Network traffic generation: a non-technical look at its characteristics and stakes

Pierre-François Gimenez, CentraleSupélec

SecGen kick-off meeting, June 21st, 2023

CentraleSupélec

CentraleSupélec

Network and security 101

Network 101

What is a network

A network is a system that connects several computers for information sharing. Internet is a network of interconnected networks.

Roles in a network

- Network equipment (routers, bridges, firewalls, etc.): here to route messages, discover computers, block unwanted traffic, etc.
- Servers: computers that offers service, such as website, email, etc.
- Consumer devices: typically "clients" that contact servers to access their services

Communications protocols

To simplify a lot, a network message is composed of two parts:

- a part required for networking (IP addresses, ports, etc.) called the "header"
- a part that's actually used by an application (website, email, etc.), called the "payload"

Flow/conversation

A flow

- A flow is a sequence of packets exchanged between two computers
- Depending of the "header protocol", some start and end packets are mandatory
- Depending of the "payload protocol", the flow can be short or long, with many or little exchanged data, etc.

A conversation

- A conversation is a sequence of sentences exchanged between two people
- Depending of the language they speak, the greetings and goodbye may be different, and phatic words as well
- Depending of the content of the conversation, it can be long or short, with complicated sentences or not, silences, etc.

Intrusion detection system 101

Intrusion detection system (IDS)

- Their goal is to detect malicious behavior
- Most of them analyzes network traffic (the attacker typically attacks one computer from another to gain control)
- Most industrial IDSs are based on signatures (patterns that match attacks), some use machine learning

IDS evaluation

- An important issue is the evaluation of IDSs because getting access to network data is difficult (privacy issue, obsolescence...)
- Synthetic data can be very useful for IDS evaluation
- Attack traffic is easy to generate, but benign traffic (i.e., users using the system) is more difficult to generate accurately $\Rightarrow$ we will only focus on benign traffic generation

Generation 101

Two kind of generation

- Dataset generation: generate a dataset from a description file ("generate data for a network with 3 computers")
- Dataset augmentation: from a sample of network traffic, generate similar traffic traces

For now, our goal is dataset augmentation of benign traffic

How to generate network traffic

- Several papers try to generate synthetic network traffic with deep learning (GAN, VAE)
- AI for generation is very popular but results for network traffic is still bad:
  - we don't have the learning dataset nor the learning hardware of GPT-3...
  - image generation techniques is bad with categorical features...
- In fact, we got better results with old-school Bayesian networks
- Adrien will present this part in detail

CentraleSupélec

The first problem: flow generation

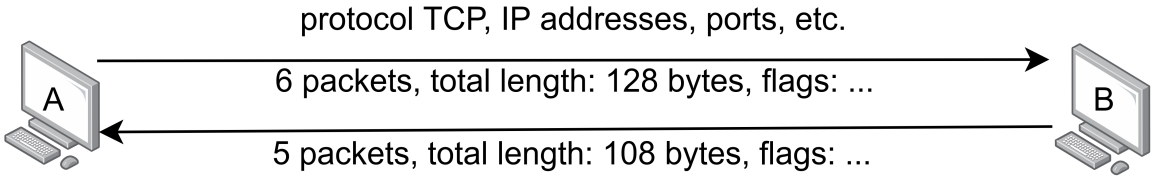
Flow description generation

Step 1: flow description

Generate some overview of the flow: number/size of packets exchanged, what protocol, etc.
⇒ many IDS only use this data
⇒ that's what we can generate right now with good results

Difficulties

- Not one description standard but many formats



Packet headers generation

Step 2: packets headers

Generate a sequence of packet headers (not their content) between two computers
⇒ the payload is generally encrypted, so it makes sense to generate only the headers
⇒ **that's our next goal**

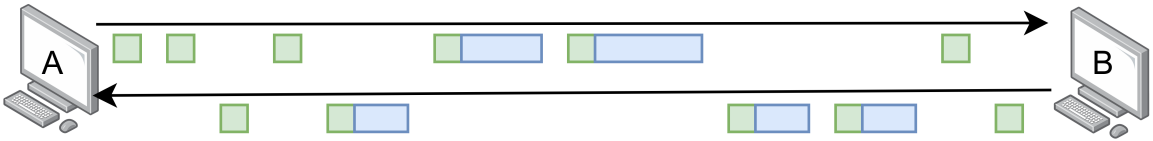
Difficulties

- Sequence learning and generation, not compatible with many ML techniques
- Must be compliant with the previously generated flow description

Complete packets generation

Step 3: complete packets

Generate a sequence of complete packets (header + payload) from a sequence of headers
⇒ some tried... and failed
⇒ not many IDS inspect the content of packets



Out of scope for now

CentraleSupélec

Packets header generation

Problem

- Available data: many packet headers sequences of similar type
- Constraint: it must be consistent with a previously generated flow description

Comments on the problem

- Packets header are described in technical papers, so an expert can verify if the generated sequence is possible or not
- The header is composed of many fields, but some of them can be filled automatically (checksum, constant fields, etc.), so we have some latitude on what to generate exactly (more data mining vs more specification-based completion)
- The available data probably contains incorrect examples

A first idea

Learning a regular expression of packets header

- One symbol = one header
- One sentence = one flow
- Most (all?) protocol rely on a finite state automata

If we can learn a regular expression of flows, it's easy to do the intersection with the grammar of constraints and generate only flows that comply with the constraints!

Challenges

- It requires strong assumptions to limit the number of different states
- Regular expression/automata/grammar exists but typically cannot deal with errors in the learning dataset
- Seems like an approach based on MDL could solve that?
- Generation probabilities?

CentraleSupélec

The second problem: flow sequence generation

The second problem

Not just one conversation...

Imagine you are at Campus Cyber, and you want to send your family a postcard of Paris during your lunch break. Here is how it may happen:

- First, you go to the welcome desk to ask where is the nearest post office
- You go there and find an employee. You are very suspicious of fake post employee, so you ask for their ID
- You give the postcard to the employee
- They check where the postcard needs to go on a map
- The post sends the postcard to your family's house
- They check their ID
- They deliver the postcard

That's a bit far fetched, but it's not very far from the journey of an email

CentraleSupélec

So what?

Generation a sequence of flows

- Many protocols are used is rapid succession (checking ID = setting up encryption, check address = get IP address, etc.)
- There is typically a specification for each protocol, but their interaction can get much more complex
- Sometimes the transmitter change, sometimes not
- Sometimes the delay is short (automatically sent) or long (a user browsing a website)
- It depends on the time of the day, on the behavior of the user, on the technologies used, etc.
- Finally: there are many concurrent users using the same services, so pattern mining is certainly not easy

We expect this to be a bigger challenge!

CentraleSupélec

Conclusion

Conclusion

Why generate traffic?

- Synthetic benign traffic is useful for evaluating intrusion detection systems
- IDS uses various network data: we do not need to generate complete data. Sometimes, data description or headers are enough

What to generate?

- Problem 1: generate the sequence of packet headers between two computers, constrained by a flow description
- Problem 2: generate the sequence of flow descriptions

In the next months...

- Joscha will soon come to Inria to work on data mining applied to network traffic
- Adrien could go to CISPA during fall to work on generation based on Joscha's findings