Synthetic Network Traffic Generation for Intrusion Detection Systems: a Systematic Literature Review

Pierre-François Gimenez, Inria

ANUBIS Workshop – September 26th

The issue of data in security

Why do we need data?

- For evaluating security tools, most notably detection
- For using machine learning in cybersecurity

The issue of data in security

Why do we need data?

- For evaluating security tools, most notably detection
- For using machine learning in cybersecurity

Current state of datasets

- We cannot access private data due to confidentiality and privacy reasons
- Public datasets are typically run in testbed with no real users
- They can suffer from mislabelling, network and attack configurations errors, etc.
- Testbeds typically have limited realism

The issue of data in security

Why do we need data?

- For evaluating security tools, most notably detection
- For using machine learning in cybersecurity

Current state of datasets

- We cannot access private data due to confidentiality and privacy reasons
- Public datasets are typically run in testbed with no real users
- They can suffer from mislabelling, network and attack configurations errors, etc.
- Testbeds typically have limited realism

⇒ we cannot confidently evaluate intrusion detection systems because of this dubious quality

The issue of data in security

Why do we need data?

- For evaluating security tools, most notably detection
- For using machine learning in cybersecurity

Current state of datasets

- We cannot access private data due to confidentiality and privacy reasons
- Public datasets are typically run in testbed with no real users
- They can suffer from mislabelling, network and attack configurations errors, etc.
- Testbeds typically have limited realism

⇒ we cannot confidently evaluate intrusion detection systems because of this dubious quality

A solution is to rely on synthetic data

Synthetic network traffic generators

What is synthetic data?

Data generated without simulation or emulation

Categories of generators

- Replay engines
- Maximum throughput generators
- Model-based generators
- High-level generators
- Special scenario generators

Synthetic network traffic generators

What is synthetic data?

Data generated without simulation or emulation

Categories of generators

- Replay engines
- Maximum throughput generators
- **Model-based generators**
- High-level generators
- Special scenario generators

Due to the surge of Generative AI, we focus on model-based generators

Systematic Literature Review

What is it?

A methodology to provide a snapshot of the research work on a certain topic while minimizing biases

Methodology steps

1. defining the scope of the study;

Systematic Literature Review

What is it?

A methodology to provide a snapshot of the research work on a certain topic while minimizing biases

Methodology steps

1. defining the scope of the study;
2. identifying precise research questions;

Systematic Literature Review

What is it?

A methodology to provide a snapshot of the research work on a certain topic while minimizing biases

Methodology steps

1. defining the scope of the study;
2. identifying precise research questions;
3. systematically collecting articles by using regular expressions on specialized search platforms;

Systematic Literature Review

What is it?

A methodology to provide a snapshot of the research work on a certain topic while minimizing biases

Methodology steps

1. defining the scope of the study;
2. identifying precise research questions;
3. systematically collecting articles by using regular expressions on specialized search platforms;
4. reading and analyzing them.

Systematic Literature Review

What is it?

A methodology to provide a snapshot of the research work on a certain topic while minimizing biases

Methodology steps

1. defining the scope of the study;
2. identifying precise research questions;
3. systematically collecting articles by using regular expressions on specialized search platforms;
4. reading and analyzing them.

The rest of the presentation follows this structure

Scope of the study

Scope

Three restrictions of the scope:

- IT environment only, with protocols based on the IP stack
    - OT, 5G, IoT, etc. are excluded
- Model-based generation only
    - a large body of work on rules-based generation is ignored
- Cybersecurity applications only
    - intrusion detection system
    - honeynet
    - red team training, etc.

Research questions

The 7 research questions we investigate

RQ1 *Which communities work on synthetic network traffic generation?*

Research questions

The 7 research questions we investigate

RQ1 *Which communities work on synthetic network traffic generation?*

RQ2 *What are the applications of synthetic network traffic generation in cybersecurity?*

Research questions

The 7 research questions we investigate

RQ1 *Which communities work on synthetic network traffic generation?*

RQ2 *What are the applications of synthetic network traffic generation in cybersecurity?*

RQ3 *What types of data are generated?*

Research questions

The 7 research questions we investigate

RQ1 *Which communities work on synthetic network traffic generation?*

RQ2 *What are the applications of synthetic network traffic generation in cybersecurity?*

RQ3 *What types of data are generated?*

RQ4 *What generation techniques are used?*

Research questions

The 7 research questions we investigate

RQ1 *Which communities work on synthetic network traffic generation?*

RQ2 *What are the applications of synthetic network traffic generation in cybersecurity?*

RQ3 *What types of data are generated?*

RQ4 *What generation techniques are used?*

RQ5 *How are generated data evaluated?*

Research questions

The 7 research questions we investigate

RQ1 *Which communities work on synthetic network traffic generation?*

RQ2 *What are the applications of synthetic network traffic generation in cybersecurity?*

RQ3 *What types of data are generated?*

RQ4 *What generation techniques are used?*

RQ5 *How are generated data evaluated?*

RQ6 *What open-source implementations exist?*

Research questions

The 7 research questions we investigate

RQ1 *Which communities work on synthetic network traffic generation?*

RQ2 *What are the applications of synthetic network traffic generation in cybersecurity?*

RQ3 *What types of data are generated?*

RQ4 *What generation techniques are used?*

RQ5 *How are generated data evaluated?*

RQ6 *What open-source implementations exist?*

RQ7 *What are the performances of the generators?*

Corpus creation

Some rules we followed

- Collection from IEEE Xplore (with regular expression search) and Google Scholar
- Only peer-reviewed articles are kept
- No filter by conference or workshop rank/prestige

Our corpus is composed of 68 articles

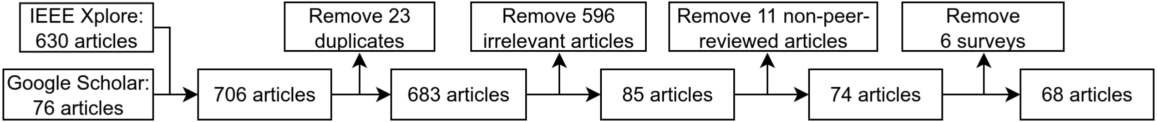
IEEE Xplore: 630 articles
Google Scholar: 76 articles
706 articles
Remove 23 duplicates
683 articles
Remove 596 irrelevant articles
85 articles
Remove 11 non-peer-reviewed articles
74 articles
Remove 6 surveys
68 articles

Figure: PRISMA diagram

Reading and analysis

RQ1

Which communities work on synthetic network traffic generation?

In our corpus:

- 37 articles have been published in networking venues
- 10 in cybersecurity venues
- 4 in AI venues
- 17 in other, generalist venues

Comments

- The cybersecurity community does not seem very interested or aware of these issues
- AI is merely seen as a tool and not a subject of study
- The "other" venues is mostly IEEE Access

RQ2

What are the applications of synthetic network traffic generation in cybersecurity?

We identified two main applications:

Data augmentation synthetic network traffic is generated to alleviate the imbalance of classes in the datasets. The goal is to improve classification results.

Dataset creation synthetic network traffic is generated as an end. Such datasets can be used for a variety of applications: background traffic, IDS evaluation, etc.

RQ2

What are the applications of synthetic network traffic generation in cybersecurity?

We identified two main applications:

Data augmentation synthetic network traffic is generated to alleviate the imbalance of classes in the datasets. The goal is to improve classification results.

Dataset creation synthetic network traffic is generated as an end. Such datasets can be used for a variety of applications: background traffic, IDS evaluation, etc.

Comments

The cybersecurity community focuses on enhancing detection, the networking community focuses building and analyzing datasets

- 50% of cybersecurity articles do data generation
- 68% of networking articles do dataset creation

RQ3

What types of data are generated?

- Traffic statistics (3 articles)
- Flow statistics (41 articles), with a wide variety a granularity
- Packet metadata (25 articles), such as TCP flags, inter-packet arrival time, etc.
- Packet payloads (16 articles), almost always alongside metadata

Applications	Data augmentation	Dataset generation
Traffic	0	3
Flow	**23**	**18**
Packet metadata only	3	8
Packet payload only	1	1
Packet metadata and payload	3	11

Table: Number of articles per types of generated data and applications

RQ4

What Generation Techniques Are Used?

Deep learning techniques:

- GAN (generative adversarial network) and variations: 47 articles
- VAE (variational auto-encoder) and variations: 8 articles
- Diffusion models: 7 articles
- LLM: 5 articles

Other techniques:

- SMOTE: 6 articles
- Bayesian networks: 2 articles

All articles (but one) include at least one deep learning technique!

RQ4

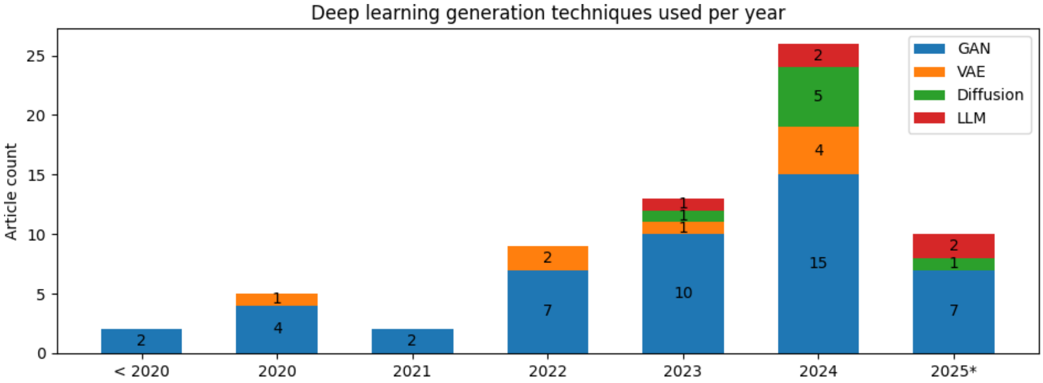

Figure: The number of occurrences of deep learning generation techniques across the years. The year 2025 is incomplete.

A recent surge of Diffusion models and LLM. GAN still very popular

RQ5

How Are Generated Data Evaluated?

A wide variety of methods have been proposed

Train on Synthetic, Test on Real (47 articles)

- TSTR (Train on Synthetic, Test on Real): learning a model on synthetic data (or synthetic + real data) and evaluating it on real data. Assess the utility of the data
- De facto standard metric for data augmentation: 97% of data augmentation articles rely on TSTR

ANUBIS

RQ5

Marginal distribution distance (19 articles)

Features are compared independently. Useful for assessing whether the diversity of the data is properly generated, and if the proportions are correct. However, the relations between features is not assessed.

Pairwise correlations (5 articles)

Pairwise correlation is rarely used. It produces many measurements (nš). It is difficult to compare numerical and category features.

Similarity distance (15 articles)

Compare real points and generated points. Distance is not trivial to define for heterogeneous data.

RQ5

Networking metrics (11 articles)

Metrics specialized for networking application: handwritten "sanity check" rules, success rate of pcap playing, etc.

Qualitative visual evaluation (16 articles)

A qualitative, visual comparison of the distributions of values. Rely on dimension reduction (t-SNE, PCA)

A few articles focus on evaluation and urge to use more metrics to assess realism and diversity

RQ6

What open-source implementations exist?

Method	Year	GitHub Repository	Replicated?
SIP-GAN	2021	amarmeddahi/sip-gan	No
STAN	2021	ShengzheXu/stan	Yes
NetShare	2022	netsharecmu/NetShare	Yes
NeCSTGen	2022	fmeslet/NeCSTGen	Yes
SyNIG	2023	Nirhoshan/SyNIG	No
PAC-GPT	2023	dark-0ne/NetworkPacketGenerator	No
NetDiffusion	2024	noise-lab/NetDiffusion_Generator	No
FlowChronicle	2024	joschac/FlowChronicleCoNEXT	No
CGAN-based-Tool	2024	Alanoud-Subahi/CGAN-based-Tool	No
NetDiffus	2024	Nirhoshan/NetDiffus	Yes
GAN-based	2024	ydataai/ydata-synthetic	Yes
PNetGPT	2025	Snail1502/PNetGPT	No
GAN-based	2025	i-am-junayed/XAI-Intrusion-Detection-System	No
GPT on the wire	2025	javieradelgado/GPT-on-the-wire	No

Table: Number of articles per types of generated data and applications

- A few open source implementations
- Rarely replicated

RQ7

What are the performances of the generators?

The difficulties in comparing methods:

- they generate different kind of data
- they use different metrics
- they are generally not compared to baselines
- they use different datasets

We will focus on the 9 articles that do use baselines

Baselines

Lower baselines

New methods should be better than these

ROS Random OverSampling: duplicates some instances

SMOTE creates new instances with linear combinations of existing ones

ADASYN a variation of SMOTE that focuses of decision boundaries

Naive sampler sample each feature independently

Higher baseline

New methods should be as close as possible to this

Reference actual data sampled from the same distribution

RQ7

Performances on data augmentation

Results per article:

1. SMOTE > CTGAN, Copula GAN, and VAE
2. most resampling strategies negatively impacted the classification performance across all models
3. the best classifiers used no data augmentation or SMOTE
4. WGANGP > SMOTE
5. SMOTE > CNN-based generation
6. PacketCGAN > ROS > GAN

Overall, SMOTE can be very effective, and the supremacy of complex, deep learning models is not proved

RQ7

Performances on dataset generation

Results per article:

- Bayesian networks $>$ CTGAN, E-WGAN-GP, and NetShare. Besides, Naive sampler $>$ NetShare
- Bayesian networks $+$ data mining $>$ TVAE, CTGAN, E-WGAN-GP, GPT2, and NetShare

These results are consistent with what we observe for data augmentation

Conclusion

- Deep learning is very popular in synthetic data generation
- But it requires a lot of data and large training times
- So far, we cannot conclude that it outperforms SMOTE

Why is that?

Some personal hypothesis

- Classical methods (Bayesian networks for example):
  - work well with a limited number of features
  - have theoretical guarantees
  - have other advantages (shorter learning time, explainability, etc.)
- Deep learning have specific issues here:
  - Deep learning assume every feature is numerical (one-shot encoding helps but brings other issues)
  - They typically require a large diversity of data to be trained properly, but network data are not so diverse
  - Empirically, we see that GAN struggle to reproduce pairwise correlation

Challenges

Synthetic network traffic generation could be more useful to the cybersecurity community. We propose three challenges:

Challenge 1

Propose a benchmark, i.e., a standard dataset, set of metrics and baselines to better evaluate and compare generation methods

Challenge 2

Rely on synthetic network traffic generation to provide the cybersecurity community with datasets with concept drift

Challenge 3

Build long-term datasets with APT-like (multi-step, long-time) attacks for the cybersecurity community