NATIONAL TECHNICAL UNIVERSITY OF ATHENS - NTUA

SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

NETWORK MANAGEMENT AND OPTIMAL DESIGN LABORATORY

INSTITUTE OF COMMUNICATION AND COMPUTER SYSTEMS - ICCS

Explainable AI for DNS Security in Privacy-aware NREN Federations

(FedXAI4DNS)

GÉANT Innovation Programme 2024
SIG-AI, April 2025 - Prague

Mary Grammatikou, NTUA
mary@netmode.ntua.gr

Introduction

Federated Learning (FL)

eXplainable Artificial Intelligence (XAI)

FL and **XAI** for:

- Training Domain Generation Algorithm (DGA) name classifiers in a privacy-aware manner
- Interpreting categorizations

Contributions

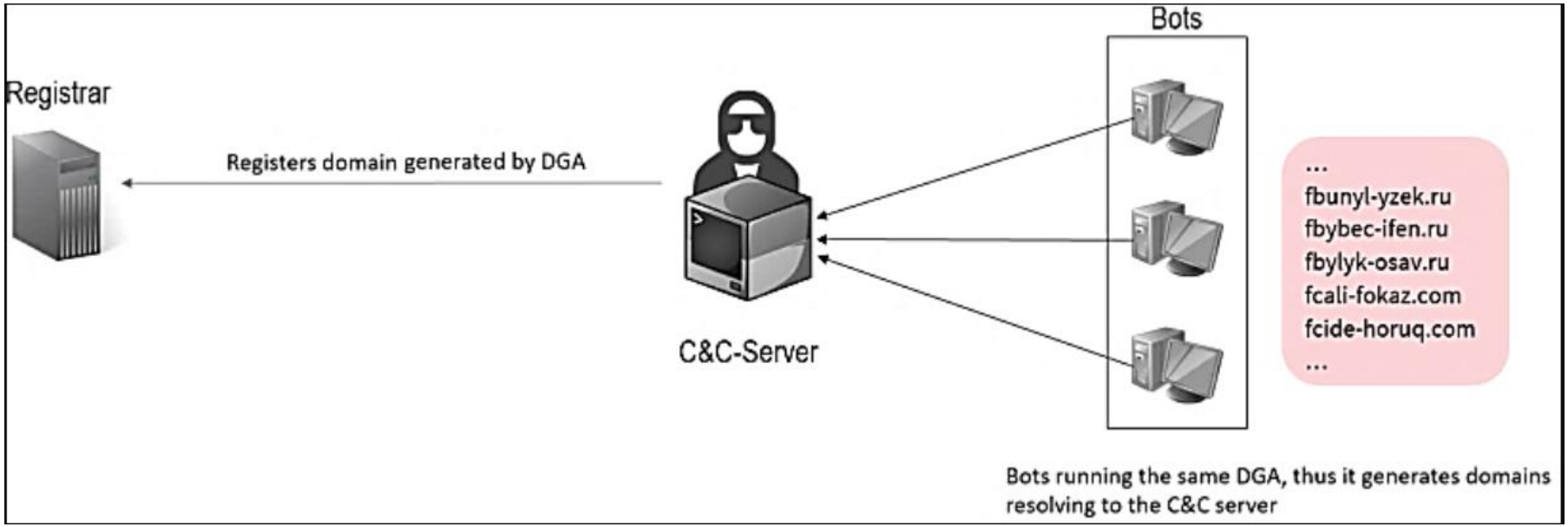
- Binary **deep neural networks** (Multi-Layer Perceptrons - MLP's) for benign vs malicious Domain Name System (DNS) traffic classifications

- **Federated Learning (FL)** for collaborative DGA name classifiers without sensitive data exchanges among participants

- **SHapley Additive exPlanation (SHAP)** for interpreting classification decisions in a model-agnostic manner

Background

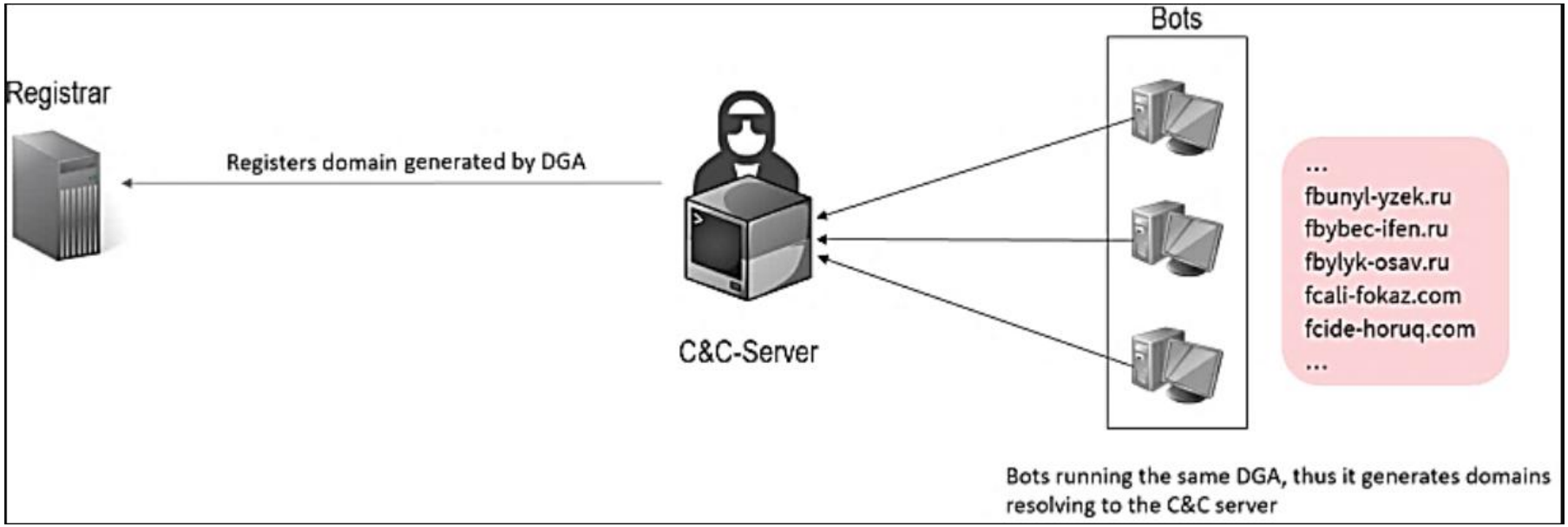
Domain Generation Algorithms - DGA's (1/2)



- Bots depend on **Command & Control (C&C)** servers to receive orders for attacks
- Bots generate large numbers of DNS requests (most invalid) based on **DGA algorithms** with a seeding technique known to C&C servers

(**Image source:** Dau Hoang – "An Enhanced Model for DGA Botnet Detection Using Supervised Machine Learning")

Domain Generation Algorithms - DGA's (2/2)

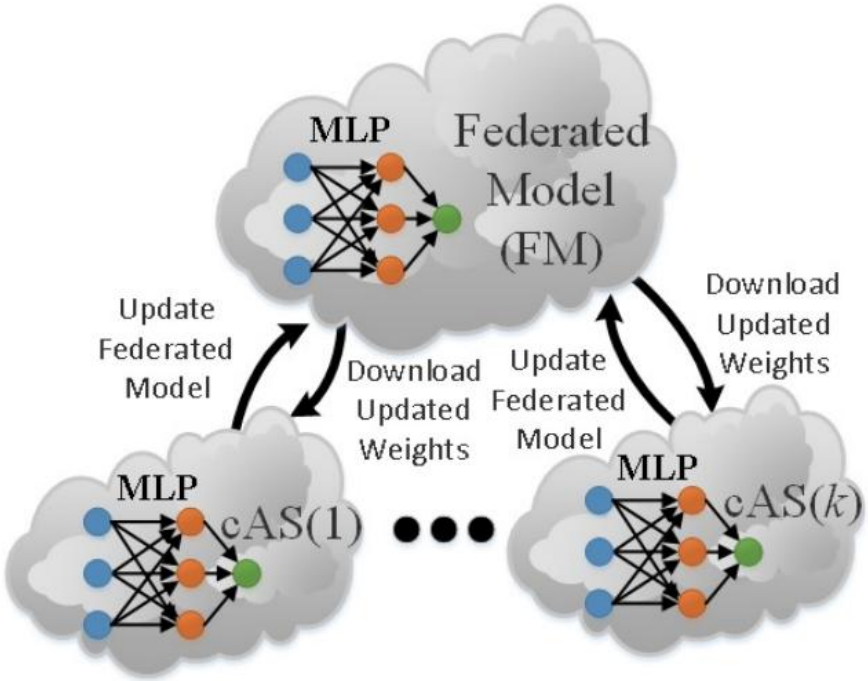


- Large (hundreds to thousands) numbers of requested names (mostly invalid) and frequent DGA seed changes render name blacklists ineffective
- DGA names can be detected via **Machine Learning (ML)** algorithms

(**Image source:** Dau Hoang – "An Enhanced Model for DGA Botnet Detection Using Supervised Machine Learning")

Federated Learning (FL)

Collaborative and **privacy-aware** ML model development



- FL participants (e.g. AS's) share **model weights** instead of raw data
  - No sensitive attack and benign data exchanges
  - Scalable model development requiring less bandwidth

- **Federated Averaging:** Participant weights are averaged by TTP into a global model
  - **FL rounds:** Total times Federated Averaging is performed
  - **Local epochs:** Each participant iteration on the local training dataset

(**Image source:** Marinos Dimolianis – "*DDoS Attack Detection via Privacy-aware Federated Learning and Collaborative Mitigation in Multi-domain Cyber Infrastructures*", IEEE CloudNet 2022)

eXplainable Artificial Intelligence (XAI)

Interpreting decisions of complex (black-box) Machine Learning (ML) models

Our approach relies on model-agnostic, post-hoc XAI methods:

- **Model-agnostic:** Applicable to many ML models (i.e. tree classifiers, neural networks) without modifications
- **Post-hoc:** Applied after model parameter tuning
- Model-specific knowledge is not required
- Selected model-agnostic method: **SHAP- SHapley Additive exPlanation**

SHapley Additive exPlanation – SHAP

SHAP operates on:

- **eXplainability Background Instances (XBI's):**
  Used to tune SHAP values (similarly to ML training data)
- **eXplainability Test Instances (XTI's):**
  Used to derive interpretations (similarly to ML testing data)

Extending SHAP values to Federated Learning schemas:

Additivity property of SHAP (Interpretable Machine Learning)

→ SHAP values can be averaged in Federated Learning schemas

Prototype Baseline

Overview

Deployed prototype steps:

- Dataset preprocessing
- Splitting dataset among FL participants using diverse methods of varying randomness
- Training participant models based on their local datasets (without FL, using FL)
- Providing SHAP-based XAI interpretations

Selected Features

Features are directly extracted from domain names:

→ Enabling real-time name classification

→ Targeting linguistic and statistical name properties

→ Avoiding time-consuming and privacy-sensitive operations

Sequence Number	Feature Name(s)	Description
1	Length	Length of the domain name
2	Max_DeciDig_Seq	Length of maximum decimal digit sequence
3	Max_Let_Seq	Length of maximum letter sequence
4 - 29	Freq_A, Freq_B, ..., Freq_Z	Frequency of letters A-Z within the domain name
30 - 39	Freq_0, Freq_1, ..., Freq_9	Frequency of digits 0-9 within the domain name
40	Spec_Char_Freq	Number of special characters (hyphens, dots) within the domain name
41	Ratio_Spec_Char	Fractional division of Spec_Char_Freq and Length
42	DeciDig_Freq	Number of decimal digits (0-9) within the domain name
43	Ratio_DeciDig	Fractional division of DeciDig_Freq and Length
44	Vowel_Freq	Number of vowels within the domain name
45	Vowel_Ratio	Fractional division of Vowel_Freq and Length
46	Max_Gap	Length of the longest domain name label
47	Reputation	Number of whitelisted N-grams (N = 3, ..., 7)
48	Words_Freq	Number of concatenated meaningful words within the domain name
49	Words_Mean	Average length of concatenated meaningful words obtained from feature 48
50	Entropy	Shannon Entropy of the domain name

50 features

Experimental Assessment

Dataset Description

Dataset for training/testing **binary MLP classifiers** including:

- Malicious (i.e. DGA) names
- Legitimate names

- **Malicious names** → DGArchive Repository

  - Popular name repository that includes names from over 100 DGA families
  - 675,000 total DGA names and 45 DGA families

- **Legitimate names** → Tranco List

  - Online service ranking websites based on their popularity
  - Top 1 million legitimate names were included in the dataset

Federated Learning – Dataset Splitting Methods

- **DGA names:**
  Three Methods (A, B, C) for distributing DGA names across Federated Learning (FL) participants (e.g. NREN's)

- **Legitimate names:**
  Randomly distributed to participants regardless of methods A, B and C

Dataset Splitting Methods – Method A (1/5)

Method A: DGA names are randomly assigned to FL participants based on a uniform distribution independently of the respective DGA family

Dataset Splitting Methods – Method A (2/5)

Method A: DGA names are randomly assigned to FL participants based on a uniform distribution independently of the respective DGA family

DGA Family 1: Corebot

rauggyguyp.com
zrkdvzjhse.com

DGA Family 2: Simda

gatyfus.com
vowydef.com

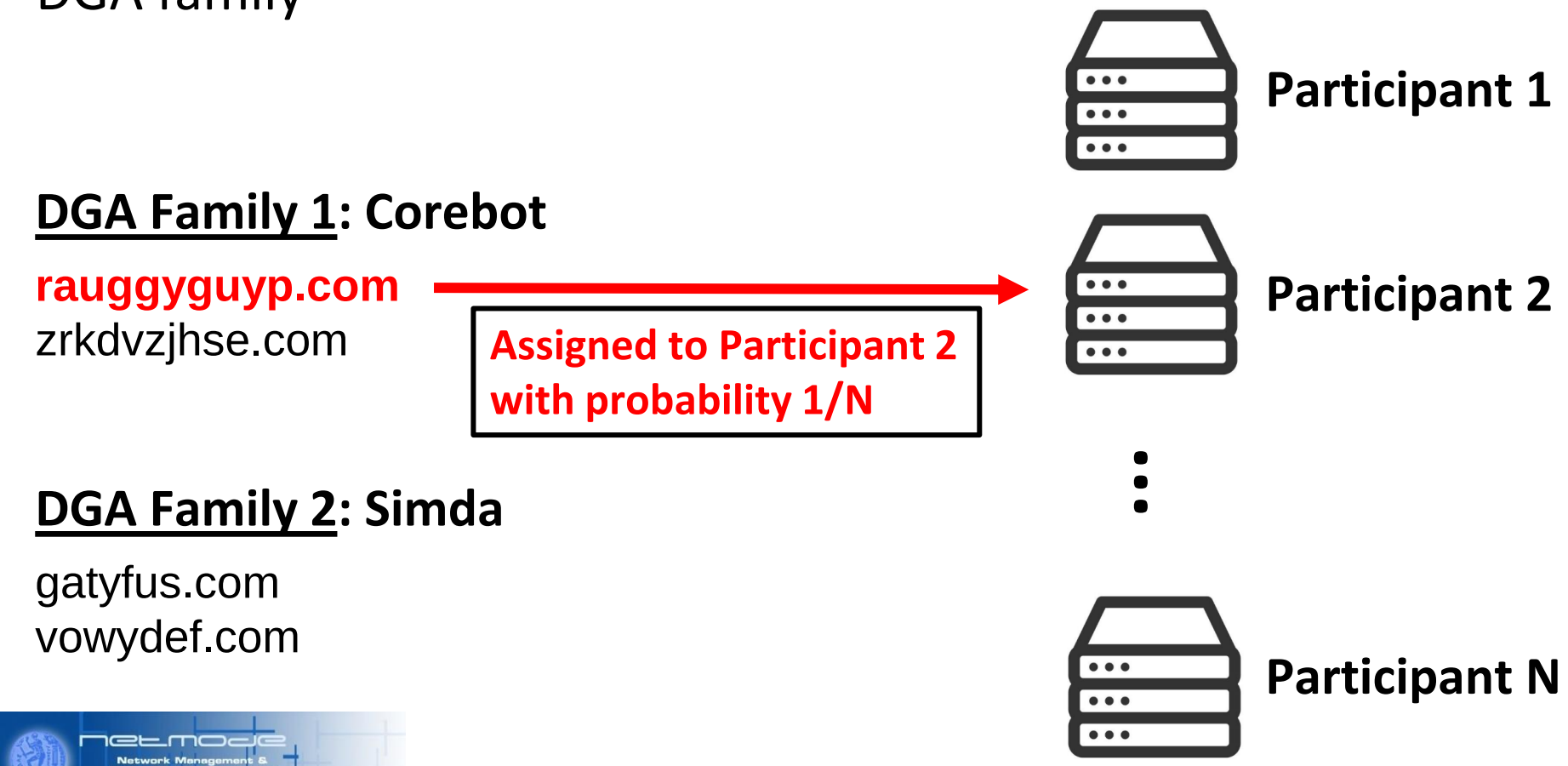
Participant 1

Participant 2

Participant N

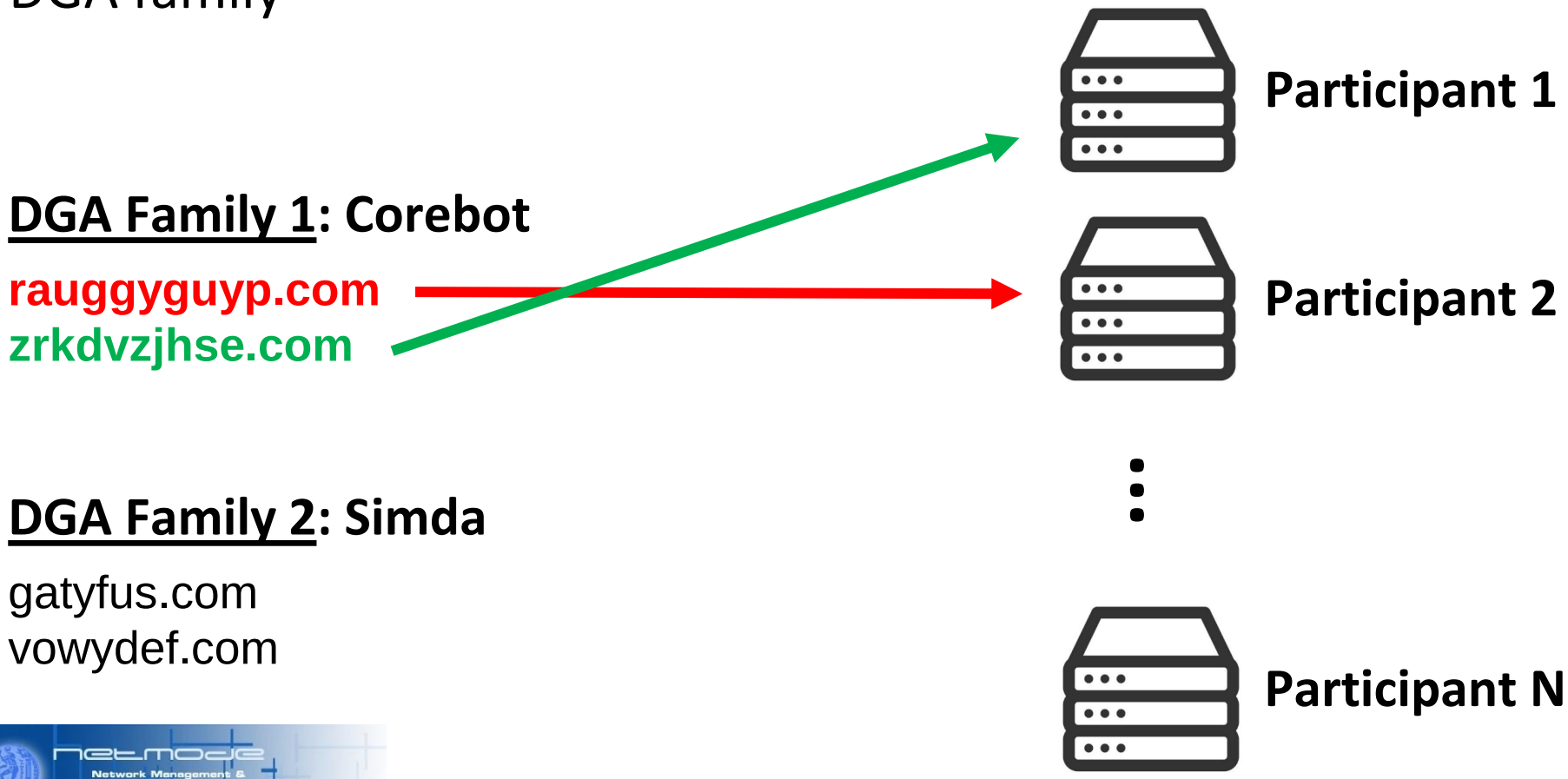
Dataset Splitting Methods – Method A (3/5)

Method A: DGA names are randomly assigned to FL participants based on a uniform distribution independently of the respective DGA family



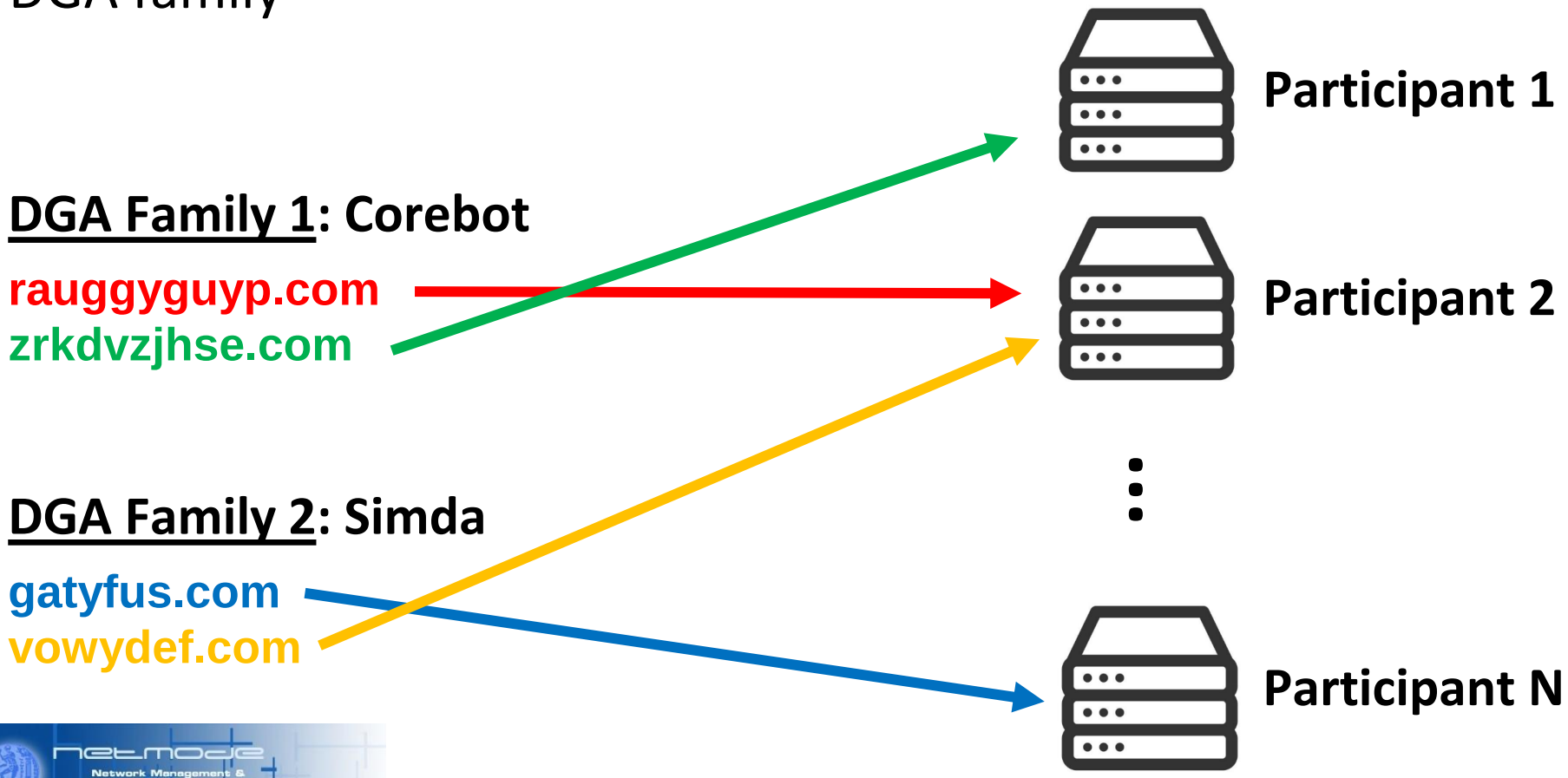
Dataset Splitting Methods – Method A (4/5)

Method A: DGA names are randomly assigned to FL participants based on a uniform distribution independently of the respective DGA family



Dataset Splitting Methods – Method A (5/5)

Method A: DGA names are randomly assigned to FL participants based on a uniform distribution independently of the respective DGA family



Dataset Splitting Methods – Method B (1/4)

Method B: A DGA family and all of its respective names are assigned to a single participant (participant selection is based on a uniform distribution)

Dataset Splitting Methods – Method B (2/4)

Method B: A DGA family and all of its respective names are assigned to a single participant (participant selection is based on a uniform distribution)

DGA Family 1: Corebot

rauggyguyp.com
zrkdvzjhse.com

DGA Family 2: Simda

gatyfus.com
vowydef.com

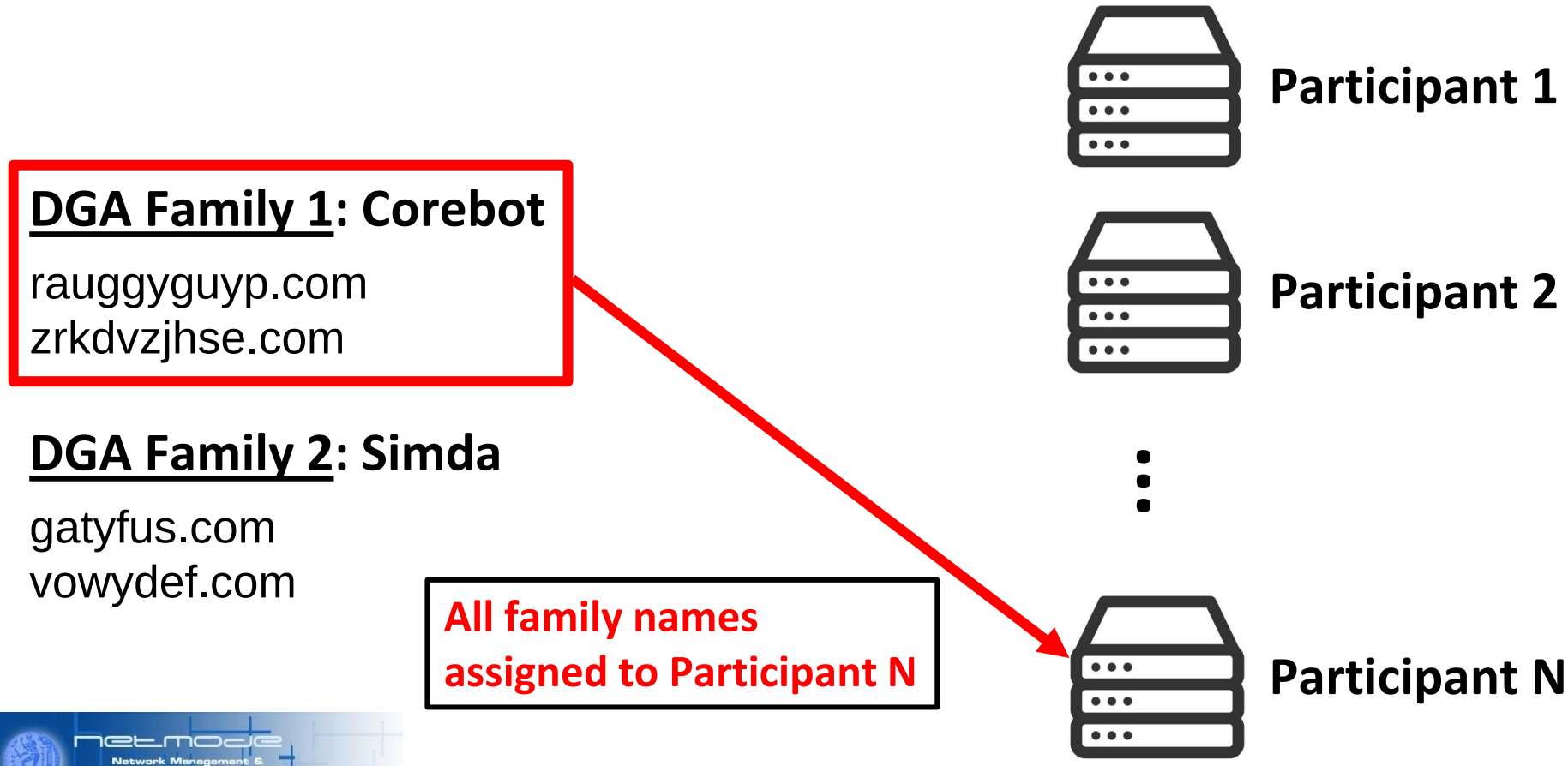
Participant 1

Participant 2

Participant N

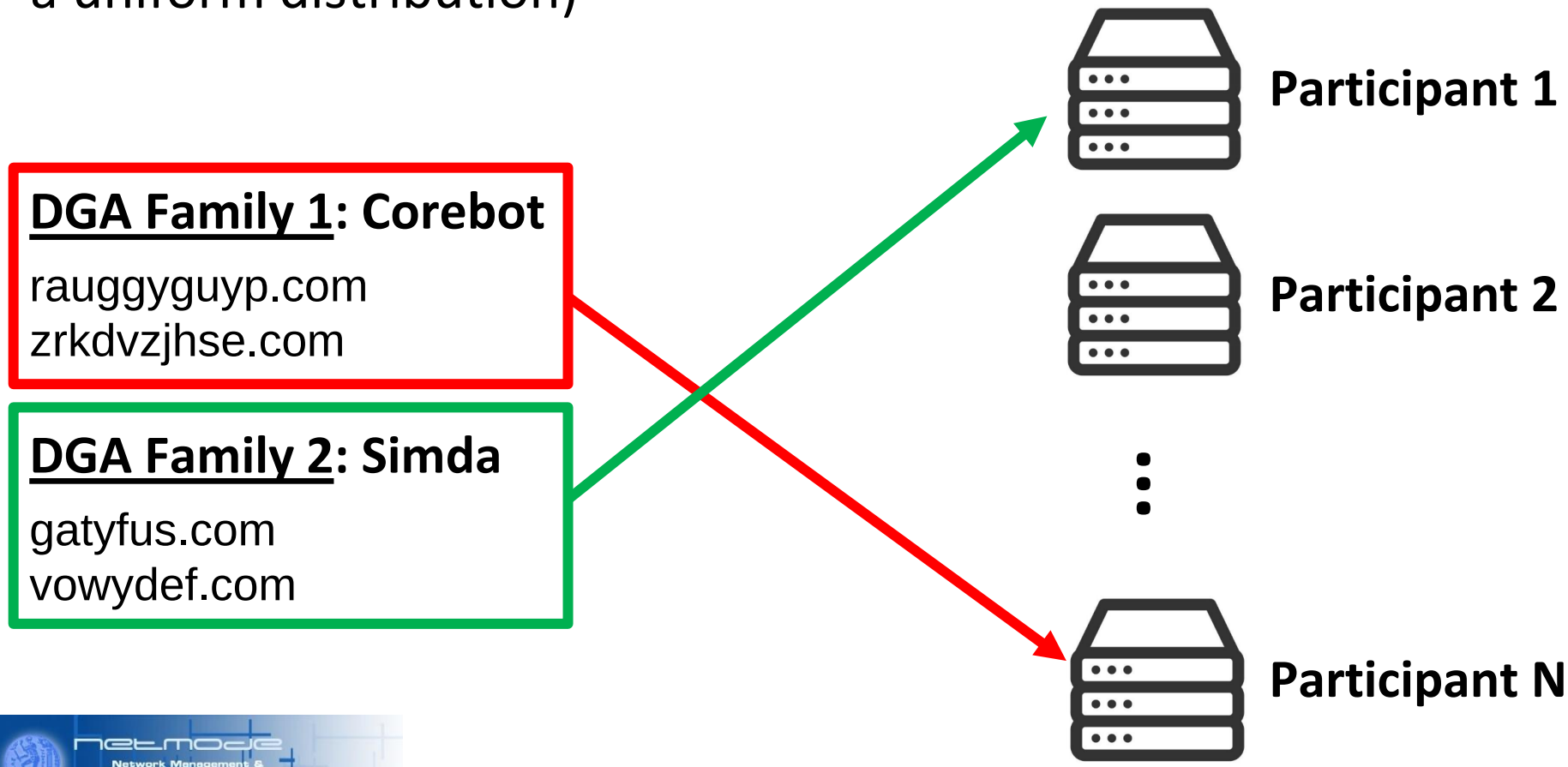
Dataset Splitting Methods – Method B (3/4)

Method B: A DGA family and all of its respective names are assigned to a single participant of the federation (participant selection is based on a uniform distribution)



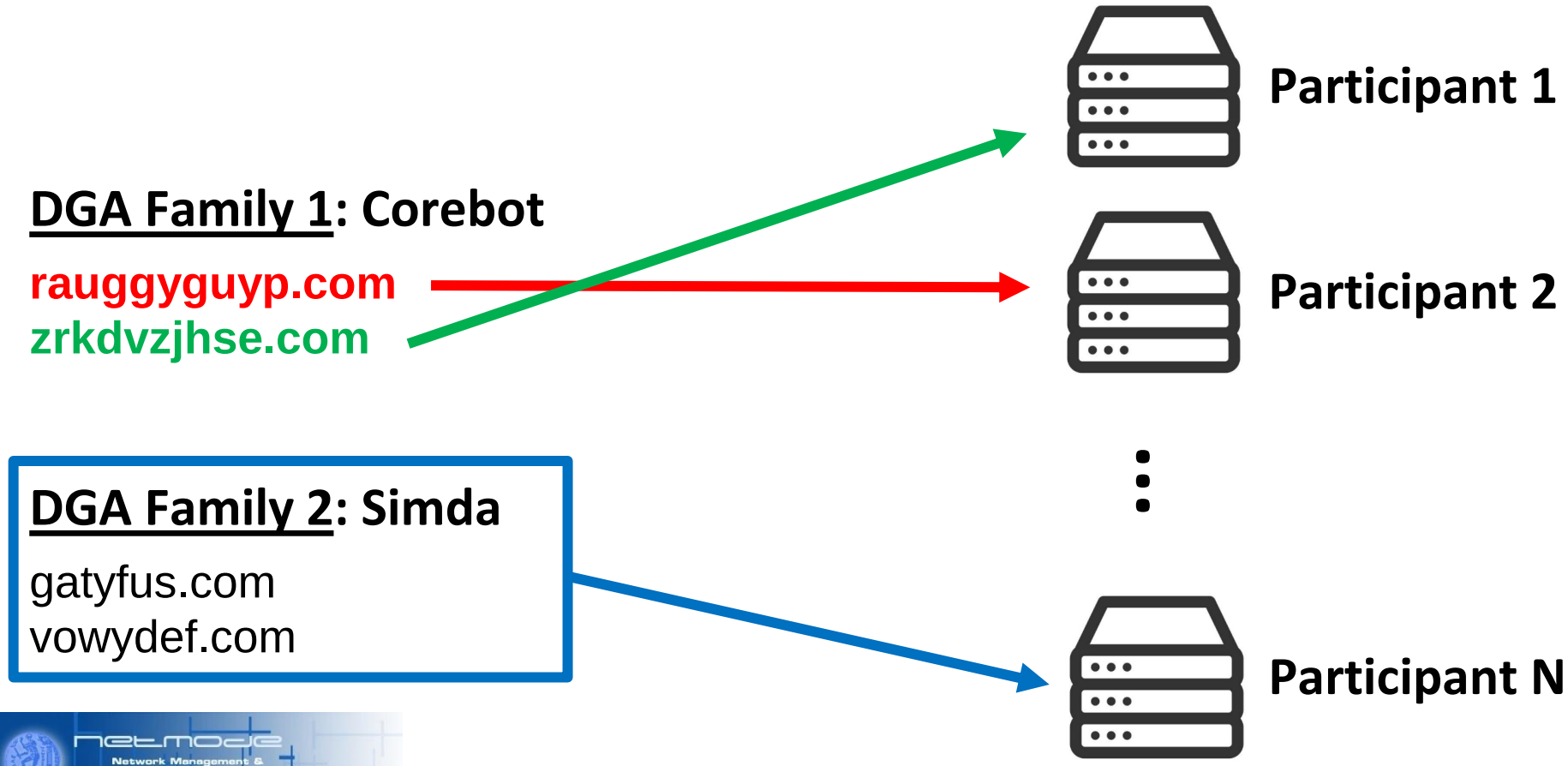
Dataset Splitting Methods – Method B (4/4)

Method B: A DGA family and all of its respective names are assigned to a single participant (participant selection is based on a uniform distribution)



Dataset Splitting Methods – Method C

Method C: Combination of methods A and B: 50% of DGA families are assigned based on method A and 50% on method B



1. Classification Performance (1/3)

Purpose:

- Evaluate Federated Model performance
- Assess if FL boosts the performance of individual participants (i.e. models trained without FL)

Performance metric → ML model **accuracy on testing dataset**

- **True Positives:** Correctly classified DGA names

- **True Negatives:** Correctly classified legitimate names

$$Accuracy = \frac{True\ Positives + True\ Negatives}{All\ Testing\ Dataset\ Names}$$

1. Classification Performance (2/3)

Purpose:

- Evaluate Federated Model performance
- Assess if FL boosts the performance of individual participants (i.e. models trained without FL)

ML model specifications

- **Deep neural networks:** Multi-Layer Perceptron (MLP)
- 3 hidden layers (300, 300, 200 neurons)
- Common architecture for the federated and individual participant models

Best classifier determined based on **Grid Search**

1. Classification Performance (3/3)

Purpose:

- Evaluate Federated Model performance
- Assess if FL boosts the performance of individual participants (i.e. models trained without FL)

Number of FL participants

- Experiments including 5 individual participants
- **Applicable to NREN collaborations**

FL specifications

- 30 federated rounds
- 5 local epochs

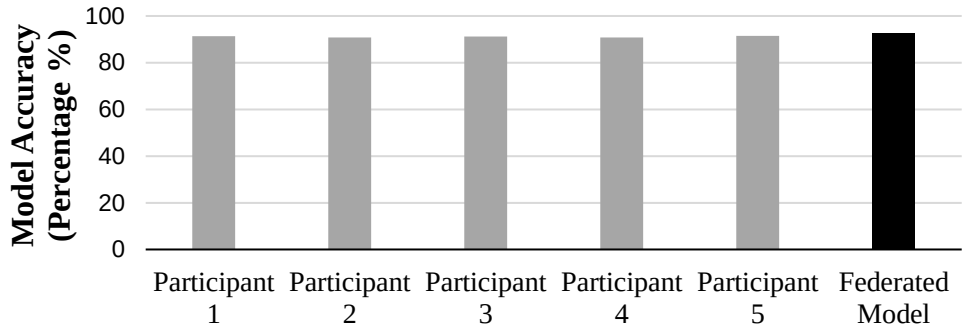
Non-FL Specifications

- Participant models trained in 100 epochs

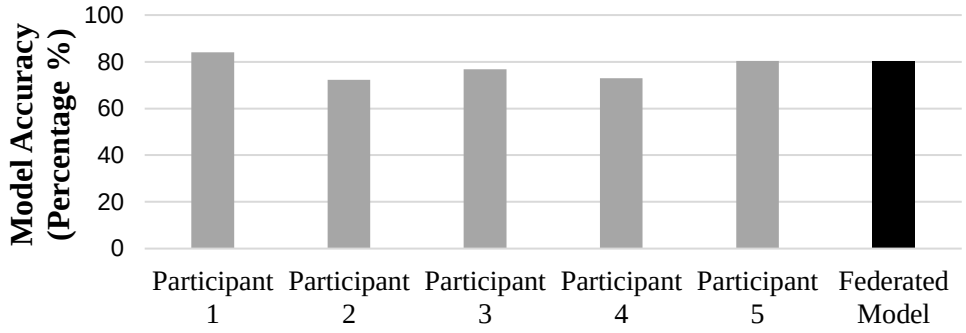
1. Accuracy Assessment – 5 FL Participants

Charts for methods A, B and C

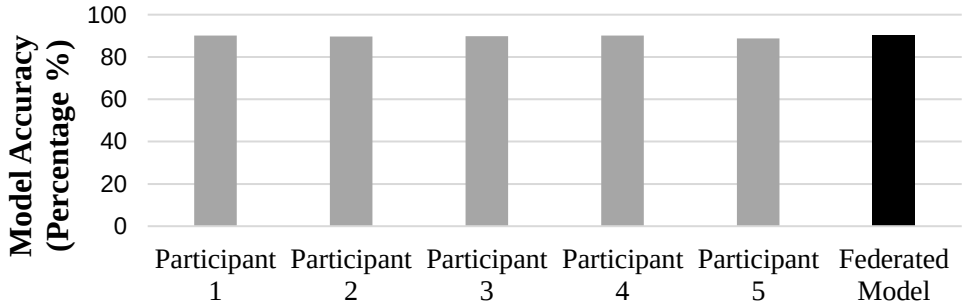
A



B



C



Black bars: With FL

Grey bars: Without FL

Federated Model Accuracy:

A: 92.65%

B: 80.32%

C: 90.43%

- FL boosts the accuracy of most participants
- **Method B:** Less accurate Federated Model than methods A and C *(Disjoint learning sets among participants)*

2. Model Interpretations (1/2)

Purpose:

Derive SHAP values for Machine Learning model to assess feature contributions to classification decisions

2. Model Interpretations (2/2)

- **SHAP values:**

  - **Negative** SHAP values point to **class 0**: Non-DGA names

  - **Positive** SHAP values point to **class 1**: DGA names

- **XAI samples:**

  - **50 eXplainability Background Instances (XBI's)**:
    Cluster centroids from K-Means execution on the training set

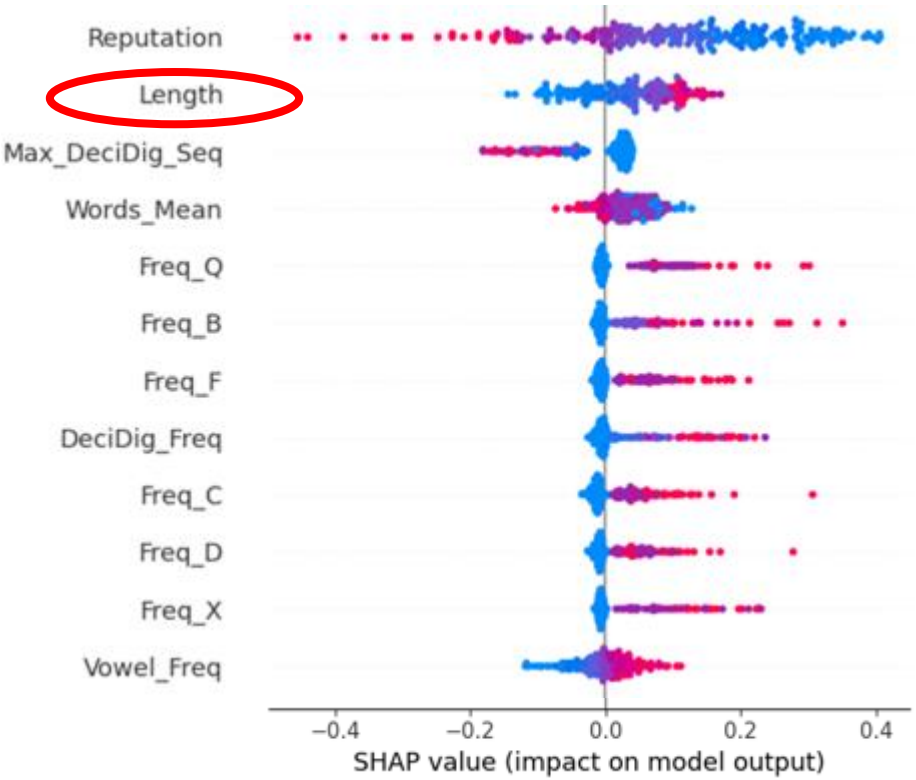
  - **250 eXplainability Test Instances (XTI's)**:
    Subsampled from the testing set sample

- Model interpretations are based on **SHAP summary plots**

2. MLP Interpretations based on Summary Plots

Summary plots:

- Top contributing features ranked in descending order
- **Plot dots:** eXplainability Test Instances (XTI's) from 45 DGA families
- **Color shades:** Low → High feature values



Indicative interpretation:

- Large *Length* feature values are **indicated by red XTI's (plot dots)**
- Red XTI's have **positive SHAP values** → they point to class 1 (class of DGA names)
- **Long names** favor DGA name classifications as expected

Key Takeaways

- **Federated Learning** is promising for collaboratively developing accurate DGA name classifiers without exchanging sensitive attack and benign data
- **SHAP** is a promising technique for interpreting the operation of ML-based security mechanisms that filter malicious DNS messages

Future Work: Federated Learning methods to be applied in NREN collaborations

fedxai4dns@netmode.ntua.gr

Project implementation available from:
https://github.com/netmode/FedXAI4DNS

THANK YOU!