Network traffic generation, between data mining and cybersecurity

Pierre-François Gimenez, CIDRE Inria team

CISPA-Inria workshop, November 6th, 2023

CentraleSupélec

Context and goal

Network

- Computers communicate over networks for information sharing
- A communication (flow) between two computers is composed of an exchange of packets
- Each packet has a payload (the data to be shared) and a header (for networking)

Intrusion detection

- Systems are under attack: DDoS, bruteforce, APT, etc.
- Network intrusion detection systems (IDS) analyze packets to identify attacks
- Commercial performances claims can be very different from actual performances
- IDS evaluation is difficult: getting network data is hard (privacy issue, obsolescence, etc.)

Our goal is to generate synthetic normal (benign) traffic to evaluate IDS

CentraleSupélec

Associated team "SecGen"

SecGen

- A formal collaboration between Inria and CISPA: the SecGen project (started in 2023)
  - CISPA brings data mining and deep learning expertise: Mario Fritz, Jilles Vreeken
  - Inria brings network and IDS expertise: Pierre-François Gimenez, Yufei Han
- Two goals:
  - generate benign network data that resemble public datasets (data augmentation)
  - evaluate synthetic traffic by using it for anomaly detection
- If you would like to know more, contact me! We are looking to expand

Ongoing work

- Generation of sequences of communication flow descriptions (not individual packets) with MDL and Bayesian networks
- Two PhD students worked in the other lab for two months, with good results

CentraleSupélec

Internship and PhD proposals

6-month internship at Inria (Rennes) from March 2024

- Goal: generate a sequence of headers, given a flow description, in two steps:
  - pattern identification with data mining and/or grammatical inference
  - traffic generation with statistical models

PhD with Inria (Rennes) from October 2024

- Generate payloads alongside headers
- Transfer generation to other network architectures
- Evaluate the performances of commercial and academic IDS with synthetic data

Expected skills

- Required: network, security, and Python programming
- Appreciated: machine learning, data mining, formal languages