Towards attack detection in traffic data based on spectral graph analysis

Majed Jaber, Pierre Parrend, Nicolas Boutry

University of Strasbourg

ICUBE

cnrs

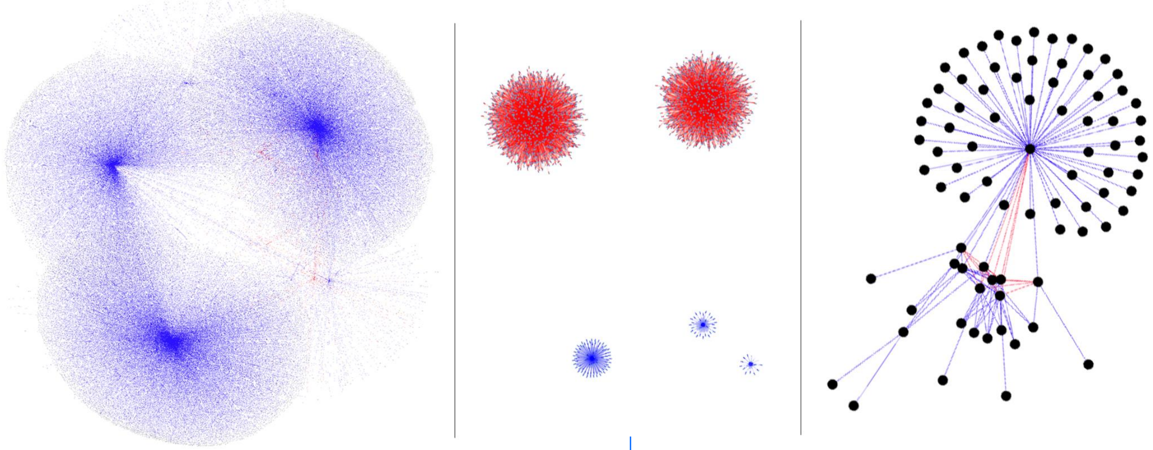
EPITA

Outline

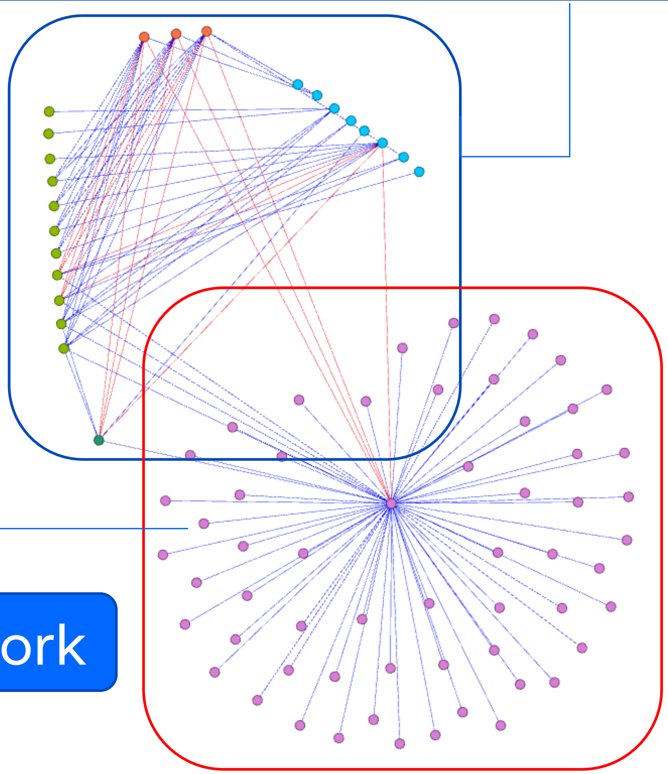
- Cybersecurity and cyberattacks
- A network can be modeled by a (dynamical) graph
- Anomaly Detection, the State-of-the-Art
- Spectral Graph Analysis, a new approach for cybersecurity
- Experiments & Evaluation
- Overlook over the Notebooks
- Future works

Cybersecurity against attacks

Graph represents a networks



Identify patterns



Decompose different patterns

Understand the network

State-of-the-Art

1

Statistical Approaches

A real-time
network anomaly-detector
(ReTiNA)

Traditional systems use elementary statistics techniques and are often inaccurate

2

ML Approaches

CAMLPAD model
anomalies are assigned an outlier score
ML-based techniques are supervised algorithms

In network security, there are not much labeled data to train efficient classifiers

3

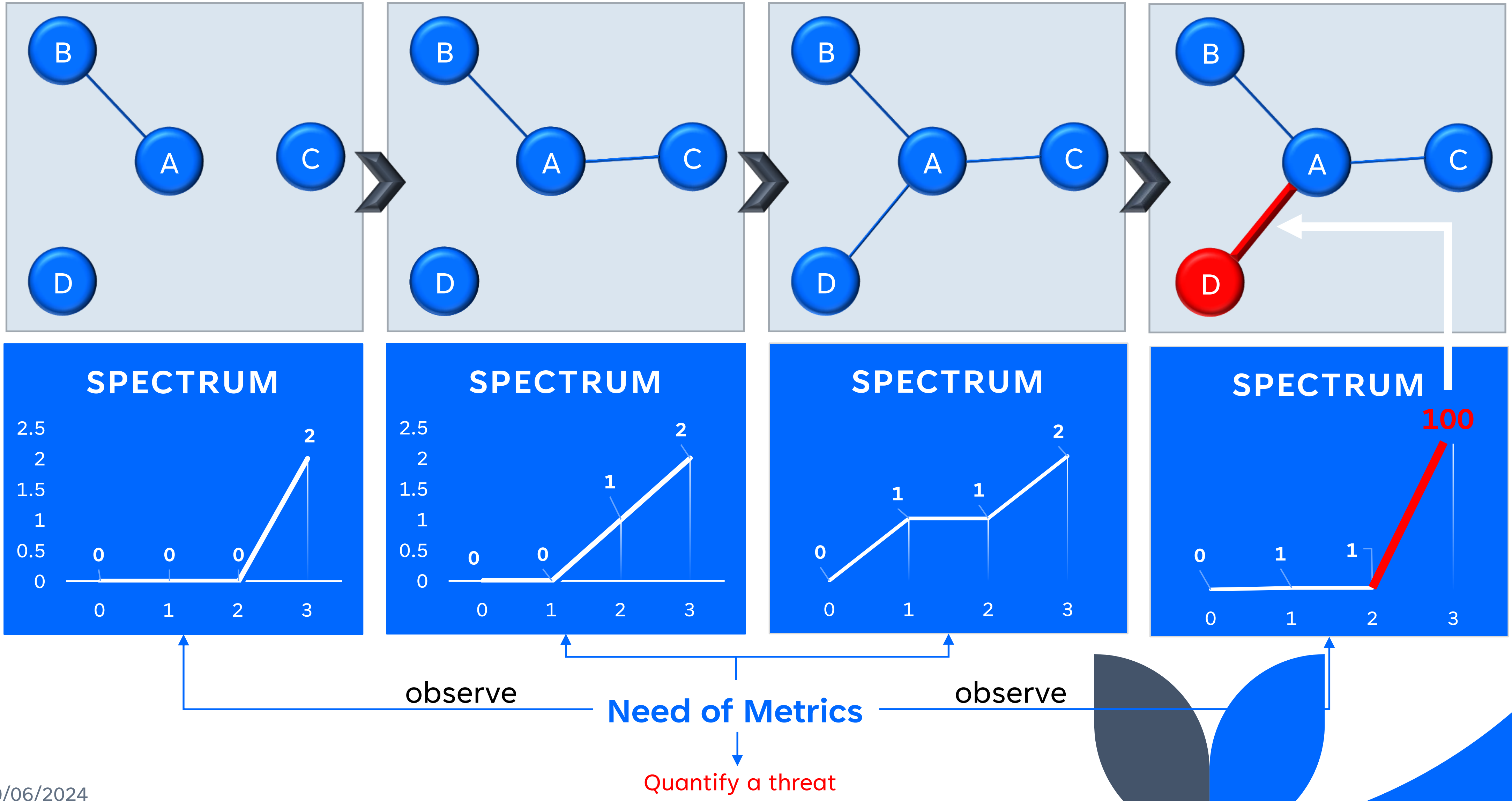
GCN Approaches

One of the best choice for graph data learning tasks

The Dynamic Graph Neural Networks (DGNNs) are known to be an interesting tool to detect anomalies in complex dynamic graphs

- Noble, J., Adams, N.: Real-time dynamic network anomaly detection. IEEE Intelligent Systems 33(2), 5–18 (2018)
- Hariharan, A., Gupta, A., Pal, T.: Camlpad: Cybersecurity autonomous machine learning platform for anomaly detection. In: Future of Information and Communication Conference. pp. 705–720. Springer (2020)
- Bowman, B., Huang, H.H.: Towards next-generation cybersecurity with graph ai. ACM SIGOPS Operating Systems Review 55(1), 61–67 (2021)
- Weifeng Liu, Sichao Fu, Yicong Zhou, Zheng-Jun Zha, and Liqiang Nie. Human activity recognition by manifold regularization based dynamic graph convolutional networks. Neurocomputing, 444:217–225, 2021.

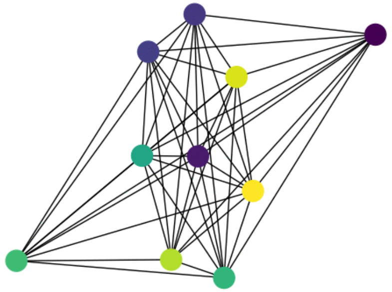
Why Spectral graph analysis?

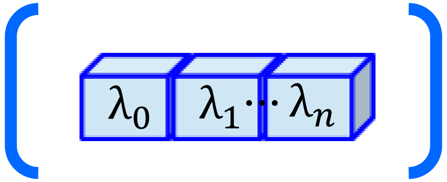


Spectral graph analysis

Studying the spectrum of
the Laplacian Matrix

Mathematical techniques

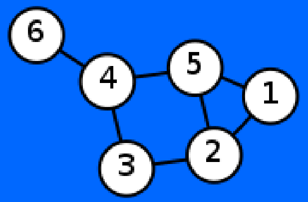




Analyze graph properties

Feature extraction

Laplacian Matrix



$$L = D - A$$

$$L = \begin{pmatrix} 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 2 & -1 & 0 & 0 & -1 & 0 \\ -1 & 3 & -1 & 0 & -1 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & -1 & 3 & -1 & -1 \\ -1 & -1 & 0 & -1 & 3 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 \end{pmatrix}$$

$$A_{i,j} := \begin{cases} 1 & if\ i \neq j \text{ and } v_i \sim v_j \\ 0 & \text{otherwise} \end{cases}$$

$$D_{i,j} := \begin{cases} \deg(v_i) & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

$$L_{i,j} := \begin{cases} \deg(v_i) & \text{if } i = j \\ -1 & \text{if } i \neq j \text{ and } v_i \text{ is adjacent to } v_j \\ 0 & \text{otherwise,} \end{cases}$$

used?

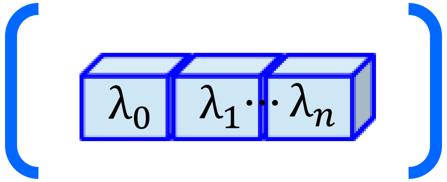
ph analysis is

other

ph structure.
atrix are used in
graph analysis,
etection, and

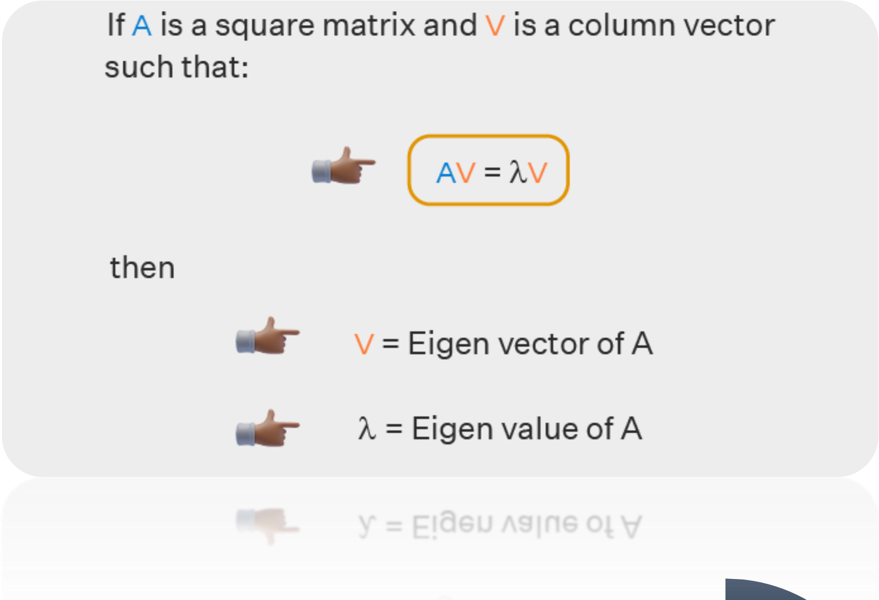
What is a spectrum?

the spectrum refers to the set of **eigenvalues** of the **Laplacian matrix**.



Spectrum

Eigenvalues



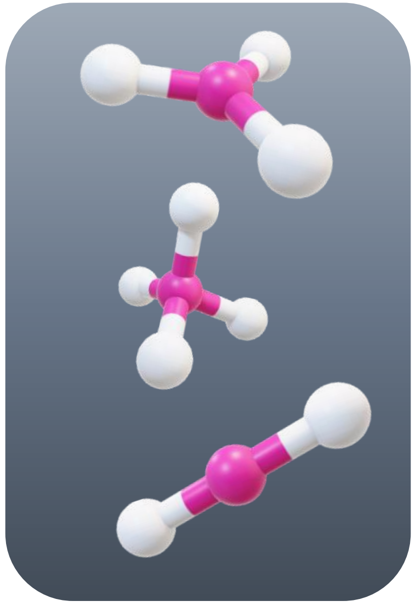
Chung, F. R. (1997). Spectral graph theory (Vol. 92). American Mathematical Soc.

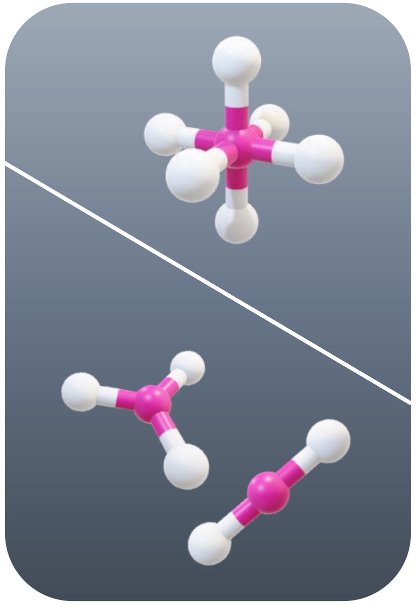
Spectrum Interesting eigenvalues

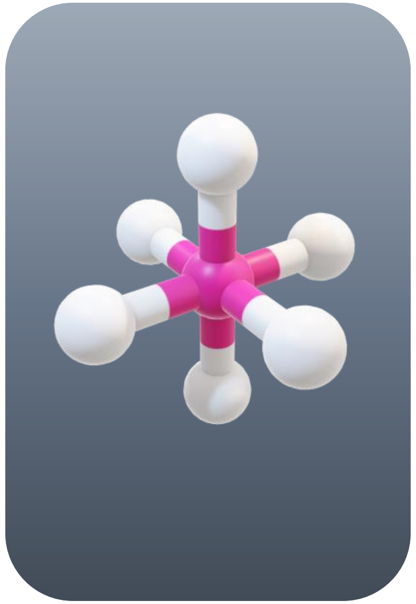
Zero eigenvalues

Algebraic connectivity

Largest eigenvalues

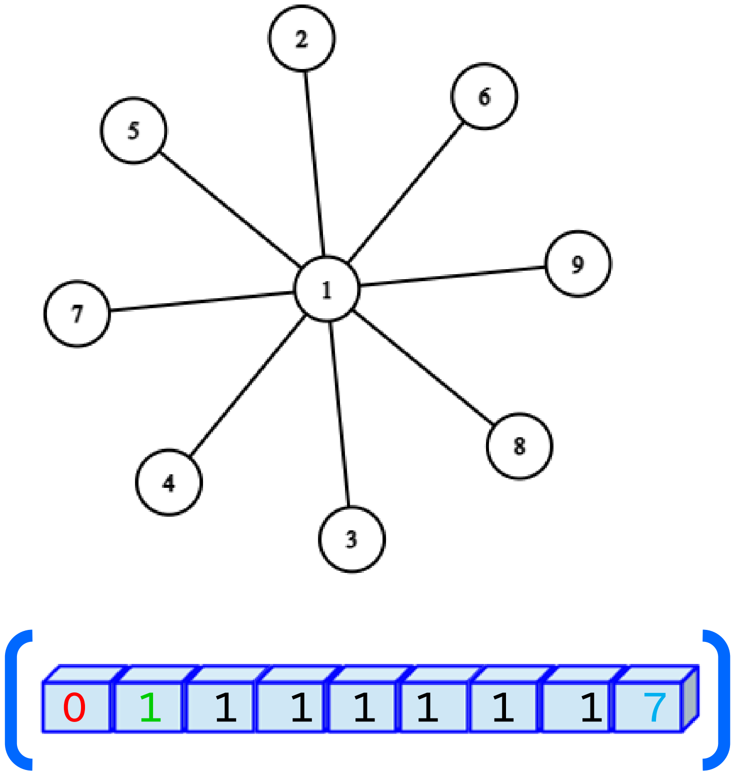


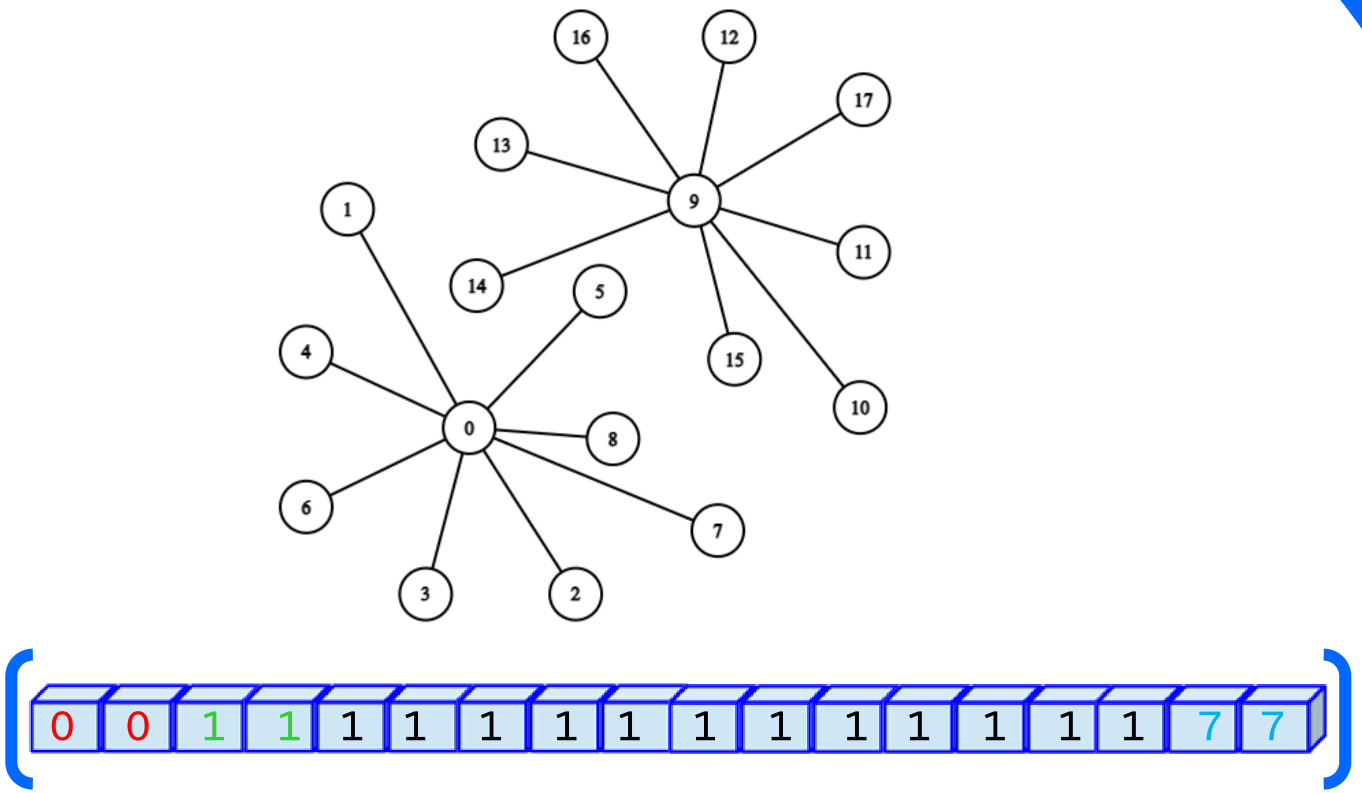




- De Abreu, N. M. M. (2007). Old and new results on algebraic connectivity of graphs. Linear algebra and its applications, 423(1), 53-73.
- Bauer, F., Jost, J.: Bipartite and neighborhood graphs and the spectrum of the normalized graph laplacian. arXiv preprint arXiv:0910.3118 (2009)

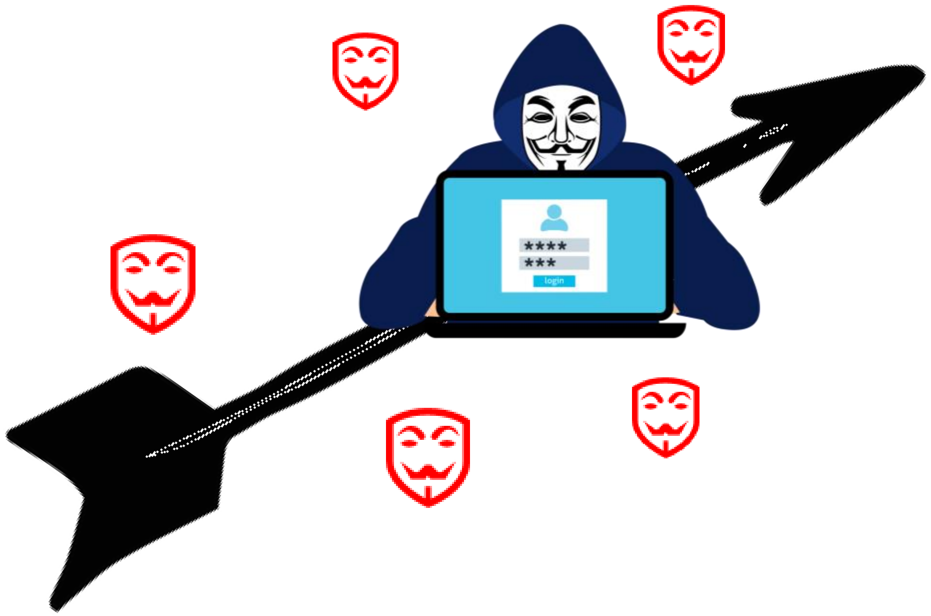
Spectrum Interesting EV - Example

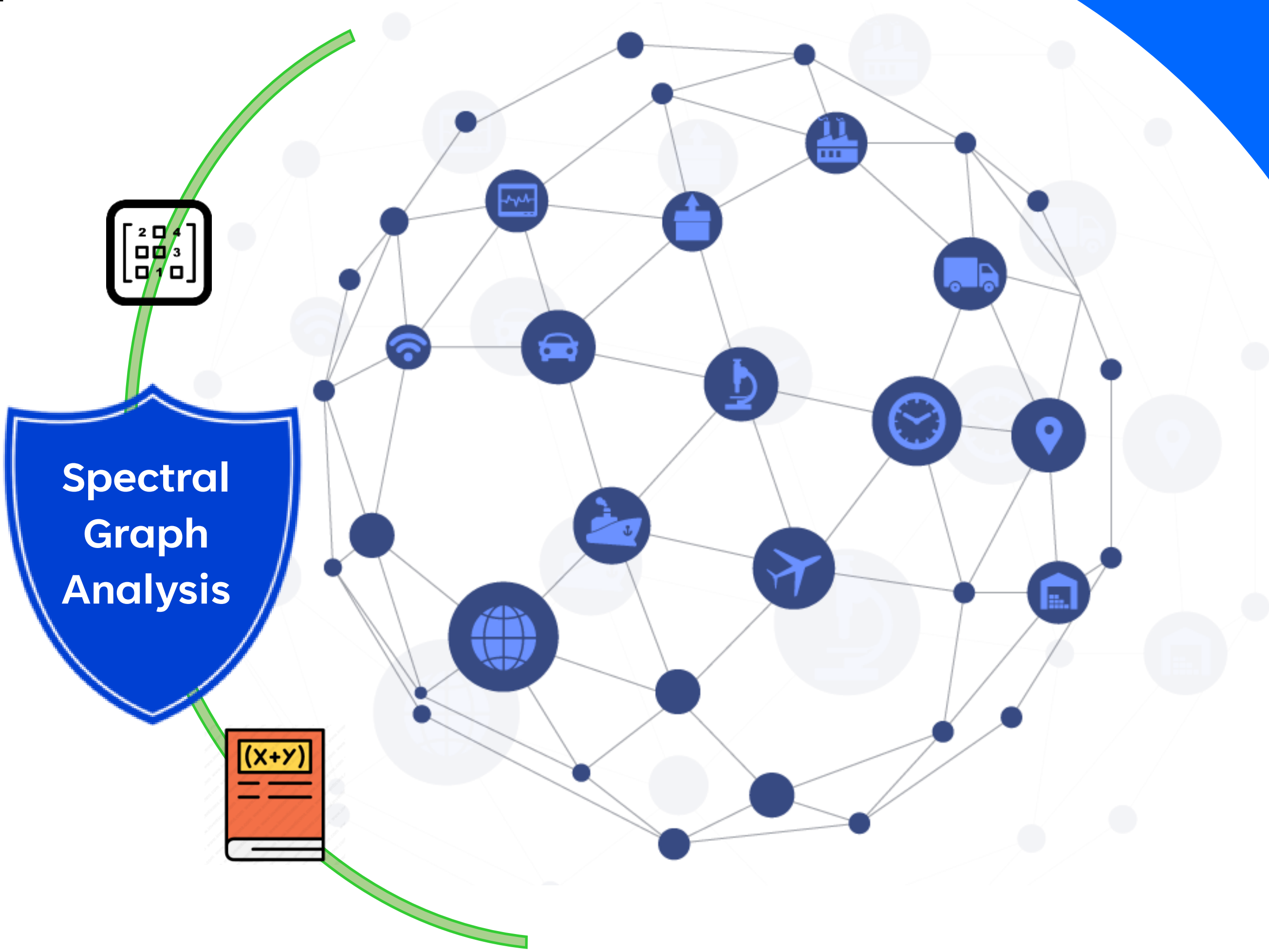




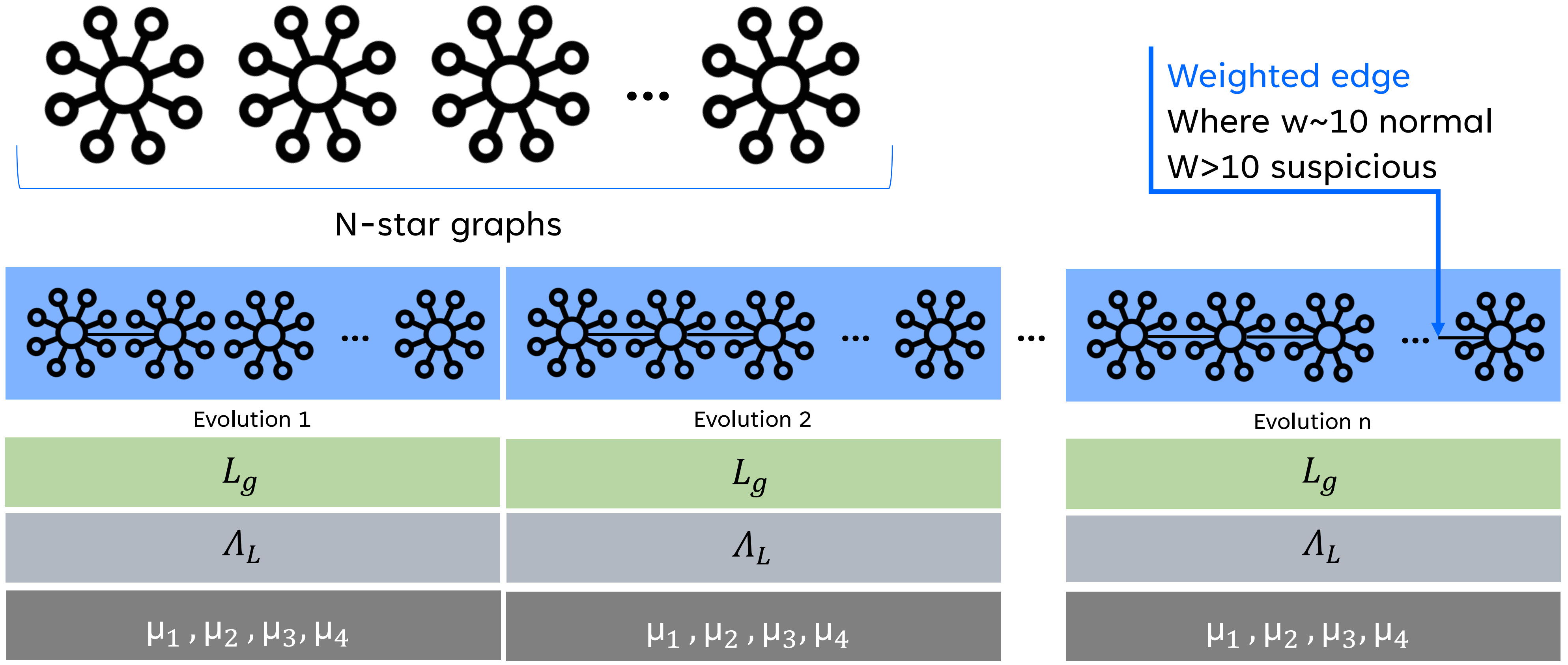
Research Question

How can we benefit from spectral graph analysis to identify and detect cyberattacks over the network?





Methodology



Dynamic Metrics

Metric 1 **Connectedness**

- Increases when interconnections occur in the network.

Metric 2 **Flooding**

- This metric is influenced by the occurrence of connections as well as the weight of those connections.

Metric 3 **Wiringness**

- It always increases when connections occur and its slope across time depends on the packets sizes.

Metric 4 **Asymmetry**

- It corresponds to the number of variations of $\Lambda(t)$ and the symmetry of the graph

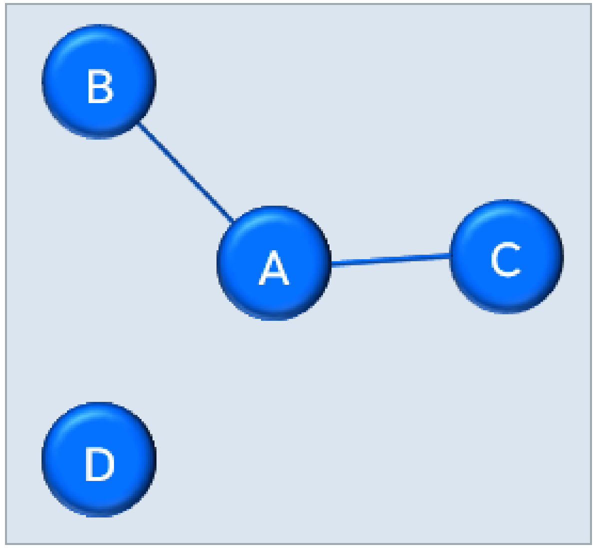
Metric 1 - Connectedness

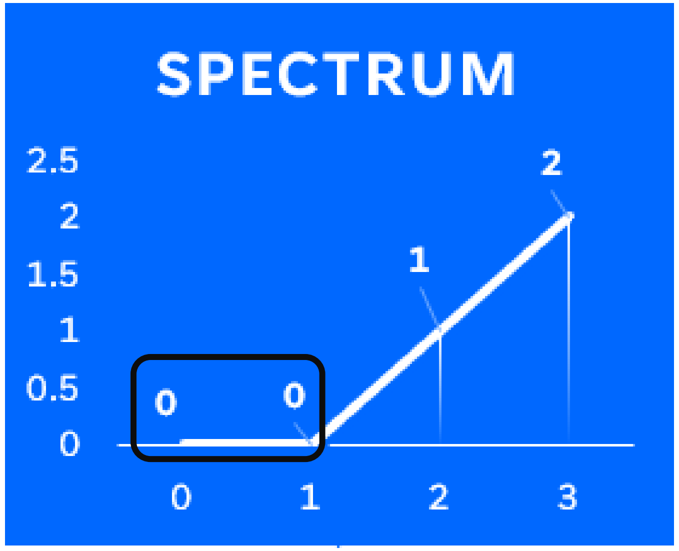
$$\mu_1(t) = \frac{\exp \frac{1}{\mathcal{Z}(t)}}{\exp(1)}$$

$\mathcal{Z}(t)$ number of zeros in the spectrum.

$$\lim_{Z(t)\to\infty} \mu_1 = e^{-1}$$

$$\lim_{Z(t)\to 1} \mu_1 = 1$$



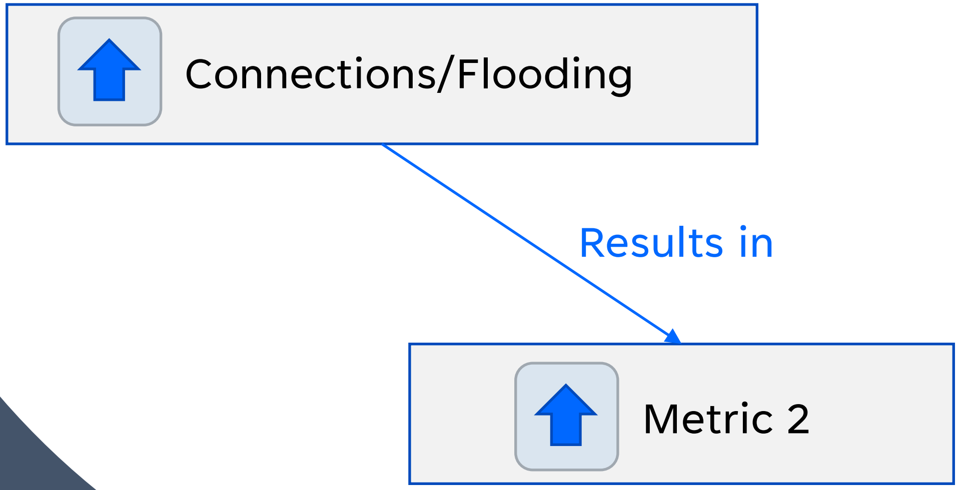


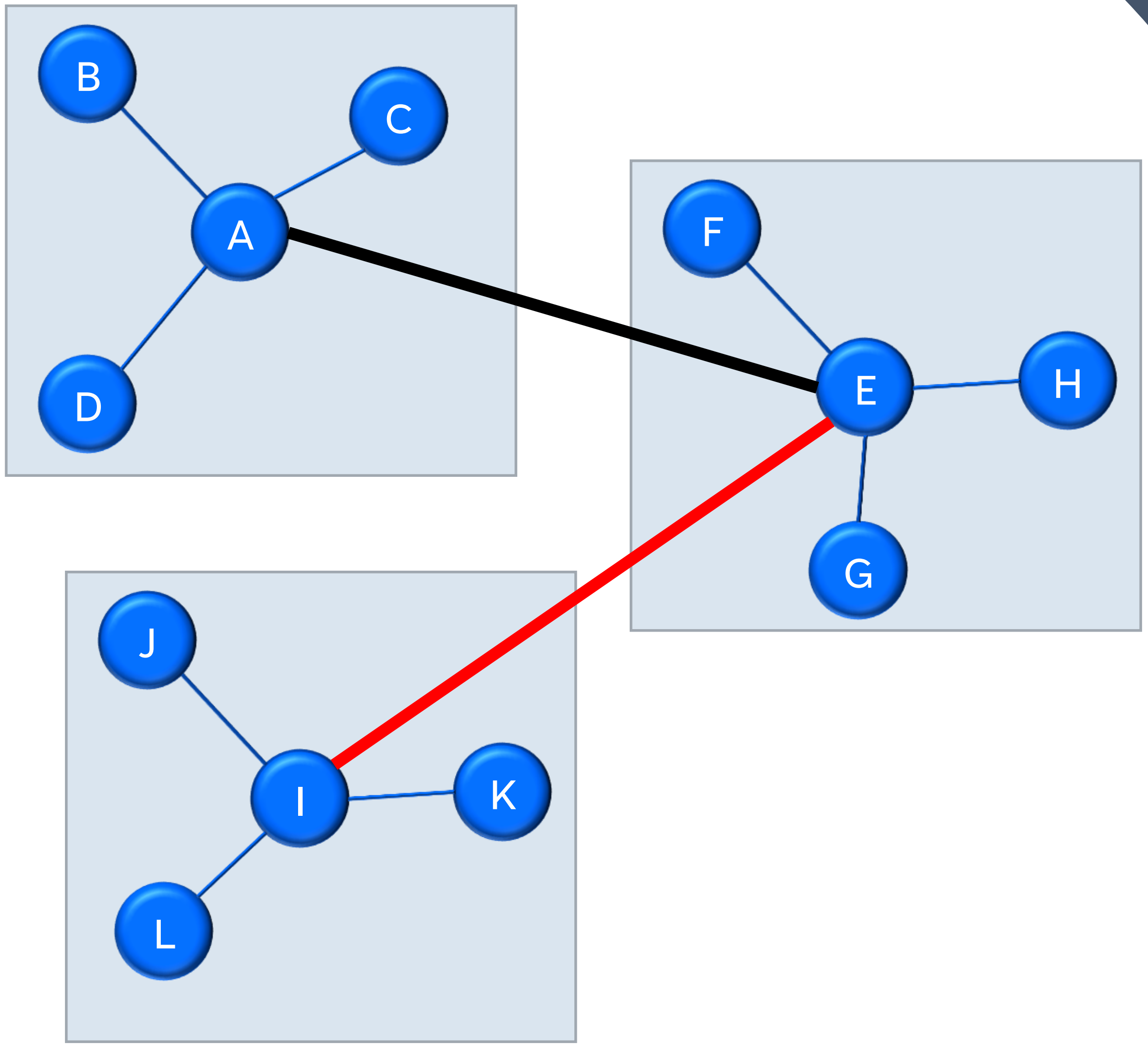
Chung, F. R. (1997). Spectral graph theory (Vol. 92). American Mathematical Soc.

Metric 2 - Flooding

$$\mu_2(t) = \left(\frac{1}{\mathcal{N}} \sum_{i=\mathcal{Z}(t)+1}^{\mathcal{Z}(t)+\mathcal{N}} \lambda_i^{(t)} \right) - 1$$

$\mathcal{N}$ is the number of servers/hubs





De Abreu, N. M. M. (2007). Old and new results on algebraic connectivity of graphs. Linear algebra and its applications, 423(1), 53-73.

Metric 3 - Wiringness

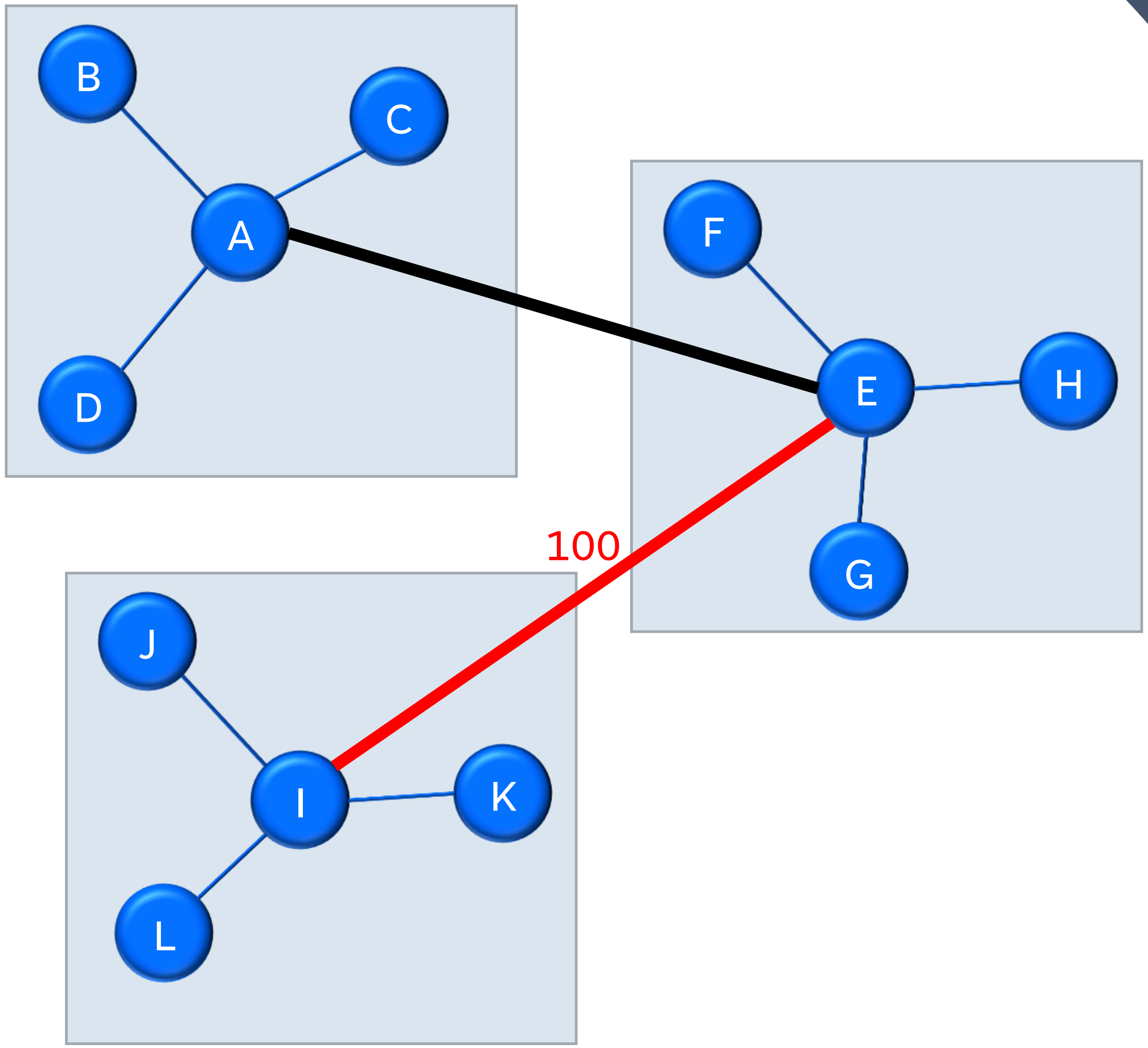
$$\mu_3(t) = \frac{1}{\mathcal{N}} \sum_{i=n-\mathcal{N}+1}^{n} \lambda_i^{(t)}$$

$\mathcal{N}$ is the number of servers/hubs

pkts/bytes

Results in

Metric 3



Bauer, F., Jost, J.: Bipartite and neighborhood graphs and the spectrum of the normalized graph laplacian. arXiv preprint arXiv:0910.3118 (2009)

Metric 4 - Asymmetry

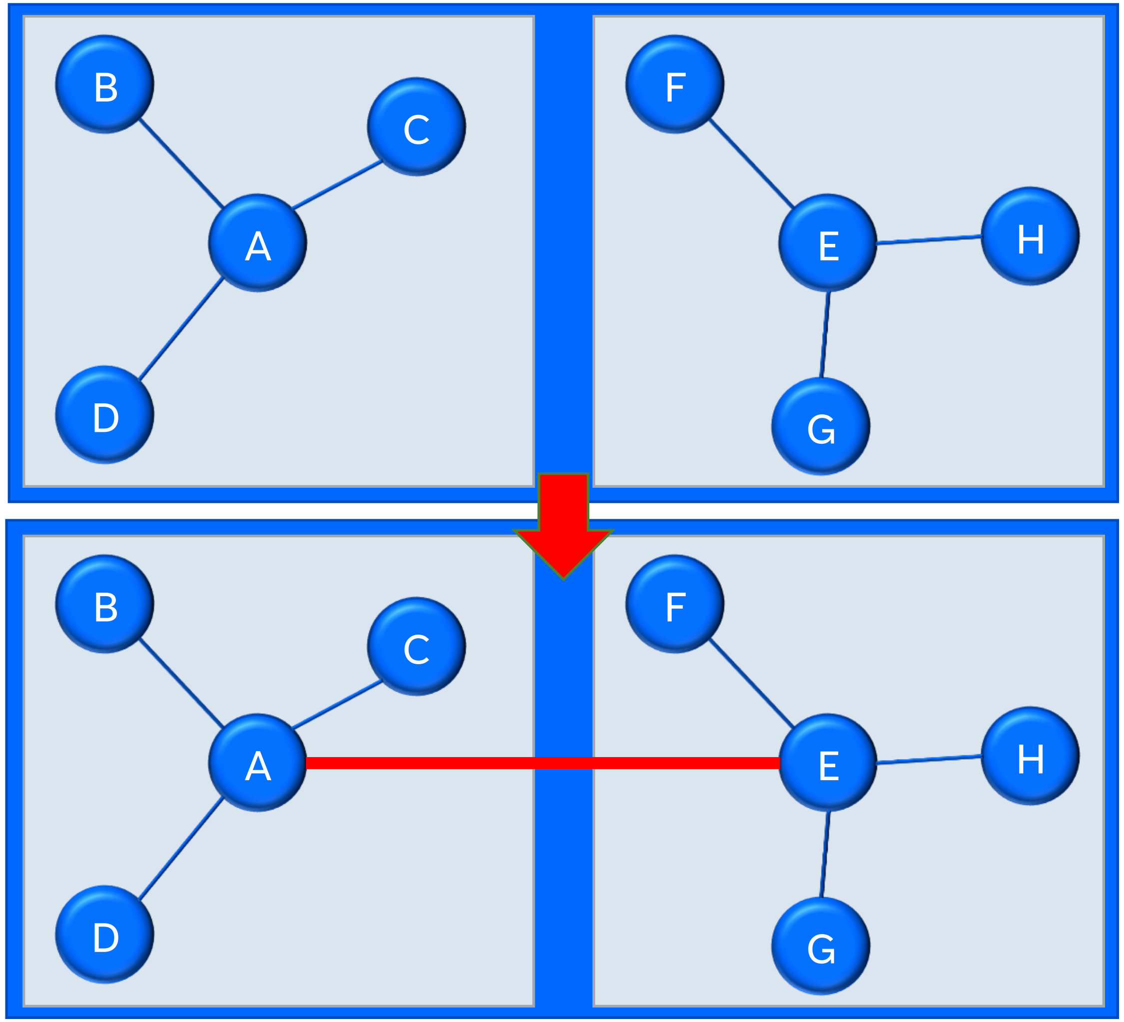
$$\mu_4(t) = Cardinality\left(\left\{i \in [2,n] \;;\; \lambda_i^{(t)} - \lambda_{i-1}^{(t)} > \varepsilon\right\}\right)$$

with $\varepsilon = 10^{-12}$

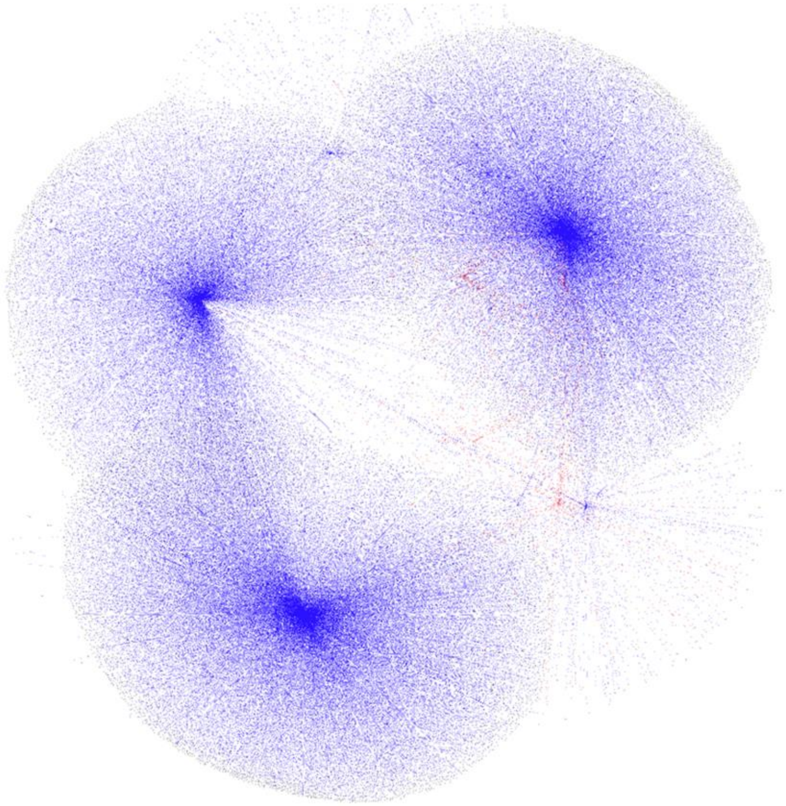
Identical patterns/symmetry

Results in

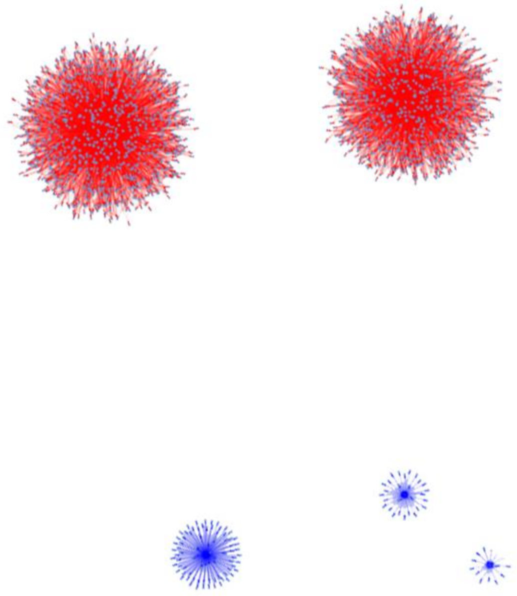
Metric 4



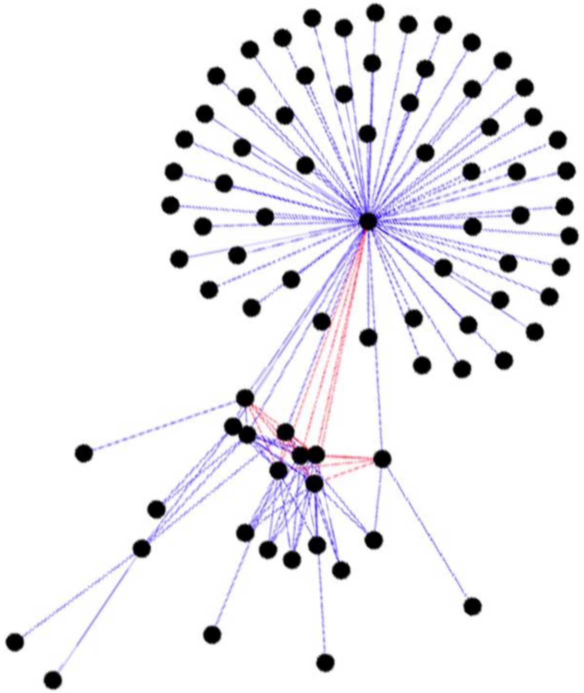
Implementation and datasets



Ton IoT



Healthcare IoT

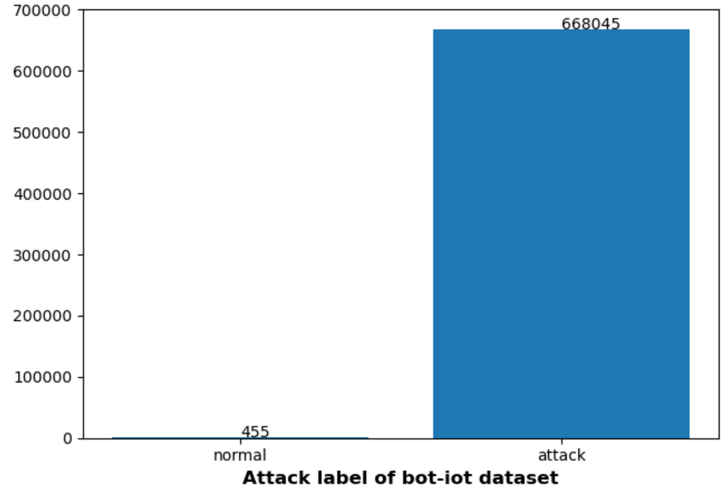


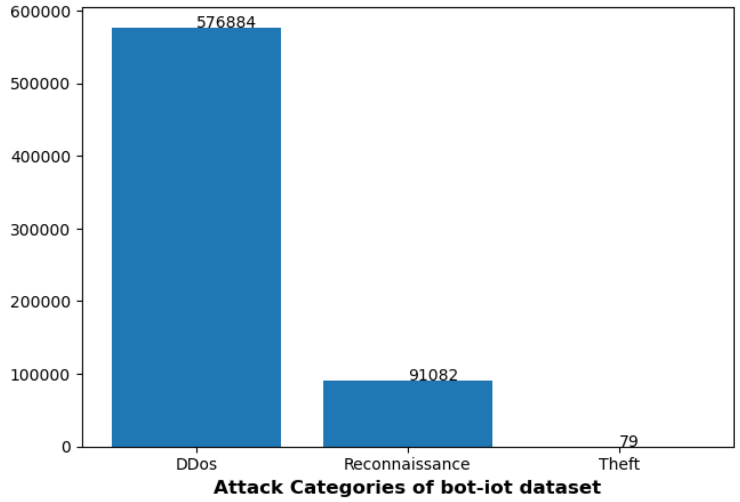
Botnet IoT

[Boo+21] Tim M Booij et al. “ToN_IoT: The role of heterogeneity and the need for standardization of features and attack types in IoT network intrusion data sets”. In: IEEE Internet of Things Journal 9.1 (2021), pp. 485–496.
[Kor+19] Nickolaos Koroniotis et al. “Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-iot dataset”. In: Future Generation Computer Systems 100 (2019), pp. 779–796.
[hussain2021iot] Hussain, F., Abbas, S. G., Shah, G. A., Pires, I. M., Fayyaz, U. U., Shahzad, F., ... & Zdravevski, E. (2021). IoT Healthcare Security Dataset. IEEE Dataport.

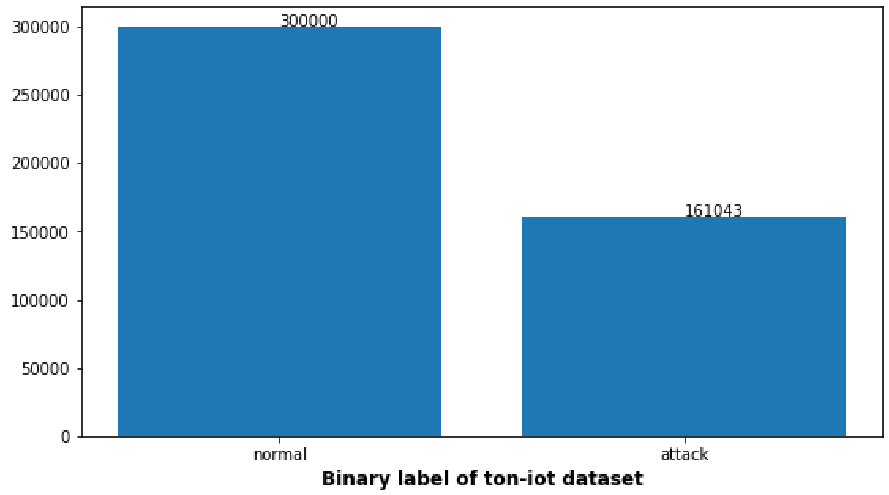
Attack analysis

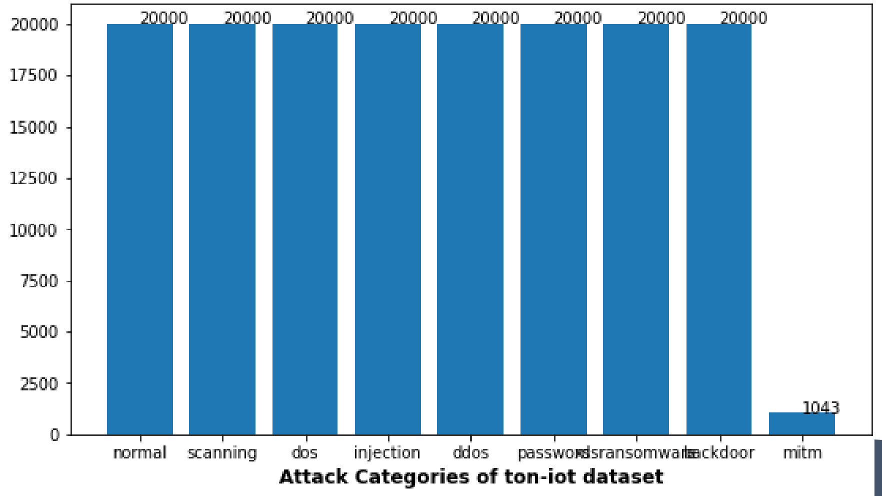
Botnet IoT dataset



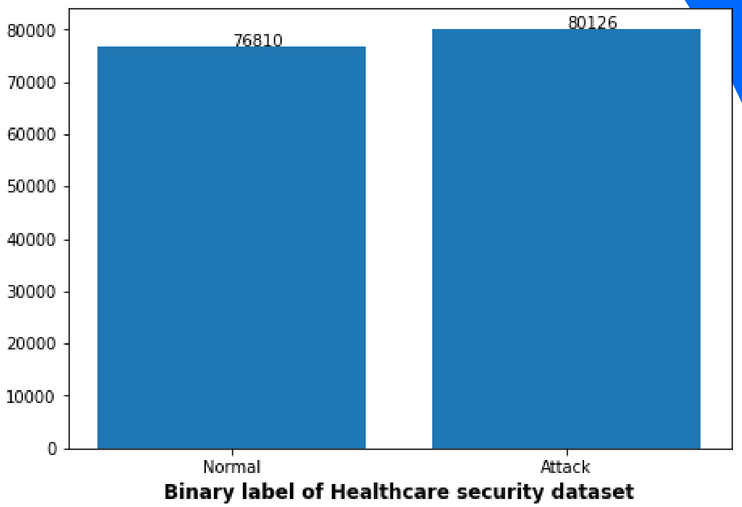


Ton IoT dataset

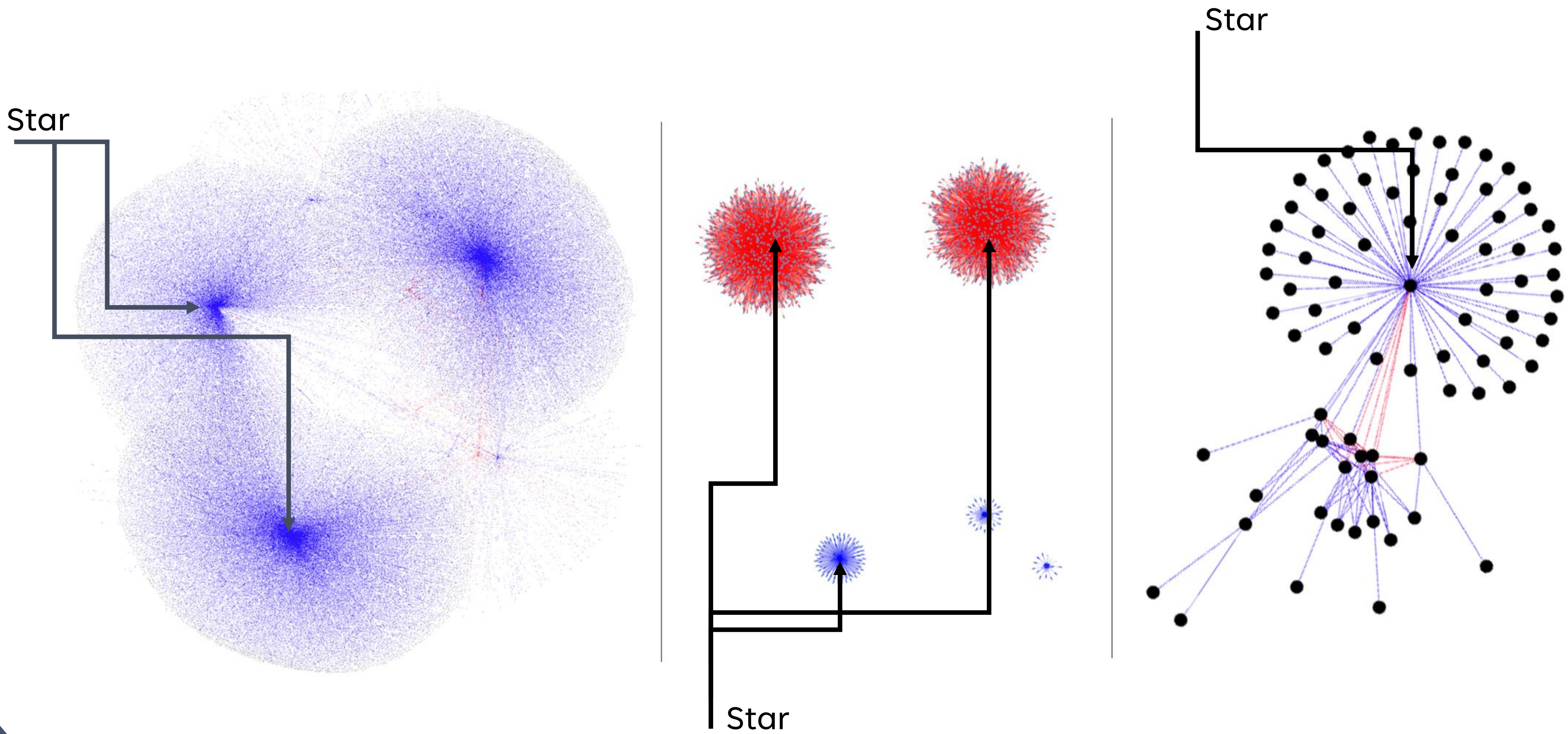




IoT Healthcare Security Dataset

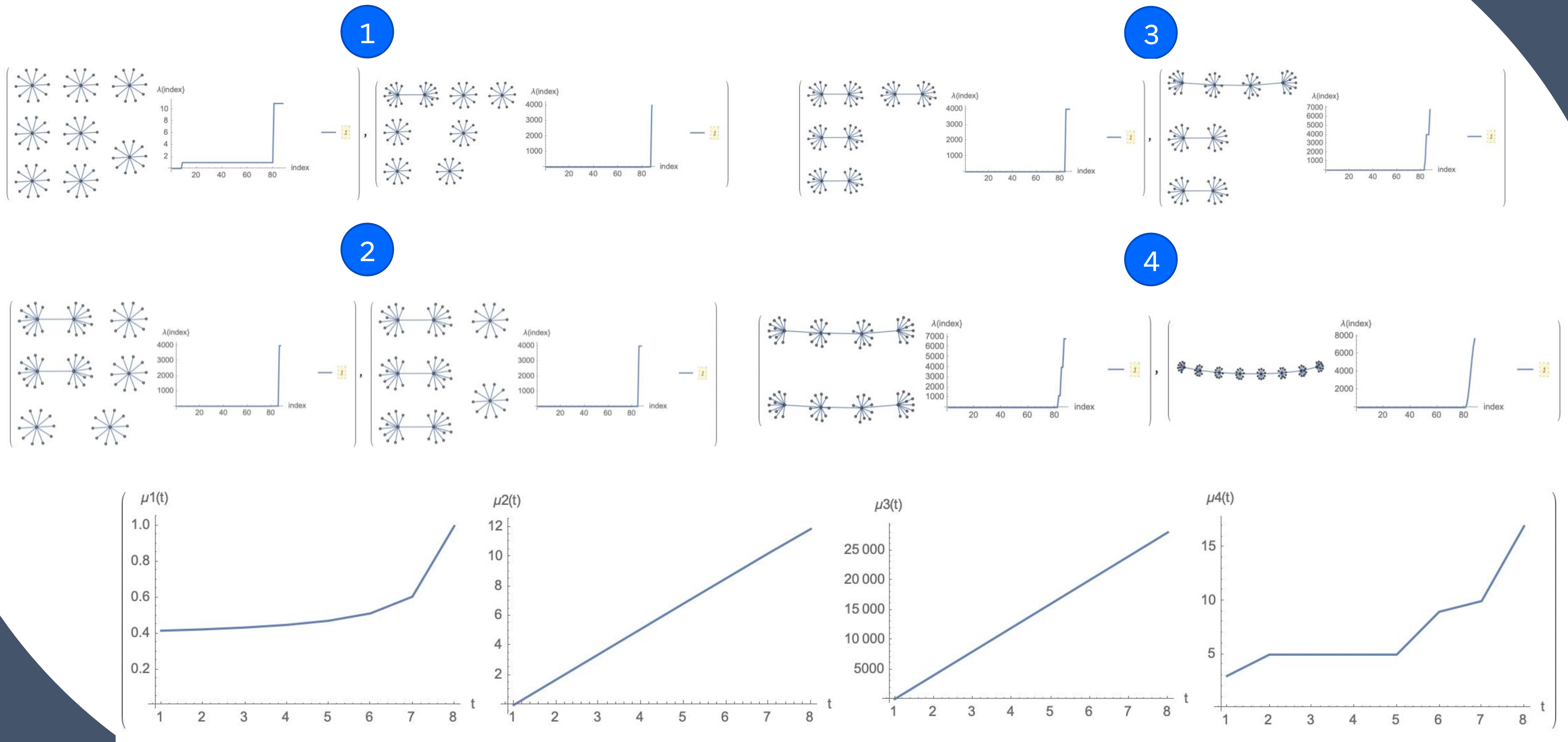


Network patterns – Star graphs

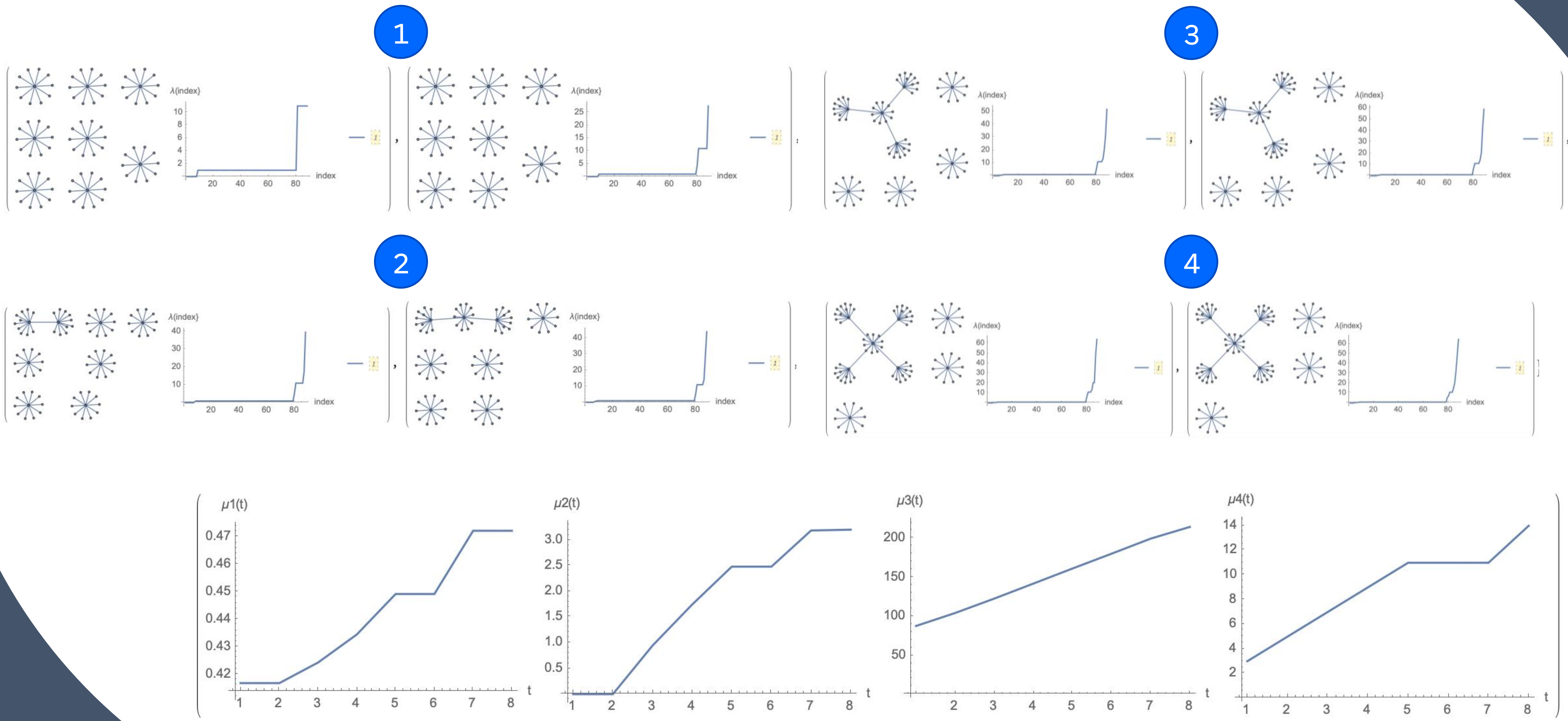


[Boo+21] Tim M Booij et al. "ToN_IoT: The role of heterogeneity and the need for standardization of features and attack types in IoT network intrusion data sets". In: IEEE Internet of Things Journal 9.1 (2021), pp. 485–496.
[Kor+19] Nickolaos Koroniotis et al. "Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-iot dataset". In: Future Generation Computer Systems 100 (2019), pp. 779–796.
[hussain2021iot] Hussain, F., Abbas, S. G., Shah, G. A., Pires, I. M., Fayyaz, U. U., Shahzad, F., ... & Zdravevski, E. (2021). IoT Healthcare Security Dataset. IEEE Dataport.

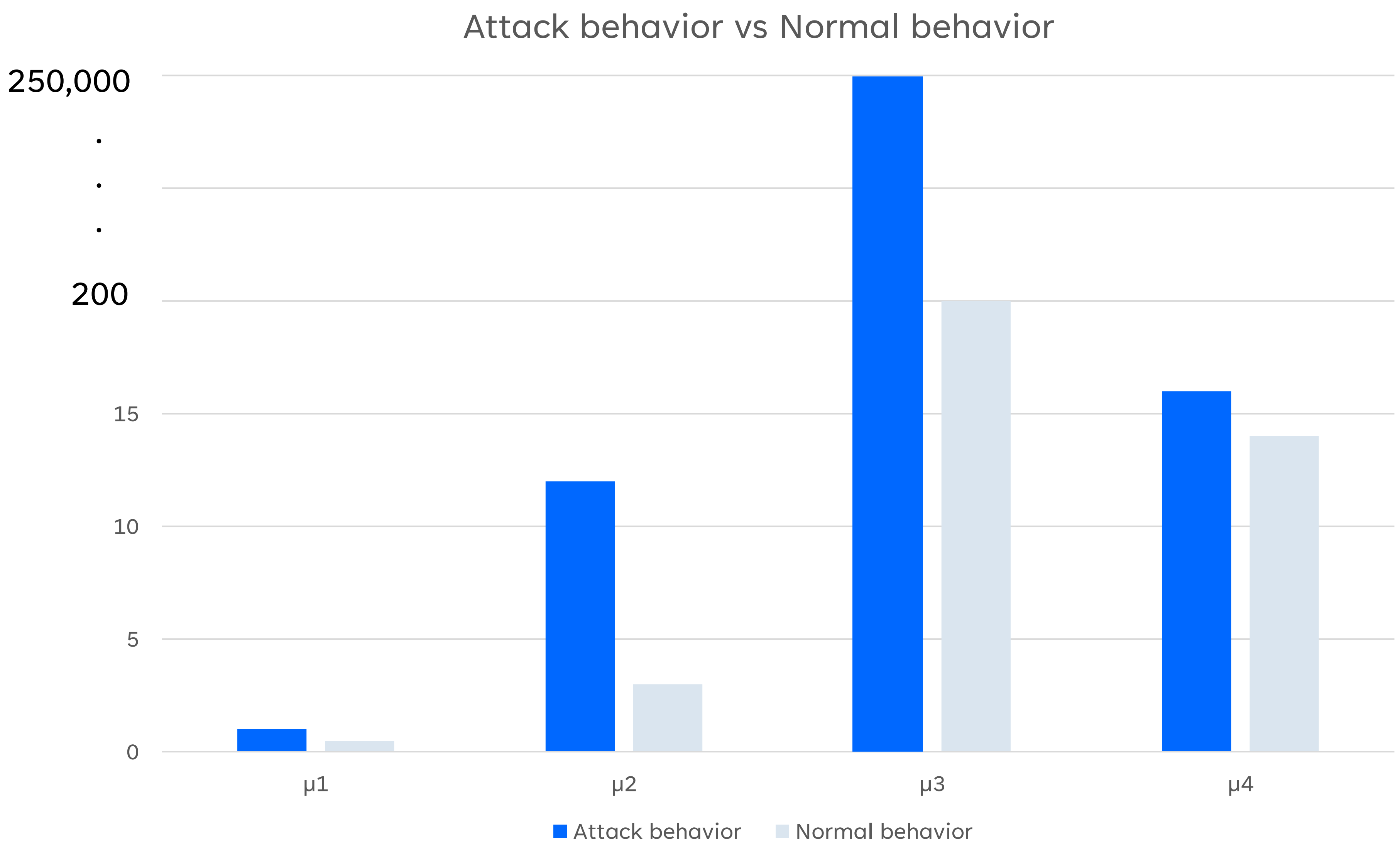
Experiments – Scenario 1 – Attack behavior



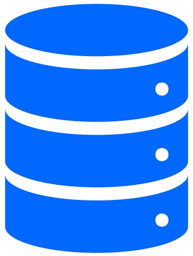
Experiments – Scenario 2 – Normal behavior



Experiments Evaluation



Metrics over real dataset

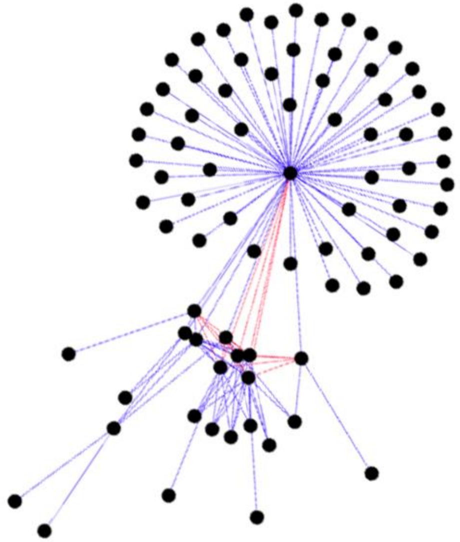


Dataset

Apply metrics on dataset

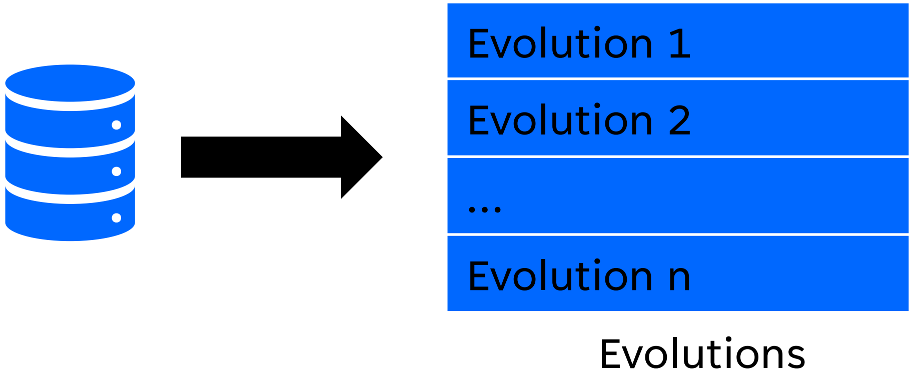
Connectedness
Flooding
Wireness
Assymetry

Detect attacks



Challenges over real datasets

	stime	saddr	daddr	pkts	label
576923	1526344032	192.168.100.46	192.168.100.5	59452	0
576917	1526344032	192.168.100.46	192.168.100.5	30157	1
576916	1526344032	192.168.100.46	192.168.100.5	29726	0
576921	1526344032	192.168.100.3	13.55.154.73	3018	0
576884	1526344121	192.168.100.1	192.168.100.3	4	0



From dataset to timeseries

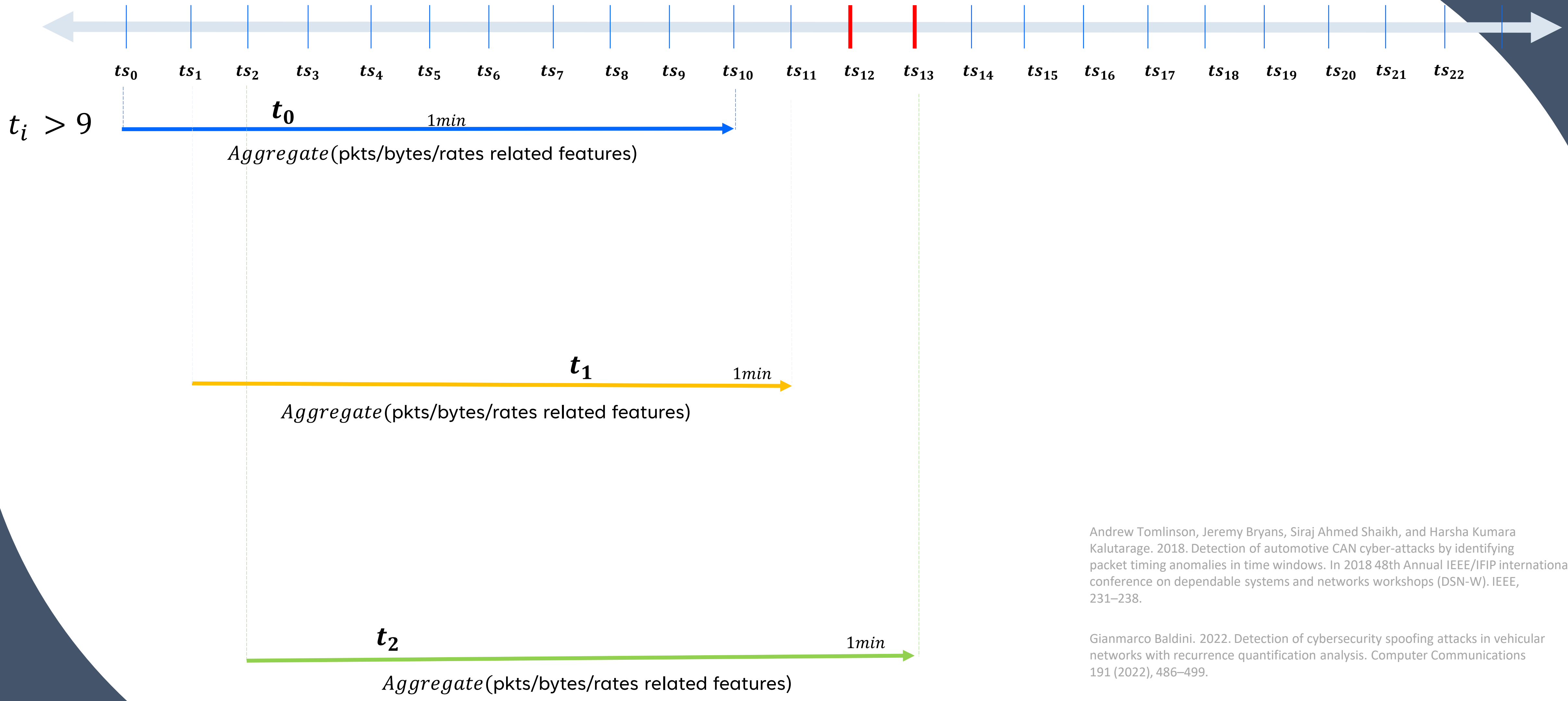
Merge

	stime	saddr	daddr	pkts	attack	weight
576923	1526344032	192.168.100.46	192.168.100.5	59452	0	1
576917	1526344032	192.168.100.46	192.168.100.5	30157	0	1
576916	1526344032	192.168.100.5	192.168.100.3	29726	0	1
576921	1526344033	192.168.100.7	13.55.154.75	3018	0	1
576884	1526344121	192.168.100.1	192.168.100.3	4	0	1

	stime	saddr	daddr	pkts	attack	requests
0	1526344032	192.168.100.46	192.168.100.5	$\sum pkts = 89{,}609$	0	$\sum weight = 2$
		192.168.100.3	13.55.154.73	$\sum pkts = 29726$	0	1
1	1526344033	192.168.100.7	13.55.154.75	$\sum pkts = 3018$	0	1
2	1526344121	192.168.100.1	192.168.100.3	$\sum pkts = 4$	0	1

Sergio Iglesias Pérez, Santiago Moral-Rubio, and Regino Criado. 2021. A new approach to combine multiplex networks and time series attributes: Building intrusion detection systems (IDS) in cybersecurity. Chaos, Solitons & Fractals 150 (2021), 111143.

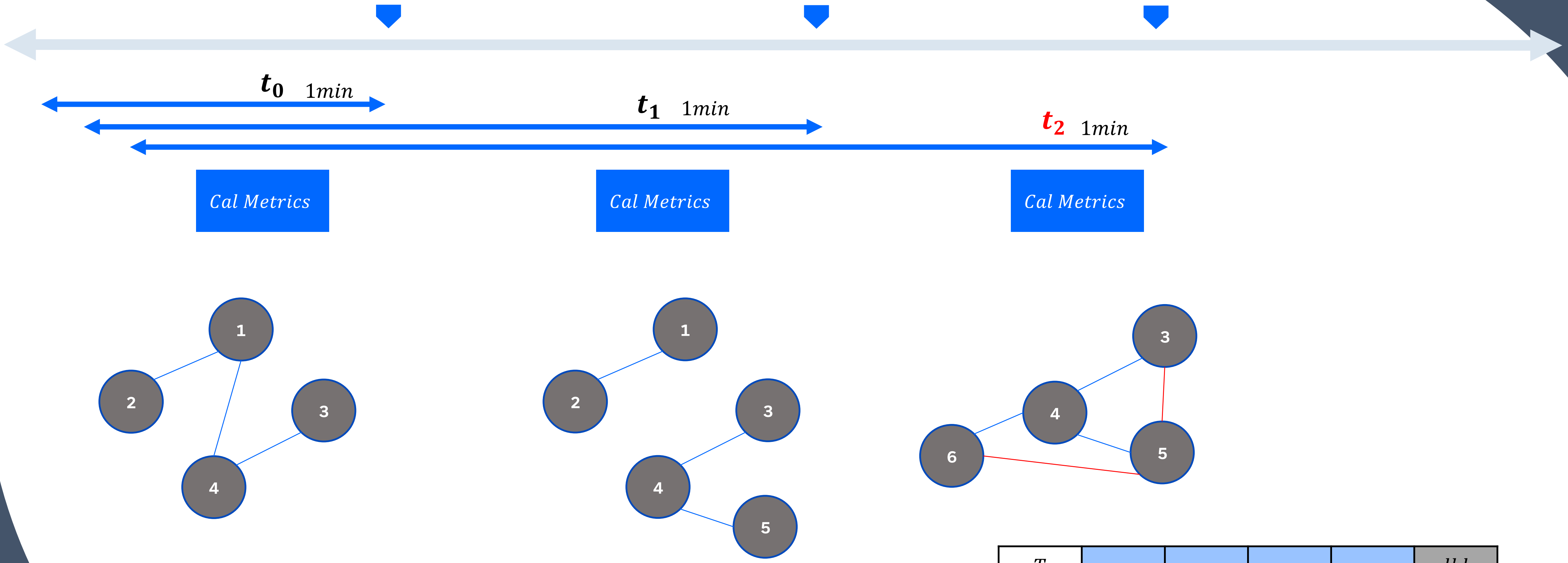
Time-windowing over time-series



Andrew Tomlinson, Jeremy Bryans, Siraj Ahmed Shaikh, and Harsha Kumara Kalutarage. 2018. Detection of automotive CAN cyber-attacks by identifying packet timing anomalies in time windows. In 2018 48th Annual IEEE/IFIP international conference on dependable systems and networks workshops (DSN-W). IEEE, 231–238.

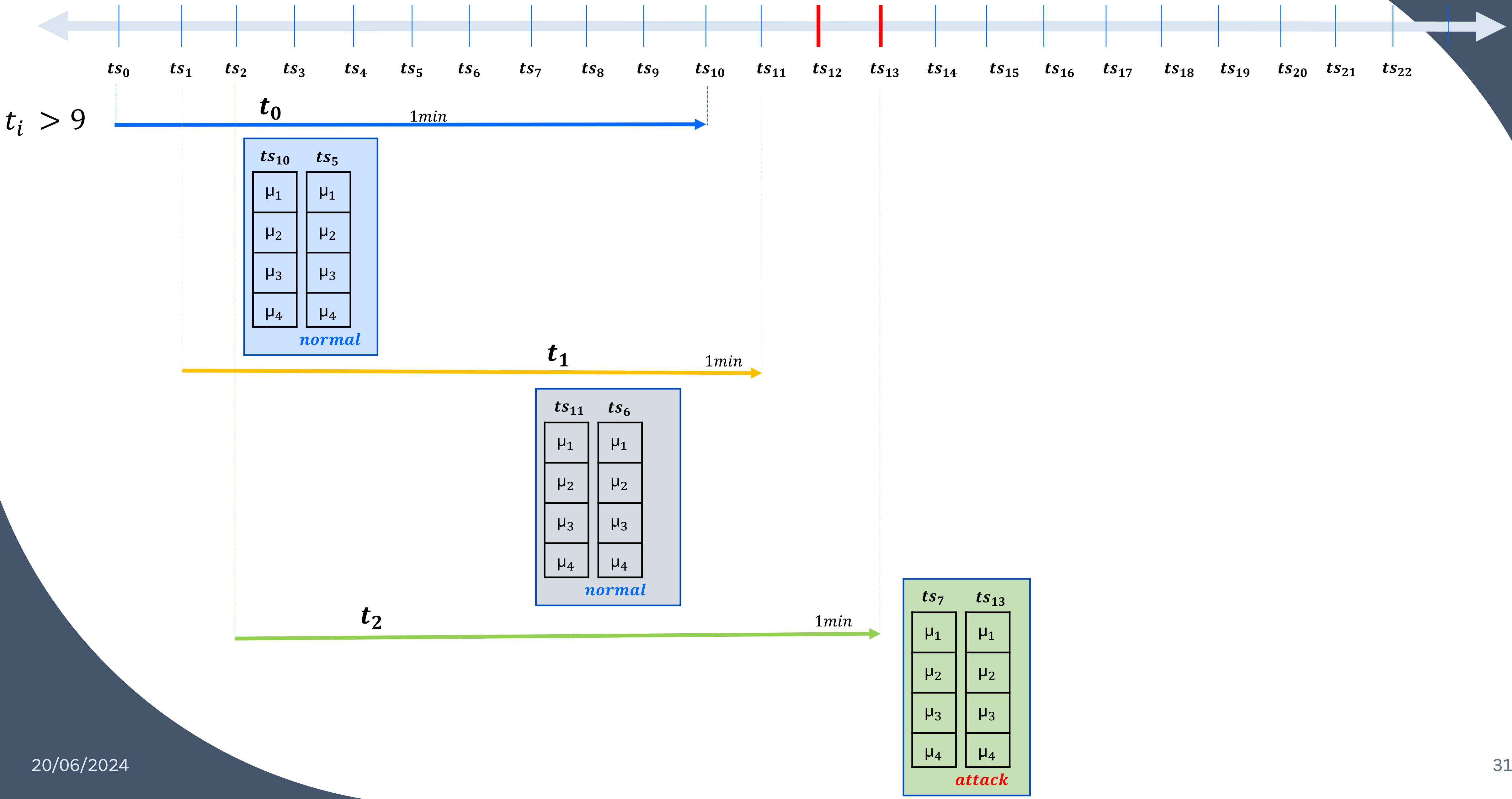
Gianmarco Baldini. 2022. Detection of cybersecurity spoofing attacks in vehicular networks with recurrence quantification analysis. Computer Communications 191 (2022), 486–499.

Time-windowing with spectral metrics



T	μ_1	μ_2	μ_3	μ_4	lbl
t_0	…	…	…	…	0
t_1	…	…	…	…	0
t_2	…	…	…	…	1

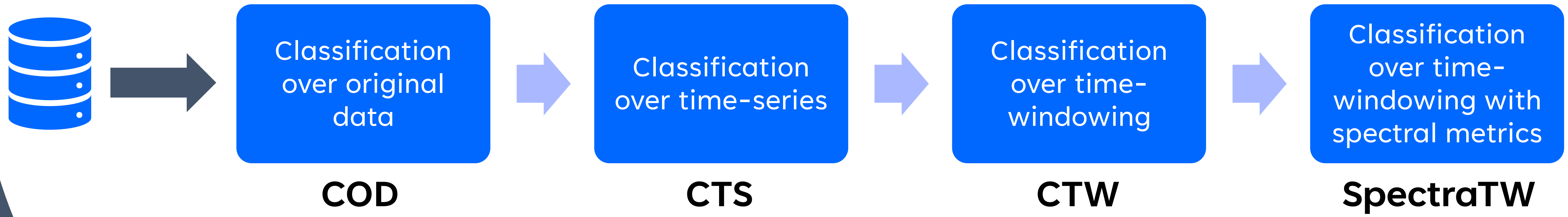
Time-windowing with spectral metrics



Phases

XGBoost is used for classification over different approaches

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 785–794.



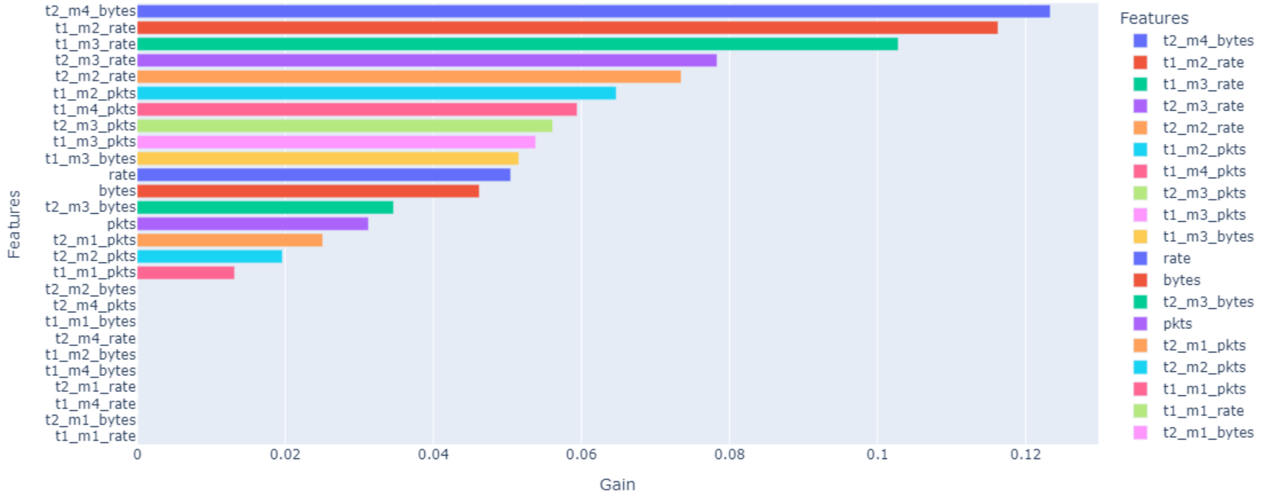
Results

Botnet dataset		COD	CTS	CTW	SM
DDoS	F1 Score	**1**	0.8750	**1**	**1**
	Balanced Acc	**1**	0.8888	**1**	**1**
	MCC	**1**	0.8816	**1**	**1**
	Precision	**1**	**1**	**1**	**1**
	Recall	**1**	0.7777	**1**	**1**
ScanService	F1 Score	0.8937	0.9966	0.9990	**0.9994**
	Balanced Acc	0.9760	0.9133	0.9834	**0.9942**
	MCC	0.8835	0.8965	0.9797	**0.9885**
	Precision	0.8225	0.9939	0.9984	**0.9994**
	Recall	0.9783	0.9993	**0.9997**	0.9994
OS Fingerprint	F1 Score	0.2617	0.8235	**0.9953**	**0.9953**
	Balanced Acc	0.5804	0.8798	**0.9953**	**0.9953**
	MCC	0.3198	0.8240	**0.9952**	**0.9952**
	Precision	0.6594	0.8974	**1**	**1**
	Recall	0.1633	0.7608	**0.9907**	**0.9907**
Keylogging	F1 Score	0.5333	0.6666	**1**	**1**
	Balanced Acc	0.7856	0.7998	**1**	**1**
	MCC	0.5344	0.6703	**1**	**1**
	Precision	0.5	0.75	**1**	**1**
	Recall	0.5714	0.6	**1**	**1**
Prediction Time (sec)		**0.24**	**0.01**	**0.03**	**0.03**

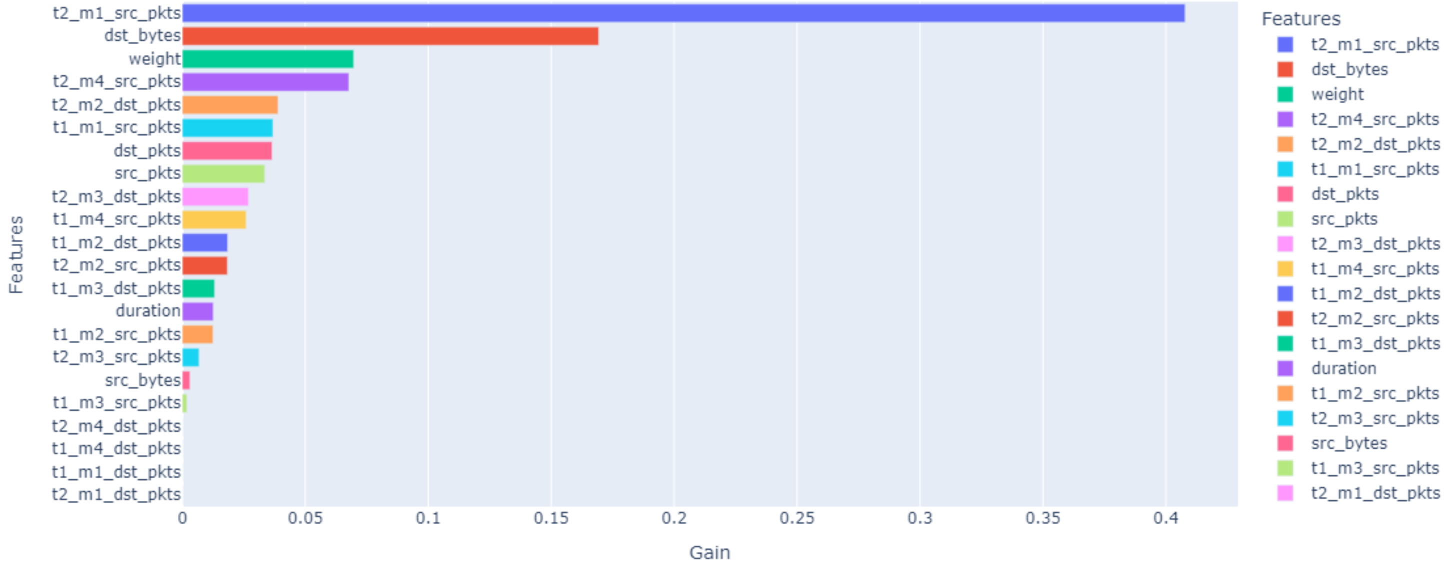
TonIoT dataset		COD	CTS	CTW	SM
DDoS	F1 Score	0.9764	0.6175	0.9943	**1**
	Balanced Acc	0.9833	0.7628	0.9971	**1**
	MCC	0.9752	0.6256	0.9943	**1**
	Precision	0.9857	0.7465	0.9943	**1**
	Recall	0.9674	0.5265	0.9943	**1**
DoS	F1 Score	0.9837	0.9054	0.9957	**1**
	Balanced Acc	0.9902	0.9403	0.9964	**1**
	MCC	**1**	0.8816	**1**	**1**
	Precision	0.9863	0.9304	0.9985	**1**
	Recall	0.9811	0.8817	0.9928	**1**
Scanning	F1 Score	0.9890	0.3363	0.9938	**1**
	Balanced Acc	0.9959	0.6102	0.9995	**1**
	MCC	0.9884	0.3631	0.9933	**1**
	Precision	0.9852	0.6411	**0.9877**	**1**
	Recall	0.9928	0.2279	**1**	**1**
Ransomware	F1 Score	0.8290	0.2978	0.9243	**0.9949**
	Balanced Acc	0.9131	0.5973	0.9526	**0.9949**
	MCC	0.8193	0.3450	0.9240	**0.9948**
	Precision	0.8218	0.6202	0.9438	**1**
	Recall	0.8364	0.196	0.9057	**0.9898**
SQL Injection	F1 Score	0.9735	0.8445	0.8428	**0.9972**
	Balanced Acc	0.9808	0.9040	0.8732	**0.9999**
	MCC	0.9721	0.8433	0.8480	**0.9972**
	Precision	0.9848	0.8827	0.9762	**0.9946**
	Recall	0.9624	0.8094	0.7468	**1**
Password	F1 Score	0.9808	0.7304	0.9148	**0.9939**
	Balanced Acc	0.9882	0.9522	0.9748	**0.9966**
	MCC	0.9798	0.7363	0.9125	**0.9937**
	Precision	0.9844	0.6022	0.8786	**0.9943**
	Recall	0.9772	0.9280	0.9541	**0.9935**
XSS	F1 Score	0.8710	0.6731	**1**	**1**
	Balanced Acc	0.9509	0.8043	**1**	**1**
	MCC	0.8645	0.6753	**1**	**1**
	Precision	0.8336	0.7513	**1**	**1**
	Recall	0.9120	0.6096	**1**	**1**
Backdoor	F1 Score	0.9985	0.8589	0.9928	**0.9995**
	Balanced Acc	0.9989	0.8928	0.9967	**0.9995**
	MCC	0.9984	0.8563	0.9923	**0.9995**
	Precision	0.999	0.9437	0.9916	**1**
	Recall	0.9980	0.7880	0.9939	**0.999**
MitM	F1 Score	0.7239	0.4542	0.7640	**1**
	Balanced Acc	0.8733	0.6747	0.8148	**1**
	MCC	0.7235	0.4742	0.7818	**1**
	Precision	0.7017	0.6458	0.9714	**1**
	Recall	0.7476	0.3502	0.6296	**1**
Prediction Time (sec)		**0.83**	**0.31**	**0.21**	**0.15**

Feature Importance

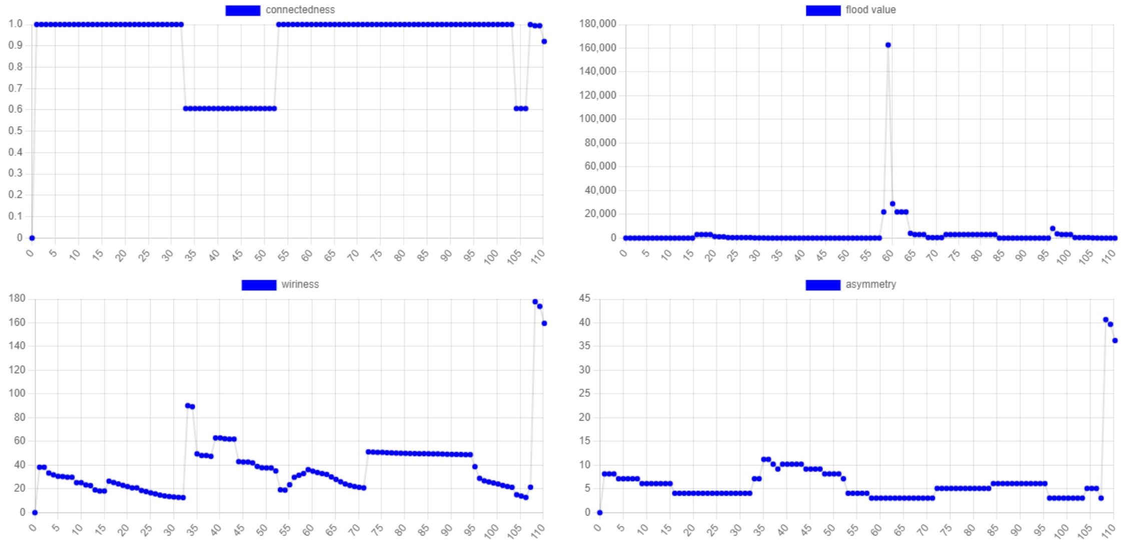
Botnet dataset



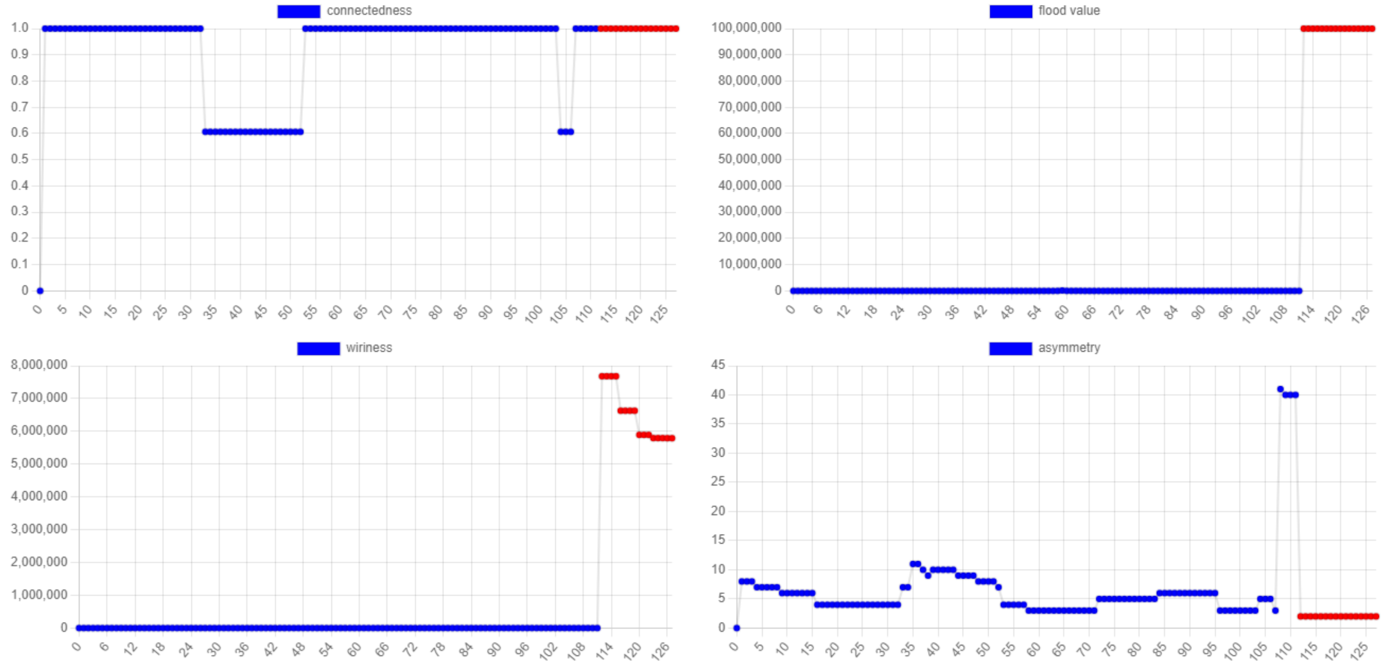
TonIoT dataset



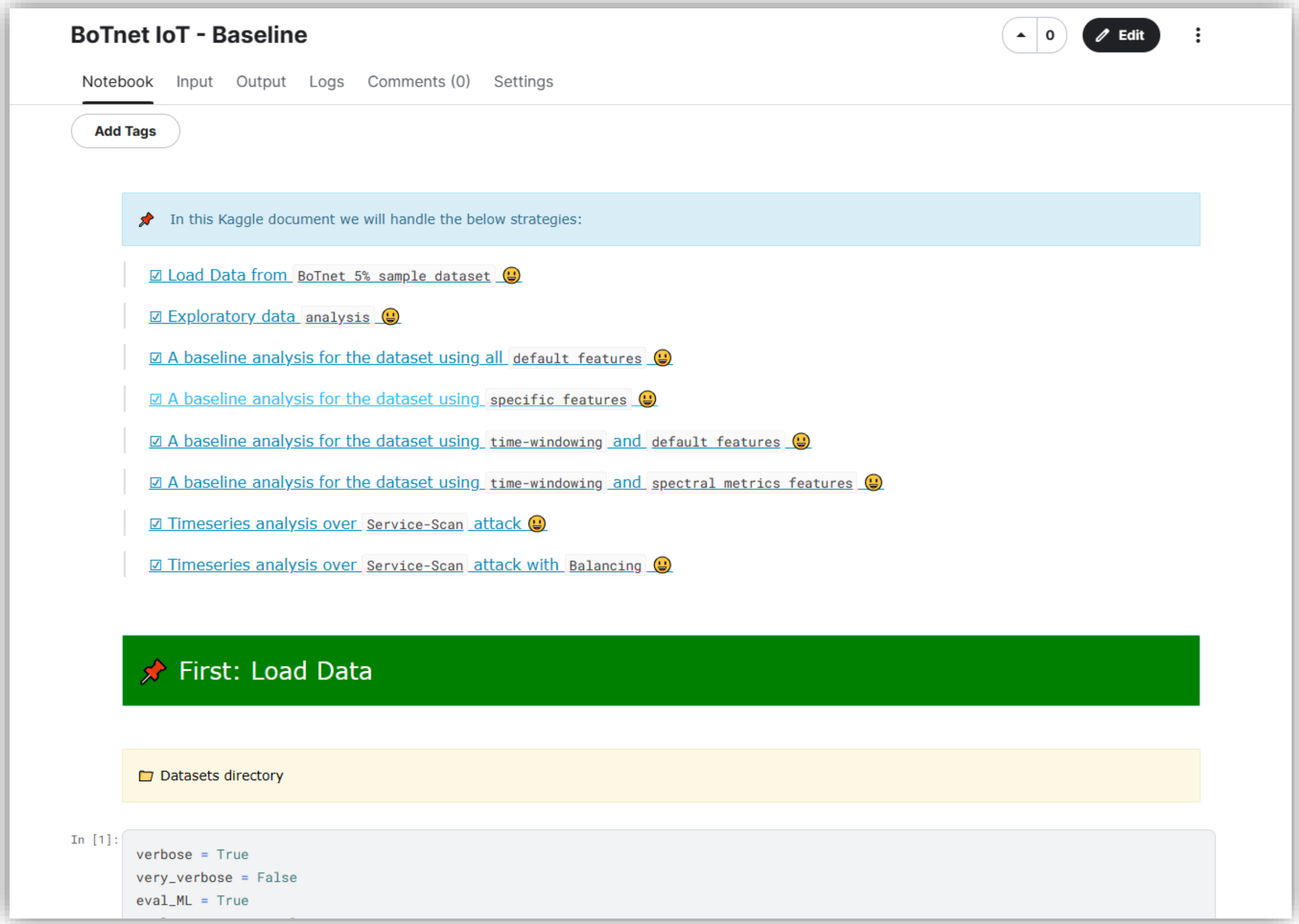
Spectral Metrics Behavior



Spectral metrics behavior before and after the attack in Botnet IoT dataset



Notebooks



BoTnet IoT - Baseline
0
Edit
Notebook
Input
Output
Logs
Comments (0)
Settings
Add Tags
In this Kaggle document we will handle the below strategies:
☑ Load Data from BoTnet 5% sample dataset
☑ Exploratory data analysis
☑ A baseline analysis for the dataset using all default features
☑ A baseline analysis for the dataset using specific features
☑ A baseline analysis for the dataset using time-windowing and default features
☑ A baseline analysis for the dataset using time-windowing and spectral metrics features
☑ Timeseries analysis over Service-Scan attack
☑ Timeseries analysis over Service-Scan attack with Balancing
First: Load Data
Datasets directory
In [1]:
verbose = True
very_verbose = False
eval_ML = True

Coming work

- Find new datasets to verify the performance of our introduced spectral metrics.
- Explain why spectral metrics works for different attacks, and different graph patterns.
- Integrate spectral metrics within the Graph processing for Machine Learning (GPML) library.

Thank you

Majed Jaber majed.jaber@epita.fr

Nicolas Boutry nicolas.Boutry@epita.fr

Pierre Parrend pierre.parrend@epita.fr

Any Questions