Poisoning Attacks on Federated Learning-based Intrusion Detection System

Thien Duc Nguyen, Phillip Rieger, Markus Miettinen, Ahmad-Reza Sadeghi
Typical IoT Devices
IoT

The S stands for Security
Mirai: Largest Disruptive Cyberattack in History

More than 145,000 infected devices

Peak bandwidth of 1156 Gbps

Source: https://www.incapsula.com/blog/malware-analysis-mirai-ddos-botnet.html
Mirai: Largest Disruptive Cyberattack in History

- More than 145,000 infected devices
- Peak bandwidth of 1156 Gbps

Source: https://www.incapsula.com/blog/malware-analysis-mirai-ddos-botnet.html
Federated Learning
Federated Learning
Federated Learning

Client

Federated Learning Aggregator

Client

Client

...
Federated Learning
Advantages of Federated Learning

• Allows all participants to profit from all data

• Privacy Preserving
  ▪ E.g.: Don’t reveal network traffic

• Distributing computation load to clients
IoT NIDS

SGW: Security Gateway (e.g., Local WiFi router)
IoT NIDS

SGW: Security Gateway (e.g., Local WiFi router)
IoT NIDS

SGW: Security Gateway (e.g., Local WiFi router)
IoT NIDS

Aggregator

SGW: Security Gateway (e.g., Local WiFi router)
Examples of Backdoor Attacks: Adversary Chosen Label

**Image classification**
Change labels, e.g.,
- Speed limit signs from 30kph to 80kph

**Word prediction**
Select end words, e.g.,
"buy phone from Google"

**IoT malware detection**
Inject malicious traffic, e.g., use compromised IoT devices

---

Our new Attack
Backdoor Attacks on FL

1. Manipulate training data
2. Manipulate local models

Nguyen et al., ICDCS 2019
Backdoor Attacks on FL

Attack Strategies:
1. Manipulate training data
2. Manipulate local models

Nguyen et. al., ICDCS 2019
Our Threat Model

Attack Goal:
- Inject Backdoor

Attacker’s Capabilities:
- Full knowledge about the targeted system
- Fully control some IoT devices

Attacker cannot:
- Control Security Gateways
- Control devices in < 50% of all networks
Our Approach – High Level Idea

• Challenge: Prevent detection of data poisoning

• Only few attack data

→ Gateway will not detect it

→ Still include malware traffic in training data

→ Neural Network learns to predict malware behavior

• Use compromised IoT devices
Our Approach

1. Compromise IoT Devices
2. Inject Malicious Data

SGW: Security Gateway (e.g., Local WiFi router)
Our Approach

1. Compromise IoT Devices
2. Inject Malicious Data

SGW: Security Gateway (e.g., Local WiFi router)
Our Approach

1. Compromise IoT Devices
2. Inject Malicious Data

SGW: Security Gateway (e.g., Local WiFi router)
Our Approach

1. Compromise IoT Devices
2. Inject Malicious Data

SGW: Security Gateway (e.g., Local WiFi router)
Experimental Setup

- 3 Real-World Datasets [1, 2]
- Consisting of traffic from 46 IoT devices
- Different stages of Mirai: infection, scanning, different DDoS attacks
- Distributed data to 100 clients
  - Approx. 2h of traffic

[1] Nguyen et.al., ICDCS 2019
Attack Parameters

• Poisoned Model Rate (PMR)
  ▪ Indicates percentage of poisoned local models
    o E.g., ratio of networks, containing compromised IoT devices

• Poisoned Data Rate (PDR)
  ▪ Indicates ratio between poisoned and benign data
    o E.g., ratio between malware and benign network traffic
Evaluation Metrics

• Backdoor Accuracy (BA)
  ▪ E.g., alerts, raised on malware traffic
  ▪ 100 % BA → No Alert for malware traffic

• Main task Accuracy (MA)
  ▪ E.g., accuracy on benign network traffic
  ▪ 100 % MA → No alert for benign traffic
Experimental Results

- Malware traffic not detected for PDR of 36.7% (± 6.5%)
Experimental Results

- Malware traffic not detected for PDR of 36.7% (± 6.5%)

- Attack successful for low number of compromised networks
  - BA 100% for PMR 25% and PDR 20%
  - Higher PMRs are successful for lower PDRs
  - Lower PMRs require higher PDRs
  - PMR 5% is too low
Experimental Results – Clustering Defense

Mechanism:
• Calculates pairwise Euclidean Distances
• Apply Clustering on them

Illustration for PDR = 30%

Experimental Results

• BA 100%
• Attack effective for PDR ≤ 20%
Experimental Results – Clustering Defense

Mechanism:
• Calculates pairwise Euclidean Distances
• Apply Clustering on them

Illustration for PDR = 20%

Experimental Results

• BA 100%
• Attack effective for PDR ≤ 20%
Experimental Results – Differential Privacy Defense

Mechanism:

• Restricts Euclidean distance of local models
• Adds gaussian noise

- Not effective for PDR >= 15%
- BA 100%
- MA reduced significantly
Conclusion

- Introduced novel backdoor attack vector
  - Requires only control of few IoT devices
  - Inject Malware Traffic Stealthily

- Evaluated on 3 real-world datasets

- Bypasses current defenses
Future Research Direction

• Improve IDS

• Filter poisoned data on clients

• Defense against these poisoning attacks