A Triad of Defenses to Mitigate Poisoning Attacks in Federated Learning

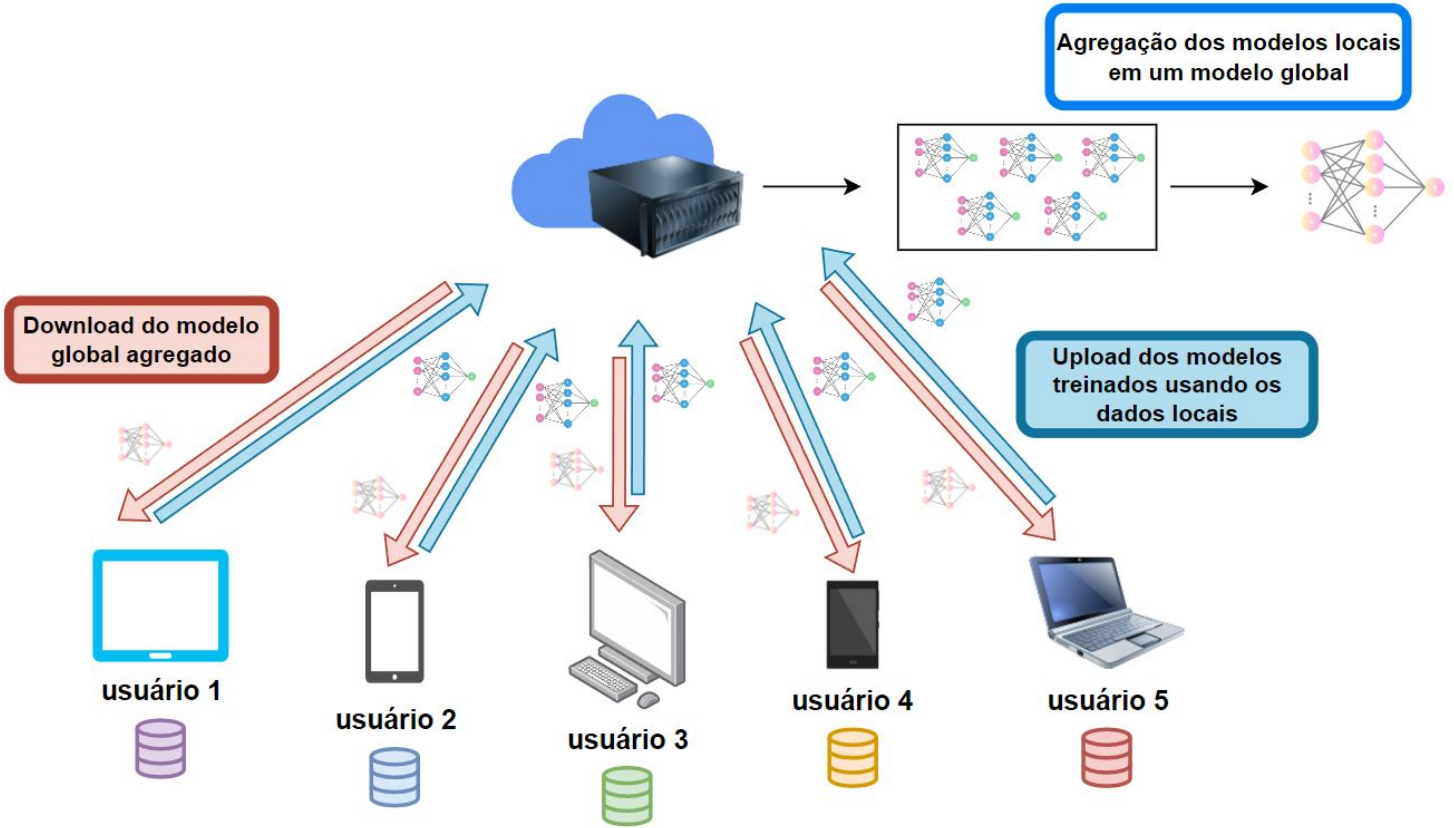
Blenda Oliveira Mazetto
Bruno Bogaz Zarpelão

Universidade Estadual de Londrina

Motivação

- O Aprendizado Federado permite que os participantes aprendam colaborativamente um modelo de aprendizado compartilhado, mantendo todos os dados no dispositivo.

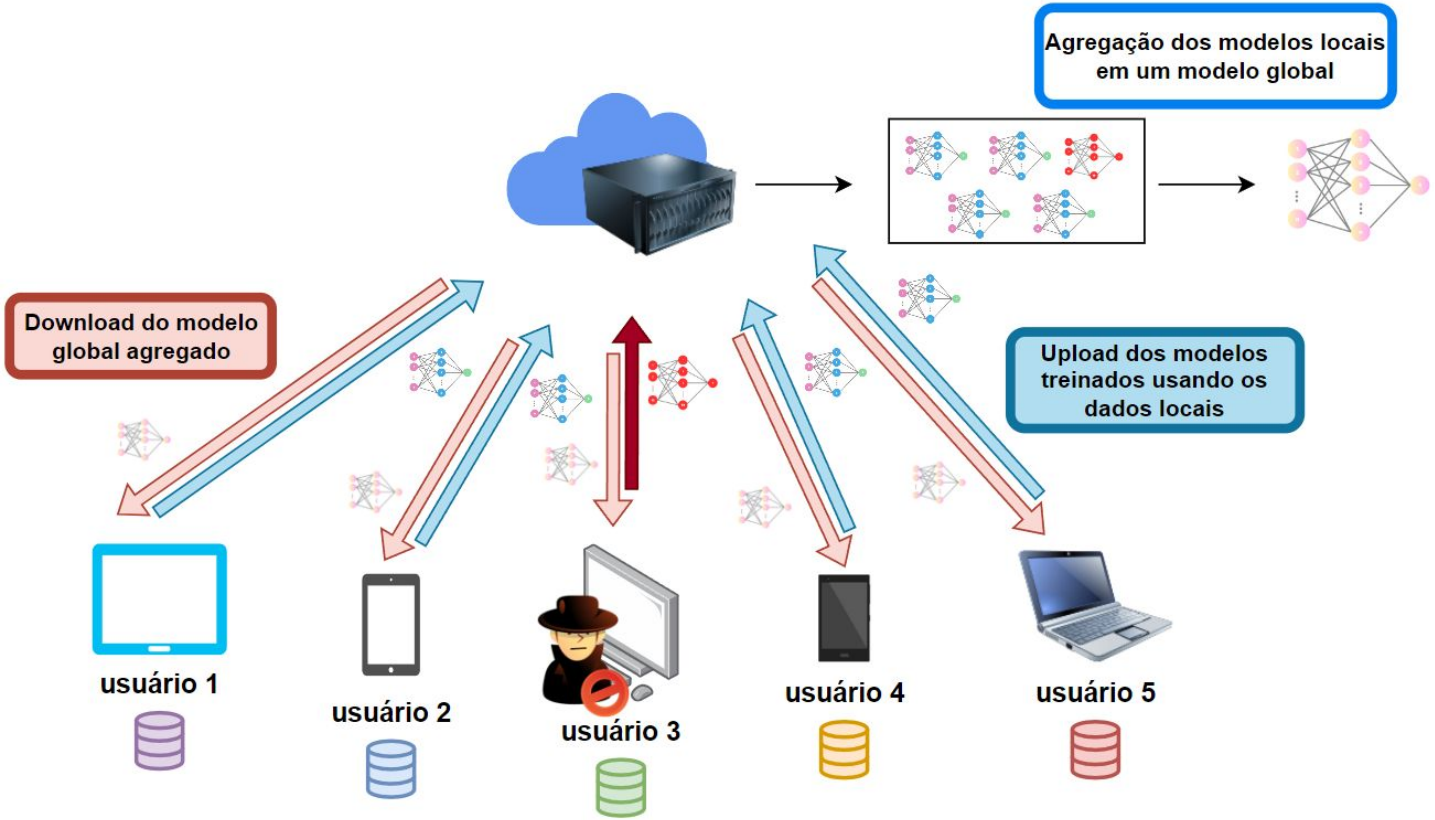
Motivação


Agregação dos modelos locais em um modelo global
Download do modelo global agregado
Upload dos modelos treinados usando os dados locais
usuário 1
usuário 2
usuário 3
usuário 4
usuário 5

Motivação

- Apesar de seus benefícios, o paradigma do Aprendizado Federado é vulnerável a ataques de envenenamento. Mais especificamente, clientes maliciosos podem se infiltrar no esquema e corromper o modelo global.

Motivação

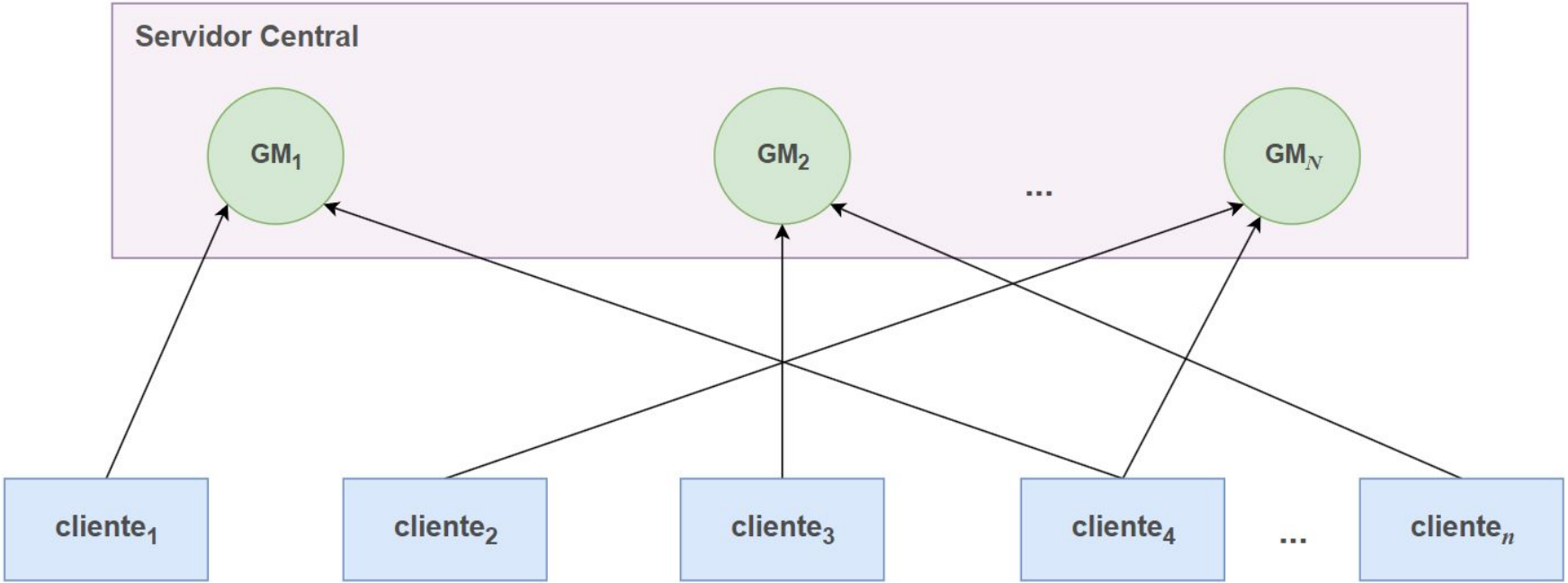

Agregação dos modelos locais em um modelo global
Download do modelo global agregado
Upload dos modelos treinados usando os dados locais
usuário 1
usuário 2
usuário 3
usuário 4
usuário 5

Objetivo

- Utilizar uma combinação de três técnicas de defesa para mitigar ataques de envenenamento no Aprendizado Federado:
  - Dividir os clientes em grupos;
  - Verificar o desempenho do modelo global;
  - Fazer inferências com base em um esquema de votação.

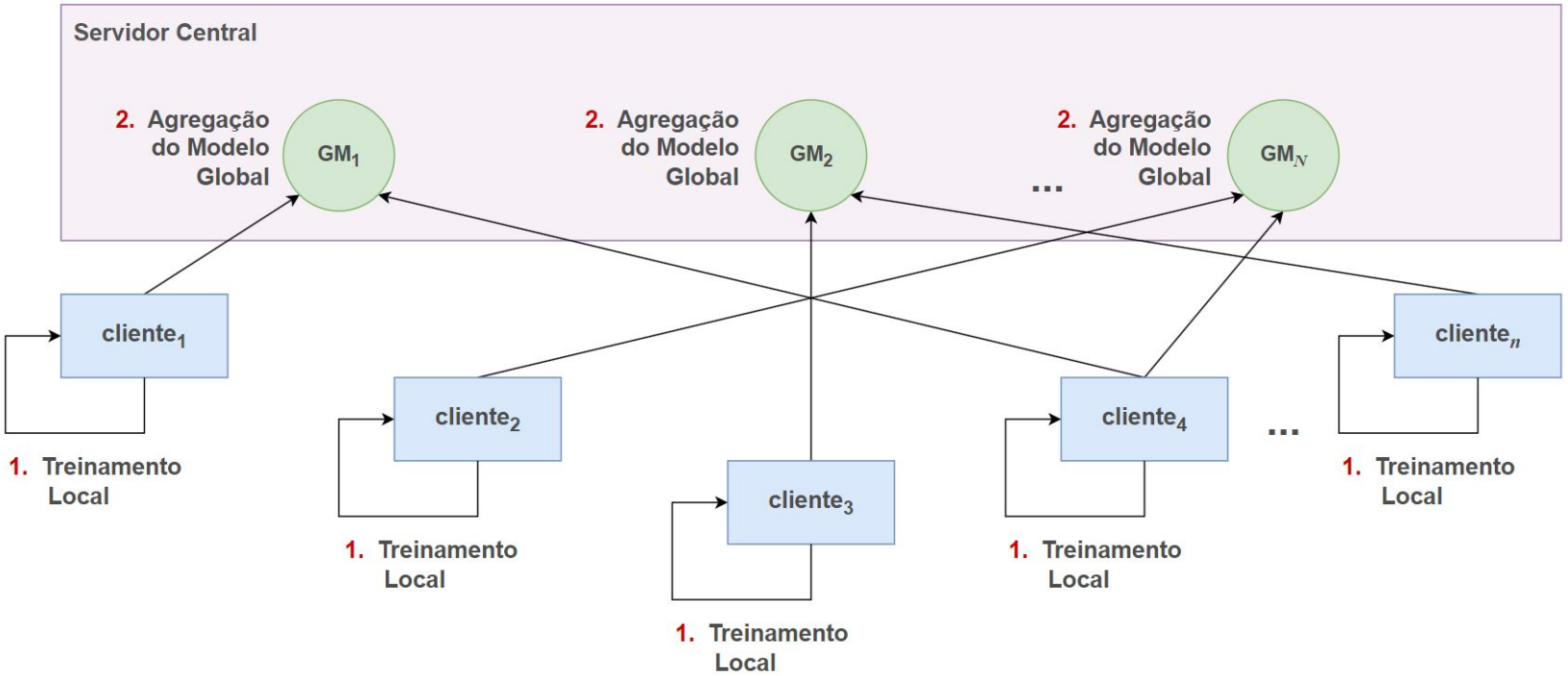
Abordagem Proposta

Divisão de grupos



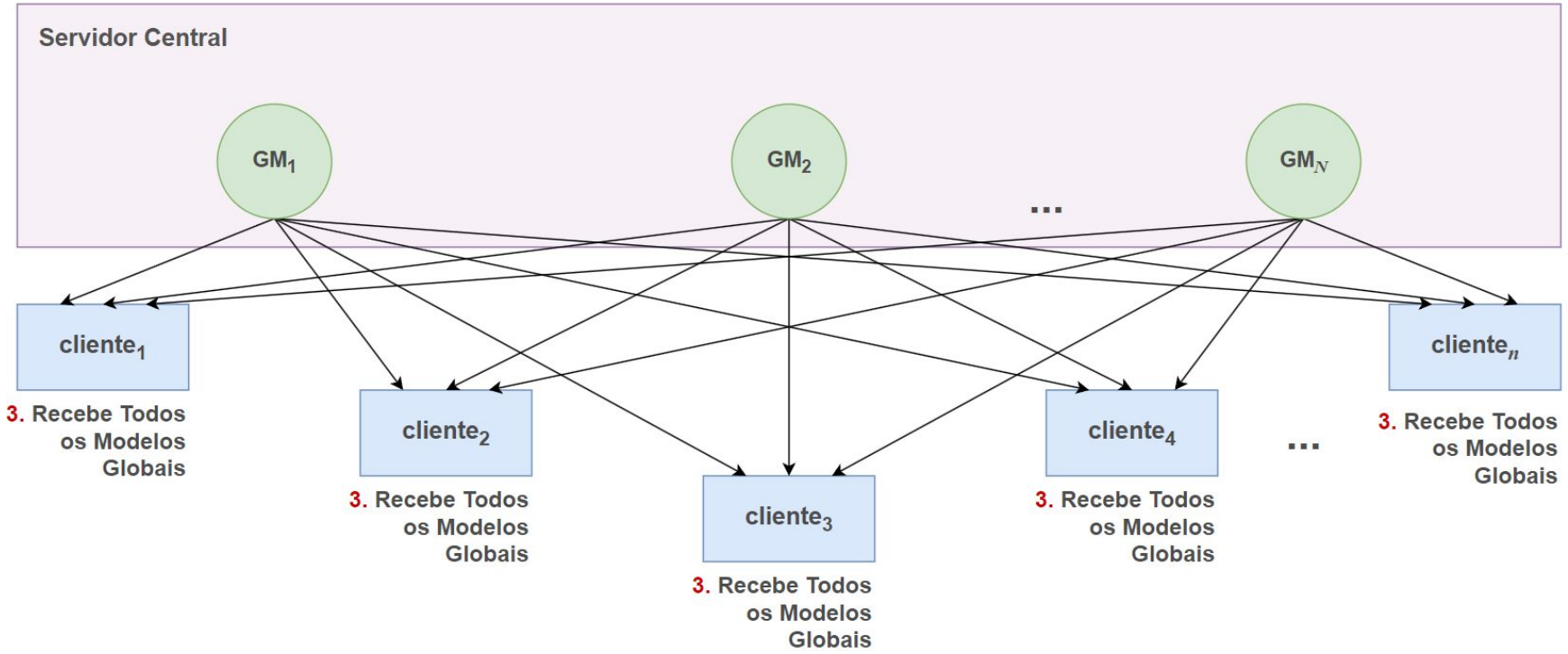
Abordagem Proposta

Treinamento local e agregação dos grupos



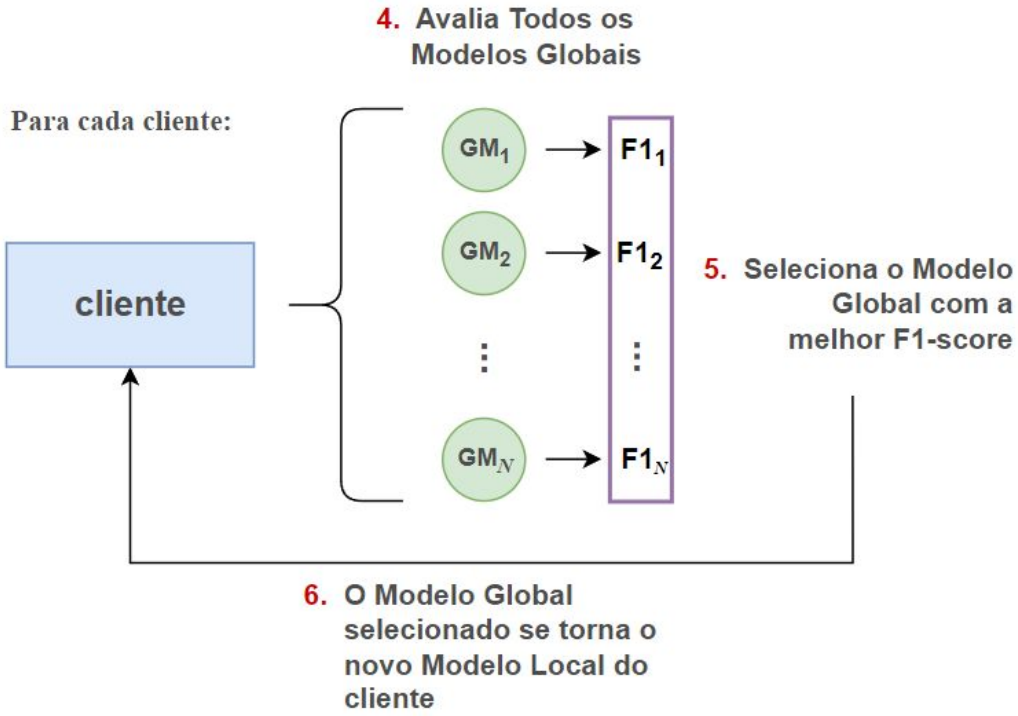
Abordagem Proposta

Clientes recebem todos os modelos globais



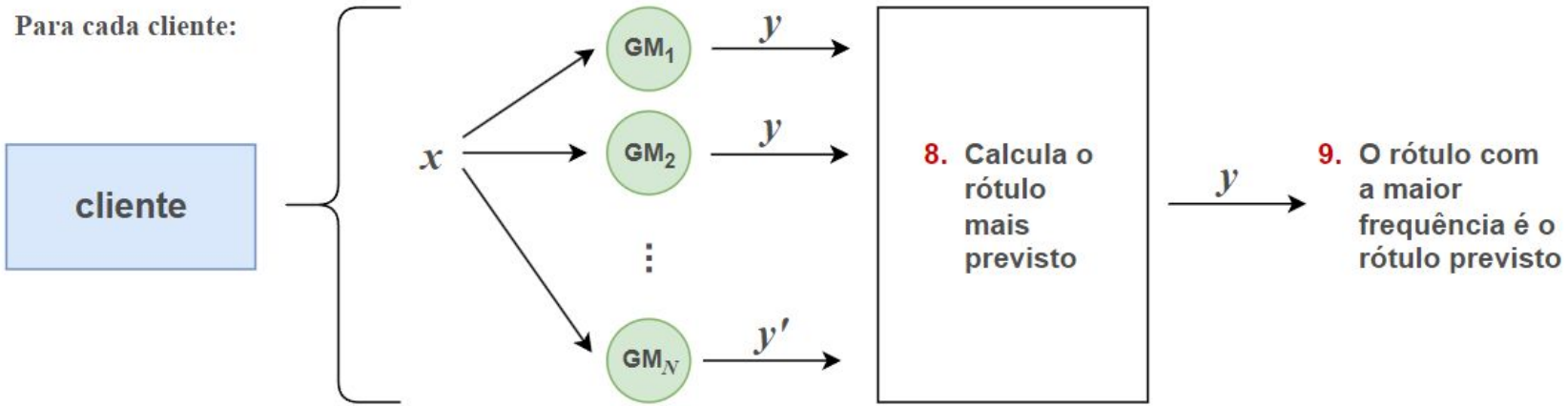
Abordagem Proposta

Verificação de desempenho dos modelos globais



Abordagem Proposta

Inferência baseada em um esquema de votação



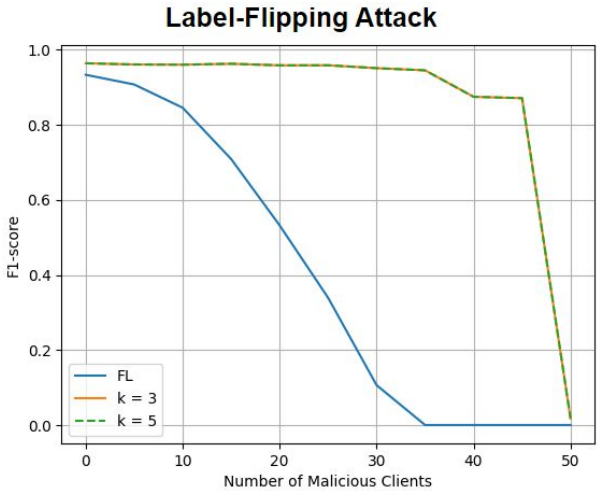
7. Para cada entrada de teste x, todos os Modelos Globais predizem o seu rótulo

Avaliação e Resultados

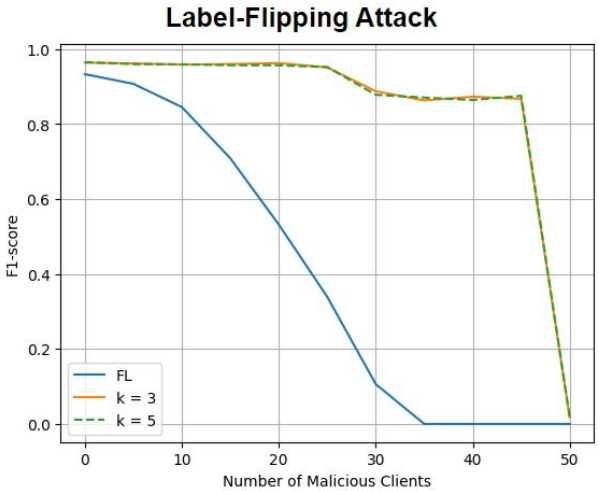
- Conjuntos de dados utilizados:
  - MNIST;
  - Human Activity Recognition Using Smartphones (HAR).
- Ataques de envenenamento:
  - Same-Value Attack;
  - Label-Flipping Attack.

Avaliação e Resultados

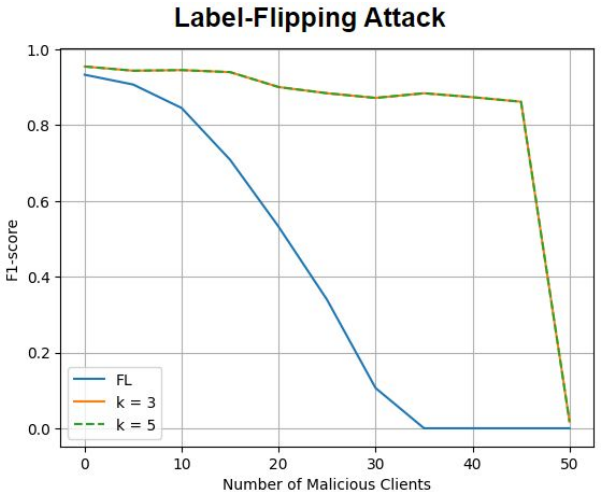
MNIST com 50 clientes



(a) $N = 15$



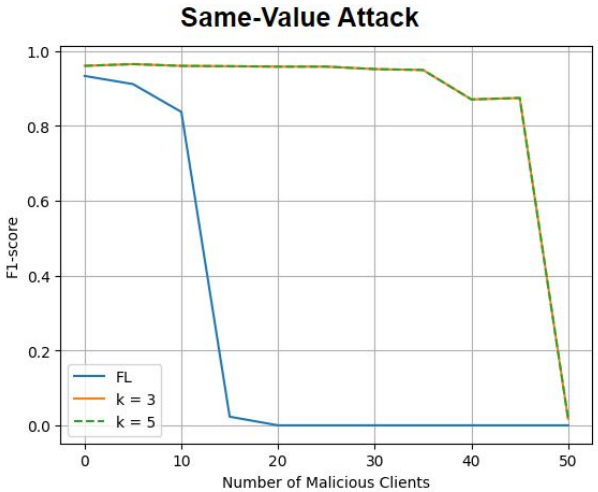
(b) $N = 25$



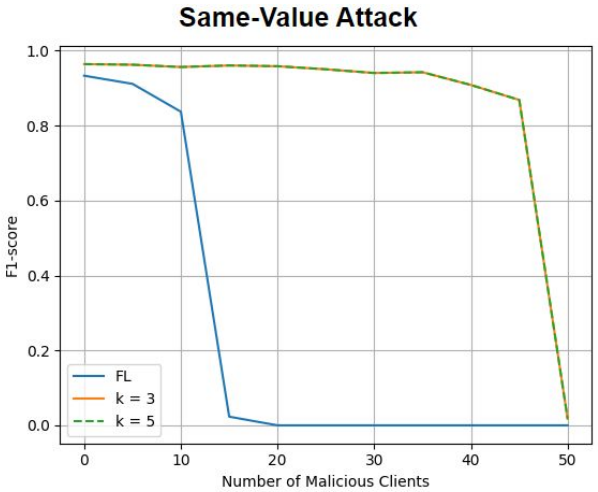
(c) $N = 35$

Avaliação e Resultados

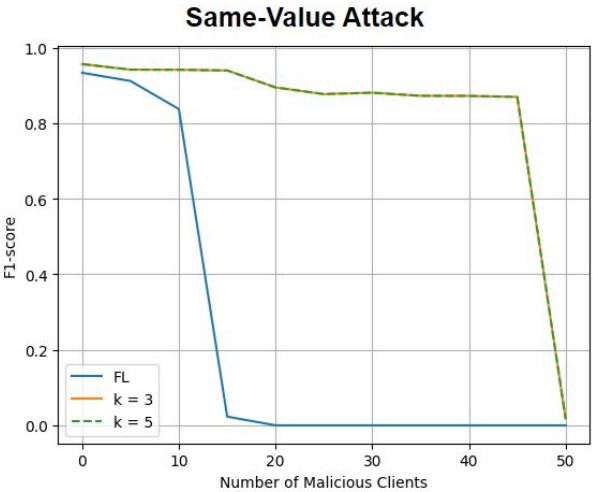
MNIST com 50 clientes



(a) $N = 15$



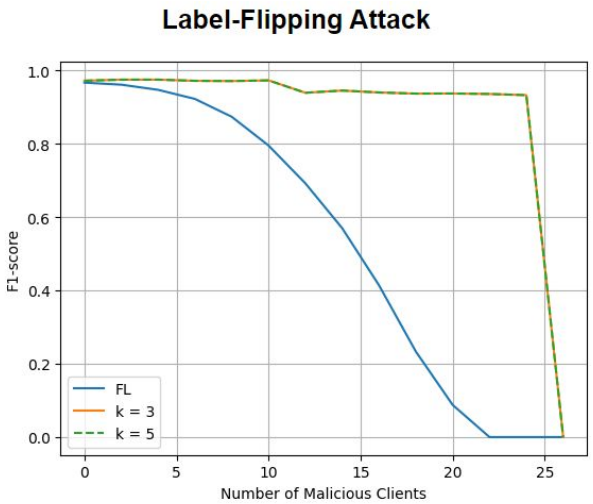
(b) $N = 25$



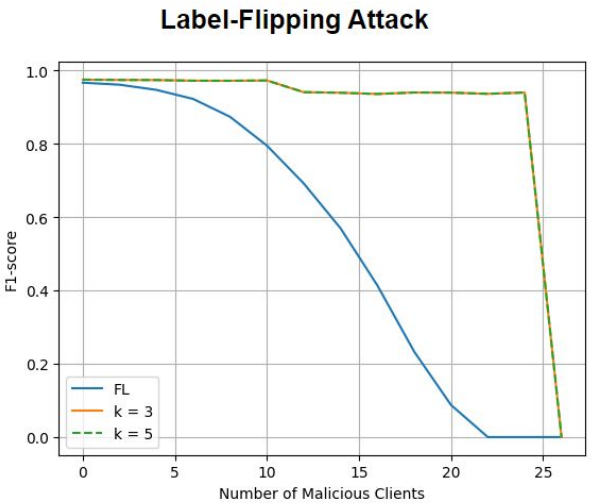
(c) $N = 35$

Avaliação e Resultados

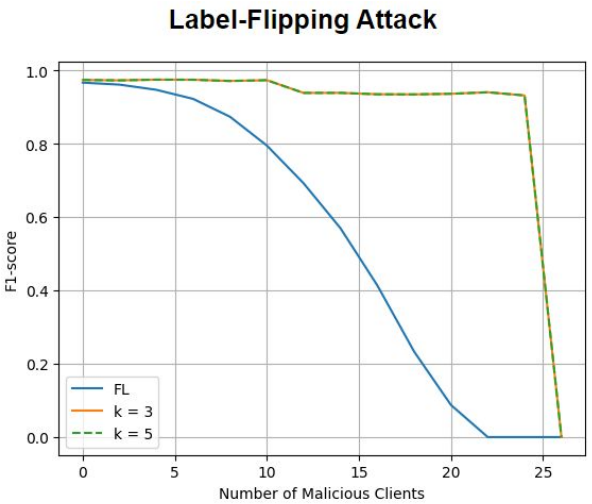
MNIST com 30 clientes



(a) $N = 7$



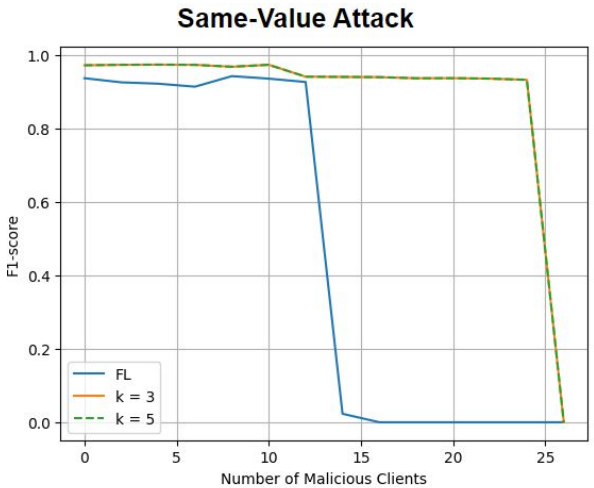
(b) $N = 13$



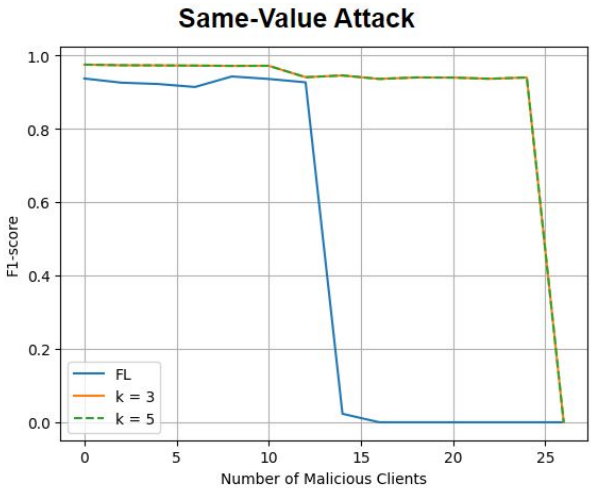
(c) $N = 17$

Avaliação e Resultados

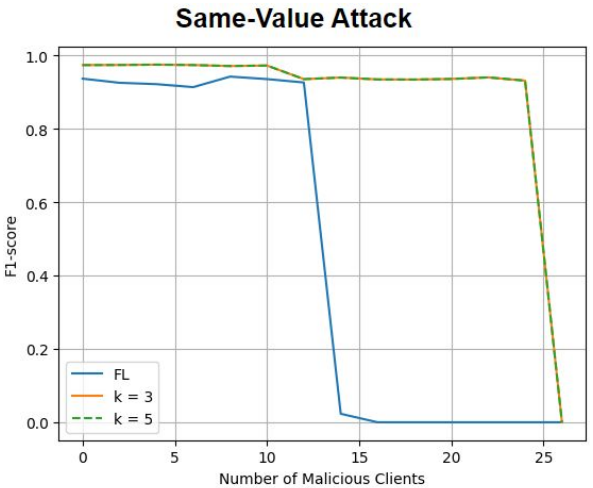
MNIST com 30 clientes



(a) $N = 7$



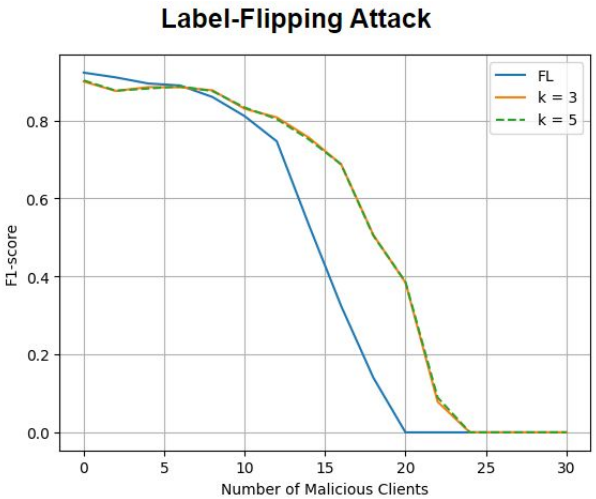
(b) $N = 13$



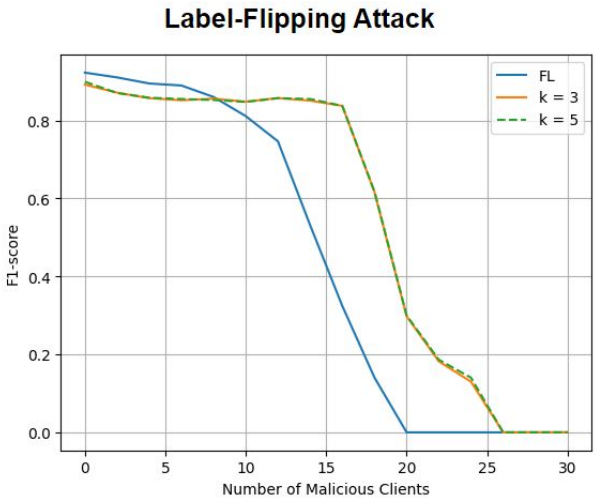
(c) $N = 17$

Avaliação e Resultados

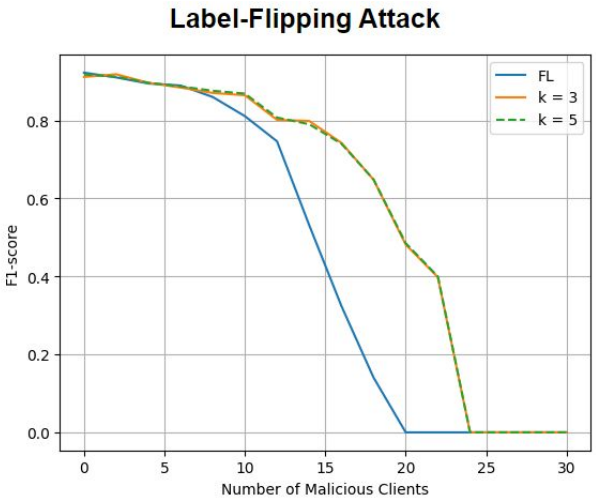
HAR com 30 clientes



(a) $N = 9$



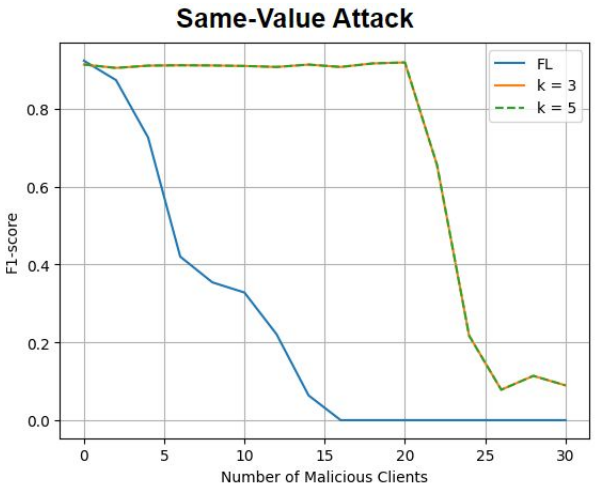
(b) $N = 15$



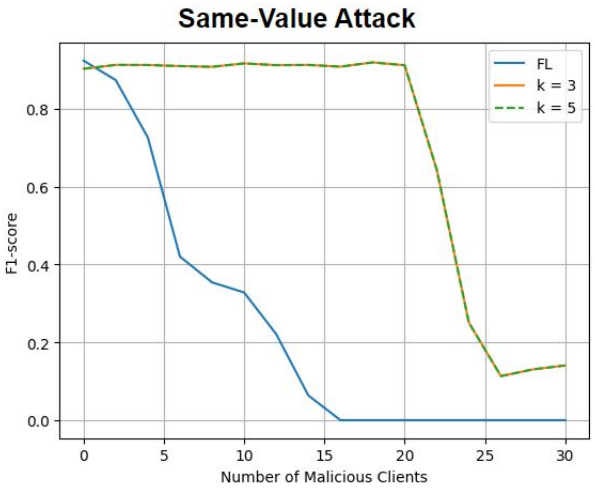
(c) $N = 21$

Avaliação e Resultados

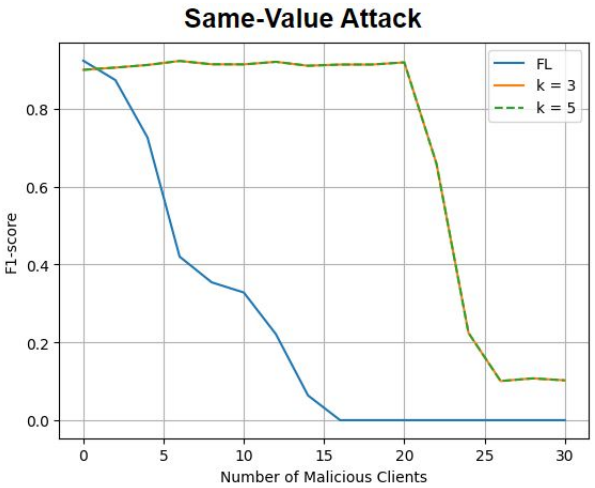
HAR com 30 clientes



(a) $N = 9$



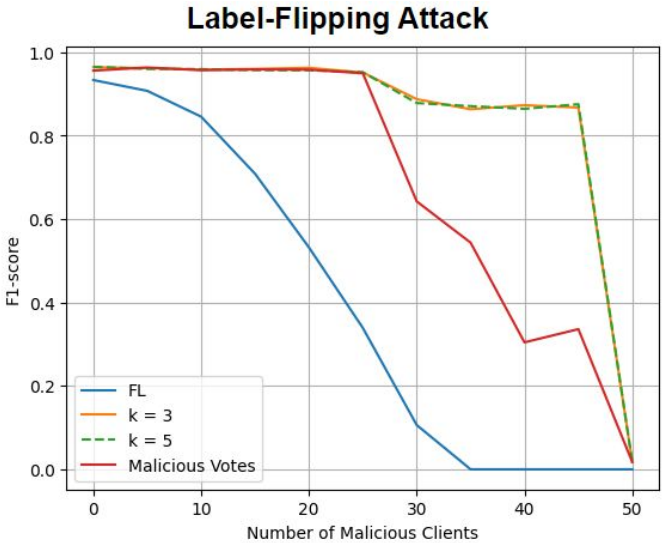
(b) $N = 15$



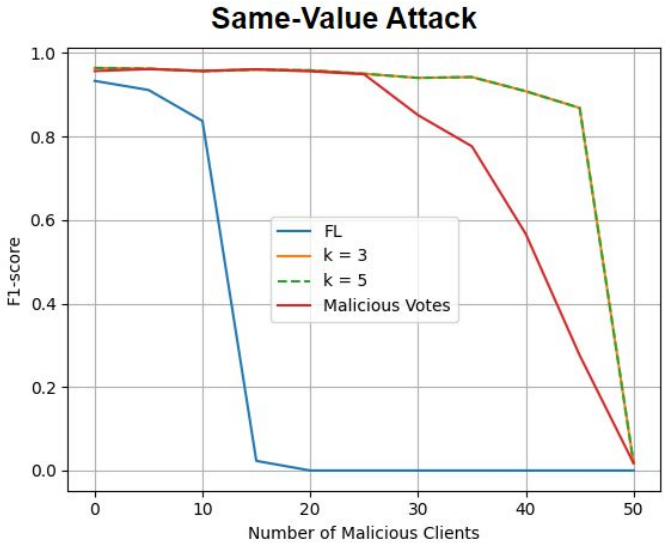
(c) $N = 21$

Avaliação e Resultados

Lidando com uma seleção maliciosa



(a) $n = 50$ and $N = 25$



(a) $n = 50$ and $N = 25$

Considerações Finais

- Para o conjunto de dados MNIST, os resultados mostraram uma F1-score acima de 0,8, mesmo com 90% de clientes maliciosos, e para o conjunto de dados HAR, os resultados da F1-score foram acima de 0,8, mesmo com 66,6% de clientes maliciosos.

Obrigado!

Contato:

- Blenda Oliveira Mazetto
- blenda.mazetto@uel.br