Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier

Turning Federated Learning Systems into Covert Channels

GABRIELE COSTA[1] (Member, IEEE), FABIO PINELLI[1], SIMONE SODERI[1] (Senior Member, IEEE), and GABRIELE TOLOMEI [2]

[1]SySMA Unit, IMT School for Advanced Studies, Lucca, Italy (e-mail: name.surname@imtlucca.it)
[2]Department of Computer Science, Sapienza University, Rome, Italy (e-mail: tolomei@di.uniroma1.it)

Corresponding author: Gabriele Costa (e-mail: gabriele.costa@imtlucca.it).

This work was partially funded by EU H2020 project "SPARTA" under Grant 830892.

ABSTRACT Federated learning (FL) goes beyond traditional, centralized machine learning by distributing model training among a large collection of edge clients. These clients cooperatively train a global, e.g., cloud-hosted, model without disclosing their local, private training data. The global model is then shared among all the participants which use it for local predictions.

This paper proves that FL systems can be turned into covert channels to implement a stealth communication infrastructure. The main intuition is that, during federated training, a malicious sender can poison the global model by submitting purposely crafted examples. Although the effect of the model poisoning is negligible to other participants and does not alter the overall model performance, it can be observed by a malicious receiver and used to transmit a sequence of bits. We mounted our attack on an FL system to verify its feasibility. Experimental evidence shows that this covert channel is reliable, efficient, and extremely hard to counter. These results highlight that our new attacker model threatens FL infrastructures.

INDEX TERMS federated learning, adversarial attacks, machine learning security, covert channel.

I. INTRODUCTION

FEDERATED learning (FL) [50], [51] has emerged as the leading technology for implementing distributed, large scale and efficient machine learning (ML) infrastructures. The main idea is that multiple clients connect to the FL system, and collaboratively train a shared, global model. Frequently FL networks consist of a centralized, e.g., cloud-hosted, server and many edge clients that iteratively run FL rounds. Each round consists of the following steps.

1) The server sends the current, global model to the clients and appoints some of them for training.
2) Each selected client locally trains its copy of the global model with its own private data. Then it sends the resulting local model back to the server.
3) The server updates the global model by applying an *aggregation function* to the local models of the clients.

FL allows clients to concurrently train a shared global model, without disclosing private training data. Hence, FL provides great benefits in terms of both scalability and privacy. Since this process smoothly integrates with ubiquitous, distributed infrastructures, it has been applied to IoT [79], Fog computing [86], autonomous vehicles [63], smartphones [83], and wearable devices [16]. Thus, nowadays, billions of devices are connected to one or more FL systems.

The growing adoption of FL also raises security concerns, for instance, about the confidentiality, integrity, and availability of FL systems. As a consequence, several authors considered attack scenarios such as *data poisoning*, where an adversary pollutes the training set with maliciously crafted examples [33], and *model poisoning*, in which the attacker directly attempts to tamper with the global model parameters [6]. Also, a large body of work deals with privacy leakage that may expose the local data of some clients [52]. However, very little research has been done for studying the emerging exploitation opportunities, i.e., new attacks carried out *by means of* FL systems.

In this paper, we discuss a recent attack scenario that we originally reported in [17]. Briefly, it consists of an adversary implementing a *covert channel* [38] over an FL system. Covert channels allow an attacker to establish illicit communication between two agents (e.g., two devices) that should stay isolated. In theory, since no trust relationship exists among the clients, FL should not be intended to support the creation of covert channels. In practice, being shared among the participants, the global model can be turned into

a communication channel. More specifically, two FL clients, i.e., a sender and a receiver, can agree on an aimed poisoning strategy that allows them to transfer one bit. In this way, they exploit the global model updates as a physical communication medium.

We start by describing our attacker model and the covert channel implementation strategy. Our attacker only requires limited capabilities and, thus, it appears very realistic. As a matter of fact, communications are established by poisoning the training set of a single client, i.e., the sender. Poisoned examples are crafted by modifying benign input examples through simple, effective and efficient heuristics.

We show how to implement our attack on an FL system, where clients collaboratively train a global model to recognize handwritten digits of the popular MNIST dataset [41]. In our FL system, private training data is represented by a subset of the MNIST dataset that is randomly assigned to each client. Such an implementation is often given as a template of a generic FL system in official tutorials.[1] Also, we demonstrate the feasibility on another image recognition task, i.e., for the CIFAR-10 dataset [37].

Our experiments highlight that FL-based covert channels are an actual threat that, to the best of our knowledge, has been neglected so far. Moreover, we show that channel performance in terms of capacity and quality can support real communications. Since parallel covert channels can exist in a single FL system, the channel bandwidth can also scale up. Finally, experiments confirm that covert channels implemented in this way are hard to detect and counter.

The main contributions of this paper are listed below.

- A novel attacker model for FL-based covert channel.
- A general attack implementation strategy.
- A prototype applied to image classification tasks.
- Experiments on the performance of the channels.
- A discussion of possible mitigation mechanisms.

The rest of the paper is organized as follows. Section II describes the main background concepts used in this work. In Section III, we revise the literature about FL security and application-level covert channels. Section IV introduces our attacker model, and Section V details the implementation of the covert channel. In Section VI, we describe the properties of our covert channel. Section VII presents our experiments. In Section VIII, we discuss detectability, countermeasures, and exploitability. Finally, Section IX concludes the paper.

II. BACKGROUND

In this section, we provide the reader with the essential context needed to understand the subject of this work.

Machine & Federated Learning. We consider the supervised learning task as the reference example of a typical ML problem. The goal of supervised learning is to estimate a function that maps an input to an output, based on a sample of observed input-output pairs, called *examples*, which is usually referred to as *training set*.

More formally, let $D = \{(x_i, y_i)\}_{i=1}^{n}$ be a training set of n examples. Each $x_i \in X \subseteq \mathbb{R}^d$ is a d-dimensional vector of *features* representing the i-th input and $y_i \in Y$ is its corresponding output value. Here, we focus on *classification* problems where $Y = \{1, \ldots, `\}$ and each y_i is known as the *class label* (as opposed to regression problems where $Y \subseteq \mathbb{R}$).

Supervised learning assumes the existence of an unknown target function $g : X \mapsto Y$ that maps any feature vector to its corresponding output. The goal is therefore to estimate a function m, namely a parametric model, that best approximates g on D. More specifically, the optimal parametric model m is the one that minimizes the value of a loss function L, which measures the cost of replacing the true g with m on the training set. In other words, learning m reduces to the following optimization problem, also known as empirical risk minimization (ERM) [76].

$$m = \operatorname{argmin}_m L(m; D) \tag{1}$$

Depending on the supervised learning task, different loss functions can be adopted. For example, cross-entropy is commonly used for classification [54], whilst mean squared error is typically employed in regression settings [30].

The standard framework above assumes that the actual training procedure, i.e., the optimizer used to solve (1), runs on a centralized location where the whole dataset D is stored. In the case of FL, instead, the learning process is distributed among several clients that collaboratively train a shared, global model with a centralized server acting as an orchestrator. Thus, the FL framework consists of a centralized server S and a set of distributed, federated clients C, such that $|C| = n_c$. Each client $c \in C$ has access to its own private training set D_c, namely the set of its local labeled examples.

The generic t-th round of FL runs the following steps.

1) S sends the current, global model $m^{(t)}$ to every client and selects a subset $C^{(t)} \subseteq C$, such that $1 \leq |C^{(t)}| \leq n_c$.
2) Each selected client $c \in C^{(t)}$ trains its local model $m_c^{(t)}$ by optimizing the same objective of (1) on its own private data D_c, starting from $m^{(t)}$; the resulting $m_c^{(t)}$ is sent to S.
3) S computes $m^{(t+1)} = \oplus(\{m_c^{(t)} \mid c \in C^{(t)}\})$ as the updated global model, where $\oplus$ is an *aggregation function* (e.g., FedAvg [50] or one of its variants [45]).

In the beginning, $m^{(0)}$ may be randomly initialized. Then, FL rounds as the one described above are iteratively executed until convergence of the global model, i.e., until $t = T$ such that $m^{(T)} \approx m$. In practice, though, many FL models are continuously trained due to the highly dynamic nature of the infrastructure (e.g., new clients joining or leaving the system and fresh local data generated over time).

To simplify the notation, in the following we refer to m as the global model and to m_c as the local model of client c.

[1] For example, see https://www.tensorflow.org/federated/tutorials/federated_learning_for_image_classification.

Binary memoryless covert channel.

Furthermore, we call $m(x) = \hat{y} \in Y$ (resp., $m_c(x) = \hat{y}_c \in Y$) the global (resp., local) model prediction on input x.

Channels and communication quality. In this paper, we are interested in *binary memoryless channels* (BMC) [49]. Brie y, an BMC has discrete input b and output b' such that $b, b' \in \{0, 1\}$. We describe the relationship between channel's input and output through the conditional probability $P(b'|b)$, where $b = \{0, 1\}$. Figure 1 depicts a generic BMC in which $P(1|0) = p_1$ and $P(0|1) = p_2$ are the probabilities for input/output bit inversion errors. Thus, their complements give the probability of receiving the correct bit, e.g., $P(0|0) = 1 - p_1$. For a BMC, the *channel capacity* C is the maximum communication rate that the sender and receiver can reach over the channel. Following [49], C is computed as

$$C = \max_{P} I(x; y) = \sum_{\substack{x \in b \\ y \in b}} P(x)P(y|x) \log \left(\frac{P(y|x)}{P(y)} \right); \quad (2)$$

where $I(x; y)$ is the mutual information between the input x and output y. From (2), we obtain the channel capacity as

$$C = 1 + \sum_{i=1;2} \frac{1}{2}(p_i + (1 - p_i) \log_2(1 - p_i)) =$$

$$= 1 - \sum_{i=1;2} \frac{H(p_i)}{2} = 1 - \frac{H(p_1)}{2} - \frac{H(p_2)}{2}; \quad (3)$$

where H is the binary Shannon entropy function.

Communication quality is typically measured in terms of *bit error rate* (BER) and *signal-to-noise ratio* (SNR). In general, BER is obtained as the number of bit inversions, e.g., due to channel noise, over the total number of transmitted bits. Instead, SNR is de ned as the ratio of signal power to the noise power, i.e., s^2/n^2, where s is the channel signal and n is the noise. In general, when signal and noise are modeled by means of two random variables, called S and N respectively, we have that $s^2 = E[S^2]$ and $n^2 = E[N^2]$, where $E[\,]$ denotes the expected value. Moreover, when N has zero mean, i.e., $E[N] = 0$, n^2 reduces to $^2_n = Var(N) = E[N^2] - E[N]^2$, that is $E[N^2]$ is equal to the variance of N.

The attacker model presented in this paper was originally put forward in our previous work [17]. Subsequently and independently, the very same attacker model was also reported in [31]. There the authors implement a spread spectrum overt channel via model poisoning. Their work con rms the relevance of our attacker model and outlines that there exist multiple exploitation techniques that real attackers can even combine.

To the best of our knowledge, our attacker model is not listed among the security challenges of FL, e.g., see [36]. In the following, we revise some related work about *adversarial attacks to ML*, and *application level covert channels*, which are closer to our proposal.

A. ADVERSARIAL ATTACKS TO ML

There exists a large body of work investigating the security of both traditional, i.e., centralized, and federated ML against so-called adversarial attacks [9], [18], [26], [32], [34], [46], [55], [62], [74]. Adversarial attacks can have different targets. For instance, model inversion attacks [22], [23], membership inference attacks [52], [64], [68], and property inference attacks [2], [25] aim to violate the con dentiality of user's private training/test data. Also, an attacker can compromise the con dentiality/intellectual property of a model provider by stealing its model parameters and hyperparameters [43], [75], [77].

From this perspective, our attacker model belongs to adversarial attacks that tamper with the integrity and the performance of a predictive model [4]. These attacks are classi ed according to the stage(s) of the ML pipeline in which they occur. In particular, attacks can happen at training time only, both at training and at test time, or at test time only. Those are called *poisoning*, *backdoor*, and *evasion* attacks, respectively.

From another viewpoint, depending on the attacker's goal, adversarial attacks can be further classi ed as untargeted (or random) [8], [33], [42], [65], [80], [82] or targeted [1], [29], [56], [67]. The former aims to reduce the overall accuracy of the learned model at inference time, regardless of what speci c testing examples get incorrectly classi ed. The latter forces the learned model to output attacker-desired labels for certain testing examples, e.g., predicting spam messages as non-spam, while not altering the output for other examples. Since targeted attacks have to do with a speci c goal, they usually require the attacker to have rather strong capabilities. In the following, we provide a detailed overview of the most prominent adversarial attacks.

Poisoning Attacks. To compromise the performance of a predictive model, poisoning attacks can target two components of the training stage, i.e., the training dataset and the learning process. The former are known as data poisoning attacks [8], [21], [32], [33], [56]. The latter are referred to as model poisoning attacks [6], [20].

Data Poisoning. These attacks pollute the training dataset by injecting it with new malicious examples or by corrupting existing ones. There are two main types of data poisoning attacks, called (i) clean-label [67] and (ii) dirty-label [28], respectively. In clean-label poisoning attacks, the adversary has no control over the labeling process. The attacker simply

injects a small number of slightly perturbed examples (whose labels remain correct) into the training set of the victim. These attacks have been proven effective only when the attacker has complete knowledge of the victim's model, i.e., underwhite-boxassumption [62]. Such knowledge is needed to craft the malicious examples [71]. More recently, clean-label poisoning attacks for unknown, i.e.,black-box[60], deep image classi ers have been explored [87].

In dirty-label poisoning, the adversary can introduce a number of training instances to be misclassi ed with a speci c label in a targeted way. An example of dirty-label poisoning is the label- ipping attack [8], [24]. Here, the labels of honest training examples of one class are ipped to another class, while the features of the data are kept unchanged. For instance, a malicious agent can poison the training set of a handwritten digit recognition system by ipping all 1s to 7s.

In the FL setting, it is common to assume that the attacker controls some clients and their training sets. Thus, dirty-label attacks have been more often considered. In [74], the authors' study targeted label- ipping attacks on FL. They nd that poisons injected late in the training process are signi cantly more effective than those injected early. Other proposals adopt a bi-level optimization approach for poisoning multi-task FL [72] and GAN-generated poisons [85].

Model Poisoning Different from data poisoning attacks, these attacks threaten the learning process directly, e.g., by changing some model parameters. Model poisoning is generally perceived as dif cult to implement in centralized ML systems as it requires the adversary to access the target model, i.e., assuming either grey-box or white-box knowledge. On the other hand, model poisoning becomes rather feasible in the case of FL, where a malicious client has direct in uence over the jointly-trained global model via its local parameters updates [6], [20], [47]. As with any poisoning attack, the adversary's goal is to cause wrong predictions of the FL model. However, she/he aims to force classi cation errors at inference time and without modifying test examples. In this respect, it is opposed to backdoor and evasion attacks (which are discussed below). In model poisoning, the misclassi cation results from the adversarial corruption of the training process, which can be achieved either by gradient or learning rule manipulation. In gradient manipulation, the adversary poisons local model gradients (or model updates), which are then sent to the central server for aggregation, thereby jeopardizing the global model performance [10]. In learning rule manipulation, instead, the attacker corrupts the actual training logic. In some cases, these model poisoning attacks have proven more effective than data poisoning, and an attacker can compromise the global model even when controlling a single client. For instance, in [6] the authors successfully achieve a stealthy targeted model poisoning attack by adding a penalty term to the objective function to minimize the distance between malicious and benign weight update distributions.

Backdoor Attacks. Backdoor attacks – also known as trojan attacks – exploit the adversary's capability of having (limited) access to input examples also at test time [26]. Hence, backdoor attacks exceed poisoning attacks, since the adversary can manipulate both training and test inputs. For this reason they are often considered more disruptive toward the victim model.

Like for standard poisoning attacks, we can distinguish between backdoor attacks affecting the data or the model. The former are referred to as backdoor data poisoning and consist of adding attacker-chosen examples to the training set. The attacker examples contain a particular trigger [15], [44], i.e., a distinguished feature that activates the backdoor. The model learned on such poisoned training set will embed a backdoor, which the attacker exploits at test time by submitting examples that contain the same trigger. Instead, backdoor model poisoning requires a stronger threat model, where the attacker can get direct access to the learning system and change the model's internals (i.e., parameters and architecture) to embed a backdoor [19], [35].

Interestingly enough, backdoor attacks have been proven ineffective under FL settings [3]. The main obstacle is that aggregation involves many clients and, assuming that the attacker only controls a minority, the effect of the adversarial updates is weakened. To overcome this limitation, the authors of [3] consider a model replacement approach, where the attacker scales up a malicious model update to increase its effect on the aggregation function.

In [81], the authors propose distributed backdoor attacks, which better exploit the decentralized nature of FL. Specifically, they decompose the backdoor pattern for the global model into multiple distributed small patterns, and inject them into training sets, used by up to 40% adversarial participants, at each round. Although more effective than global backdoor trigger injection, this approach comes at the price of controlling a signi cant subset of the FL clients.

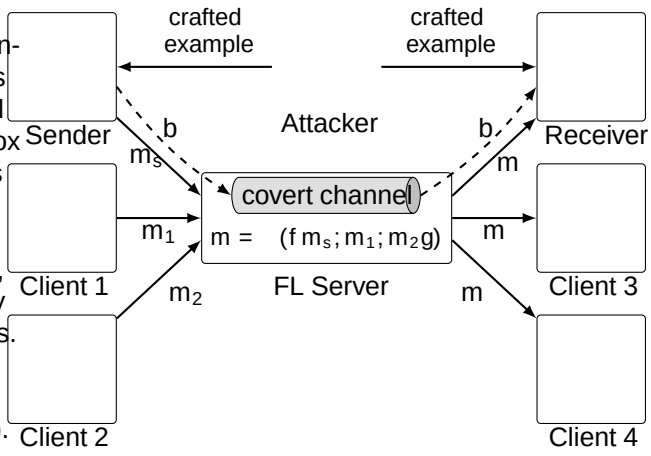
Evasion Attacks. Adversarial attacks that occur only at test time are called evasion attacks. Here, the goal of the adversary is still to fool an ML model yet without tampering it at training time. In fact, evasion attacks use so-called adversarial examples [4], [73], i.e., crafted (minimal) perturbations of test instances that cause prediction errors (either targeted or untargeted) when input to a legitimately trained model.

Evasion attacks may look similar to backdoor poisoning attacks [7]. However, the key difference between the two is that evasion attacks exploit the decision boundaries learned by an uncorrupted model to construct adversarial examples that are misclassi ed by the model. In contrast, backdoor attacks intentionally shift these decision boundaries as a result of a jeopardized training process, so that certain examples get eventually misclassi ed [26].

Several works have explored evasion attacks in the context of computer vision [12], where adversarial examples are obtained by adding random noise to test images. Even though such images look legitimate to a human, they are wrongly classi ed by the image recognition system. Also, more recent works investigate the applicability of evasion attacks to malware classi cation [70].

In the FL setting, the global model maintained by the server suffers from the same evasion attacks as in the conventional ML setting when the target model is deployed as a service. Moreover, at each FL training round the global model sent to the federated clients is exposed as a white-box to any malicious participant. Thus, FL requires extra efforts to defend against white-box evasion attacks [46].

In this work, we consider an attack scenario partially related to the one presented in [3]. As a matter of fact, our covert channel is implemented by means of maliciously crafted examples that carry a trigger as in backdoor attacks. In particular, our adversary (i) crafts malicious examples carrying a trigger (see Section V-A), and (ii) poisons the federated model to transmit one bit per trigger (see Section V-B).



Overview of the attacker model.

B. APPLICATION LEVEL COVERT CHANNELS

In a general sense, a covert channel is any communication channel that is not intended for information transfer [38]. Although we are not aware of FL-based covert channels, some authors already investigated the implementation of covert channels at the application level. Most authors considered encapsulation of hidden communications in application-layer network protocols. Being the main application-level protocol, HTTP is the primary target for covert channel implementations, e.g., see [5]. Nevertheless, the entire TCP/IP ecosystem can be at risk, and we refer the interest reader to [53] for a survey. More recently, also web applications were proposed for the implementation of covert channels. For instance, in [66] the author considers social networks such as Facebook and Twitter. However, since these kinds of covert channels rely on already existing communications between devices, in practice, they usually do not break any sandbox policy.[2] Also, as the authors of [84] point out, most of these covert channels can be detected and some effective countermeasures exist, e.g., packet inspection can be used to detect illegal traf c. Possibly for these reasons, application-level covert channels are rare in the literature.

Loosely speaking, also our proposal relies on a sort of encapsulation mechanism. However, here we do not wrap information inside protocol messages. Rather, we embed information inside FL models, which prevents standard detection techniques based on traf c inspection.

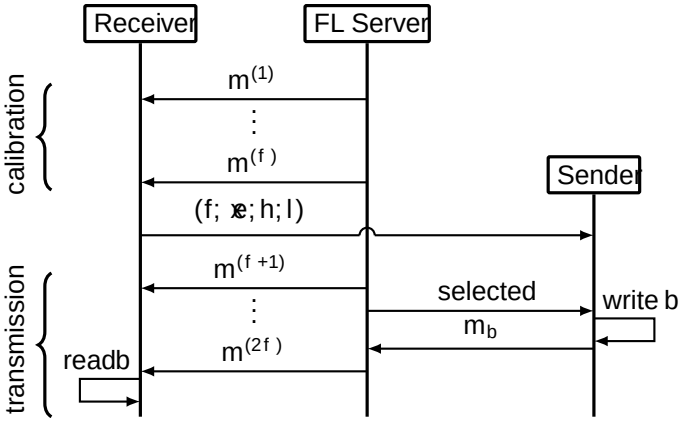
In this section, we present our attacker model. The attacker's goal is to establish a covert channel between two clients, namely *Sender* and *Receiver*, of an FL infrastructure. In terms of capabilities, our adversary resembles that of [3]. Here, we assume that both Sender and Receiver are controlled by the attacker. For instance, think of Sender as a malware-compromised device and the Receiver as the malware command and control server. Although they are compromised, we assume the attacker does not tamper with the standard FL infrastructure behavior, i.e., Sender and Receiver follow the FL client protocol. Furthermore, Sender and Receiver do not need to inspect nor jeopardize their local models to set up a covert channel. More precisely, Sender is only allowed to poison its local dataset and Receiver can only classify examples using its own local model. Intuitively, these assumptions hold for most FL systems.

The overall attacker model is schematically depicted in Figure 2. The FL server randomly selects a subset of clients at each federated round. Selected clients work as expected, i.e., they use their own private datasets to train their local models starting from the last global model received by the server (see Client 1 and Client 2 in Figure 2). Then, FL clients upload their newly trained local models to the server, aggregating them into an updated global model through the aggregation function. At the end of each federated round, all the clients receive a copy of . When selected, Sender (top left of Figure 2) poisons its local model by training it with some malicious, attacker-provided examples. Its goal is to transmit a bit by inducing a perturbation of the global model that the Receiver can test. On the other hand, Receiver (top right of Figure 2) uses to classify test examples, e.g., the same malicious examples used by Sender. According to the classi cation outcome, the Receiver deduces whether 0 or 1 was sent. Implementing such a covert channel is non-trivial and depends on the underlying FL system. We discuss the implementation details in the next section.

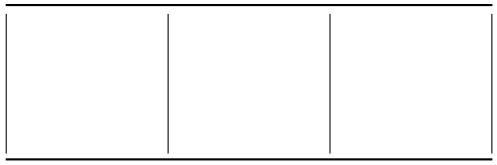
In this section, we detail the implementation strategy for creating the previously described covert channel. Without loss of generality, every implementation is based on the abstract protocol schematically depicted in Figure 3.

A transmission starts with a calibration phase during which Receiver observes the global model updates $m^{(1)}; : : : ; m^{(f)}$ (for f FL rounds). Eventually, Receiver computes channel parameters (see Section V-A) to be shared

[2] On the contrary, they are very relevant, for instance, when considering inter-process channels as in [14].



Communication protocol phases.



Linear transformation $_e$ of an example with $= 0:3$ (middle) and $= 0:5$ (rightmost).

Algorithm 1: Edge example binary search algorithm.

Input: $x_1; \ldots; x_k, " > 0$
$_{high} := H$
$_{low} := L$
repeat
  $x_{high} := (x_1; \ldots; x_k; _{high})$
  $\hat{y}_{high} := m_r(x_{high})$
  $x_{low} := (x_1; \ldots; x_k; _{low})$
  $\hat{y}_{low} := m_r(x_{low})$
  if $\hat{y}_{high} = \hat{y}_{low}$ then Failure: cannot be equal
  $_{mid} := (_{low} + _{high})=2$
  $x_{mid} := (x_1; \ldots; x_k; _{mid})$
  $\hat{y}_{mid} := m_r(x_{mid})$
  if $\hat{y}_{high} \neq \hat{y}_{mid}$ then $_{low} := _{mid}$ else $_{high} := _{mid}$
until $_{low} - _{high} < "$
Output: $x_e := x_{high}$, labels $h := \hat{y}_{high}$ and $l := \hat{y}_{low}$

with Sender. For instance, these parameters can be provided through a secondary channel or hard-coded in Sender before its deployment. Then, Sender and Receiver synchronize on transmission frames of size f to send a bit b. During each frame, if selected, Sender trains its local model according to the channel parameters and the bit to be sent. In the meanwhile, Receiver monitors the global model updates and, at the end of the frame, it tests the received bit.

Below, we discuss the implementation of both the calibration and transmission phases.

A. CALIBRATION OF CHANNEL PARAMETERS

The first parameter to be determined is the size of the transmission frame f, i.e., the number of FL rounds used to transmit a single bit. Intuitively, too small values for f would increase transmission errors (e.g., Sender being never selected within a frame may result in a bit transmission error). On the opposite, if f is too large, channel throughput will be reduced. Finding optimal values for f is non-trivial as discussed in Section VII. Clearly, when the attacker knows the details of the target FL system, e.g., client selection probability (p_c), the desired value of f can be obtained analytically. For instance, when $p_c = 0:1$ and attacker wants Sender to be selected at least once with probability grater than $0:9$, f is computed so that $1 - (1 - p_c)^f > 0:9$, i.e., $f = 22$. Otherwise, Receiver must estimate f, e.g., by taking advantage of its selection notifications. For instance, it can set f as the number of rounds that it takes to be selected T times (for a constant T).

Another channel parameter to be determined is the crafted example x_e carrying the trigger that Sender will use to poison its local training set. To generate x_e, Receiver applies a linear transformation function $_e$[3] to a subset of k randomly selected examples from its local training set. Briefly, $x_e = (x_1; \ldots; x_k;)$ means that, starting from $k > 0$ examples $x_1; \ldots; x_k$, $_e$ returns a new example x_e, where $2 [0; 1]$ is a parameter controlling the transformation.

To provide an intuition of this process, we put forward an example taken from the MNIST dataset. Consider x to be the representation of a generic MNIST input image, i.e., a 28×28 matrix of pixels flatten into a 784-dimensional vector. We define $_e(x;)$ as the function erasing, from left to right, a fraction $$ of a single example image x (i.e., $k = 1$). For instance, when $= 0:3$, $_e$ sets to 0 the 235 values associated with the leftmost pixels of the target image vector. More formally, $_e(x;) = x \quad x^{( )}$ where $x^{( )} = [x^1; \ldots; x^{b \ 784c}; 0; \ldots; 0]$ and x^i is the i-th element of x. Figure 4 shows the behavior of $_e$, where we highlighted the erased part of the example.

Intuitively, the attacker can select a function $$ based on some semantic property of the classification domain. For instance, in the previous example, $_e$ encodes the simple fact that the right half of a handwritten 8 looks like a 3. In the experiments of Section VII we consider two slightly different functions. Others can be found, e.g., in [13].

Receiver repeatedly applies $$ to the selected examples until it identifies an edge example x_e. Formally, given Receiver's local model m_r and an arbitrarily small $" > 0$, we look for $x_e = (x_1; \ldots; x_k;)$ such that $m_r(x_1; \ldots; x_k;)) = h$ and $m_r(x_1; \ldots; x_k; + ")) = l$, where $h; l 2 Y$ and $l \neq h$. Interestingly, given $x_1; \ldots; x_k$, Receiver can efficiently compute via binary search, as sketched in Algorithm 1.

The search procedure starts from the predefined interval

[3] Although we do not explicitly prove it, the reader can easily check that all the transformation functions presented in the following are linear since they reduce to finite sums of matrices.

Algorithm 2: Sender transmission algorithm.

```
Input: f, x_e, h, l
repeat
    if r = 1 then
        b := nextBit()
        v := m(x_e)
        v_: := if v = h then l else h
    if selected by server for training then
        v_r := m(x_e)
        switch b do
            case 0 do if v_r ≠ v then train(m_s, x_e, v)
            case 1 do if v_r ≠ v_: then train(m_s, x_e, v_:)
        upload(m_s)
    r := (r % f) + 1
until transmission completed
```

Sender's model poisoning cases.

b = 0		b = 1	
$v_r = v$	$v_r = v_:$	$v_r = v$	$v_r = v_:$
do nothing	train(m_s, x_e, v)	train$(m_s, x_e, v_:)$	do nothing

Algorithm 3: Receiver bit test algorithm.

```
Input: f, x_e, h, l
repeat
    if r = 1 then
        v_1 := m(x_e)
    else if r = f then
        v_f := m(x_e)
        b := if v_1 = v_f then 0 else 1
        received(b)
    r := (r % f) + 1
until transmission completed
```

[H, L].[4] At each iteration, two examples, i.e., x_{high} and x_{low}, are generated and classified with Receiver's model m_f, so obtaining $\hat{y}_{high}$ and $\hat{y}_{low}$, respectively. If $\hat{y}_{high} = \hat{y}_{low}$, the algorithm terminates with a failure and no edge example is returned. Otherwise, the current search interval is split in half by computing $_{mid}$, x_{mid}, and $\hat{y}_{mid}$. Then, if $\hat{y}_{high} \neq \hat{y}_{mid}$, the algorithm iterates on the first half of the current interval. Otherwise, if $\hat{y}_{low} \neq \hat{y}_{mid}$, the search procedure continues on the second half. The loop terminates when the interval width goes under a threshold δ, which represents the granularity of a single position in the example feature vector, e.g., a single pixel in the case of Figure 4. Eventually, the algorithm returns the edge example x_{high}, as well as the two class labels h and l associated with the last, smallest interval.

It is worth noticing that, since it does not explore the entire feature space, Algorithm 1 might fail to generate edge examples for some inputs. In general, the effectiveness of this heuristic method depends on the choice of δ. We empirically show that Algorithm 1 can generate millions of edge examples in our experimental settings with only two linear transformation functions (see Section VII-A).

At the end of the calibration phase, channel parameters generated by Receiver amount to the tuple (f, x_e, h, l).

B. BIT TRANSMISSION

The transmission of one bit is based on the variations, during f FL rounds, of $m(x_e)$ between h and l. In particular, by poisoning its local model, Sender drives $m(x_e)$ to assume the desired value, while Receiver monitors it to read the transmitted bit. Transmissions are organized in consecutive frames of size f. Sender and Receiver are synchronized through an FL round counter $r \in \{1, \ldots, f\}$. Sender follows the procedure given in Algorithm 2.

At the beginning of each transmission frame, i.e., when $r = 1$, Sender sets the next bit b and uses the last received global model $m_r = m$ to classify x_e, thus obtaining $v \in \{h, l\}$. Also, Sender sets $v_: \in \{h, l\}$ so that $v \neq v_:$. At each round r, if selected by the FL server, Sender classifies x_e with the global model m, so obtaining v_r. Then, depending on b, Sender trains its local model m_s according to Table 1.

Intuitively, the purpose of the operation above is to keep the label assigned to x_e by the global model when sending 0 and to flip it when sending 1. This channel implementation amounts to a Differential Manchester encoding [49]. Eventually, Sender uploads its local model to the FL server.

Concurrently, Receiver executes Algorithm 3. Receiver uses the global model m to classify x_e both when $r = 1$ and $r = f$, so obtaining v_1 and v_f, respectively. Finally, Receiver reads 0 if $v_1 = v_f$ and 1 otherwise.

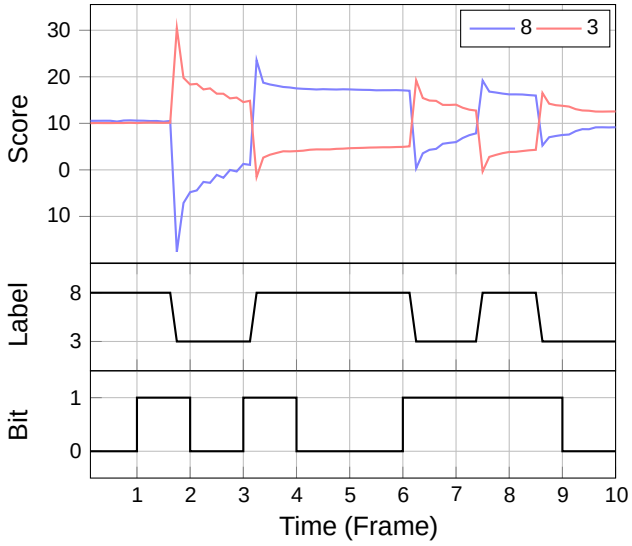
In Figure 5, we show as an example the transmission of 10 bits for the MNIST scenario discussed above, using f of Figure 4, $h = 8$, and $l = 3$. The top diagram shows the internal scores assigned by the global model to the edge example x_e during the transmission. By classifying x_e, Receiver observes the alternation of labels 8 and 3 (center). Then, Receiver interprets it as the sequence of bits 0101001110.

For each edge example, our implementation provides a digital, broadcast channel supporting half duplex[5] transmissions. In Section VII we will show that an attacker can even create several channels of this type to enlarge the communication bandwidth. Below, we detail the features of a single channel.

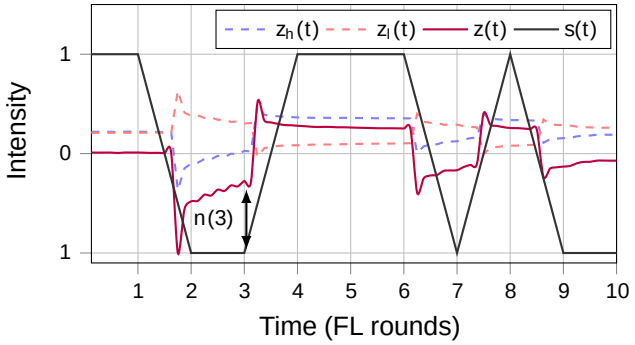
In terms of capacity, we treat the covert channel as a standard BMC (see, Section II). We just notice that the binary input of the channel is m_b, with $b \in \{0, 1\}$. Intuitively, inputs m_b represent Sender's local models m_s uploaded to the FL server. More precisely, we use m_b to distinguish between the two model poisoning cases used to transmit b (see Table 1).

[4] In our experiments we use $H = 0.5$ and $L = 1/2$.

[5] Intuitively, Sender and Receiver cannot transmit at the same time, but they can alternate their roles.



Transmission of bits 0101001110 over a channel.



Signals and noise for the channel of Figure 5.

In terms of communication quality, we consider BER and SNR, as discussed in Section II. While BER is straightforward, de ning SNR in our context requires more attention. In general, SNR is computed by periodically sampling the channel signal s and noise n at the end of each transmission frame t. Here, de ning s and n is non-trivial since our channel does not rely on a physical medium. Indeed, a covert channel consists of two prediction labels, namely l_h and l_l, and a perturbed example x. For each label $l \in \{l_h, l_l\}$, at time t we can measure the prediction score assigned by the global model to l when classifying x at each frame's end. Also, assuming that prediction scores range within the interval $[-Z, Z]$,[6] we normalize the prediction scores by linearly scaling them in $[-1, 1]$. Thus, for each label l, the label signal $z_l(t)$ amounts to the normalized score described above. We de ne the overall received signal as $z(t) = z_h(t) - z_l(t)$, i.e., as the differential, normalized signal.

Since a direct measure of the transmitted signal $s(t)$ cannot be computed, we approximate it to the differential signal that switches its intensity between 1 and -1. Intuitively, this is equivalent to stating that Sender attempts to transmit 0 by setting $z(t) - z(t+1) = 1$ and 1 by setting $z(t) - z(t+1) = -1$.[7] Thus, for each frame t, we set

$$s(t) = \frac{1 - 2b}{s(t-1)}; \tag{4}$$

where b is the bit transmitted during frame t, and, by construction, $s(0) = 1$. Then, we de ne the noise at time t as $n(t) = z(t) - s(t)$.

The overall intuition behind our de nitions of $z(t)$, $s(t)$ and $n(t)$ is given in Figure 6. There, red and blue dashed lines denote z_h and z_l signals, i.e., the normalized version of the scores of Figure 5. Instead, the purple line denotes $z(t)$ and the gray line denotes $s(t)$. Also, the vertical arrow shows the value of $n(t)$ at $t = 3$. From this and (4), we de ne the SNR of the covert channel as

$$\text{SNR} = \frac{z(t)^2}{\sigma_n^2}; \tag{5}$$

where σ_n is the standard deviation of the normalized noise $n(t)$ de ned as $n(t) = \frac{n(t)}{\max_{t' \le t} |n(t')|}$.

In this section, we present our experimental results. The implementation of our covert channel, called FedExp, is available at https://github.com/fpinell/sec_federated_learning.

A. EXPERIMENTAL SETTING

All the tests described in this section were executed on Docker containers running on a dedicated Intel® Xeon® Gold 5218 2.30GHz CPU with 64 GB of memory. The implementation is based on the popular ML framework PyTorch [61]. We implement an FL system for handwritten digit classi cation in our testing environment.

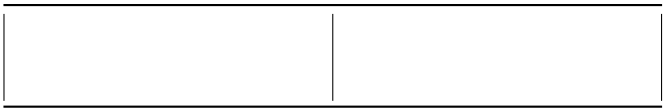
Dataset description. The MNIST dataset [41] is one of the most popular datasets used as a benchmark for training and testing image classi ers. It contains a total of 70,000 greyscale images of handwritten, single digits. Digit images are taken from American Census Bureau employees and American high school students. Usually, 60,000 images are used for training and the remaining 10,000 for testing. Images are represented by 28×28 matrices of bytes, each byte representing a single pixel (where 0 is for background color, i.e., white, and 255 for the foreground color, i.e., black).

FL system parameters. Our FL system is con gured according to three parameters, i.e., the number of honest FL clients (n_c), the client selection probability $p_c()$, and the neural network architecture $\mathcal{N}()$.

At startup, both the training and the test portions of the full MNIST dataset are uniformly distributed randomly among the n_c honest clients. At each round, the server sends the current global model to all the clients; then, it randomly selects

[6] Z can be estimated as the maximum score observed during a transmission.

[7] Notice that this assumption is more restrictive w.r.t. the actual implementation, since Receiver only requires $z(t) - z(t+1) < 0$ and $z(t) - z(t+1) > 0$ to read 0 and 1, respectively.



Applications of $_v$ to examples #22242 and #7596 (left) and $_h$ to examples #32481 and #18198 (right).

p_c n_c clients. Thus, selected clients use their portion of the MNIST dataset to train their own local models. Local models are then sent back to the server, which aggregates them and updates the global model using the standard federated averaging function FedAvg [50].

The FL system we implemented supports two different types of neural network architectures. The first one is a fully connected Neural Network (NN) [39], composed of three layers with 200 neurons each. The second one is a Convolutional Neural Network (CNN) [40], [57] with two convolutional layers, each one of size 3x3, and one fully connected output layer. After each convolutional layer, we use a Max Pool 2D layer with kernel size 2x2. Both networks use a ReLU activation function and are trained by minimizing cross entropy loss function via stochastic gradient descent [27].

Attacker parameters. Receiver is added to the FL system after 200 training rounds. The frame size can be manually configured. Otherwise, during the calibration phase, Receiver estimates as the number of rounds it takes to select 4 times by the server. Another parameter is the number of parallel covert channels to be established. Receiver generates k edge examples, one for each channel, to be used during the communication. Each edge example is generated by applying one of the following two linear transformation functions to the randomly selected MNIST examples and .

- $_v(x_1; x_2;)$ combines the upper fragment of with the lower (1) fragment of .
- $_h(x_1; x_2;)$ combines the leftmost fragment of with the rightmost (1) fragment of .

Functions $_v$ and $_h$ resemble the example function of Section V. Figure 7 shows two edge examples generated with Algorithm 1 when executing $_v$ and $_h$ on MNIST pairs (#22242, #7596) and (#32481, #18198), respectively. In the first case, the two original examples are both classified as 4, and the resulting edge example is classified between 9 and 4. In the second case, the two images are classified as 2, and the edge example is classified between 2 and 4.

For the generation of edge examples, we randomly selected pairs of pictures from the MNIST dataset, and applied Algorithm 1 with both $_h$ and $_v$. Over $6,000$ considered pairs, we obtained 101 edge examples. Thus, among all possible pairs of MNIST examples, approximately 1.7% might be used to create edge examples. This amounts to millions of possible channel implementations with $_v$ and $_h$ over the whole MNIST dataset.

When Receiver completes the calibration phase, Sender is configured with the generated channel parameters. Then, Sender joins the FL client network, consisting of n_c + 2 clients, and starts the transmission. Transmission parameters include n_b, i.e., the number of bits to be sent on each channel. By default, Sender automatically generates random bit sequences. Optionally, instead of random bits, one can specify a fixed bit pattern w. Finally, when selected, Sender trains its local model in the same way, e.g., using the same training epochs as all the other clients. The only difference resides in the training examples. That is, we assume Sender to have the same capabilities as honest clients.

Table 2 summarizes the experiment parameters for the FL system (top) and the attacker (bottom).

Parameters of FL system (top) and attacker (bottom).

Par.	Description	Range	Example	Default
n_c	Number of honest clients	N	10, 50	10
p_c	Client selection probability	[0; 1]	0.1, 0.5	0.5
	Neural network type	NN, CNN	NN	NN
f	Frame size	N	6, 8	auto
k	Number of channels	N	10, 20	1
n_b	Transmission length (bits)	N	10, 100	10
w	Transmission pattern	(0j1)	10, 101	auto

B. RESULTS

We assess channel quality in terms of BER and SNR. The box diagram[8] of Figures 8 shows single channel BER measured on NN (top) and CNN (bottom) under different settings. In particular, we consider n_c 2 f 10; 50g, p_c 2 f 0:1; 0:2; 0:3; 0:4; 0:5g, n_b = 100, and f estimated by Receiver as detailed in Section VII-A. Under the same settings described above, Figures 9 and 10 show the average BER for 10-bit transmission slots.

Channel BER for NN (top) and CNN (bottom).

Results show that, when p_c > 0:1, BER tends to stay below 6% already with our simple heuristics for estimating f. Nevertheless, the attacker can achieve better performance by searching for optimal values of f (as discussed in Section V, BER is mainly affected by choice of f). Figures 11 shows

[8]Each box represents 5 distinct experiments.

Average BER on NN for 10-bit transmission slots.

Frame length estimation on NN (top) and CNN (bottom).

Average BER on CNN for 10-bit transmission slots.

BER vs. frame size (= NN, $n_b = 100$, $n_c = 10$, $p_c = 0.5$).

the frame size estimated at calibration time for the NN (top) and CNN (bottom) experiments presented above. Instead, Figures 12 and 13, depict the cumulative BER for transmissions with increasing values of, from 6 to 15. Interestingly enough, NN and CNN exhibit different behaviors, which confirms our expectation that finding the optimal value for is non-trivial. In particular, under the same settings, CNN tends to behave monotonically, i.e., increasing reduces BER, while NN does not exhibit a consistent behavior. The experiments suggest that the optimal frame size in the considered interval is = 7 for NN and = 12 for CNN.

Transmitted bit sequences may also impact communication quality. An example of BER for different transmission patterns is shown in Figure 14. There, we considered five different patterns of transmitted bytes, i.e., 00, 0F, 55, FF, and Random, for both NN and CNN. In both cases, we set to the optimal values above, i.e., 7, and 12 for NN and CNN, respectively. As one might expect, the 00 pattern results in the minimum BER. The reason is that, since the FL system is well-trained, in most cases, Sender does not need to poison the global model at all. However, different patterns slightly affect BER in NN, which still stays below 4%, while have almost no effect on CNN. These results highlight that different types of networks may change the behavior of the channel. In general, finding the optimal channel parameter may require the attacker to carry out transmission tests, and this operation is allowed under our attacker model

The number of parallel covert channels an FL system can support is another crucial aspect, as it multiplies the total transmission bandwidth. Intuitively, we expect that an FL system can only host a finite number of channels before the model is saturated and the communication quality decays. Reasonably, we expect that saturation occurs first in simpler, shallow models. Figure 15 shows the increase of BER over 20 randomly generated channels in NN (top) and CNN (bottom), where the average BER increases with

We tested our definition of SNR (see Section VI) as an indicator of the covert channel quality. To this aim, we compared SNR versus BER and capacity. In both cases, for all the simulations, (i) we computed the SNR, and (ii) we binned SNR in 5 dB intervals. Then, we computed the average BER (capacity, respectively) for each bin. Figure 16 illustrates the BER when SNR increases, for both NN and CNN, under the settings presented above. Under the same settings, Figure 17

BER vs. frame size (= CNN, $n_b = 100$, $n_c = 10$, $p_c = 0:5$).

BER variation over the number of parallel covert channels k for NN and CNN (with $n_c = 10$, $p_c = 0:5$).

BER for patterns w 2 {00, 0F, 55, FF, Random} for both = NN and = CNN ($n_b = 100$, $n_c = 10$, $p_c = 0:5$, f = 7 and f = 12, for NN and CNN, respectively).

shows the relation between the SNR and channel capacity C. Our experiments show that higher values of SNR result in better communications, both in terms of lower BER and higher successful transmission rate. This trend follows the expected behavior of a channel in the classical information theory [49]. Thus, our experiments con rm that (5) is an appropriate de nition of SNR.

Figure 18 shows the average BER of 5 tests transmitting 100 bits with a different number of clients, ranging from 100 to 500. The results con rm that average BER stays below 10% as it does with a lower number of clients.

To show the feasibility of our attack on other tasks, we also mounted it on a CIFAR-10 [37] FL classi ers. Such a classi er relies on a neural network implementing (a simpli ed version of) VGGNet [69]. As for MNIST, we tested our method by measuring the BER for 100 bits transmissions using different frame sizes f). This experiment, reported in Figure 19, con rms that our covert channel attack is general w.r.t. the used dataset and classi cation algorithm.

SNR versus BER for NN and CNN.

The main goal of this paper is to prove the feasibility of a new, covert channel attack on FL systems. Our attack opens a number of research questions that are worth being investigated in the future. We brie y discuss them below.

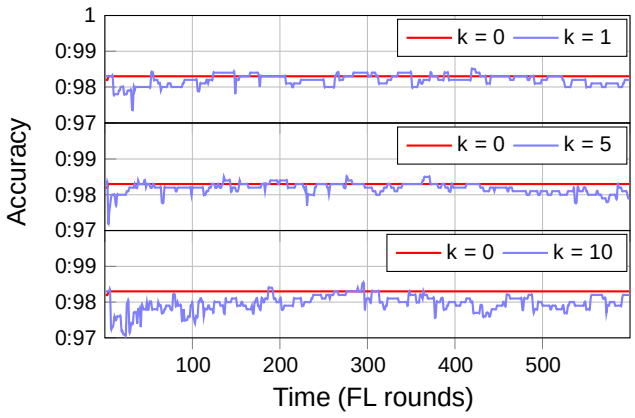
Channel throughput. Our experiments highlight that covert channels can be instantiated on both NN and CNN networks, although NN provides better performances in terms of channel quality, transmission rate f), and bandwidth. To better understand the actual throughput, consider a system where FL rounds occur every hour. If the attacker can implement 20 parallel covert channels with f = 12, a 128-bit key can be transmitted in 128 12=20 = 76:8 hours, i.e., less than 3 days. Although the transmission rate may seem low, there can be practical exploitation, e.g., by Advanced Persistent Threats. Also, our covert channel can be combined with others, thus implementing hybrid covert channels. Further investigation is necessary to nd techniques to maximize the channel performance.

Impact on classi cation accuracy. Although the goal of the attacker is not to compromise the accuracy of the classi er, the covert channel might degrade the performance of the FL system. We compared the accuracy measured by an honest client when Sender is not transmitting k = 0) against that measured during transmissions with 1, 5 and, 10 channels. The results are shown in Figures 20 and 21.

Note that the honest client accuracy when k = 0 is stable. This happens because the network is well-trained by using

SNR versus channel capacity (C) for NN and CNN.

BER for CIFAR-10 dataset, 100 bits and different frame sizes.



Average BER over 100 bits with 100, 200, 300, 400, and 500 clients using NN for the classi cation task.

Accuracy for k 2 f 1; 5; 10g, = NN, n_c = 10 , p_c = 0 :5.

the entire MNIST dataset. In real FL systems, this could not happen, since fresh training examples can appear over time.

For NN, when k = 0 the average accuracy is 0.983. When transmitting, the average accuracy slightly degrades, from 0:982 (k = 1) to 0:980 (k = 10). Similarly for CNN, the accuracy goes from 0:992 (k = 1) to 0:988 (k = 10). These results con rm that classi cation accuracy is minimally affected by transmissions. In larger FL systems, trained with fresh examples, accuracy is likely to be affected even less.

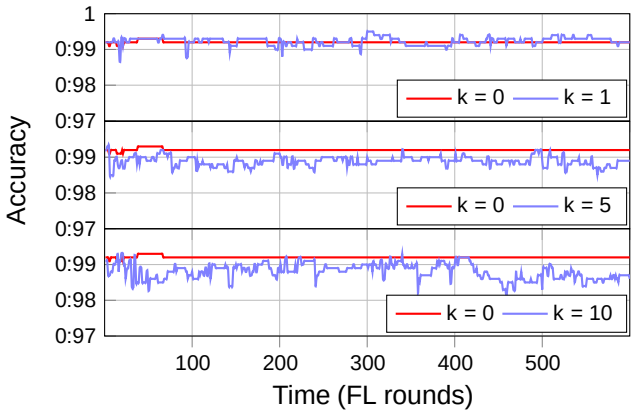
Possible mitigation mechanisms. Although the attacker model is new, some existing techniques might mitigate its effectiveness. For instance, anomaly detection techniques might identify the sender. Although designed for countering byzantine adversaries, Krum [11] aggregation function skips outlier models and could block or degrade the transmission. In our setting, the FL server can measure the distance between two models by using matrix norm. With L2-norm, the distance of a local model m_c w.r.t. the global model m is given by $d_c = km_c \quad mk_2$.[9] Figure 22 shows the maximum d_c, i.e., the outlier, detected by the server during 200 FL rounds for both NN (top) and CNN (bottom).

Red bars correspond to true positives (TP), i.e., rounds where Sender's model was detected. Instead, blue bars denote false positives (FP), i.e., when an honest client was tagged. Detection precision is 12% and 17% for NN and CNN, respectively. These values are quite low in a 10-client network since 10% is the accuracy of a random strategy. Moreover, to remain undetected, the attacker may try to minimize d_c, e.g., using poisoning data rate control [59].

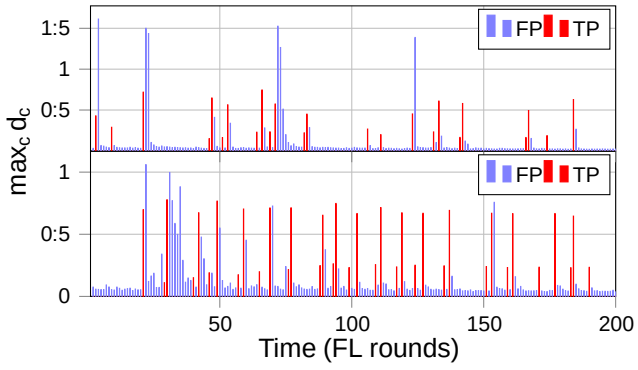
Another countering approach consists in adopting differential privacy (DP) [78]. Brie y, in DP mechanisms for FL both the honest clients and the server inject random noise in their models to prevent a malicious client from leaking private information. DP typically does not work against collusion, i.e., Sender wants to leak information. Nevertheless, at the price of degrading model accuracy, noise added by the server may reduce the channel quality.

In this paper, we introduced a new covert channel leveraging FL systems. Our attack allows a malicious agent to establish stealth communications between FL clients that should rather stay isolated. We discuss a prototype implementation and we empirically assess its performance. Our experiments con rm that the attack is feasible and the covert channel supports good-quality communications. Since the attack is new, no speci c countermeasures exist, however some existing countermeasures for poisoning attack, e.g., [11], [48], [58] , might

[9]More precisely, this amounts to computing the L2-norm of the difference between the vectors of parameters of m_c and m.



Accuracy for k 2 f 1; 5; 10g, = CNN, n_c = 10 , p_c = 0 :5.



Anomalous local models detected over 200 rounds for = NN (top) and = CNN (bottom), with k = 1 , n_c = 10 , p_c = 0 :5 and f = 6 .

affect our covert channel and we plan to study them in the future. Further research directions include those related to the actual exploitability in real-world systems and to the costs for an attacker to implement our channel.

[1] Hojjat Aghakhani, Thorsten Eisenhofer, Lea Schönherr, Dorothea Kolossa, Thorsten Holz, Christopher Kruegel, and Giovanni Vigna. VENOMAVE: clean-label poisoning against speech recognition. CoRR, abs/2010.10682, 2020.

[2] Giuseppe Ateniese, Luigi V. Mancini, Angelo Spognardi, Antonio Villani, Domenico Vitali, and Giovanni Felici. Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classi ers. International Journal of Security and Networks, 10(3):137–150, September 2015.

[3] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How To Backdoor Federated Learning. In Silvia Chiappa and Roberto Calandra, editors, Proceedings of the Twenty Third International Conference on Arti cial Intelligence and Statistics, volume 108 of Proceedings of Machine Learning Research, pages 2938–2948, Online, 26–28 Aug 2020. PMLR.

[4] Marco Barreno, Blaine Nelson, Russell Sears, Anthony D. Joseph, and J. D. Tygar. Can machine learning be secure? In Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security, ASIACCS '06, page 16–25, New York, NY, USA, 2006. Association for Computing Machinery.

[5] Matthias Bauer. New Covert Channels in HTTP: Adding Unwitting Web Browsers to Anonymity Sets. In Proceedings of the 2003 ACM Workshop on Privacy in the Electronic Society, WPES '03, page 72–78, New York, NY, USA, 2003. Association for Computing Machinery.

[6] Arjun Nitin Bhagoji, Supriyo Chakraborty, Prateek Mittal, and Seraphin Calo. Analyzing federated learning through an adversarial lens. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, Proceedings of the 36th International Conference on Machine Learning, volume 97 of Proceedings of Machine Learning Research, pages 634–643. PMLR, 09–15 Jun 2019.

[7] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In Hendrik Blockeel, Kristian Kersting, Siegfried Nijssen, and Filip elezný, editors, Machine Learning and Knowledge Discovery in Databases, pages 387–402, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[8] Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. In Proceedings of the 29th International Coference on International Conference on Machine Learning, ICML'12, page 1467–1474, Madison, WI, USA, 2012. Omnipress.

[9] Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition, 84:317 – 331, 2018.

[10] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, page 118–128, Red Hook, NY, USA, 2017. Curran Associates Inc.

[11] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent. In Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, page 118–128, Red Hook, NY, USA, 2017. Curran Associates Inc.

[12] Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. In 2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017, pages 39–57. IEEE Computer Society, 2017.

[13] Daniel C. Castro, Jeremy Tan, Bernhard Kainz, Ender Konukoglu, and Ben Glocker. Morpho-MNIST: Quantitative assessment and diagnostics for representation learning. Journal of Machine Learning Research, 20(178), 2019.

[14] Swarup Chandra, Zhiqiang Lin, Ashish Kundu, and Latifur Khan. Towards a systematic study of the covert channel attacks in smartphones. In Jing Tian, Jiwu Jing, and Mudhakar Srivatsa, editors, International Conference on Security and Privacy in Communication Networks, pages 427–435, Cham, 2015. Springer International Publishing.

[15] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. CoRR, abs/1712.05526, 2017.

[16] Y. Chen, X. Qin, J. Wang, C. Yu, and W. Gao. FedHealth: A Federated Transfer Learning Framework for Wearable Healthcare. IEEE Intelligent Systems, 35(4):83–93, 2020.

[17] Gabriele Costa, Fabio Pinelli, Simone Soderi, and Gabriele Tolomei. Covert channel attack to federated learning systems. CoRR, abs/2104.10561, 2021.

[18] Nilesh N. Dalvi, Pedro M. Domingos, Mausam, Sumit K. Sanghai, and Deepak Verma. Adversarial classi cation. In SIGKDD Conference on Knowledge Discovery and Data Mining, pages 99–108. ACM, 2004.

[19] Jacob Dumford and Walter J. Scheirer. Backdooring convolutional neural networks via targeted weight perturbations. CoRR, abs/1812.03128, 2018.

[20] Minghong Fang, Xiaoyu Cao, Jinyuan Jia, and Neil Zhenqiang Gong. Local model poisoning attacks to byzantine-robust federated learning. In Srdjan Capkun and Franziska Roesner, editors, 29th USENIX Security Symposium, USENIX Security 2020, August 12-14, 2020, pages 1605–1622. USENIX Association, 2020.

[21] Minghong Fang, Guolei Yang, Neil Zhenqiang Gong, and Jia Liu. Poisoning attacks to graph-based recommender systems. In Proceedings of the 34th Annual Computer Security Applications Conference, ACSAC '18, page 381–392, New York, NY, USA, 2018. Association for Computing Machinery.

[22] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit con dence information and basic countermeasures. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, CCS '15, page 1322–1333, New York, NY, USA, 2015. Association for Computing Machinery.

[23] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In 23rd USENIX Security

Symposium (USENIX Security 14), pages 17–32, San Diego, CA, August 2014. USENIX Association.

[24] Clement Fung, Chris J. M. Yoon, and Ivan Beschastnikh. Mitigating Sybils in Federated Learning Poisoning. arXiv e-prints, August 2018.

[25] Karan Ganju, Qi Wang, Wei Yang, Carl A. Gunter, and Nikita Borisov. Property inference attacks on fully connected neural networks using permutation invariant representations. In Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, CCS '18, page 619–633, New York, NY, USA, 2018. Association for Computing Machinery.

[26] Micah Goldblum, Dimitris Tsipras, Chulin Xie, Xinyun Chen, Avi Schwarzschild, Dawn Song, Aleksander Madry, Bo Li, and Tom Goldstein. Dataset Security for Machine Learning: Data Poisoning, Backdoor Attacks, and Defenses. arXiv e-prints, page arXiv:2012.10544, December 2020.

[27] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep Learning. The MIT Press, 2016. ISBN: 0262035618.

[28] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. CoRR, abs/1708.06733, 2017.

[29] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. IEEE Access, 7:47230–47244, 2019.

[30] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The Elements of Statistical Learning. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001. ISBN: 0387848576.

[31] Dorjan Hitaj, Giulio Pagnotta, Briland Hitaj, Fernando Perez-Cruz, and Luigi V. Mancini. Fedcomm: Federated learning as a medium for covert communication. CoRR, abs/2201.08786, 2022.

[32] Ling Huang, Anthony D. Joseph, Blaine Nelson, Benjamin I.P. Rubinstein, and J. D. Tygar. Adversarial machine learning. In Proceedings of the 4th ACM Workshop on Security and Arti cial Intelligence, AISec '11, page 43–58, New York, NY, USA, 2011. Association for Computing Machinery.

[33] Matthew Jagielski, Alina Oprea, Battista Biggio, Chang Liu, Cristina Nita-Rotaru, and Bo Li. Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. In 2018 IEEE Symposium on Security and Privacy (SP), pages 19–35. IEEE Computer Society, 2018.

[34] M. S. Jere, T. Farnan, and F. Koushanfar. A taxonomy of attacks on federated learning. IEEE Security Privacy, 19(2):20–28, 2020.

[35] Yujie Ji, Xinyang Zhang, Shouling Ji, Xiapu Luo, and Ting Wang. Model-reuse attacks on deep learning systems. In Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, CCS '18, page 349–363, New York, NY, USA, 2018. Association for Computing Machinery.

[36] Peter Kairouz et al. Advances and Open Problems in Federated Learning. Foundations and Trends in Machine Learning, 14(1), 2021.

[37] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

[38] Butler W. Lampson. A note on the con nement problem. Communications of the ACM, 16(10):613–615, October 1973.

[39] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. Neural Computation, 1:541–551, 1989.

[40] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In Proceedings of the IEEE, volume 86, pages 2278–2324, 1998.

[41] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010.

[42] Bo Li, Yining Wang, Aarti Singh, and Yevgeniy Vorobeychik. Data poisoning attacks on factorization-based collaborative ltering. In Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16, page 1893–1901, Red Hook, NY, USA, 2016. Curran Associates Inc.

[43] Bin Liang, Miaoqiang Su, Wei You, Wenchang Shi, and Gang Yang. Cracking classi ers for evasion: A case study on the google's phishing pages lter. In Proceedings of the 25th International Conference on World Wide Web, WWW '16, page 345–356, Republic and Canton of Geneva, CHE, 2016. International World Wide Web Conferences Steering Committee.

[44] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. In 25th Annual Network and Distributed System Security Symposium, NDSS 2018, San Diego, California, USA, February 18-21, 2018. The Internet Society, 2018.

[45] Yanyang Lu and Lei Fan. An ef cient and robust aggregation algorithm for learning federated cnn. In Proceedings of the 2020 3rd International Conference on Signal Processing and Machine Learning, SPML 2020, pages 1–7, New York, NY, USA, 2020. Association for Computing Machinery.

[46] Lingjuan Lyu, Han Yu, and Qiang Yang. Threats to Federated Learning: A Survey. arXiv e-prints, March 2020.

[47] Saeed Mahloujifar, Mohammad Mahmoody, and Ameer Mohammed. Universal multi-party poisoning attacks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, Proceedings of the 36th International Conference on Machine Learning, volume 97 of Proceedings of Machine Learning Research, pages 4274–4283. PMLR, 09–15 Jun 2019.

[48] Yunlong Mao, Xinyu Yuan, Xinyang Zhao, and Sheng Zhong. Romoa: Robust model aggregation for the resistance of federated learning to model poisoning attacks. In European Symposium on Research in Computer Security, pages 476–496. Springer, 2021.

[49] P. Massoud Salehi and J. Proakis. Digital Communications 5th Edition. McGraw-Hill Education, 2007. ISBN: 9780072957167.

[50] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Agüera y Blaise Arcas. Communication-ef cient learning of deep networks from decentralized data. AISTATS, pages 1273–1282, 2017.

[51] Brendan McMahan and Daniel Ramage. Federated Learning: Collaborative Machine Learning without Centralized Training Data. https://ai.googleblog.com/2017/04/federated-learning-collaborative.html, 2017. Accessed: 2017-04-06.

[52] L. Melis, C. Song, E. De Cristofaro, and V. Shmatikov. Exploiting unintended feature leakage in collaborative learning. In 2019 IEEE Symposium on Security and Privacy (SP), pages 691–706, 2019.

[53] Aleksandra Mileva and Boris Panajotov. Covert channels in TCP/IP protocol stack - Extended version -. Central European Journal of Computer Science, 4:45–66, 06 2014.

[54] Kevin P. Murphy. Machine Learning: A Probabilistic Perspective. The MIT Press, 2012. ISBN: 0262018020.

[55] Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In 2019 IEEE Symposium on Security and Privacy, SP 2019, San Francisco, CA, USA, May 19-23, 2019, pages 739–753. IEEE, 2019.

[56] Blaine Nelson, Marco Barreno, Fuching Jack Chi, Anthony D. Joseph, Benjamin I. P. Rubinstein, Udam Saini, Charles Sutton, J. D. Tygar, and Kai Xia. Exploiting machine learning to subvert your spam lter. In Proceedings of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats, LEET'08, USA, 2008. USENIX Association.

[57] Jiquan Ngiam, Zhenghao Chen, Daniel Chia, Pang Koh, Quoc Le, and Andrew Ng. Tiled convolutional neural networks. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, Advances in Neural Information Processing Systems, volume 23. Curran Associates, Inc., 2010.

[58] Thien Duc Nguyen, Phillip Rieger, Huili Chen, Hossein Yalame, Helen Möllering, Hossein Fereidooni, Samuel Marchal, Markus Miettinen, Azalia Mirhoseini, Shaza Zeitouni, et al.f FLAMEg: Taming backdoors in federated learning. In 31st USENIX Security Symposium (USENIX Security 22), pages 1415–1432, 2022.

[59] Thien Duc Nguyen, Phillip Rieger, Markus Miettinen, and Ahmad-Reza Sadeghi. Poisoning Attacks on Federated Learning-based IoT Intrusion Detection System. In Workshop on Decentralized IoT Systems and Security, DISS, San Diego, California, USA, February 23, 2020. The Internet Society, 2020.

[60] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, ASIA CCS '17, page 506–519, New York, NY, USA, 2017. Association for Computing Machinery.

[61] Adam Paszke et al. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach et al., editors, Advances in Neural Information Processing Systems 32, pages 8024–8035. Curran Associates, Inc., 2019.

[62] Nikolaos Pitropakis, Emmanouil Panaousis, Thanassis Giannetsos, Eleftherios Anastasiadis, and George Loukas. A taxonomy and survey of attacks against machine learning. Computer Science Review, 34:100–199, 2019.

[63] S. R. Pokhrel and J. Choi. A Decentralized Federated Learning Approach for Connected Autonomous Vehicles. In 2020 IEEE Wireless Communi-

cations and Networking Conference Workshops (WCNCW), pages 1–6, 2020.

[64] Apostolos Pyrgelis, Carmela Troncoso, and Emiliano De Cristofaro. Knock knock, who's there? membership inference on aggregate location data. In 25th Annual Network and Distributed System Security Symposium, NDSS 2018, San Diego, California, USA. The Internet Society, 2018.

[65] Benjamin I.P. Rubinstein, Blaine Nelson, Ling Huang, Anthony D. Joseph, Shing-hon Lau, Satish Rao, Nina Taft, and J. D. Tygar. Antidote: Understanding and defending against poisoning of anomaly detectors. In Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement, IMC '09, page 1–14, New York, NY, USA, 2009. Association for Computing Machinery.

[66] Jose Selvi. Covert channels over social networks. Technical report, SANS Institute, June 2012. https://www.sans.org/reading-room/whitepapers/engineering/paper/33960.

[67] Ali Shafahi, W. Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS'18, page 6106–6116, Red Hook, NY, USA, 2018. Curran Associates Inc.

[68] 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), pages 3–18, 2017.

[69] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 2015, Conference Track Proceedings, 2015.

[70] O. Suciu, S. E. Coull, and J. Johns. Exploring adversarial examples in malware detection. In 2019 IEEE Security and Privacy Workshops (SPW), pages 8–14, 2019.

[71] Octavian Suciu, Radu Mărginean, Yiğitcan Kaya, Hal Daumé, and Tudor Dumitraş. When does machine learning fail? generalized transferability for evasion and poisoning attacks. In Proceedings of the 27th USENIX Conference on Security Symposium, SEC'18, page 1299–1316, USA, 2018. USENIX Association.

[72] Gan Sun, Yang Cong, Jiahua Dong, Q. Wang, and J. Liu. Data poisoning attacks on federated machine learning. ArXiv, abs/2004.10020, 2020.

[73] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In Yoshua Bengio and Yann LeCun, editors, 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings, 2014.

[74] Vale Tolpegin, Stacey Truex, Mehmet Emre Gursoy, and Ling Liu. Data poisoning attacks against federated learning systems, 2020.

[75] Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction apis. In Proceedings of the 25th USENIX Conference on Security Symposium, SEC'16, page 601–618, USA, 2016. USENIX Association.

[76] V. Vapnik. Principles of risk minimization for learning theory. In Proceedings of the 4th International Conference on Neural Information Processing Systems, NIPS'91, page 831–838, San Francisco, CA, USA, 1991. Morgan Kaufmann Publishers Inc.

[77] Binghui Wang and Neil Zhenqiang Gong. Stealing hyperparameters in machine learning. In 2018 IEEE Symposium on Security and Privacy, SP 2018, Proceedings, 21-23 May 2018, San Francisco, California, USA, pages 36–52. IEEE Computer Society, 2018.

[78] Kang Wei, Jun Li, Ming Ding, Chuan Ma, Howard H. Yang, Farhad Farokhi, Shi Jin, Tony Q. S. Quek, and H. Vincent Poor. Federated Learning With Differential Privacy: Algorithms and Performance Analysis. IEEE Transactions on Information Forensics and Security, 15:3454–3469, 2020.

[79] Q. Wu, K. He, and X. Chen. Personalized Federated Learning for Intelligent IoT Applications: A Cloud-Edge Based Framework. IEEE Open Journal of the Computer Society, 1:35–44, 2020.

[80] Huang Xiao, Battista Biggio, Gavin Brown, Giorgio Fumera, Claudia Eckert, and Fabio Roli. Is feature selection secure against training data poisoning? In Francis Bach and David Blei, editors, Proceedings of the 32nd International Conference on Machine Learning, volume 37 of Proceedings of Machine Learning Research, pages 1689–1698, Lille, France, 07–09 Jul 2015. PMLR.

[81] Chulin Xie, Keli Huang, Pin-Yu Chen, and Bo Li. Dba: Distributed backdoor attacks against federated learning. In International Conference on Learning Representations, 2020.

[82] Guolei Yang, Neil Zhenqiang Gong, and Ying Cai. Fake co-visitation injection attacks to recommender systems. In 24th Annual Network and Distributed System Security Symposium, NDSS 2017, San Diego, California, USA, February 26 - March 1, 2017. The Internet Society, 2017.

[83] Timothy Yang, Galen Andrew, Hubert Eichner, Haicheng Sun, Wei Li, Nicholas Kong, Daniel Ramage, and Françoise Beaufays. Applied Federated Learning: Improving Google Keyboard Query Suggestions. CoRR, abs/1812.02903, 2018.

[84] Sebastian Zander, Grenville Armitage, and Philip Branch. Covert channels and countermeasures in computer network protocols. IEEE Communications Surveys and Tutorials, 9:44–57, 09 2007.

[85] Jiale Zhang, Junjun Chen, Di Wu, Bing Chen, and Shui Yu. Poisoning attack in federated learning using generative adversarial nets. In 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications / 13th IEEE International Conference On Big Data Science And Engineering, TrustCom/BigDataSE 2019, Rotorua, New Zealand, August 5-8, 2019, pages 374–380. IEEE, 2019.

[86] C. Zhou, A. Fu, S. Yu, W. Yang, H. Wang, and Y. Zhang. Privacy-Preserving Federated Learning in Fog Computing. IEEE Internet of Things Journal, 7(11):10782–10793, 2020.

[87] Chen Zhu, W. Ronny Huang, Hengduo Li, Gavin Taylor, Christoph Studer, and Tom Goldstein. Transferable clean-label poisoning attacks on deep neural nets. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, Proceedings of the 36th International Conference on Machine Learning, volume 97 of Proceedings of Machine Learning Research, pages 7614–7623. PMLR, 09–15 Jun 2019.

GABRIELE COSTA is associate professor at the SySMA Group of the IMT School for Advanced Studies. He received his M.Sc. in Computer Science in 2007 and his Ph.D. in Computer Science in 2011. He was a member of the cybersecurity group of the Istituto di Informatica e Telematica (IIT) of the CNR. His appointments include a period as visiting researcher at ETH Zurich in 2016-2017. He is the co-founder, and CRO of Talos, a spin-off of DIBRIS focused on Cybersecurity. He focuses on studying and applying formal methods for the automatic verification and testing of mobile and modular systems.

FABIO PINELLI is an assistant professor at the SySMA Group of the IMT School for Advanced Studies. He received his M.Sc. in Computer Science in 2005 and his Ph.D. in Information Engineering in 2010 from the University of Pisa. He was Research Scientist at IBM-Research Ireland and Senior Data Scientist at Vodafone Italia. His main research interests focus on data mining and machine learning methods applied to different domains, from economics to urban environments.