poisoning Attack

Target

Sybil-based Virtual Untarget Data Poisoning Attack

* 使用Gradient Matching產生高質量的毒Data

* 降低攻擊成本

* 在non-IID的情況下能具有高效的攻擊

* Untargeted Poisoning 降低模型準確率+並使用Sybil 放大Untargeted Poisoning 成效

Threat Model

- 參與者：系統共有 $K = N + M \cdot v$ 個節點，其中 N 為Honest users， M 為Malicious users，惡意者為每個客戶端生成 v 個 Sybil 節點
- 攻擊目標：降低全局模型的整體準確率
- 攻擊能力：僅可修改本地訓練數據，不繞過安全協議，可完全獲取並利用當前全局模型參數
- 數據分佈：允許 Non-IID

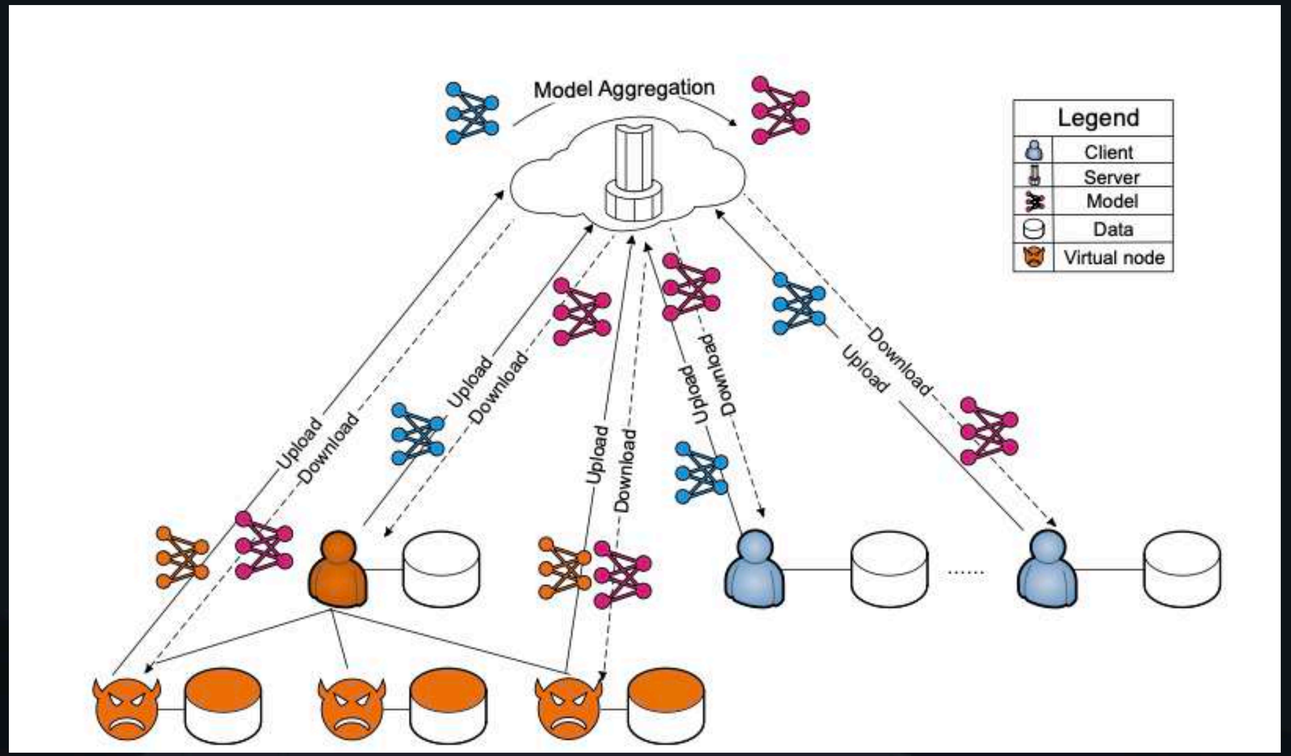
Setup

Experimental

- DataSet
  - MNIST
  - FMNIST
  - CIFAR-10
- Network
  - Fully Connected Neural Network (FC)
  - Convolutional Neural Network (CNN)
  - ResNet18
- FL
  - FedAvg

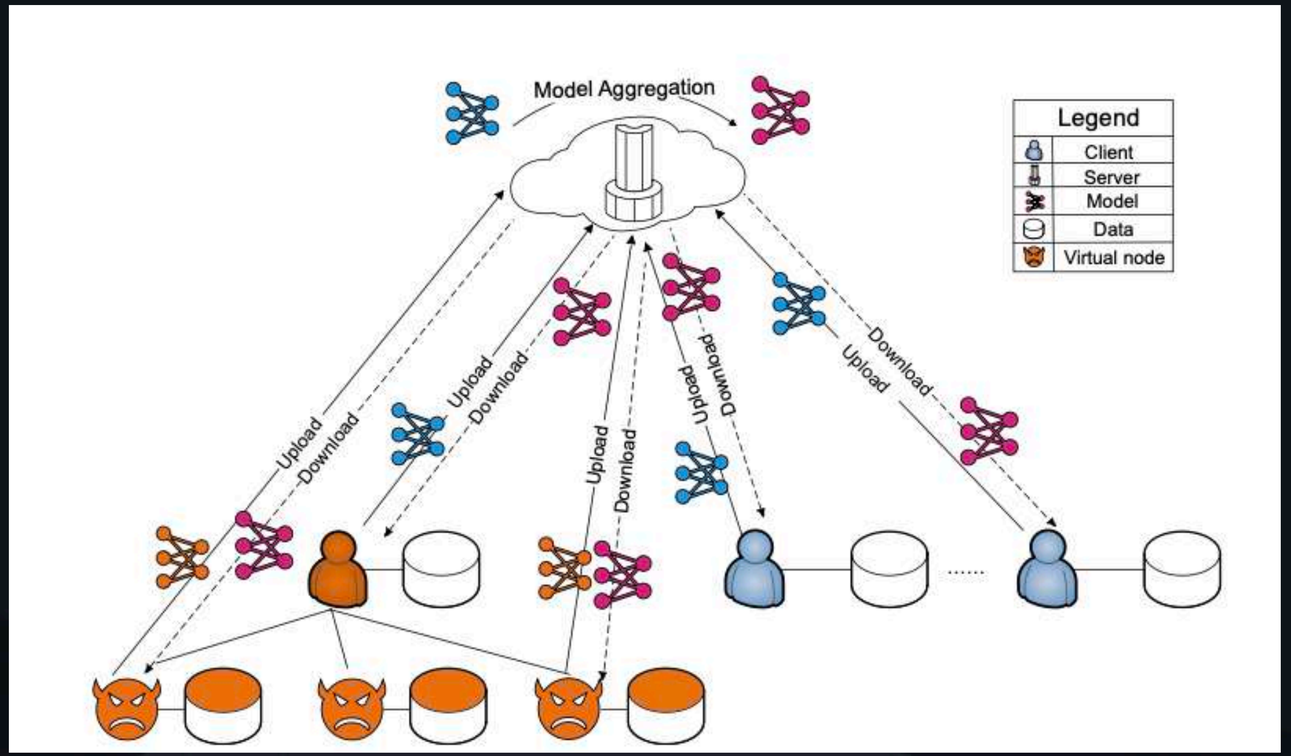
Process

- Server
  - 決定全局模型w
  - 從所有Client 選n個人出來
  - 將當前全局模型下放給這n個Client
  - 按FedAvg聚合所有w_i，更新得到新的全局模型w2



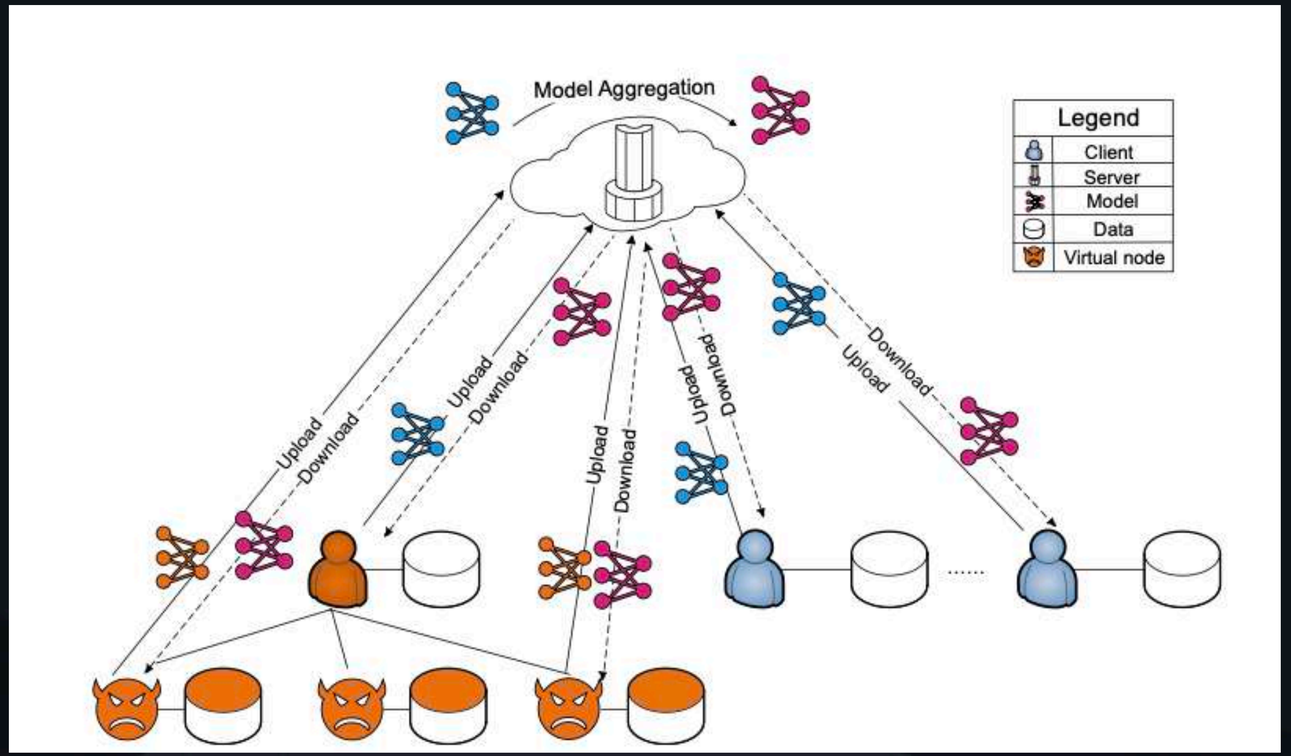
Process

- Client
  - Honest users
    - 使用本地資料進行E輪的SGD ，更新得到w_i
  - Malicious users
    - 得到目標模型
    - 中毒數據生成
  - Sybil 節點
    - 收到中毒數據後，進行E輪的訓練，得到w_i



Process

- 中毒數據生成
  - 選擇得到目標模型的隨機n筆資料(x,y)
  - 使用Gradient Matching計算Δ擾動，並進行T次
  - 將取出來的資料加上擾動(x+Δ,y)
  - 傳給Sybil 節點



Result

- 使用Main Task Accuracy 以及 Poisoning Success Rate 驗證此攻擊效果
- 比較2023 SPoiL: Sybil-Based Untargeted Data Poisoning Attacks in Federated Learning 的方式

Thanks