NSS-SocialSec 2023

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

Authors: Zhuotao Lian, Chen Zhang, Kaixi Nan, Chunhua Su

School of Computer Science and Engineering,
The University of Aizu

Outline

01 Introduction

02 Background

03 System Design

04 Experiments

05 Conclusion

1 Introduction

Federated learning has been widely used in many fields 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 sybils and increase the damage of poisoning.

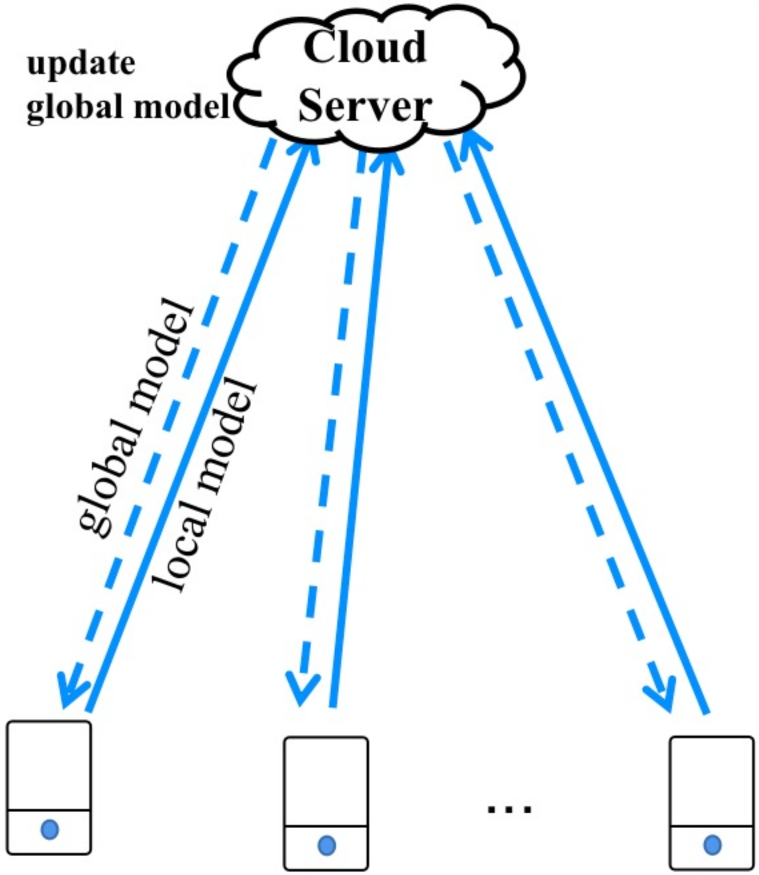


Figure 1. Federated Learning

1 Introduction

Motivation:

1. There are limited studies focusing on sybil-based untargeted poisoning attacks, which are more prevalent and practical in real-world scenarios compared to targeted attacks.
2. Through our inductive analysis, we have identified untargeted attacks as having higher prevalence and research value.

Therefore, this paper aims to propose a **s**ybil-based untargeted **poi**soning attack in federated **l**earning (SPoiL) that leverages virtual node forgery by malicious users and manipulates local training data to undermine the performance of the global model.

2 Background

1. Federated Learning

- Federated Learning (FL) is a distributed approach to machine learning that allows large-scale training on devices with decentralized data while preserving the privacy of sensitive data held by device owners.

Key Steps:

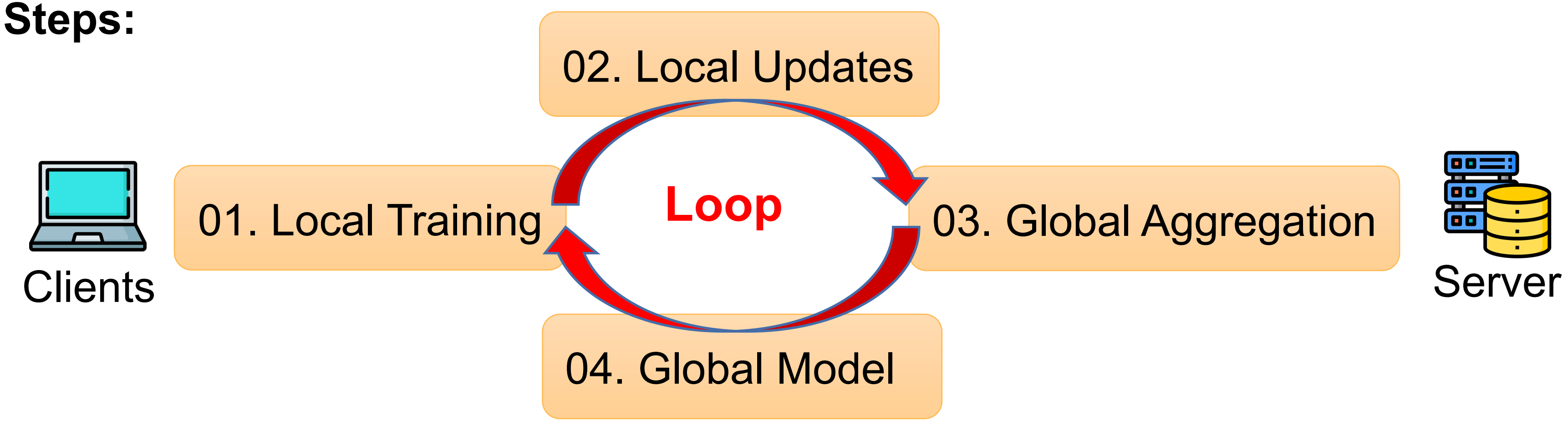


Figure 2. Key Steps of Federated Learning

2. Poisoning Attacks in Federated Learning

- Poisoning attacks are malicious attempts by participants to inject harmful information into the training data or uploaded model parameters, with the intention of disrupting the accuracy of federated learning.

Types of Poisoning Attacks:

1. Target-based Classification:
- Data Poisoning
- Model Poisoning

2. Goal-based Classification:
- Targeted Attacks
- Backdoor Attacks
- Untargeted Attacks

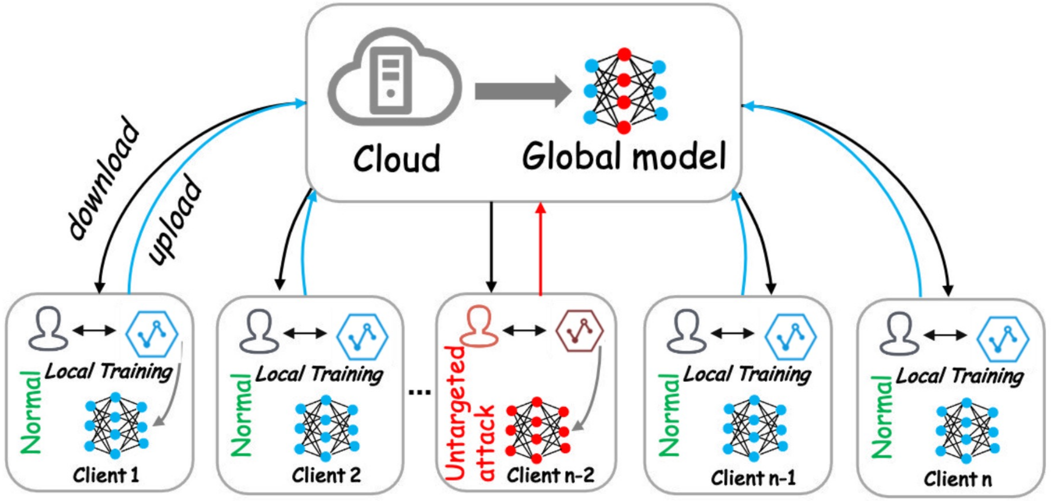


Figure 3. Untargeted Attacks

3. Sybil-based Attacks in Federated Learning

Definition of Sybil-based Attack: Sybil-based attack is a term used to describe the act of adversaries creating multiple virtual identities, accounts, or nodes in order to disrupt the balance of a global system.

Similarities and Differences with Distributed Poisoning Attacks:

Sybil-based attacks share similarities with distributed poisoning attacks, such as controlling multiple clients to interfere with the global model. However, sybil-based attacks are often carried out by a single adversary, while distributed poisoning attacks involve multiple adversaries.

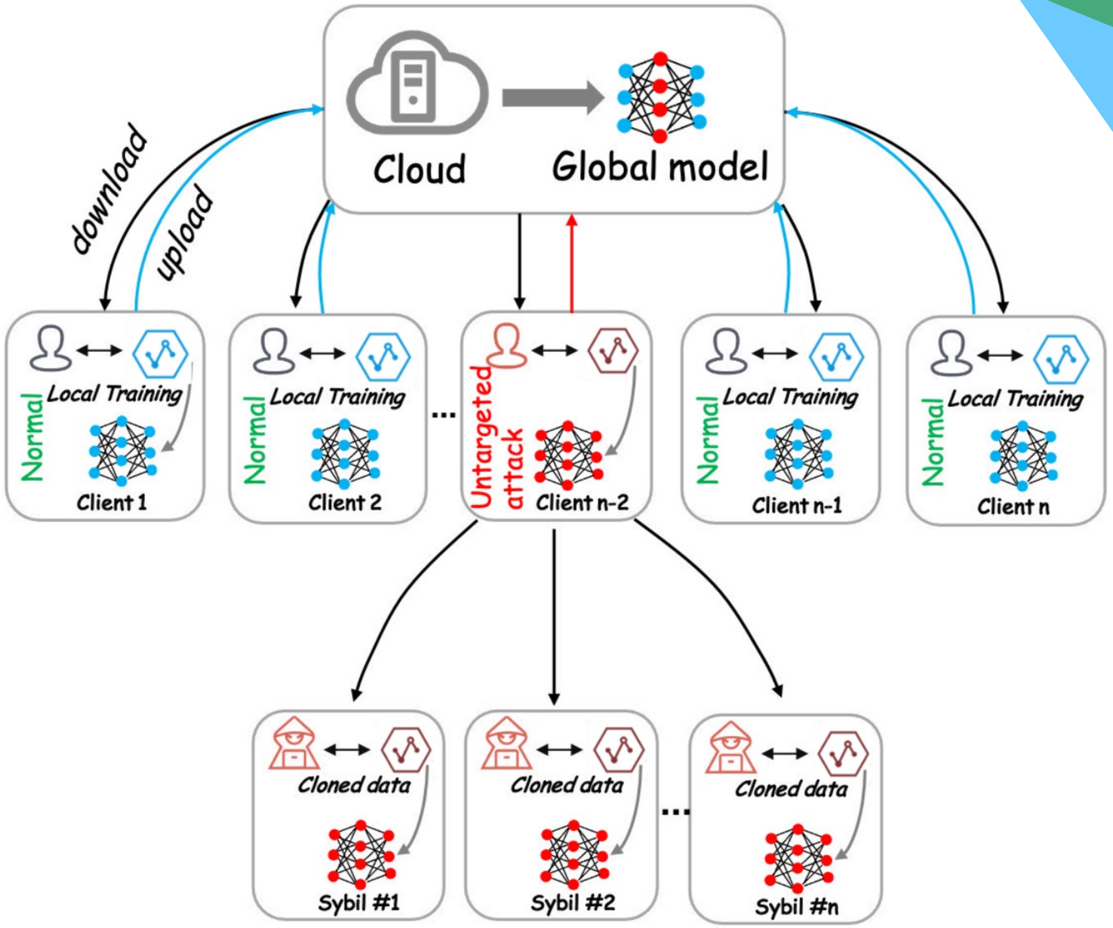


Figure 4. Sybil-based Untargeted Attacks

3 System Design

Threat Model:

- Adversary's Objective:

In SPoiL, the adversaries aim to cause misclassification of any test input without targeting a specific class. The objective of the attack is indiscriminate, and the specific incorrect class label assigned to the misclassified sample is not a concern for the attacker.

- Adversary's Knowledge

We assume that the adversary has complete access to the global model information as the global model is shared among all participants. Regarding data access, we assume that the adversary can only access the local initial data of compromised users.

- Adversary's Capabilities

In SPoiL, we consider a more generalized assumption where the adversary can only modify local data to perform data poisoning attacks. It is unrealistic to expect the adversary to directly manipulate the updated data uploaded by users by bypassing the security protocols. This assumption is not practical in large-scale federated learning systems.

3 System Design

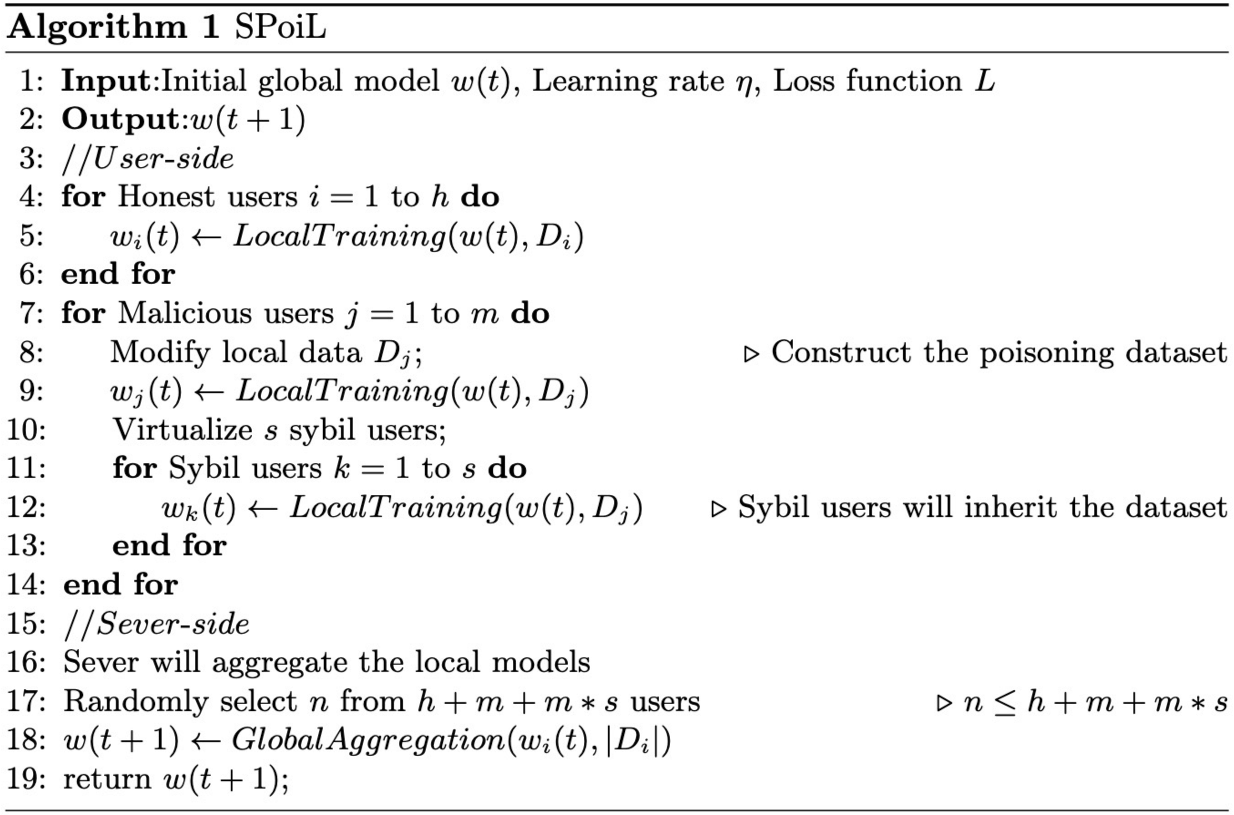
In this system, there are three types of users, including benign users, malicious users, and sybil users, collectively forming the participant pool of federated learning.

- Benign users employ gradient descent to train their local models on their respective local data and update them accordingly.

- Malicious users, on the other hand, start by modifying their local data. In this context, we consider the construction of a poisoned dataset using random label flipping.

- Subsequently, malicious users create multiple sybil users, as depicted in line 10 of the algorithm. The sybil users inherit the poisoned data from the original malicious users.

Algorithms:



Algorithm 1 SPoiL

```
1: Input:Initial global model $w(t)$, Learning rate $\eta$, Loss function $L$
2: Output:$w(t+1)$
3: //User-side
4: for Honest users $i = 1$ to $h$ do
5:     $w_i(t) \leftarrow LocalTraining(w(t), D_i)$
6: end for
7: for Malicious users $j = 1$ to $m$ do
8:     Modify local data $D_j$;                          ▷ Construct the poisoning dataset
9:     $w_j(t) \leftarrow LocalTraining(w(t), D_j)$
10:    Virtualize $s$ sybil users;
11:    for Sybil users $k = 1$ to $s$ do
12:        $w_k(t) \leftarrow LocalTraining(w(t), D_j)$  ▷ Sybil users will inherit the dataset
13:    end for
14: end for
15: //Sever-side
16: Sever will aggregate the local models
17: Randomly select $n$ from $h+m+m*s$ users              ▷ $n \leq h+m+m*s$
18: $w(t+1) \leftarrow GlobalAggregation(w_i(t), |D_i|)$
19: return $w(t+1)$;
```

3 System Design

It is important to note that data poisoning occurs during specific rounds, which is referred to as static poisoning.

If the poisoned data is not repaired afterward, allowing for long-term attacks, the attack becomes persistent. Additionally, if data poisoning is performed in multiple rounds with different techniques, it can be considered a dynamic poisoning attack.

In this paper, we primarily focus on a typical static poisoning approach.

4 Experiments

1. Settings

- We conducted our distributed virtual experiments on a single machine running Ubuntu 18.04. The machine was equipped with 32GB of RAM and an Nvidia GTX 3070 GPU, which provided the necessary computational resources for our experiments.

- In this study, we selected the Kaggle fake news dataset for our experiments. The dataset consists of five attributes: "ID," "Title," "Text," "Author," and "Label." For our experiments, we focused on two attributes: "Text" and "Label," which were used to train our model.

- Our neural network model consisted of three hidden layers, totaling 163,570 trainable parameters. The activation function used for the first three dense layers was Rectified Linear Unit (ReLU), while the activation function for the last layer was softmax.

4 Experiments

2. Results and Analysis

We conducted poison attacks initiated by the malicious users in the 9th epoch, where they randomly flipped the corresponding labels of the training data.

We tested different scenarios with varying proportions of malicious users: 0%, 5%, 10%, and 20%.

The presence of even a small proportion of malicious users can lead to a considerable decrease in classification accuracy.

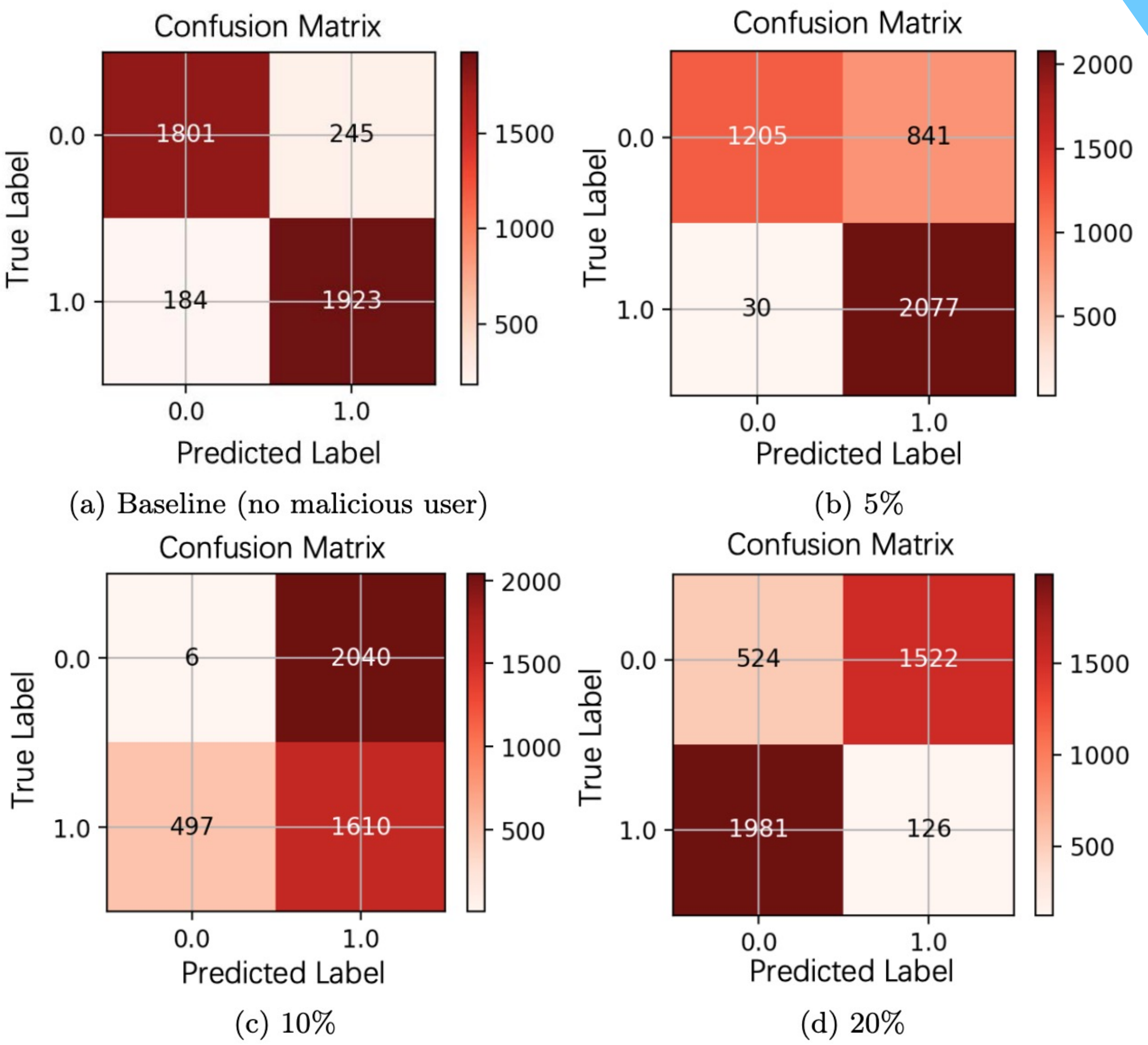


Figure 5. Confusion matrix with respect to the proportion of malicious users.

4 Experiments

Next, we investigated the influence of the number of sybils each malicious user can create on the model. In this scenario, we considered a general case where 10% of the users are malicious. We used s to represent the number of sybil users created by each malicious user.

A larger number of sybil users leads to a greater impact on the global model. Since the poisoned data is not rectified, the influence of malicious users on the model's training persists for a long time.

Even if the majority of users are benign, the presence of malicious users still hampers the performance of the model.

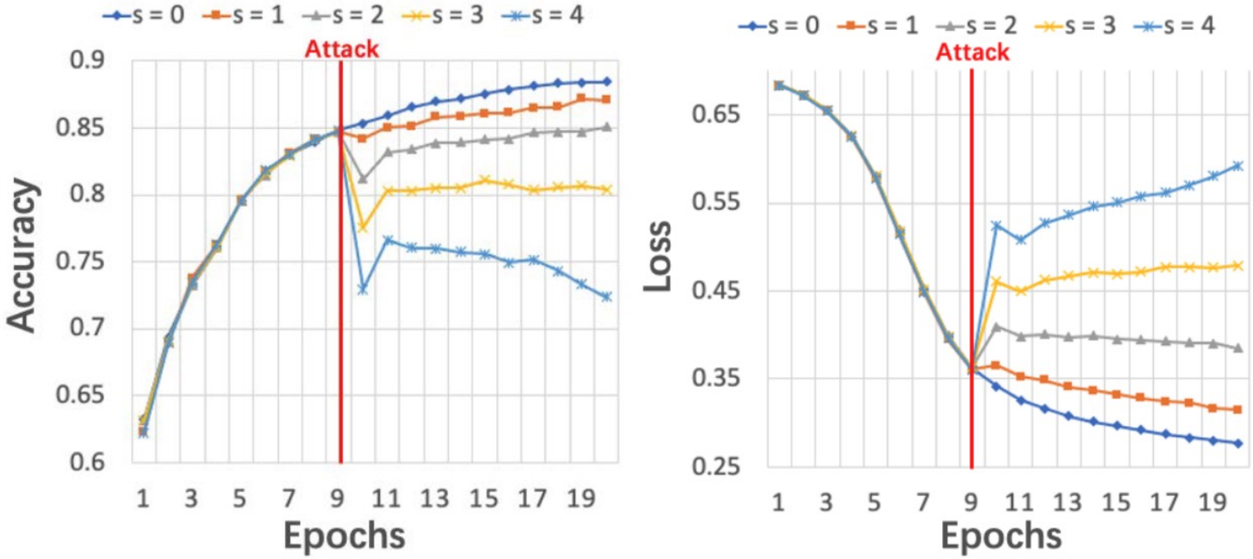


Figure 6. Accuracy and loss versus epochs.

4 Experiments

Finally, we investigated the combined impact of the number of Sybil users created, denoted as "s", and the proportion of poisoned data, denoted as "P_d", on the model's accuracy.

The deepest blue region in the plot corresponds to the lowest accuracy, indicating that higher values of P_d and a larger number of s result in a more pronounced negative impact on the model.

Furthermore, we observed that the decrease in accuracy is not linear but rather exhibits a gradual acceleration as P_d and s increases.

This observation provides inspiration and raises considerations for our future work, particularly in the design of defenses that can achieve better results while minimizing costs.

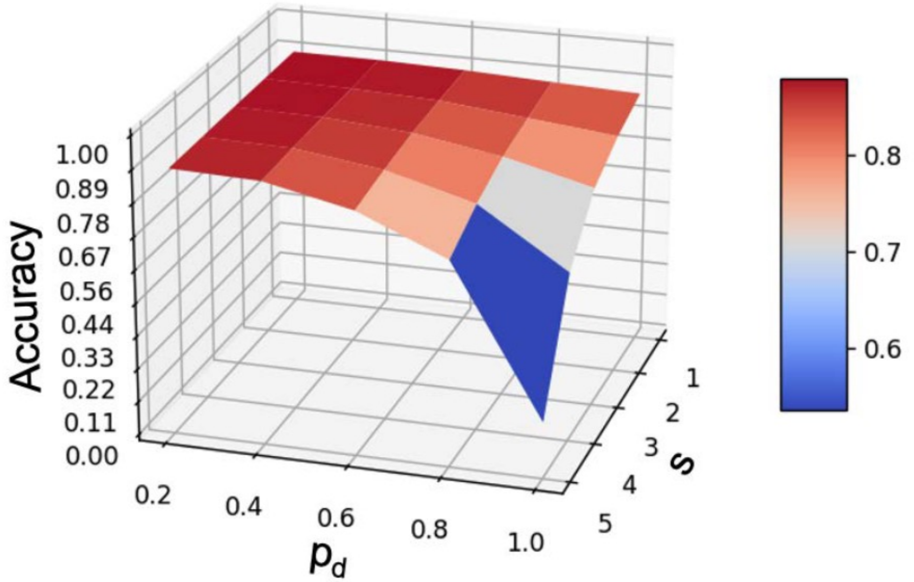


Figure 7. Test accuracy with different Pd and s. Pd refers to the proportion of poisoned data while s refers to the number of sybils created by each malicious user.

5 Conclusion

- In this paper, we present SPoiL, a sybil-based untargeted poisoning attack.
- Our approach involves manipulating local data by flipping labels and creating virtual sybil nodes to simulate participants during training and global model updates.
- We conduct experiments using a fake news detection dataset to evaluate the feasibility of our method. The preliminary results confirm the effectiveness of sybil-based untargeted attacks and investigate the influence of sybil node count and poisoned data proportion on attack performance.

Thanks for listening!

Q&A