GANs vs Bayesian Networks for Network Traffic Data Generation

Student : Adrien Schoen
Supervisor : Ludovic Mé
Co-supervisor : Yufei HAN, Gregory Blanc, Frédéric Majorczyk, Pierre-François Gimenez

Table of contents

01

Introduction

The need for benign traffic

Network Intrusion Detection System (NIDS)

NIDS : devices/applications that **monitor traffic inside a network** in order to **detect malicious activities**/policy violations.[Burton2003]

Anomaly-based NIDS : Compute (statistical) models for normal network traffic and **generate alarms when there is a large deviation from the normal model**. [Wang2004]

Many **false positives**.

The evaluation of anomaly-based NIDS

To evaluate those NIDS, we need benign traffic in particular.

The need for artificial traffic

Real benign traffic is **hard to collect and to share** for various reasons [Ring2018] :

- Recording is tedious,
- Sharing threatens privacy
- Data becomes obsolete fast
- Labeling is not certain

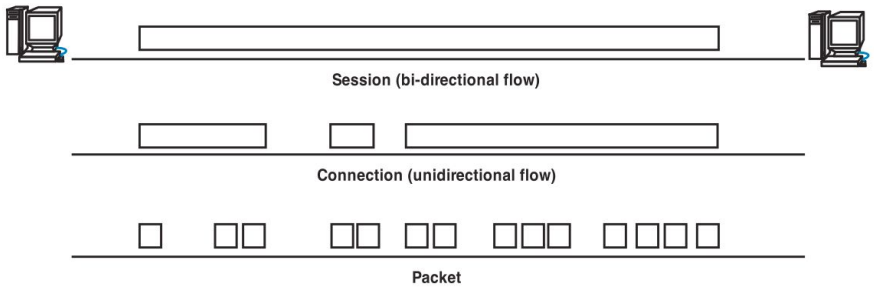
A solution is to use **synthetic** benign traffic instead.

Synthetic : Not a direct recording of user activities.

What is traffic anyway...

Different scales of traffic

- Flow : unidirectional sequence of packets with some common properties that pass through a network device. Netflow format
- Packet : formatted unit of data carried by a packet-switched network. Packet repartition inside a flow.
- Binary : Actual content of the packets of the flow.



02

GANs for generating Network Traffic

First try

Generative Adversarial Networks [Goodfellow2014]

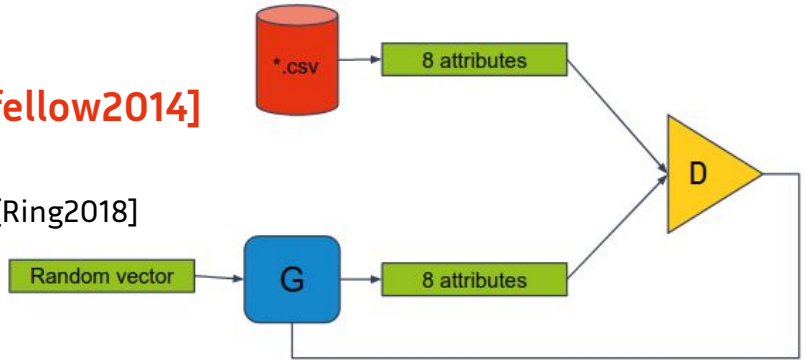
GAN have been proposed for generating benign traffic[Ring2018]

A generative model composed of two neural networks.

The two networks train concurrently.

We have tried multiple global GAN architectures and multiple layer structures.

We can see that the result is incoherent (port issue).



	Duration	Proto	Src IP Addr	Src Pt	Dst IP Addr	Dst Pt	Packets	Bytes
0	0.000921	TCP	192.168.220.15	80	192.168.220.15	80	2	54
1	0.002552	TCP	192.168.200.9	40289	96.76.60.29	39151	4	1581
2	0.000017	TCP	192.168.220.4	443	192.168.220.6	445	3	740

Evaluation of Generation

Evaluating data generation is difficult

Question that was risen in other domains.

We extract 3 criteria from these domains:

- Realism : The produced synthetic network flows should be close to the real network flows
- Diversity : The network flows should have the same variability as the real flow
- Authenticity : A generated traffic flow should not be a simple mere copy of a real traffic flow

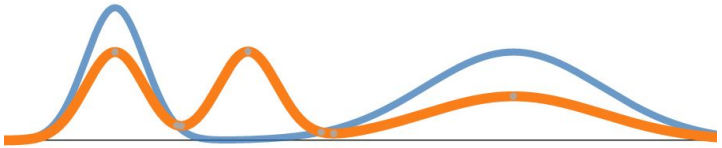
From Network generation, we extract another criterion:

- Compliance : A generated flow should be conformed to network protocol specifications.

Evaluation of Generation

Illustration of realism

Real
Generated



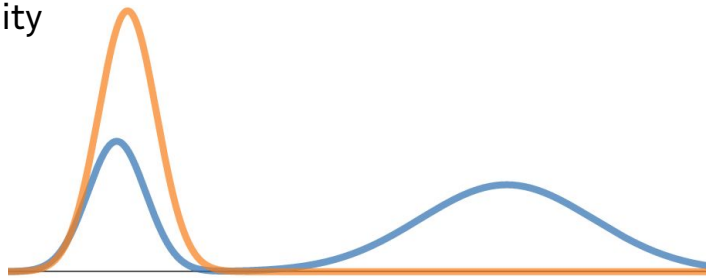
Duration	Protocol	Src IP Addr	Src Pt	Dst IP Addr	Dst IP Pt	Packets	Bytes
0.003	TCP	192.168.220.5	443	192.168.100.8	44870	2	174
0	TCP	192.168.100.8	44870	192.168.220.5	443	1	108
0.004	TCP	192.168.220.13	8080	192.168.100.7	59628	2	174
0	TCP	192.168.100.7	59628	192.168.220.13	8080	1	108
0.003	TCP	192.168.220.13	8080	192.168.100.7	59628	2	174
0.002	TCP	192.168.100.7	59628	192.168.220.13	8080	1	108
0	TCP	192.168.220.5	443	192.168.100.8	44870	1	108
0.01	TCP	192.168.220.13	8080	192.168.100.7	59628	1	108
0	TCP	192.168.100.8	44870	192.168.220.5	443	1	108
0.006	TCP	192.168.220.5	443	192.168.100.8	44870	3	286
0.004	TCP	192.168.100.8	44870	192.168.220.5	443	2	174

Duration	Protocol	Src IP Addr	Src Pt	Dst IP Addr	Dst IP Pt	Packets	Bytes
0.003	TCP	192.168.220.5	443	192.168.100.8	44870	2	174
0	TCP	192.168.100.8	44870	192.168.220.5	443	1	108
0.004	TCP	192.168.220.13	8080	192.168.100.7	59628	2	174
0	TCP	192.168.100.7	59628	192.168.220.13	8080	1	108
0.003	TCP	192.168.237.65	80	192.168.91.74	62933	2	174
0.002	TCP	192.168.91.74	62933	192.168.237.65	80	1	108
0	TCP	192.168.220.5	443	192.168.100.8	44870	1	108
0.01	TCP	192.168.220.13	8080	192.168.100.7	59628	1	108
0	TCP	192.168.100.8	44870	192.168.220.5	443	1	108
0.006	TCP	192.168.220.5	443	192.168.100.8	44870	3	286
0.004	TCP	192.168.100.8	44870	192.168.220.5	443	2	174

Evaluation of Generation

Illustration of diversity

Real
Generated



Duration	Protocol	Src IP Addr	Src Pt	Dst IP Addr	Dst IP Pt	Packets	Bytes
0.003	TCP	192.168.220.5	443	192.168.100.8	44870	2	174
0	TCP	192.168.100.8	44870	192.168.220.5	443	1	108
0.004	TCP	192.168.220.13	8080	192.168.100.7	59628	2	174
0	TCP	192.168.100.7	59628	192.168.220.13	8080	1	108
0.003	TCP	192.168.220.13	8080	192.168.100.7	59628	2	174
0.002	TCP	192.168.100.7	59628	192.168.220.13	8080	1	108
0	TCP	192.168.220.5	443	192.168.100.8	44870	1	108
0.01	TCP	192.168.220.13	8080	192.168.100.7	59628	1	108
0	TCP	192.168.100.8	44870	192.168.220.5	443	1	108
0.006	TCP	192.168.220.5	443	192.168.100.8	44870	3	286
0.004	TCP	192.168.100.8	44870	192.168.220.5	443	2	174

Duration	Protocol	Src IP Addr	Src Pt	Dst IP Addr	Dst IP Pt	Packets	Bytes
0.003	TCP	192.168.220.5	443	192.168.100.8	44870	2	174
0	TCP	192.168.100.8	44870	192.168.220.5	443	1	108
0.003	TCP	192.168.220.5	443	192.168.100.8	44870	2	174
0	TCP	192.168.100.8	44870	192.168.220.5	443	1	108
0.003	TCP	192.168.220.5	443	192.168.100.8	44870	2	174
0	TCP	192.168.100.8	44870	192.168.220.5	443	1	108
0	TCP	192.168.220.5	443	192.168.100.8	44870	1	108
0	TCP	192.168.100.8	44870	192.168.220.5	443	1	108
0.006	TCP	192.168.220.5	443	192.168.100.8	44870	3	286
0.004	TCP	192.168.100.8	44870	192.168.220.5	443	2	174

Evaluation of Generation

Illustration of authenticity

Real

	srcport	dstport	proto	pkt	byt
96659	35046	53	udp	2.0	86.0
44168	51073	80	tcp	1.0	0.0
100181	42847	53	udp	2.0	178.0
68375	57820	80	tcp	3.0	233.0
12856	37396	80	tcp	18.0	1735.0
20671	44140	53	udp	2.0	116.0
52870	48028	443	tcp	3.0	0.0
80140	53910	80	tcp	7.0	1359.0
103118	44474	53	udp	2.0	78.0
43116	57783	53	udp	2.0	98.0
10140	50275	53	udp	2.0	116.0
90343	55992	80	tcp	10.0	1318.0
60605	44494	80	tcp	7.0	0.0
32768	57839	53	udp	2.0	86.0
13514	40909	53	udp	2.0	108.0
8430	60654	80	tcp	3.0	0.0

Generated

	srcport	dstport	proto	pkt	byt
576280	44140	53	udp	2.0	116.0
310049	53910	80	tcp	7.0	1359.0
1040927	57783	53	udp	2.0	98.0
881965	57839	53	udp	2.0	86.0
201389	42847	53	udp	2.0	178.0
943518	36153	80	tcp	6.0	604.0
303795	47369	18325	tcp	3.0	624.0
1019824	35046	53	udp	2.0	86.0
751289	55992	80	tcp	10.0	1318.0
358804	40909	53	udp	2.0	108.0
816450	44474	53	udp	2.0	78.0
728132	57820	80	tcp	3.0	233.0
112527	34221	443	tcp	157.0	27512.0
641611	50275	53	udp	2.0	116.0
320719	56276	443	tcp	6.0	654.0
171456	25847	15401	tcp	25.0	1494.0

Evaluation of Generation

Illustration of compliance

Duration	Protocol	Src IP Addr	Src Pt	Dst IP Addr	Dst IP Pt	Packets	Bytes
0.003	TCP	192.168.220.5	443	192.168.100.8	44870	2	174
0	TCP	192.168.100.8	44870	192.168.220.5	443	1	108
0.004	TCP	192.168.220.13	53	192.168.100.7	59628	2	174
0	TCP	192.168.100.7	59628	192.168.220.13	49326	1	108
0.003	TCP	192.168.237.65	42256	192.168.91.74	62933	2	174
0.002	TCP	192.168.91.74	62933	192.168.237.65	53	1	108
0	TCP	192.168.220.5	443	192.168.100.8	44870	1	32245
0.01	TCP	192.168.220.13	8080	192.168.100.7	59628	1	108
-0.01	TCP	192.168.100.8	44870	192.168.220.5	443	1	108
0.006	TCP	192.168.220.5	443	192.168.100.8	44870	3	286
0.004	TCP	192.168.100.8	44870	192.168.220.5	443	2	174

03

Bayesian Network for generating netflows

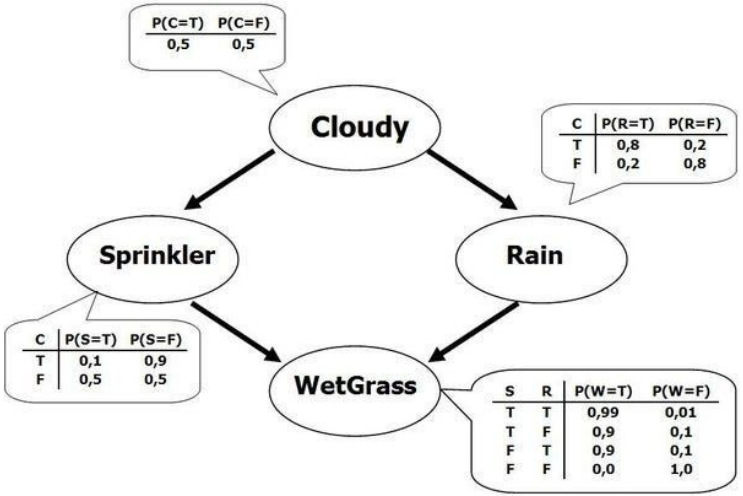
Generating Netflow traffic with Bayesian Networks (BNs)

Netflow traffic is similar to tabular data

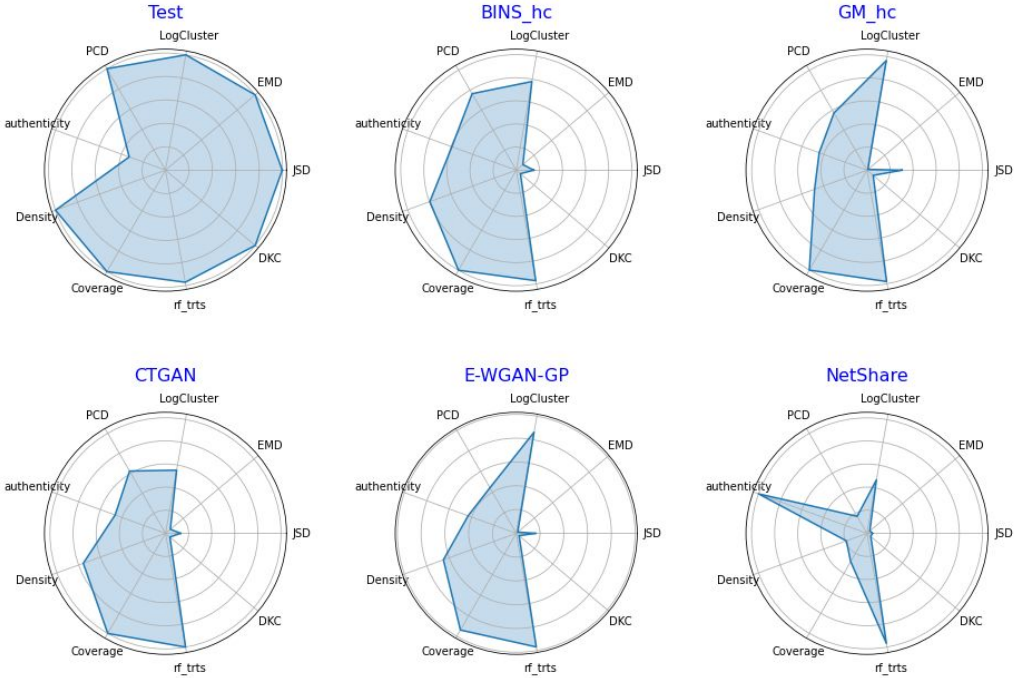
Graphical models representing probabilistic relationships among variables[Heckermn1998]

Consists of nodes and directed edges in a directed acyclic graph

Each node has a conditional probability table (CPT)

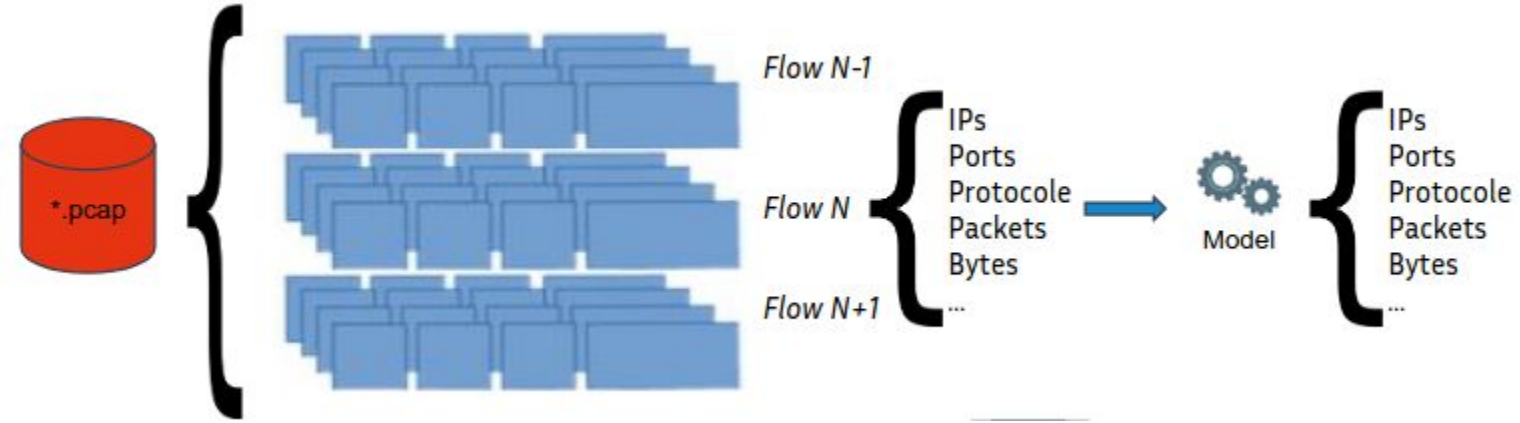


Comparing BNs and GANS



Where are we

We can generate independent flow descriptor



04

Proposals

Generating sequences of Flow

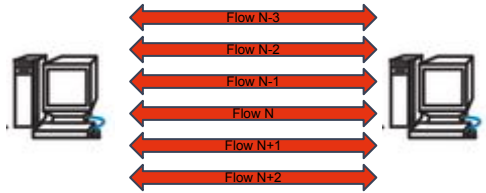
Generate flow sequentially and compare the results with the real sequence

For the moment, we generate flow independently

This is quite unrealistic in a network context (DNS request before HTTP)

We would like to generate flow sequentially

We would have the challenge to see if temporal dependency is well-preserved



Generating the sequence of headers

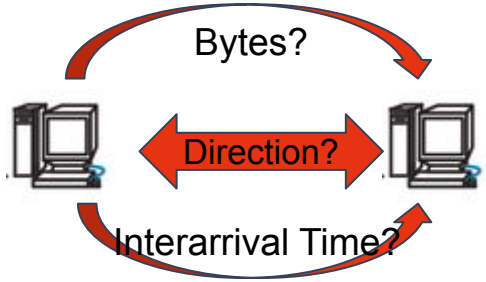
Exhibit patterns in a time series and generate those conditionally

For the moment, we have flow descriptor

If we want to go beyond, we need the ordering of the packet of the flow

We need to extract patterns of how packets are ordered inside a flow

We also need to study the repartition of some properties among packets



Generating the payload of packet

Generate payload of packets.

Generate hex of the packet constituting the flow

Maybe using a GPT model, to generate a sequence of hex conditioned on the size

Not entire packets, because most of it is encrypted

Less interesting for you, I think

05

Conclusion

To conclude

I'm working on generating synthetic benign traffic

I'm using machine learning for this task

I'm currently able to generate independent Netflows

There is room for improvement :

- Including temporal dependencies in the generation
- Generating packets inside a flow
  - Generating their ordering or their content

References

Chapter 1 - introduction to intrusion detection systems. In James Burton, Ido Dubrawsky, Vitaly Osipov, C. Tate Baumrucker, and Michael Sweeney, editors, Cisco Security Professional's Guide to Secure Intrusion Detection Systems, pages 1–38. Syngress, Burlington, 2003. ISBN 978-1-932266-69-6. doi: https://doi.org/10.1016/B978-193226669-6/50021-5. URL https://www.sciencedirect.com/science/article/pii/B9781932266696500215

Wang, Ke & Stolfo, Salvatore. (2004). Anomalous Payload-Based Network Intrusion Detection. 3224. 203-222. 10.1007/978-3-540-30143-1_11.

Heckerman, David. A tutorial on learning with Bayesian networks. Springer Netherlands, 1998.

Naeem, Muhammad Ferjad, et al. "Reliable fidelity and diversity metrics for generative models." International Conference on Machine Learning. PMLR, 2020.

Ring, Markus & Schlör, Daniel & Landes, Dieter & Hotho, Andreas. (2018). Flow-based Network Traffic Generation using Generative Adversarial Networks. Computers & Security. 82. 10.1016/j.cose.2018.12.012.

Goncalves, Andre, et al. "Generation and evaluation of synthetic patient data." BMC medical research methodology 20.1 (2020): 1-40.

Ashrapov, Insaf. "Tabular GANs for uneven distribution." arXiv preprint arXiv:2010.00638 (2020).

Yin, Yucheng, et al. "Practical gan-based synthetic ip header trace generation using netshare." Proceedings of the ACM SIGCOMM 2022 Conference. 2022.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville et Yoshua Bengio, « Generative Adversarial Networks », dans Advances in Neural Information Processing Systems 27, 2014

Sharafaldin, Iman & Habibi Lashkari, Arash & Ghorbani, Ali. (2018). Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization. 108-116. 10.5220/0006639801080116.

Merci !

Suivez-nous sur www.inria.fr