Data and Network Security

(Master Degree in Computer Science and Cybersecurity)

Lecture 4

Outline for today

- Recap last lecture
- Theory behind covert communication in FL
- Advanced persistent threats

Outline for today

- **Recap last lecture**
- Theory behind covert communication in FL
- Advanced persistent threats

Covert Channel

Indirect communication channel between unauthorized parties that violates some security policy by using **shared resources** in a way in which these resources are not initially designed.

Covert channel types

- Storage
- Timing

Storage based covert channel

Covert channels that exploit storage resources to conceal data, often utilizing file attributes or reserved storage space.

- Data hidden within file (such as steganography)
- Modifying header fields

Timing based covert channel

Covert channels that exploit variations in timing or delays within a standard communication channel to conceal data.

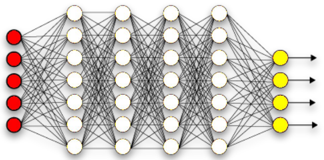
Modulating inter-packet delays

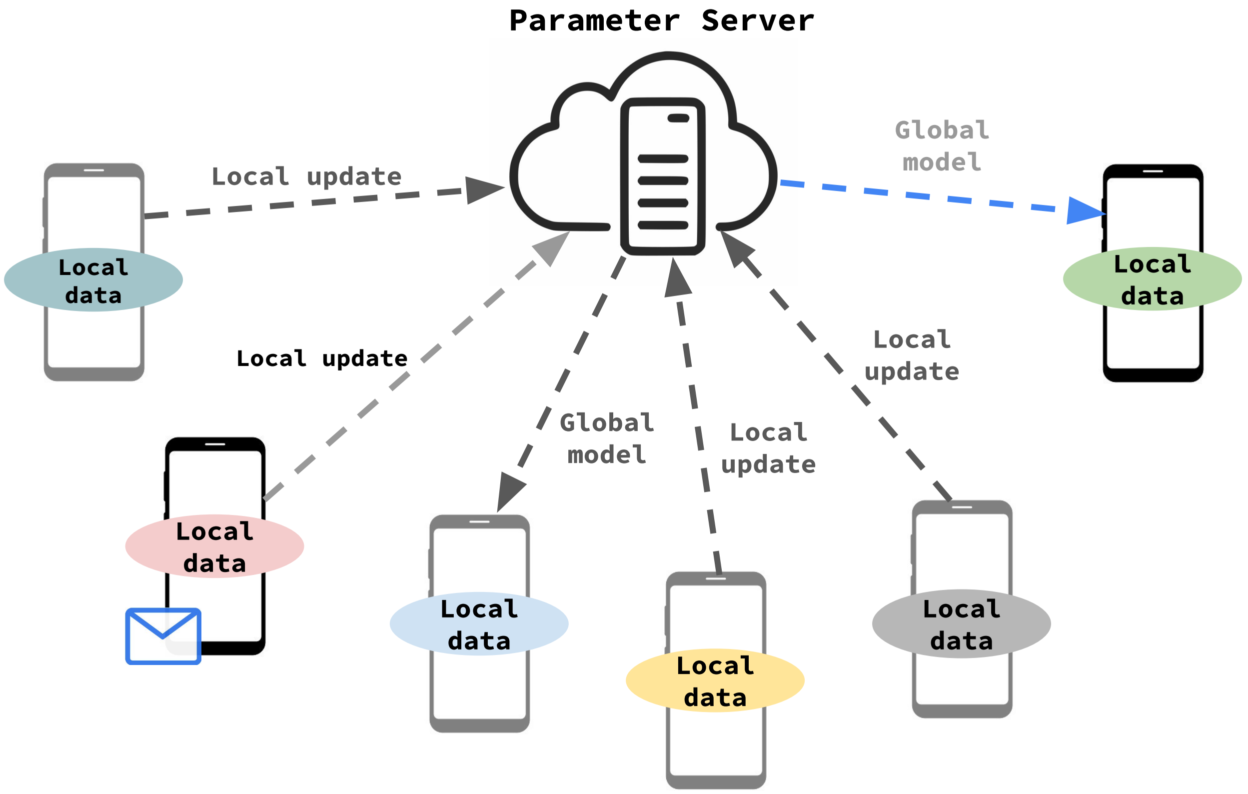
- Large delay - bit 1
- Small delay - bit 0

Covert Channels in Collaborative (federated) learning

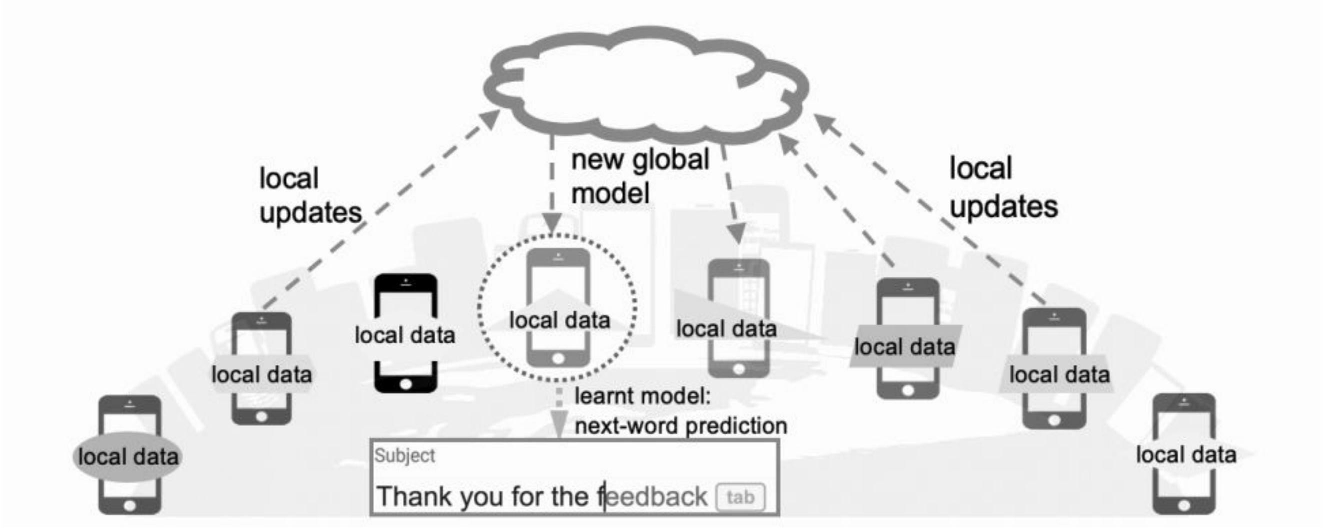
Collaborative Learning

Collaborative Learning





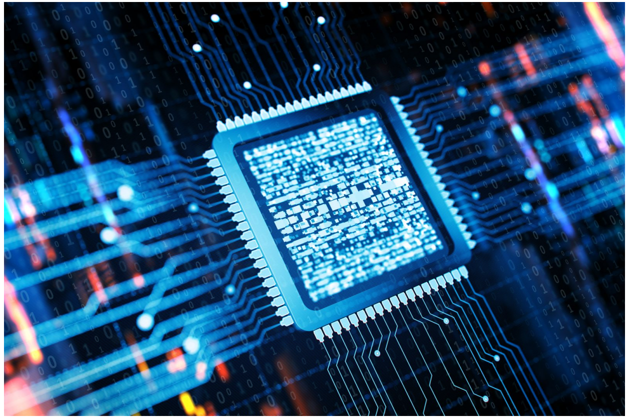
What is Federated Learning?



Federated learning (FL) (also known as **collaborative learning**) is a machine learning technique that trains an algorithm across multiple decentralized edge devices or servers holding local data samples, without exchanging them.

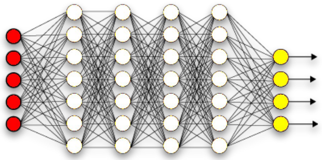
Why federated Learning?

- Data Privacy
- Low individual computing power
- Large collaborative computing power

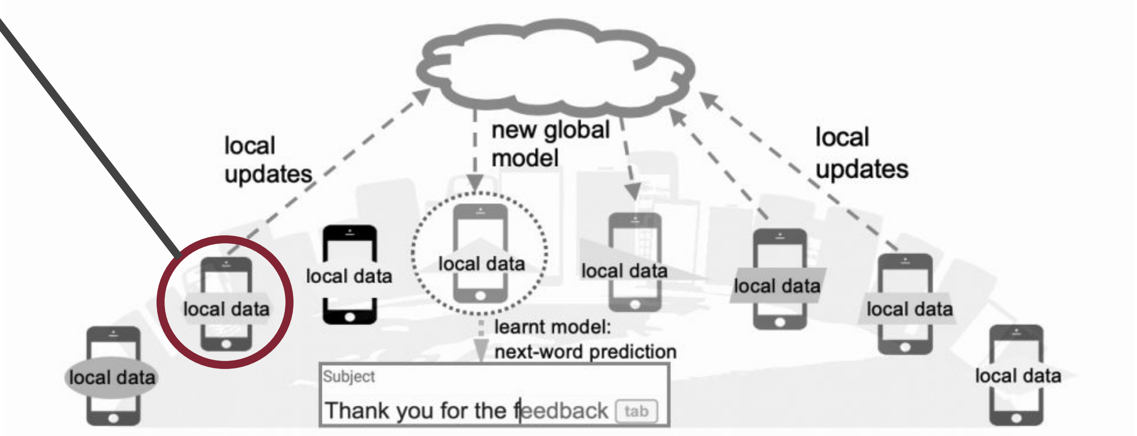


How does FL work? (local update)

$$\mathbf{W}_{t+1}^{k} = \mathbf{W}_t + \alpha \nabla \mathbf{W}_t^{k}$$

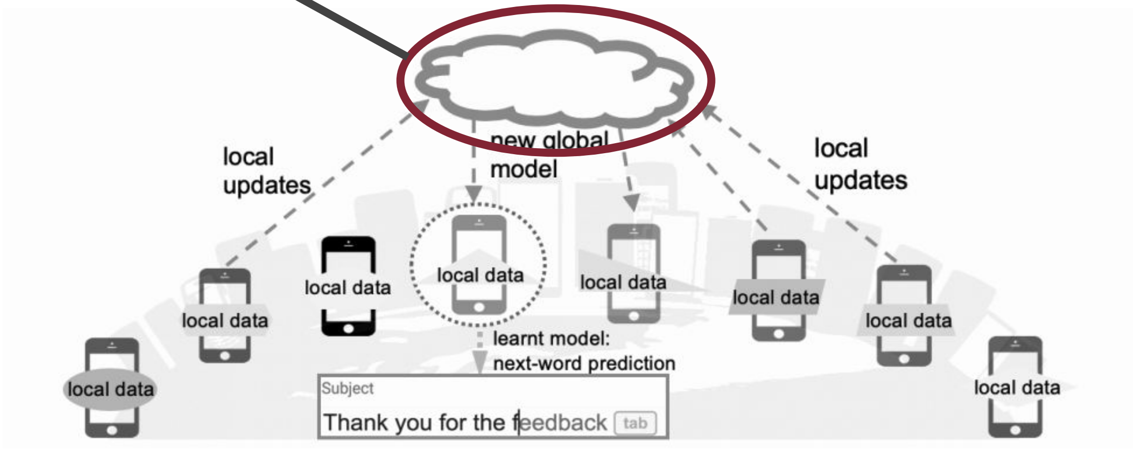


W



How does FL work? (global update)

$$\mathbf{W}_{t+1} = \mathbf{W}_t + \frac{\alpha}{n'} \sum_{k=1}^{n'} \nabla \mathbf{W}_t^k$$



CDMA - example

```
Binary sequence: [0, 1, 1]

PSK              [-1, 1, 1]

Spreading code: [-1, 1, -1, -1, 1]

Chip sequence:  [1, -1, 1, 1, -1, -1, 1, -1, -1, 1, -1, 1, -1, -1, 1]
```

−5 **+5** **+5**

HowTo (cont.)

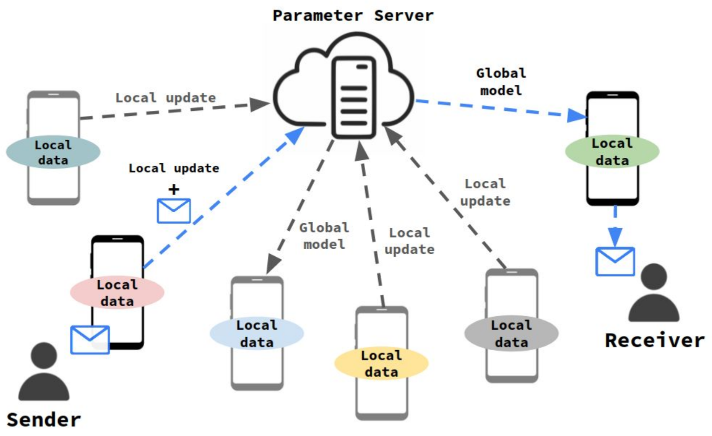
Payload -> P bits b = $[b_0, \ldots, b_{P-1}]$

- **C** is an **R** by **P** matrix that collects all the codes.

$$
\begin{array}{cc}
\mathbf{R} & \begin{array}{ccccc}
\mathbf{+1} & \mathbf{+1} & \mathbf{-1} & & \mathbf{+1} \\
\mathbf{-1} & \mathbf{-1} & \mathbf{-1} & & \mathbf{-1} \\
\mathbf{+1} & \mathbf{-1} & \mathbf{+1} & & \mathbf{+1} \\
\vdots & \vdots & \vdots & \cdots & \vdots \\
\mathbf{-1} & \mathbf{-1} & \mathbf{-1} & & \mathbf{+1}
\end{array} \\
 & \mathbf{P}
\end{array}
$$

“message” embedding

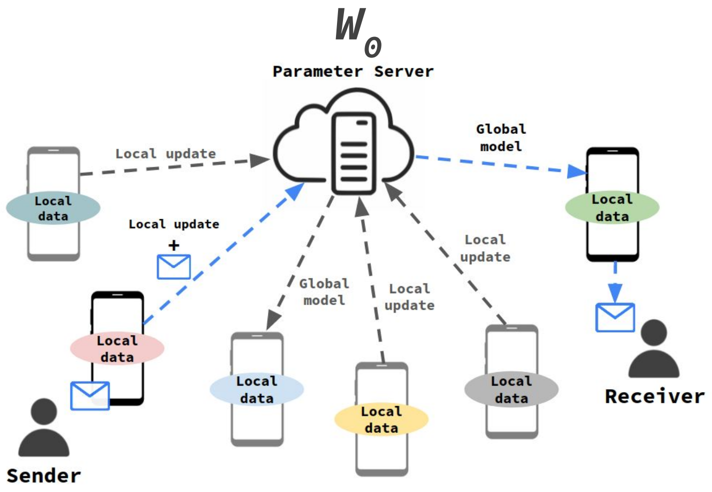
- **n** participants
- Parameter server proposes a set of weights W_0
- At each iteration, the participants use their local data to compute the gradient ∇W_t^k k = 0, …, n-1; t = 0, . . . T − 1



“message” embedding (cont.)

$$\widehat{\nabla \mathbf{W}_t^0} = \beta \nabla \mathbf{W}_t^0 + \gamma \mathbf{C}\mathbf{b}$$

The gradient update of the sender



$$\widehat{\nabla \mathbf{W}_t^0} = \beta \nabla \mathbf{W}_t^0 + \gamma \mathbf{C}\mathbf{b}$$

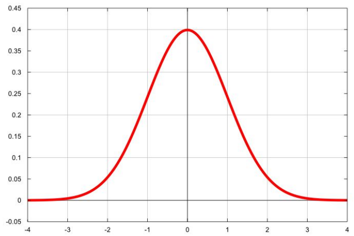
γ and **β** are two gain factors to ensure that the message cannot be detected and that the power of the modified gradient is like the unmodified gradient for the other users.

“message” embedding (cont.)

How do we choose **γ** and **β** ?

$$\widehat{\nabla \mathbf{W}_t^0} = \beta \nabla \mathbf{W}_t^0 + \gamma \mathbf{C} \mathbf{b}$$

“message” embedding (cont.)



How do we choose **γ** and **β** ?

$$\widehat{\nabla \mathbf{W}_t^0} = \beta \nabla \mathbf{W}_t^0 + \gamma \mathbf{C}\mathbf{b}$$

β=0 and **γ=σ/√P**

- Our gradient would have the same power as the original.
- A hypothesis testing looking for a binomial or a Gaussian distribution will be able to detect that our gradient is not a true gradient.

"message" embedding (cont.)

How do we choose **γ** and **β** ?

$$\widehat{\nabla \mathbf{W}_t^0} = \beta \nabla \mathbf{W}_t^0 + \gamma \mathbf{C}\mathbf{b}$$

β=1 and **γ=0.1σ/√P**

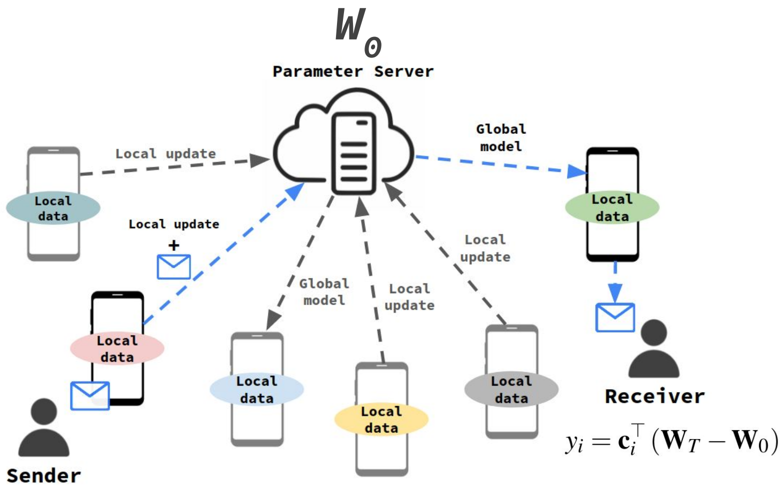
- Our gradient will have the same distribution as the original gradient.
- The signal will be undetectable.

“message” extraction

To recover bit i of the payload:

$$y_i = \mathbf{c}_i^\top (\mathbf{W}_T - \mathbf{W}_0)$$

1. Bit i signal.
2. Noise from the gradients.
3. Noise from the other bits of the payload.



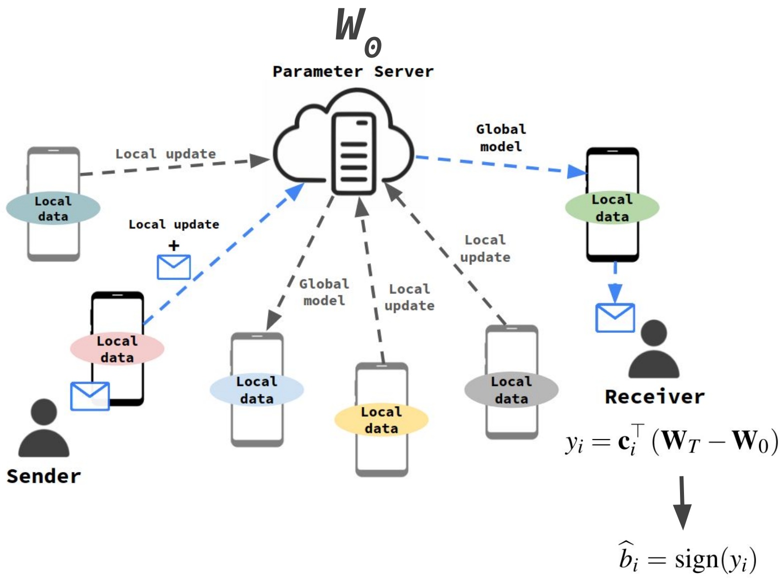
“message” extraction

To recover bit i of the payload:

$$y_i = \mathbf{c}_i^\top (\mathbf{W}_T - \mathbf{W}_0)$$

$$\downarrow$$

$$\widehat{b}_i = \text{sign}(y_i)$$



How we recover b_i ?

How we recover b_i ?

$$y_i = \mathbf{c}_i^\top (\mathbf{W}_T - \mathbf{W}_0)$$

Assume $\nabla \mathbf{W}_t^k$ is a zero-mean with a variance $\boldsymbol{\sigma}^2$

How we recover b_i ?

$$
\begin{aligned}
y_i =& \mathbf{c}_i^\top (\mathbf{W}_T - \mathbf{W}_0) \\
=& \frac{\alpha}{n}\mathbf{c}_i^\top \left(\sum_{t=0}^{T-1} \left(\widehat{\nabla \mathbf{W}_t^0} + \sum_{k=1}^{n-1} \nabla \mathbf{W}_t^k \right) \right) \\
=& \frac{\alpha}{n} \left(T\gamma \mathbf{c}_i^\top \mathbf{C}\mathbf{b} + \mathbf{c}_i^\top \sum_{t=0}^{T-1} \left(\beta \nabla \mathbf{W}_0^k + \sum_{k=0}^{n-1} \nabla \mathbf{W}_t^k \right) \right) \\
=& \frac{\alpha}{n} \left(T\gamma \mathbf{c}_i^\top \mathbf{c}_i b_i + T\gamma \mathbf{c}_i^\top \mathbf{C}_{\neg i} \mathbf{b}_{\neg i} + \mathbf{c}_i^\top \widetilde{\mathbf{w}} \right) \\
=& \frac{\alpha}{n} \left(T\gamma \mathbf{c}_i^\top \mathbf{c}_i b_i + T\gamma \mathbf{c}_i^\top \widetilde{\mathbf{c}} + \mathbf{c}_i^\top \widetilde{\mathbf{w}} \right) \\
=& \frac{\alpha}{n} \left(T\gamma R b + \varepsilon_i^{\mathbf{C}} + \varepsilon_i^{\mathbf{W}} \right) = \frac{\alpha}{n} (T\gamma R b + \varepsilon_i)
\end{aligned}
$$

How we recover b_i ?

$$
\begin{aligned}
y_i =& \mathbf{c}_i^\top (\mathbf{W}_T - \mathbf{W}_0) \\
=& \frac{\alpha}{n} \mathbf{c}_i^\top \left(\sum_{t=0}^{T-1} \left(\widehat{\nabla \mathbf{W}_t^0} + \sum_{k=1}^{n-1} \nabla \mathbf{W}_t^k \right) \right) \\
=& \frac{\alpha}{n} \left(T\gamma \mathbf{c}_i^\top \mathbf{C}\mathbf{b} + \mathbf{c}_i^\top \sum_{t=0}^{T-1} \left(\beta \nabla \mathbf{W}_0^k + \sum_{k=0}^{n-1} \nabla \mathbf{W}_t^k \right) \right) \\
=& \frac{\alpha}{n} \left(T\gamma \mathbf{c}_i^\top \mathbf{c}_i b_i + T\gamma \mathbf{c}_i^\top \mathbf{C}_{\neg i} \mathbf{b}_{\neg i} + \mathbf{c}_i^\top \widetilde{\mathbf{w}} \right) \\
=& \frac{\alpha}{n} \left(T\gamma \mathbf{c}_i^\top \mathbf{c}_i b_i + T\gamma \mathbf{c}_i^\top \widetilde{\mathbf{c}} + \mathbf{c}_i^\top \widetilde{\mathbf{w}} \right) \\
=& \frac{\alpha}{n} \left(T\gamma R b + \varepsilon_i^{\mathbf{C}} + \varepsilon_i^{\mathbf{W}} \right) = \frac{\alpha}{n} \left(T\gamma R b + \varepsilon_i \right)
\end{aligned}
$$

- Each component ~c is a symmetric binomial distribution between **±(P - 1).**
- Multiplying it by c_i we get a binomial distribution with values between **±R(P-1).**
- For large **R** can be approximated by a zero-mean Gaussian with variance $\mathbf{T^2\gamma^2R(P-1)}$

How we recover b_i ?

$$
\begin{aligned}
y_i =& \mathbf{c}_i^\top (\mathbf{W}_T - \mathbf{W}_0) \\
=& \frac{\alpha}{n} \mathbf{c}_i^\top \left(\sum_{t=0}^{T-1} \left(\widehat{\nabla \mathbf{W}_t^0} + \sum_{k=1}^{n-1} \nabla \mathbf{W}_t^k \right) \right) \\
=& \frac{\alpha}{n} \left(T\gamma \mathbf{c}_i^\top \mathbf{C}\mathbf{b} + \mathbf{c}_i^\top \sum_{t=0}^{T-1} \left(\beta \nabla \mathbf{W}_0^k + \sum_{k=0}^{n-1} \nabla \mathbf{W}_t^k \right) \right) \\
=& \frac{\alpha}{n} \left(T\gamma \mathbf{c}_i^\top \mathbf{c}_i b_i + T\gamma \mathbf{c}_i^\top \mathbf{C}_{\neg i} \mathbf{b}_{\neg i} + \mathbf{c}_i^\top \widetilde{\mathbf{w}} \right) \\
=& \frac{\alpha}{n} \left(T\gamma \mathbf{c}_i^\top \mathbf{c}_i b_i + T\gamma \mathbf{c}_i^\top \widetilde{\mathbf{c}} + \mathbf{c}_i^\top \widetilde{\mathbf{w}} \right) \\
=& \frac{\alpha}{n} \left(T\gamma R b + \varepsilon_i^{\mathbf{C}} + \varepsilon_i^{\mathbf{W}} \right) = \frac{\alpha}{n} \left(T\gamma R b + \varepsilon_i \right)
\end{aligned}
$$

- Each component of $\widetilde{\mathbf{w}}$ adds up **T(n-1+β)** of these values
- By CLT (large enough **T(n-1+β)**) each one of these variables would be zero-mean Gaussian with a variance **$Tn\sigma^2$**. (**β=1**)
- When we multiply this vector by $\mathbf{c}_i$ and add all the components together, we end up with a zero-mean Gaussian with a variance **$RTn\sigma^2$**, because the components of $\mathbf{c}_i$ are ±1.

How we recover b_i?

How we recover b_i?

$$
\begin{aligned}
y_i = \mathbf{c}_i^\top (\mathbf{W}_T - \mathbf{W}_0) &= \frac{\alpha}{n}\left(T\gamma\mathbf{c}_i^\top \mathbf{c}_i b_i + T\gamma\mathbf{c}_i^\top \mathbf{C}_{\neg i}\mathbf{b}_{\neg i} + \mathbf{c}_i^\top \widetilde{\mathbf{w}}\right) \\
&= \frac{\alpha}{n}\left(T\gamma\mathbf{c}_i^\top \mathbf{c}_i b_i + T\gamma\mathbf{c}_i^\top \widetilde{\mathbf{c}} + \mathbf{c}_i^\top \widetilde{\mathbf{w}}\right) \\
&= \frac{\alpha}{n}\left(T\gamma R b + \varepsilon_i^{\mathbf{C}} + \varepsilon_i^{\mathbf{W}}\right) = \frac{\alpha}{n}\left(T\gamma R b + \varepsilon_i\right)
\end{aligned}
$$

- **C**, **b** and $\nabla \mathbf{W}_t^k$ are mutually independent, thus $\boldsymbol{\varepsilon}_i$ is zero mean with a variance that it is the sum of the variances of $\boldsymbol{\varepsilon}^C_i$ and $\boldsymbol{\varepsilon}^W_i$ and also Gaussian distributed.
- The distribution of $\mathbf{y}_i$ is given by $\mathbf{y}_i \sim \mathcal{N}$ **($T\gamma R b_i$, $T^2\gamma^2 R(P-1) + RTn\sigma^2$)**, and if normalize by **$T\gamma R$:**

$$
y_i \sim \mathcal{N}\left(b_i, \frac{T^2\gamma^2 R(P-1) + TRn\sigma^2}{T^2\gamma^2 R^2}\right)
$$

$$
\mathcal{N}\left(b_i, \frac{P-1}{R} + \frac{n\sigma^2}{TR\gamma^2}\right)
$$

How we recover b_i?

$$y_i \sim \mathcal{N}\left(b_i, \frac{T^2\gamma^2 R(P-1) + TRn\sigma^2}{T^2\gamma^2 R^2}\right)$$

$$\mathcal{N}\left(b_i, \frac{P-1}{R} + \frac{n\sigma^2}{TR\gamma^2}\right)$$

β=1 and **γ=0.1σ/√P**

$$\frac{P-1}{R} + \frac{n\sigma^2}{TR\gamma^2} = \frac{P-1}{R} + \frac{n\sigma^2}{TR\left(\frac{0.1\sigma}{\sqrt{P}}\right)^2}$$

$$= \frac{P-1}{R} + \frac{100nP\sigma^2}{TR} \approx \frac{(T+100n)P}{TR}$$

- At least **T>100nP/(R-P)** rounds before the message can be decoded.

How we recover b_i faster?

- We incorporate multiple senders

How we recover b_i faster?

- We incorporate multiple senders

    - The y_i distribution will become:

$$\mathcal{N}(MT\gamma R b_i, M^2T^2\gamma^2 R(P-1) + RTn\sigma^2)$$

How we recover b_i faster?

- We incorporate multiple senders

  - The y_i distribution will become:

$$\mathcal{N}(MT\gamma Rb_i, M^2T^2\gamma^2R(P-1)+RTn\sigma^2)$$

T>nP/M²(R-P) M^2 times faster

The pillars of evaluation

Stealthiness of the communication

Impact on model performance

“Message” delivery time

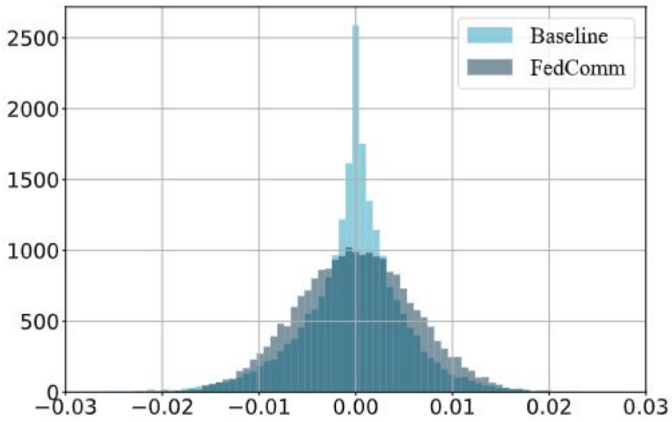
The pillars of evaluation

Stealthiness of the communication

Impact on model performance

“Message” delivery time

Stealthiness of communication



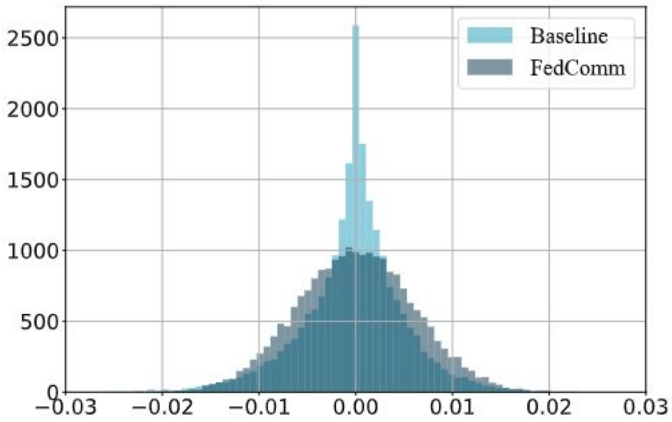
Regular gradient update

vs.

Non-Stealthy gradient update

In **non-stealthy** we are not sending a gradient that is useful for learning, instead we send our signal with the same power as our gradient would have.

Stealthiness of communication



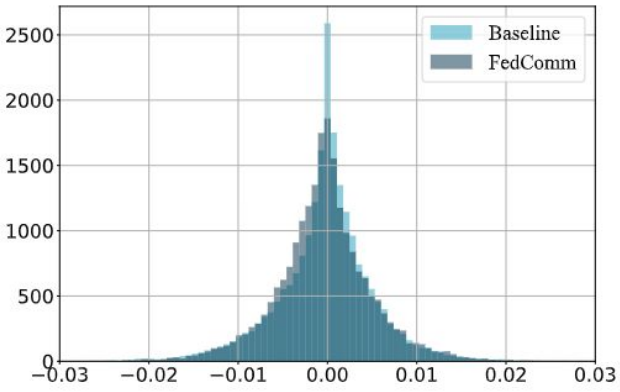
Regular gradient update

vs.

Non-Stealthy gradient update

NON-STEALTHY **might** be detectable in cases where the global parameter server can observe individual gradient updates.

Stealthiness of communication



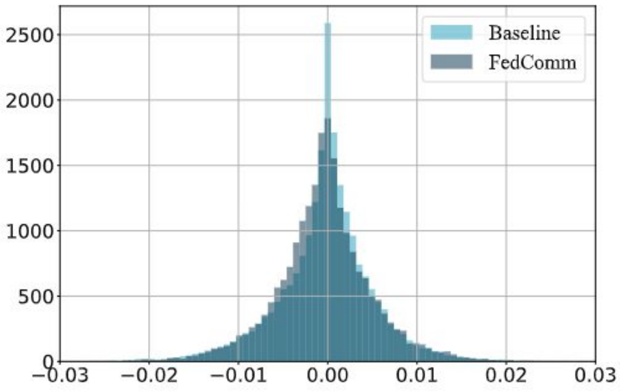
Regular gradient update

vs.

Full-Stealthy gradient update

In **full-stealthy** we are sending a gradient that is useful for learning, and buried into it we put also our signal.

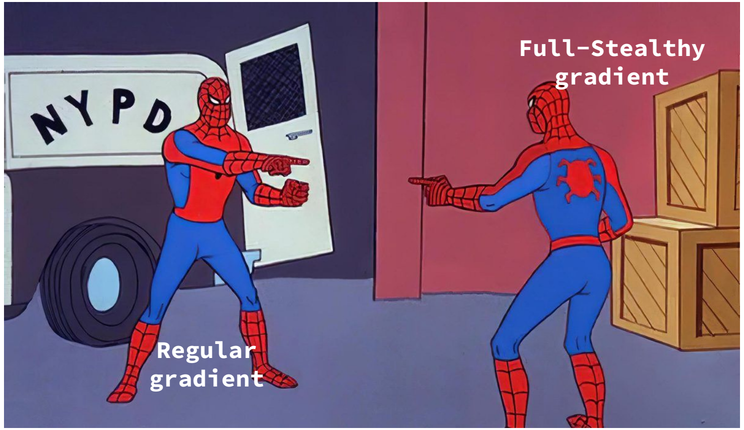
Stealthiness of communication



Regular gradient update

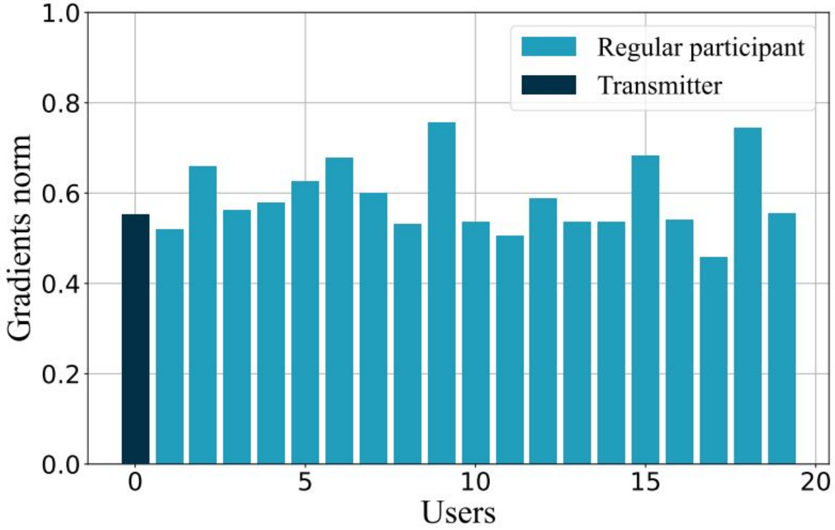
vs.

Full-Stealthy gradient update



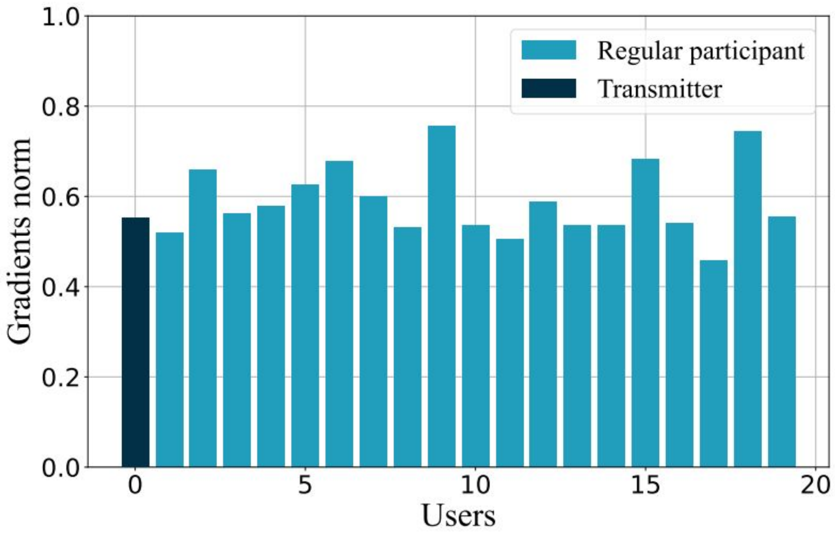
Gradients are statistically indistinguishable

Stealthiness of communication (cont.)



Parameter server observes individual gradient updates

Stealthiness of communication (cont.)



Parameter server observes individual gradient updates

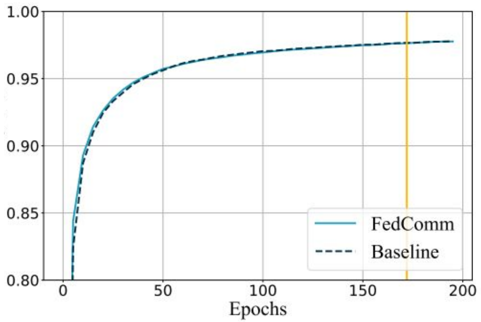
The pillars of evaluation

Stealthiness of the communication

Impact on model performance

“Message” delivery time

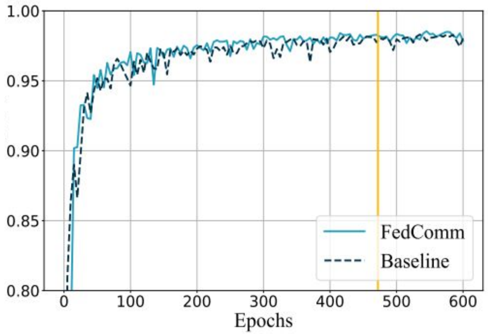
Impact on the model performance



100 participants

10 senders FULL Stealthy

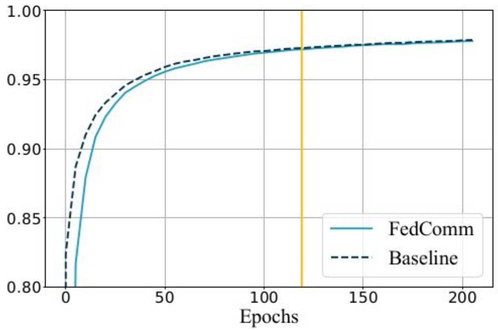
100% aggregated per round



100 participants

10 senders FULL Stealthy

20% aggregated per round



100 participants

1 sender NON Stealthy

100% aggregated per round

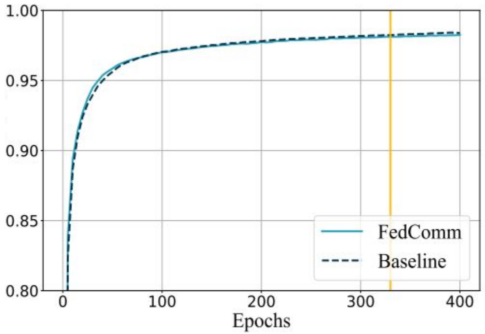
The pillars of evaluation

Stealthiness of the communication

Impact on model performance

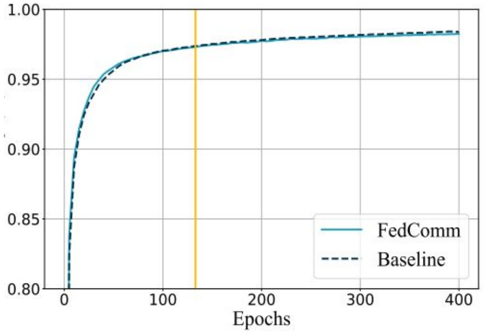
“Message” delivery time

“message” delivery time



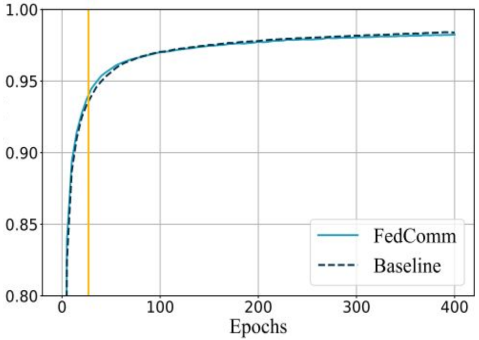
1 sender

100% aggregated per round



2 senders

100% aggregated per round

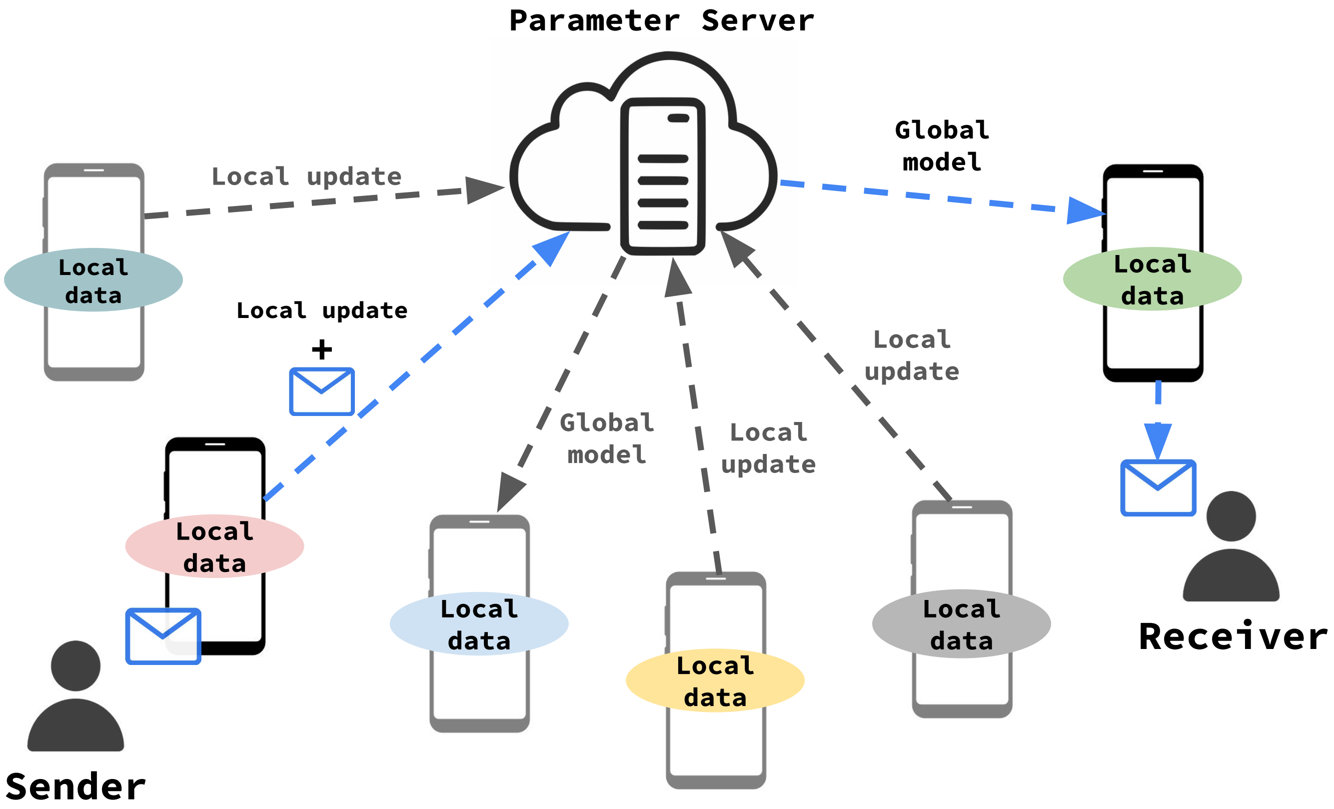


4 senders

100% aggregated per round

- Everything else held equal.

Remarks



Parameter Server
Local update
Local data
Local update
+
Local data
Sender
Global model
Local data
Local update
Local data
Local update
Local data
Global model
Local data
Receiver

Outline for today

- Recap last lecture
- Theory behind covert communication in FL
- **Advanced persistent threats**

Advanced Persistent Threats

Sophisticated, targeted cyberattack in which an unauthorized entity gains access to a network and remains undetected for an extended period.

Advanced Persistent Threats

Sophisticated, targeted cyberattack in which an unauthorized entity gains access to a network and remains undetected for an extended period.

- APT attacks are characterized by:
  - advanced tactics,
  - stealthy infiltration methods,
  - persistent presence within the targeted network.

APTs vs. Common attacks

Opportunistic (common) attacks:

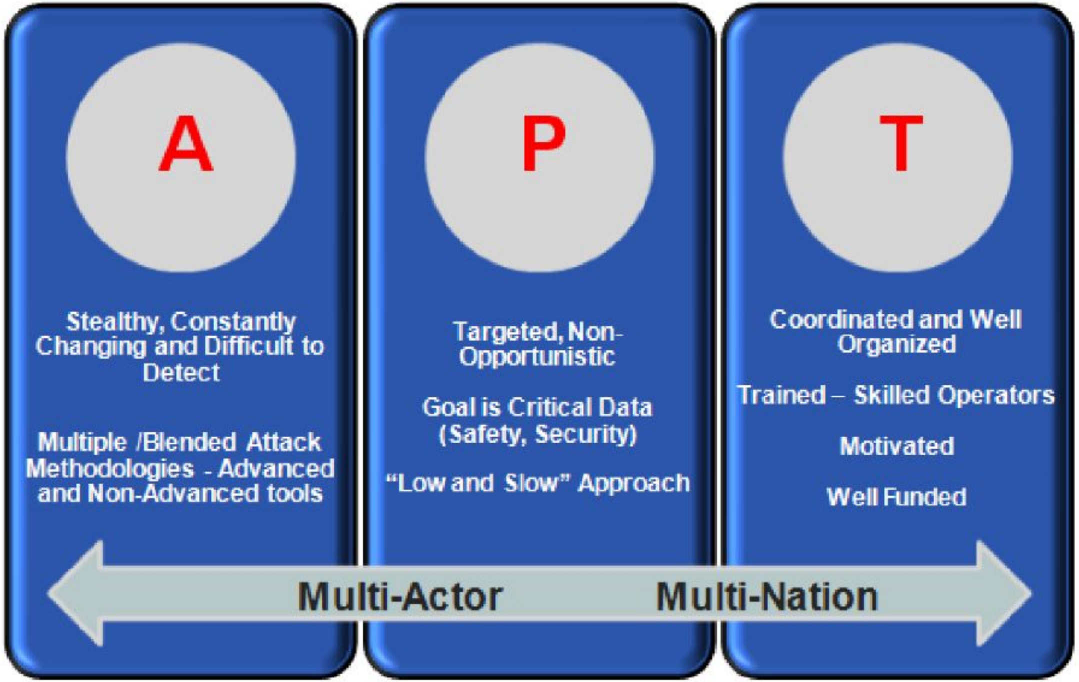
- short-lived
- indiscriminate

APTs vs. Common attacks

APT attacks:

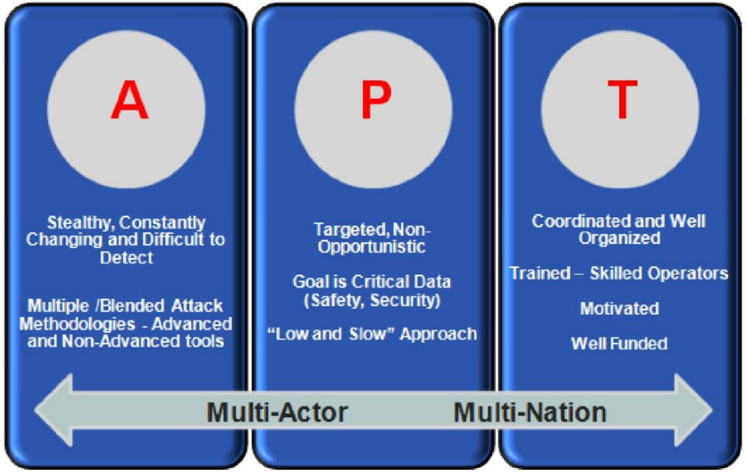
- Carefully planned,
- Well-funded,
- Tailored to target high-value assets, such as sensitive data, intellectual property, or strategic information.

APT


A
Stealthy, Constantly Changing and Difficult to Detect
Multiple /Blended Attack Methodologies - Advanced and Non-Advanced tools
P
Targeted, Non-Opportunistic
Goal is Critical Data (Safety, Security)
"Low and Slow" Approach
T
Coordinated and Well Organized
Trained – Skilled Operators
Motivated
Well Funded
Multi-Actor
Multi-Nation

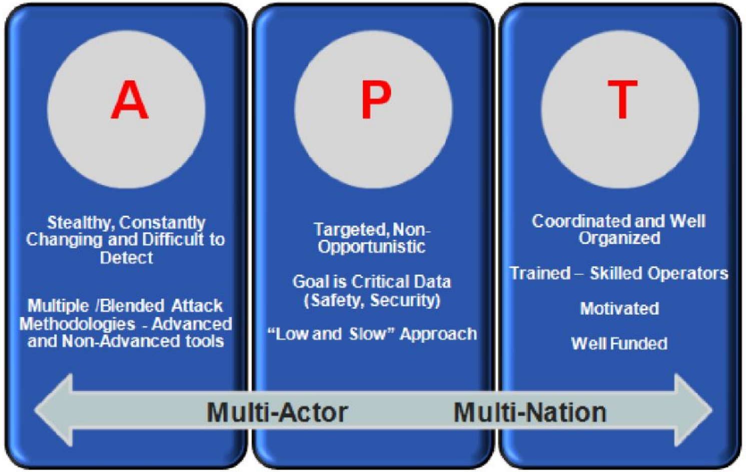
ADVANCED

The attack team has significant levels of expertise and significant resources, allowing the use of multiple and elaborated different attack vectors.



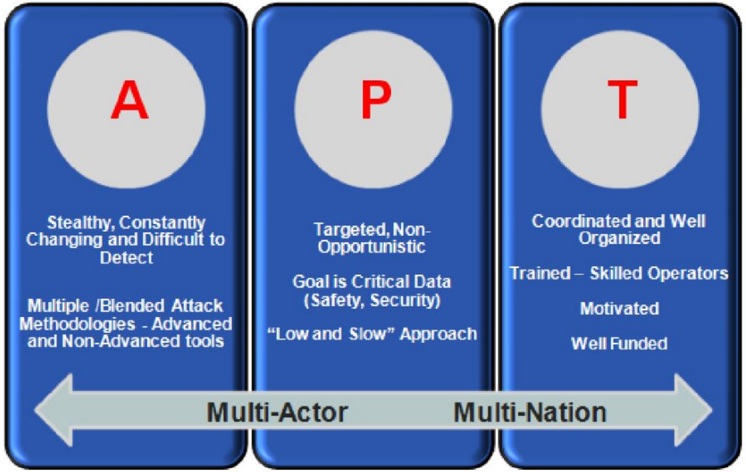
PERSISTENT

The attack team operates in order to remain present and undetected within the organization as long as possible

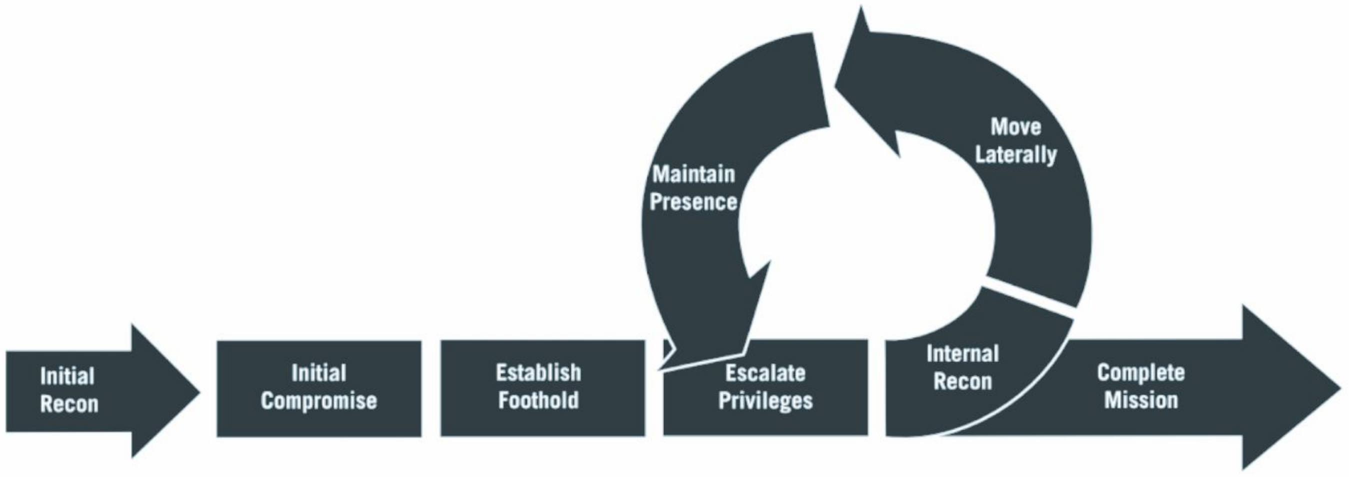


THREAT

Potential to adversely impact organizational operations, their assets, or individuals.

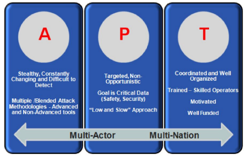


Life Cycle



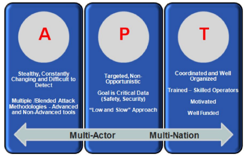
Initial Recon
Initial Compromise
Establish Foothold
Escalate Privileges
Internal Recon
Move Laterally
Maintain Presence
Complete Mission

Why APTs?



- Economic espionage
- Political espionage
- Ideological motivations

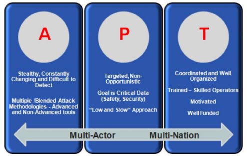
Why APTs?



- Economic espionage

  Seek to steal valuable intellectual property, trade secrets, or proprietary information from targeted organizations.

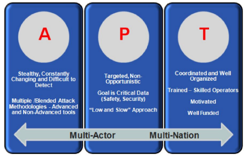
Why APTs?



- Political espionage

  Nation-state actors may target government agencies, diplomatic organizations, political parties, or foreign entities to gain insights into:
  - geopolitical developments,
  - national security strategies,
  - diplomatic matters (e.g negotiations).
  - …

Why APTs?



- Ideological motivations

  Groups or individuals with specific ideological agendas may target organizations or entities that they perceive as adversaries or opponents to advance their ideological goals or raise awareness about social or political issues.

Reading Material

1. Covert Channels (Concepts and definitions): Link
2. Covert communication in collaborative learning: Link-1, Link-2 (research papers).
3. Advanced Persistent Threats: Link-1, Link-2