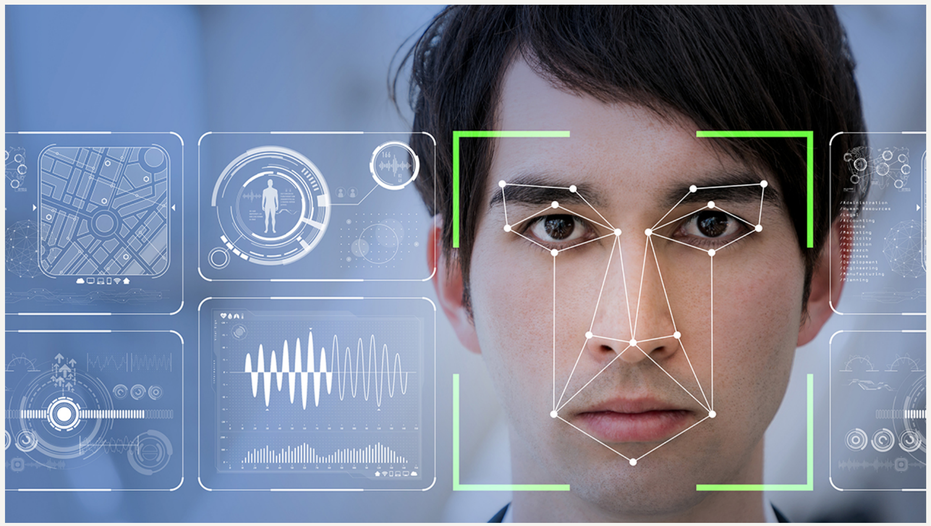
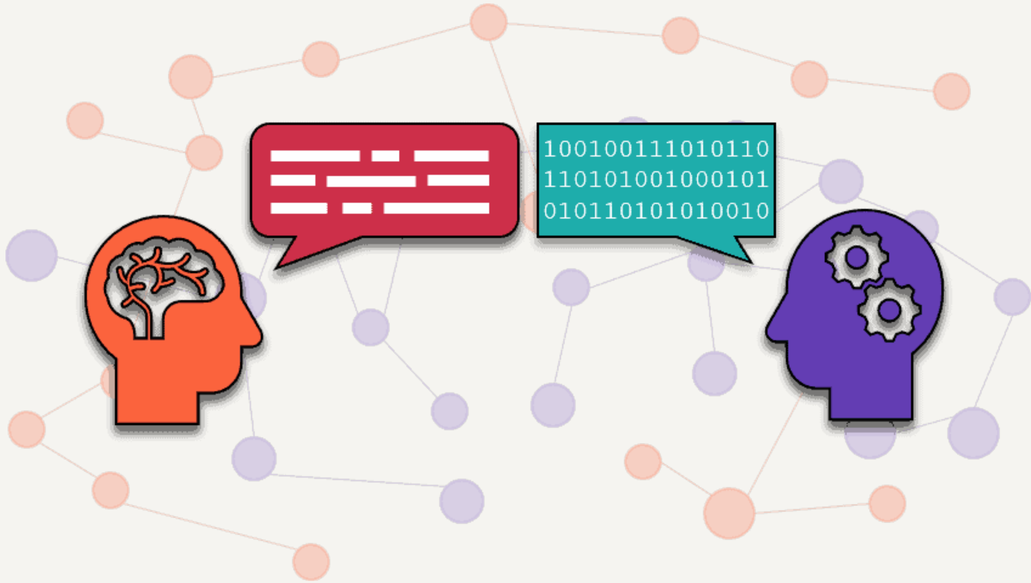
Poisoning Attacks on Deep Learning based Wireless Traffic Prediction

Tianhang Zheng, Baochun Li
University of Toronto

Advances in Deep Learning



Computer Vision



Natural Language Processing

Wireless Traffic Prediction (WTP)

Traditional Methods: Autoregressive integrated moving average (ARIMA) and support vector regression (SVR)

LSTM-based WTP [Wang et al. INFOCOM17]: An autoencoder + long short memory units (LSTM) model for learning spatial and temporal wireless traffic data

Federated Learning for WTP [Zhang et al. INFOCOM21]: LSTM with a dual attention based model aggregation mechanism

However, deep learning is vulnerable in an adversarial environment!

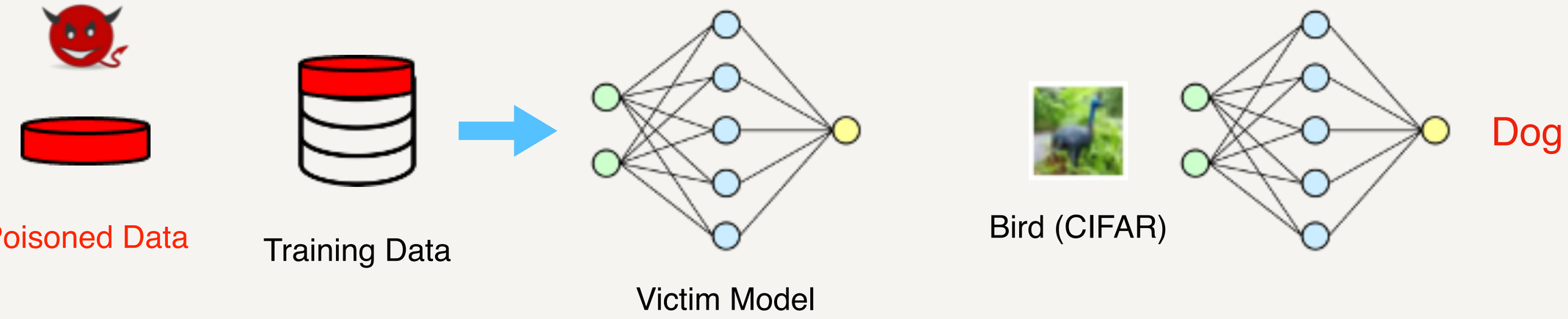
Training Stage Attacks (Poisoning Attacks)

Inference Stage Attacks (Adversarial Attacks)

Privacy Attacks (Reconstruction, Attribute Inference)

Poisoning Attacks

(Training) Data Poisoning

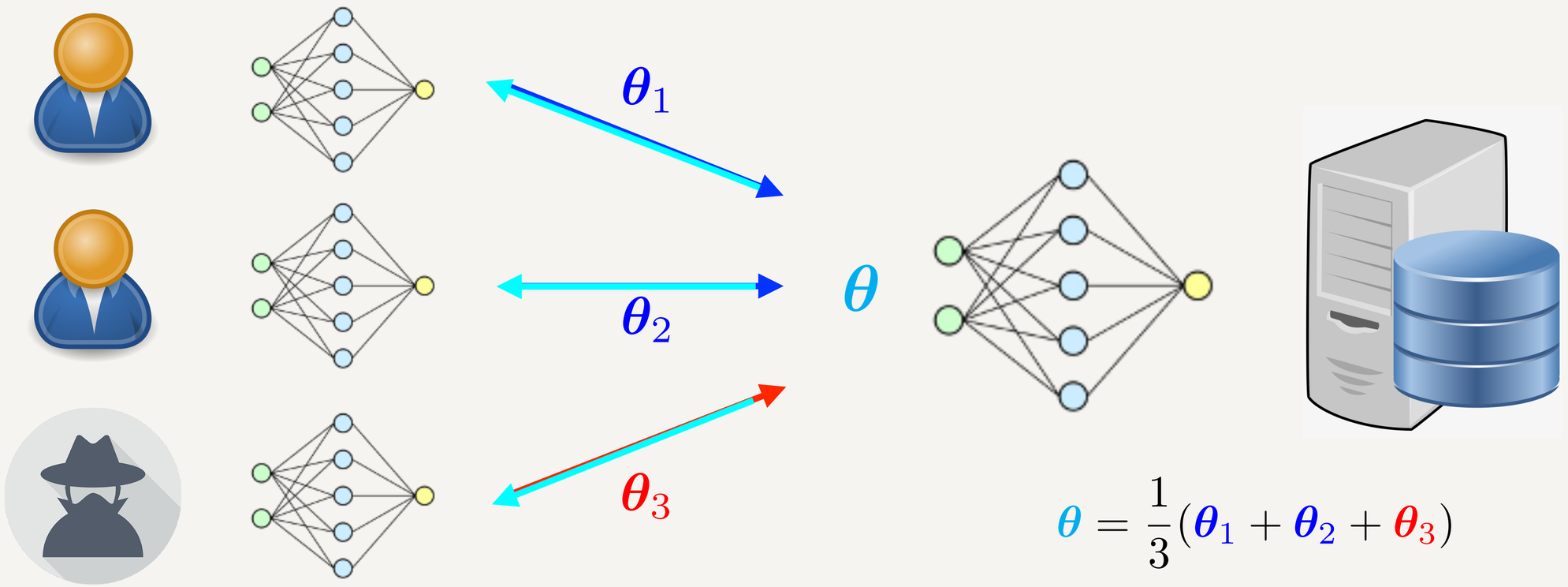


Training Stage

Inference Stage

Poisoning Attacks

Model Poisoning (Federated Learning)



$$\theta = \frac{1}{3}(\theta_1 + \theta_2 + \theta_3)$$

Model Aggregation

WTP as Time Series Forcasting

In general, wireless traffic prediction can be formulated as a time series forcasting problem:

$$\mathbf{v} = \{v_1, v_2, ..., v_T\}$$ traffic volumes with a total of T time points

sliding window

$$\boldsymbol{x} = \{v_{t-1}, v_{t-2}, ...v_{t-p}, v_{t-\phi 1}..., v_{t-\phi q}\}$$

Recent traffic volumes

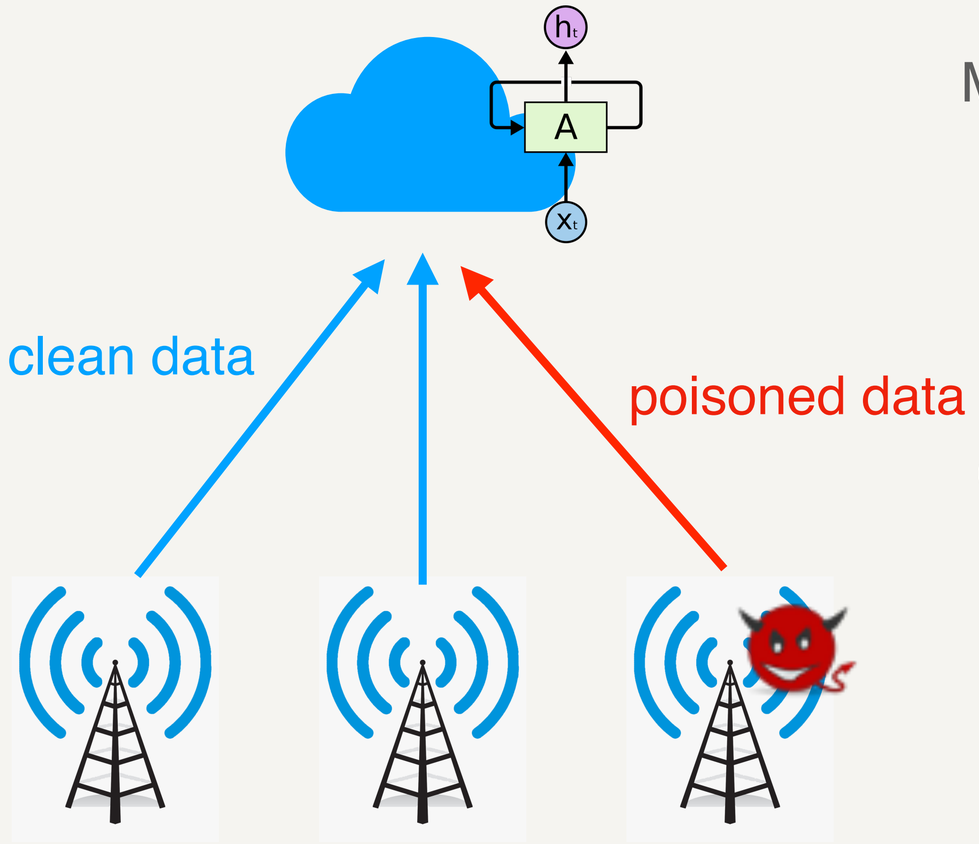
$$\boldsymbol{y} = v_t$$

current traffic volume

model training

$$\hat{y}_t = f_{\boldsymbol{\theta}}(\mathbf{x}_t)$$ wireless traffic prediction model

Centralized Training Scenario



Maximize the mean square error

$$\max_{\boldsymbol{\delta}_x, \delta_y} \frac{1}{|\mathcal{D}|} \sum_{\mathbf{x}, y \in \mathcal{D}} \| f_{\hat{\boldsymbol{\theta}}}(\mathbf{x}) - y \|_2^2$$

Constraints

$$s.t. \quad \hat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta}} \frac{1}{|\mathcal{D}|} \sum_{\mathbf{x}, y \in \mathcal{D}} \| f_{\boldsymbol{\theta}}(\mathbf{x} + \boldsymbol{\delta}_x) - (y + \delta_y) \|_2^2$$

$$\| \boldsymbol{\delta}_{\boldsymbol{x}} \|_p \leq \epsilon_1, \delta_y \leq \epsilon_2$$

Centralized Training Scenario

Threat Model and Challenges:

- The adversary does not have access to the other clients' data or data distribution
- The adversary can poison all of its data, which is still a small part of all the training data
- An intuitive attack method is to optimize the owned data with the existing data poisoning methods (sub-optimal)

Centralized Training Scenario

sample a data batch

$$\{\mathbf{x}_n, y_n, \boldsymbol{\delta}_{\mathbf{x}_n}, \delta_{y_n}\}_{n=1}^{N}$$

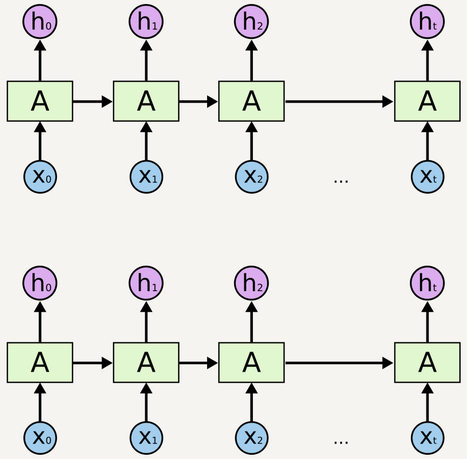
sample random masks

$$\xi_n \sim Bern(0.2)$$

$$\{\boldsymbol{x}_n + \xi_n \boldsymbol{\delta}_{\boldsymbol{x_n}}, \boldsymbol{y}_n + \xi_n \delta_{y_n}\}$$

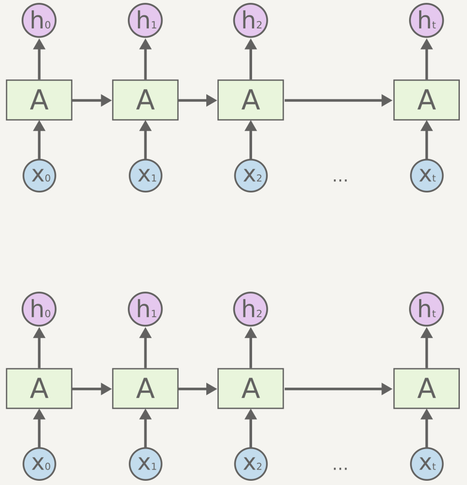
20% poisoned data

M surrogate models



SGD

Intermediate models $\tilde{\boldsymbol{\theta}}$

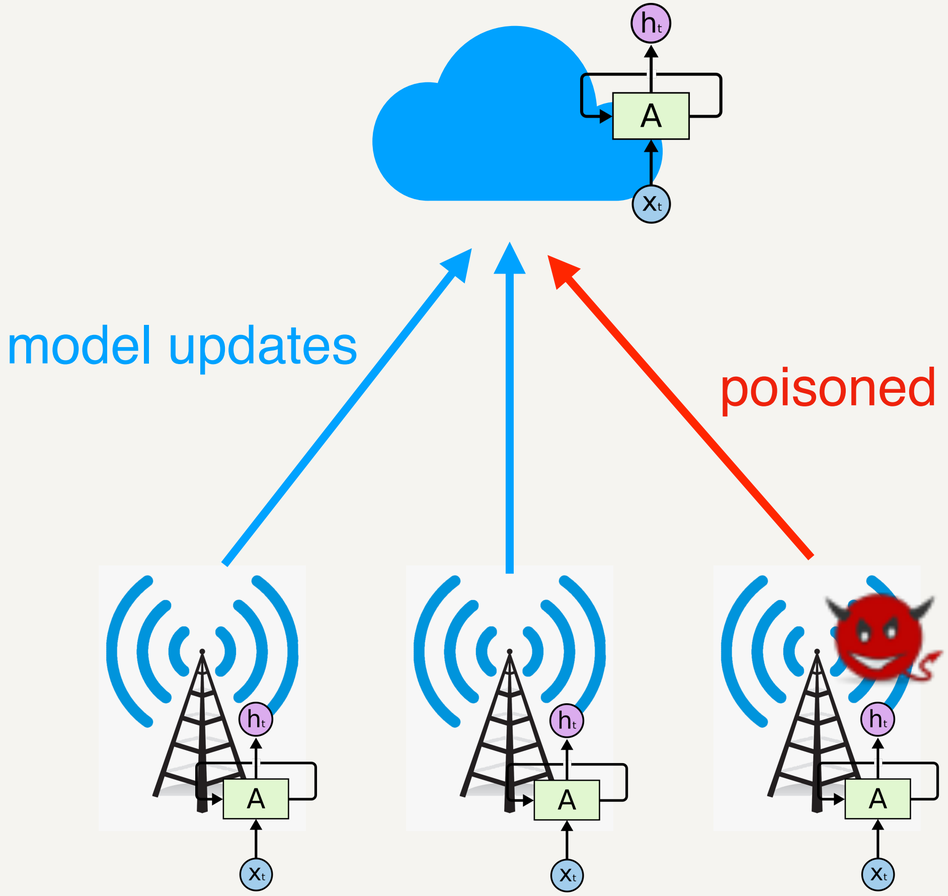


Grad for updating the perturbations

$$\mathbf{g}_{n,m}(\boldsymbol{\delta}_{\mathbf{x}_n}) = -\nabla_{\boldsymbol{\delta}_{\mathbf{x}_n}} \frac{1}{N} \sum_{n=1}^{N} \| f_{\tilde{\boldsymbol{\theta}}}(\mathbf{x}_n) - y_n \|_2^2$$

$$\mathbf{g}_n(\boldsymbol{\delta}_{\mathbf{x}_n}) = \frac{1}{M} \sum_{m=1}^{M} \mathbf{g}_{n,m}(\boldsymbol{\delta}_{\mathbf{x}_n})$$

Distributed Training Scenario



Maximize the error

$$\max_{\Delta\tilde{\boldsymbol{\theta}}_k^t} \frac{1}{|\mathcal{D}|} \sum_{\mathbf{x},y\in\mathcal{D}} \|f_{\boldsymbol{\theta}^{t+1}}(\mathbf{x}) - y\|_2^2$$

Model aggregation

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t + \frac{1}{|S_t|}\Big(\sum_{k\in S_t/\tilde{S}_t} \Delta\boldsymbol{\theta}_k^t + \sum_{k\in\tilde{S}_t} \Delta\tilde{\boldsymbol{\theta}}_k^t\Big)$$

Distributed Training Scenario

Threat Model and Challenges:

- The adversary does not know other clients' model updates
- The adversary does not have the other clients' data/data distribution
- The overall effect of malicious updates may be weakened since the malicious updates have different directions.

Distributed Training Scenario

$\boldsymbol{\theta}_{t-1}$

$\boldsymbol{\theta}_{t-1}$

Local Update

Fine-tune by maximization with a small learning rate

$$\max_{\boldsymbol{\theta}} \frac{1}{|\mathcal{D}|} \sum_{\mathbf{x},y \in \mathcal{D}} \|f_{\boldsymbol{\theta}}(\mathbf{x}) - y\|_2^2$$

Model Aggregation

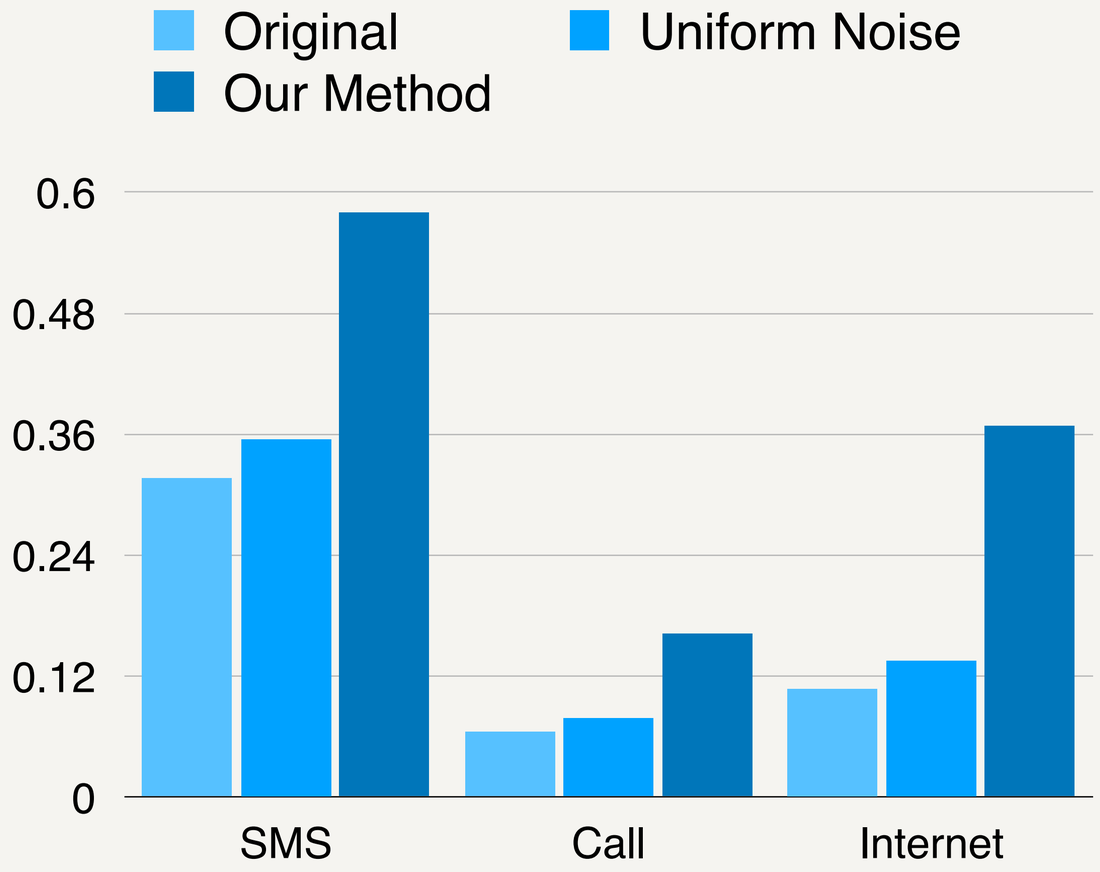
$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t + \frac{1}{|S_t|}\Big(\sum_{k \in S_t / \tilde{S}_t} \Delta\boldsymbol{\theta}_k^t + \sum_{k \in \tilde{S}_t} \Delta\tilde{\boldsymbol{\theta}}_k^t\Big)$$

Robust aggregation?

Experimental Results

	LSTM	Conv+LSTM	FedAvg	FedDA
SMS	0.3171	0.3081	0.3744	0.3411
After Attack	0.5802	0.5206	1.6e+09	1.6e+11
Call	0.0653	0.0639	0.0776	0.0742
After Attack	0.1624	0.1302	2.7e+09	9.2e+09
Internet	0.1083	0.1051	0.1096	0.1061
After Attack	0.3681	0.3133	3.5e+07	6.7e+10

The testing MSE (With/Without Attack) in centralized and decentralized scenarios

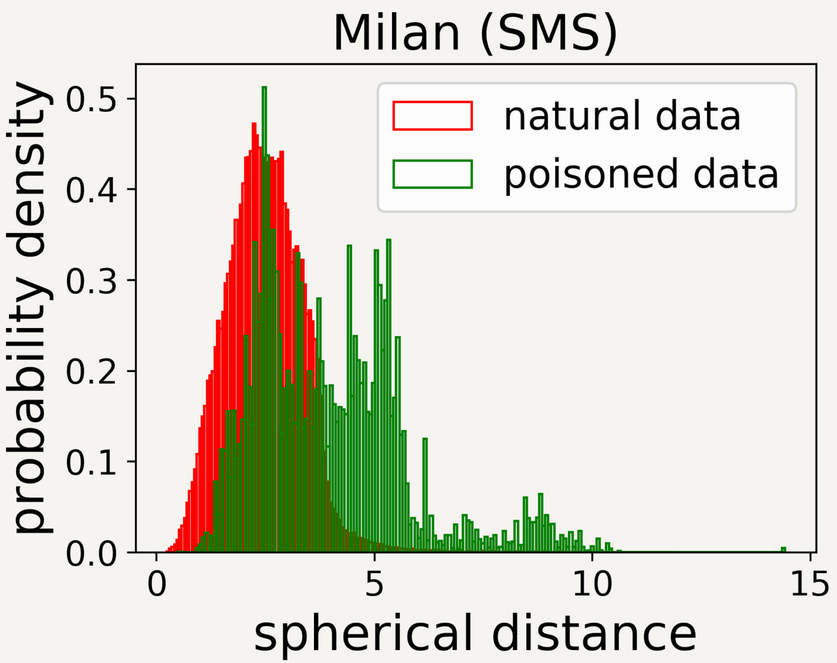


Potential Defenses

Data Sanitization: remove the outliers according to certain metric

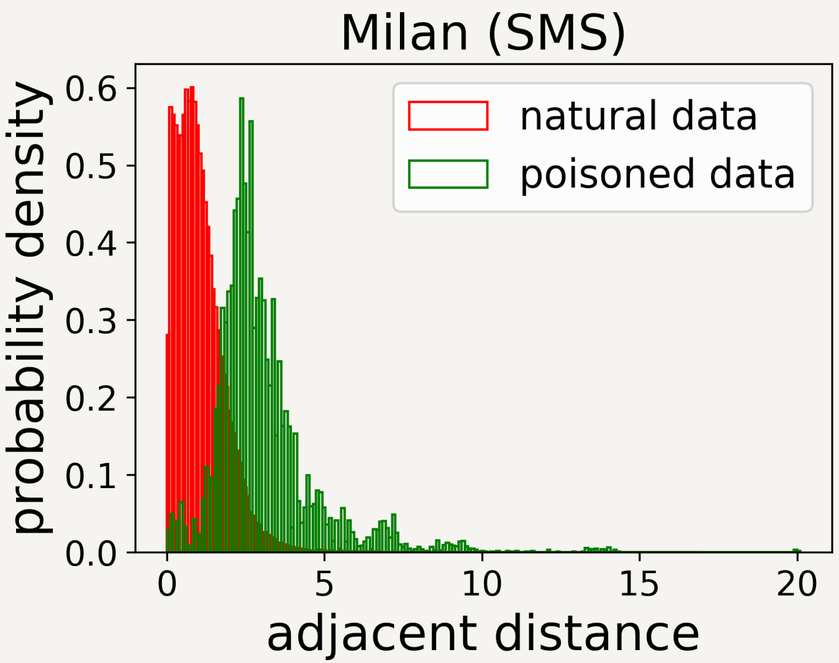
$$\sqrt{\| \mathbf{x} - \overline{\mathbf{x}} \|_2^2 + (y - \overline{y})^2}$$

Spherical distance



$$|y_i^k - \mathbf{x}_i^k[0]| + \sum_{j=0}^{J-1} |\mathbf{x}_i^k[j] - \mathbf{x}_i^k[j+1]|$$

Adjacent distance



Defenses against Model Poisoning

Byzantine-Robust Aggregation:

- Multi-Krum: Select the m model updates with relatively small distances from the other model updates
- Trimmed Median: Sort the values of every dimension of the model update and only use the median update value

	SMS	Call	Internet
FedAvg + MKrum	8.9e+05	14247	1.4e+08
FedAvg + TMean	3.5e+05	2.7e+06	8e+08

The probability that the 10 clients sampled by the server include at least 5 malicious clients in at least one round of the 100 training rounds is approximately over 0.9. Multi-drum will choose at least one malicious client (with m=6)

Defenses against Model Poisoning

Anomaly Detection (Dynamic)

- Criterion 1: The maximum L2 norm of all model updates should not be larger than $c_1 \mu_t$
- Criterion 2: The maximum L2 norm of all model updates should not be larger than $\mu_t + c_2 \sigma_t$ (robust estimations of mean and deviation)

	SMS	Call	Internet
FedAvg + AD	4.933e-01	0.3408	1.562e-01
FedAvg + AD + MKrum	4.342e-01	8.16e-02	1.158e-01

The adversary can not upload model updates with very large L2 norm. The adversary has to clip its model updates, which limits the negative effect of the poisoned model updates.

Takeaways

- DL based wireless traffic prediction is vulnerable to poisoning attacks in centralized and distributed scenario
- Data sanitization with adjacent distance metric can mitigate the negative effect of data poisoning
- Anomaly detection is essential in federated learning for wireless traffic prediction

th.zheng@mail.utoronto.ca