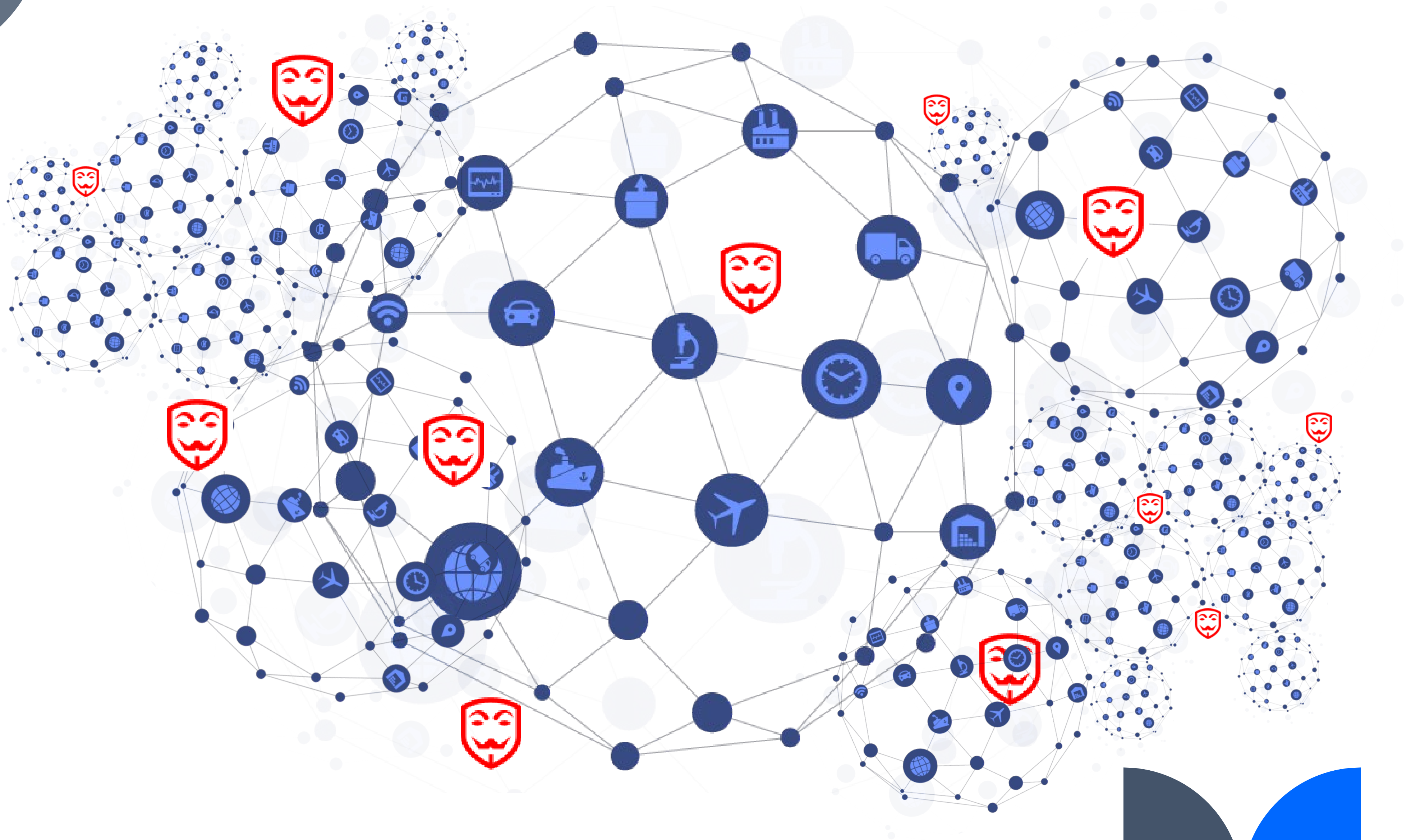
Towards attack detection in traffic data based on spectral graph analysis

Majed Jaber, Pierre Parrend, Nicolas Boutry

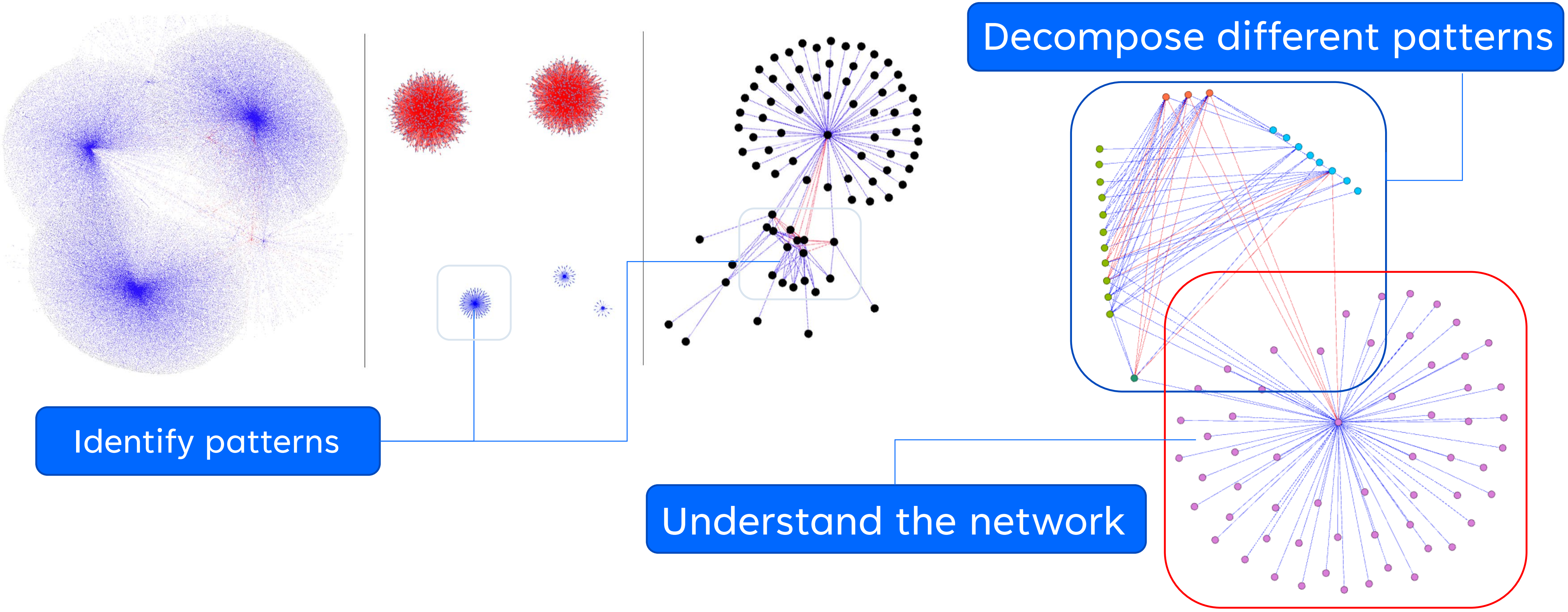
Outline

- Introduction of the team and my career
- 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
- Future works

Cybersecurity against attacks



Graph represents a networks



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

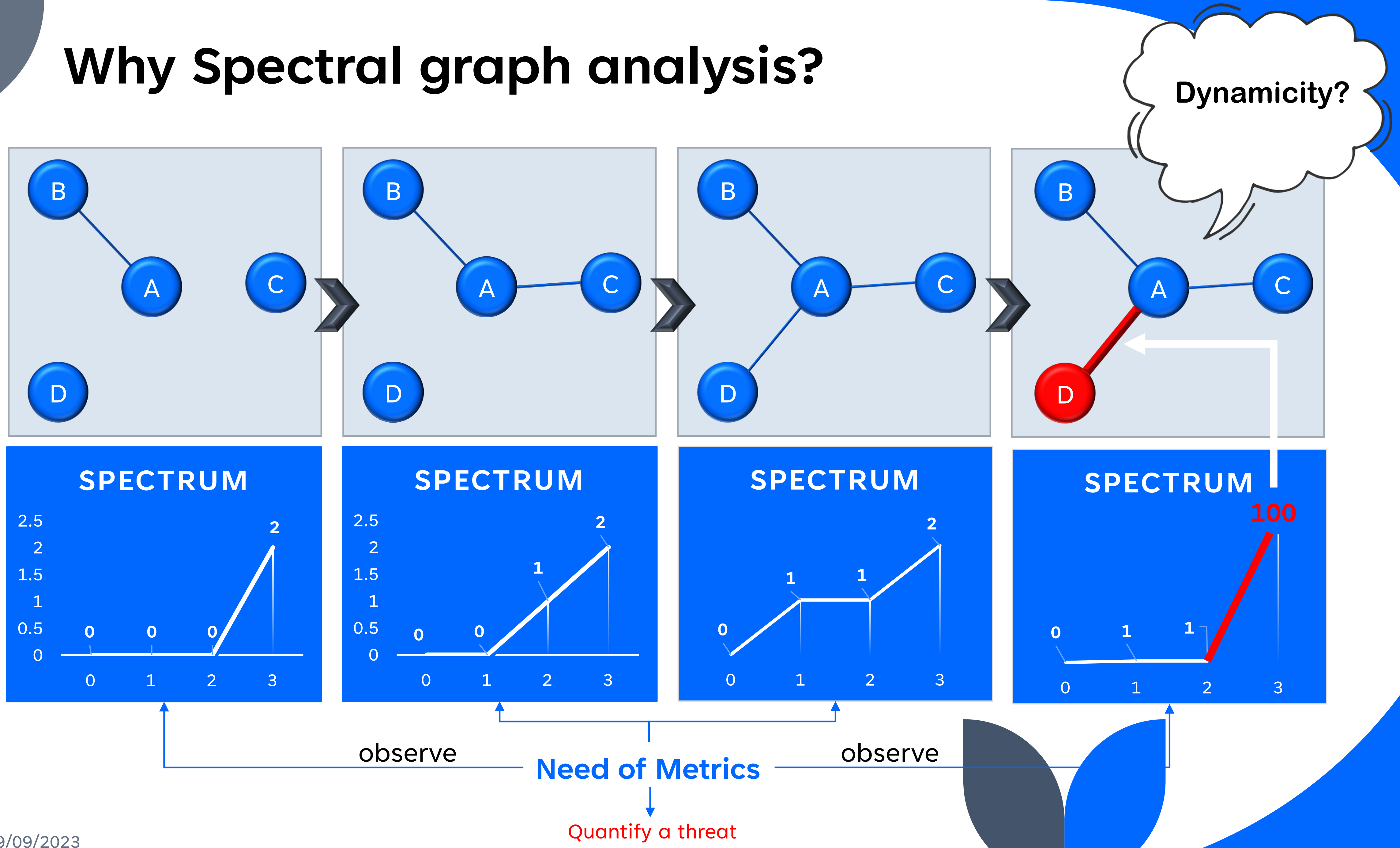
GCN Approaches

One of the best choice for graph data learning tasks

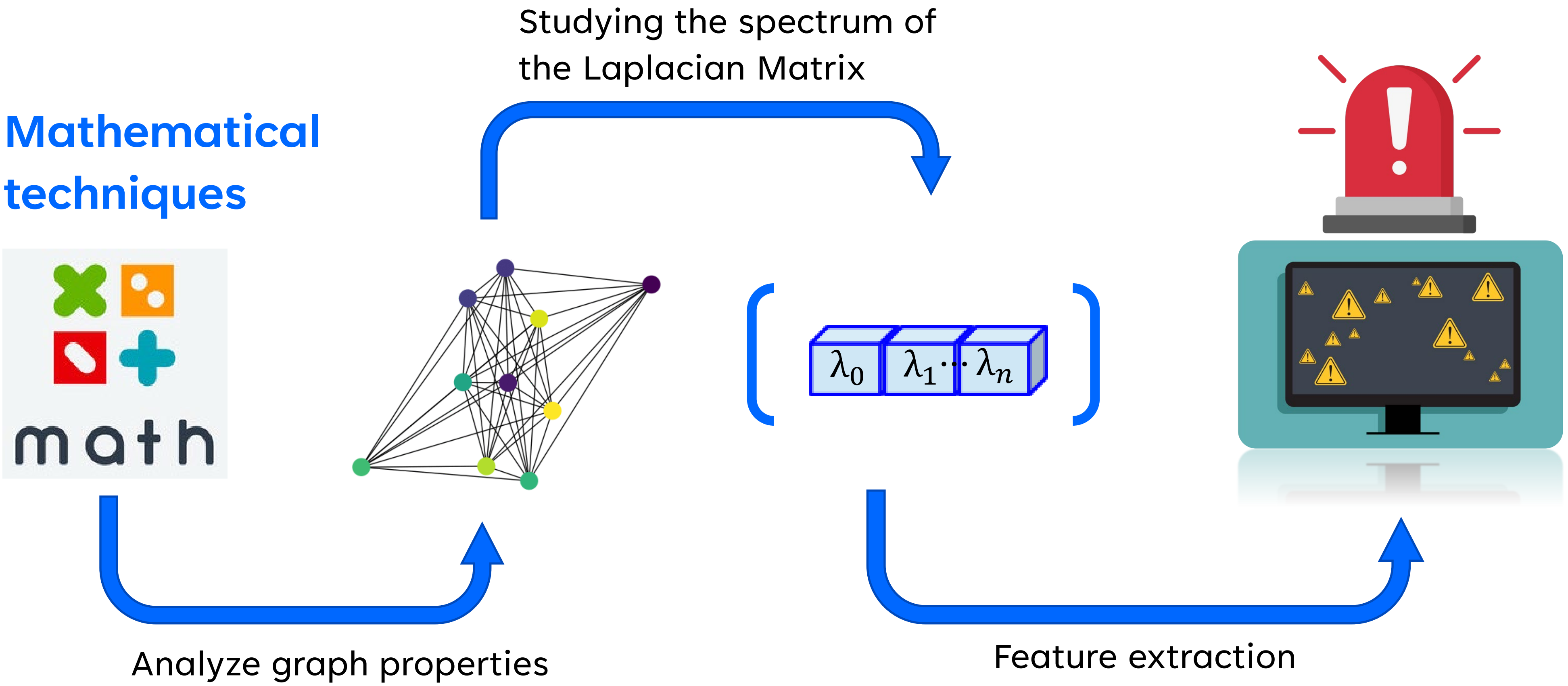
The Dynamic Graph Neural Networks (DGNNs) are known to be an interesting tool to detect anomalies in complex dynamic graphs

- 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.

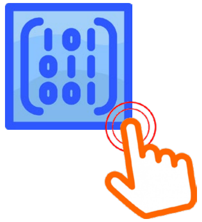
Why Spectral graph analysis?



Spectral graph analysis



What type of matrix used?



The most commonly used matrix in spectral graph analysis is the **Laplacian matrix**.

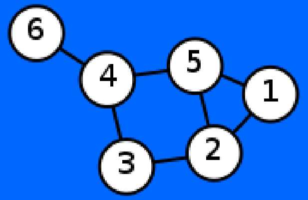
Laplacian Matrix

Why Laplacian rather than other matrixes?

- Better spectral properties
- More robust to changes in the graph structure.
- The spectrum of the Laplacian matrix are used in various applications of spectral graph analysis, such as clustering, community detection, and graph partitioning.

Laplacian Matrix

used?



$$L = D - A$$

ph analysis is

other

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

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

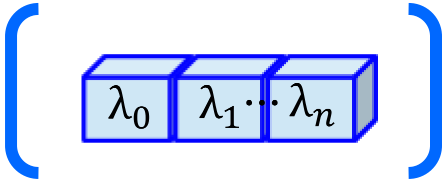
$$A_{i,j} := \begin{cases} 1 & if\ i \neq j \text{ and } v_i \sim v_j \\ 0 & \text{otherwise} \end{cases}$$

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

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

What is a spectrum?

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



Spectrum

Eigenvalues

If A is a square matrix and V is a column vector such that:

$$\boxed{AV = \lambda V}$$

then

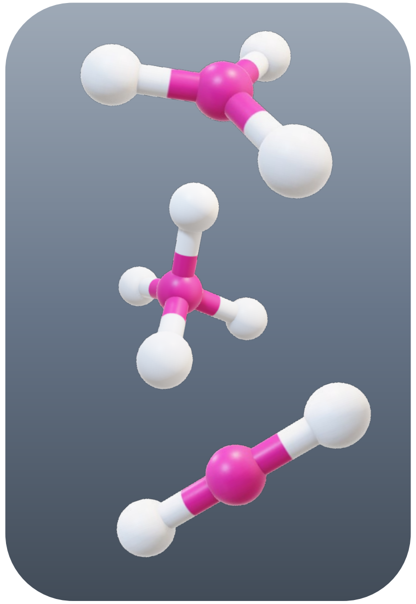
V = Eigen vector of A

λ = Eigen value of A

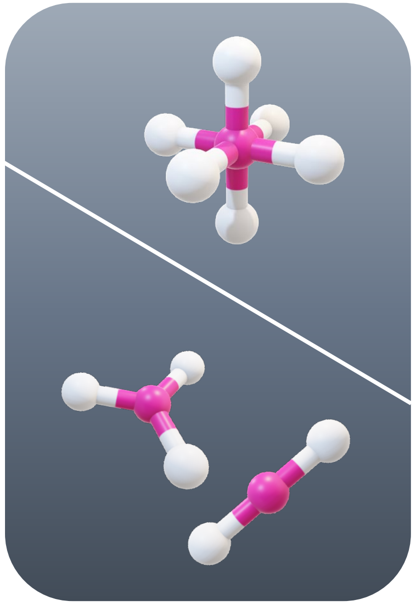
Chung, F. R. (1997). Spectral graph theory (Vol. 92). American Mathematical Soc.

Spectrum Interesting eigenvalues

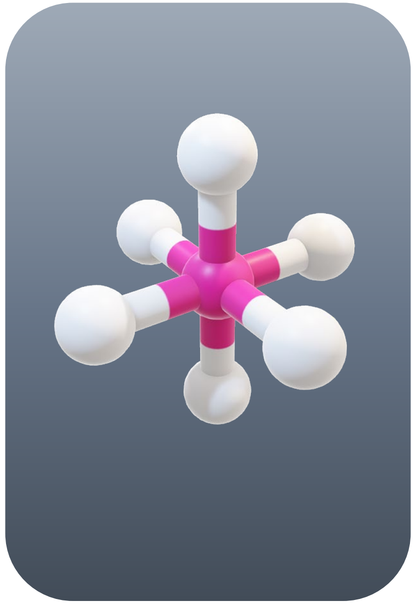
Zero eigenvalues



Algebraic connectivity

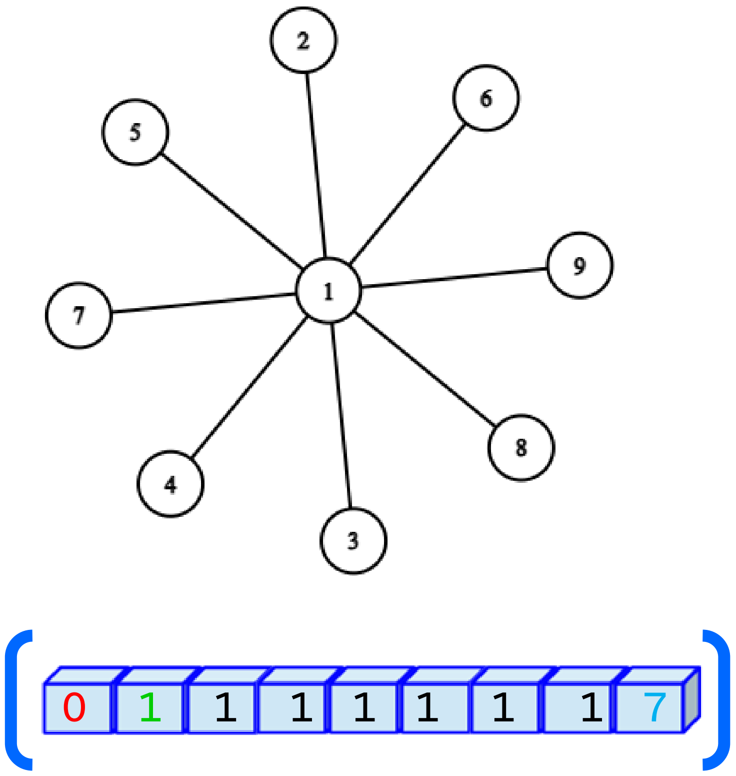


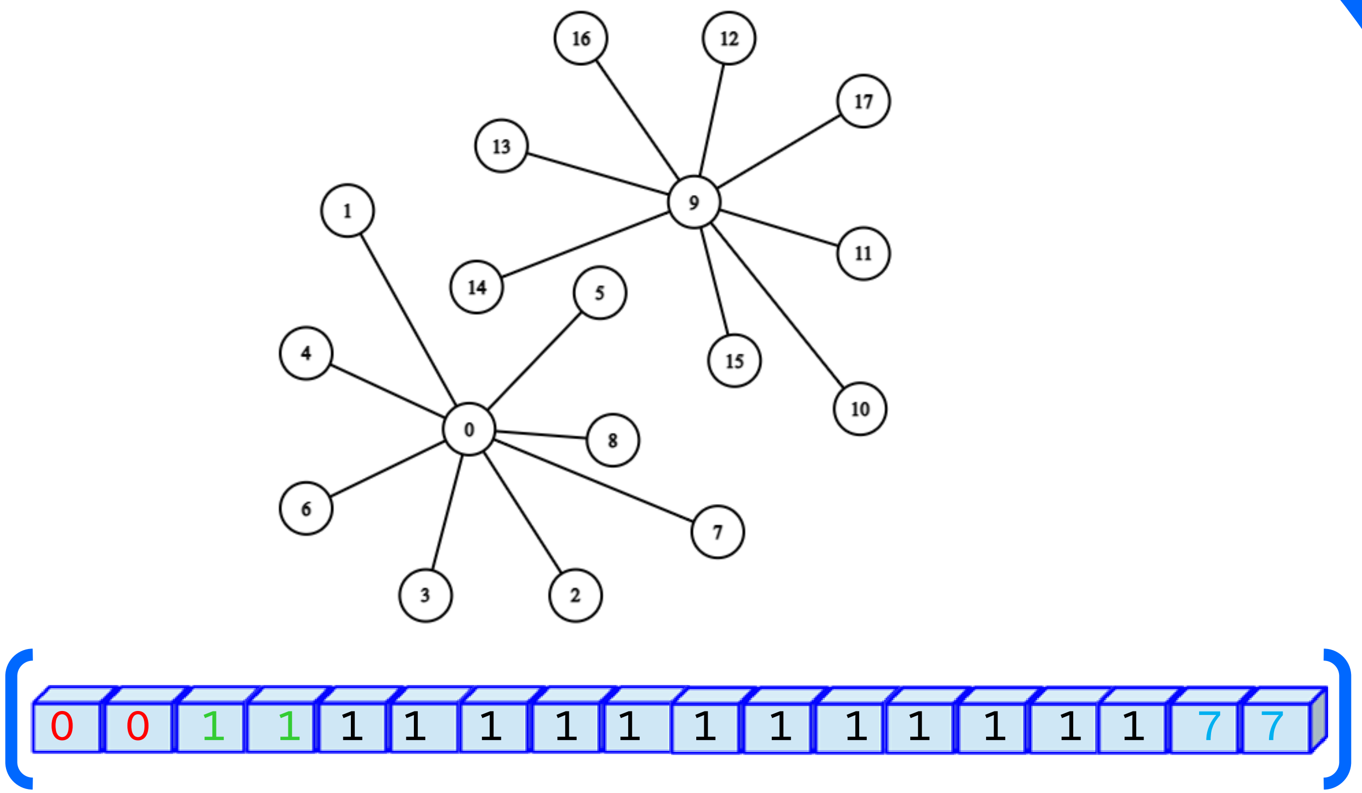
Largest eigenvalues



- 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)

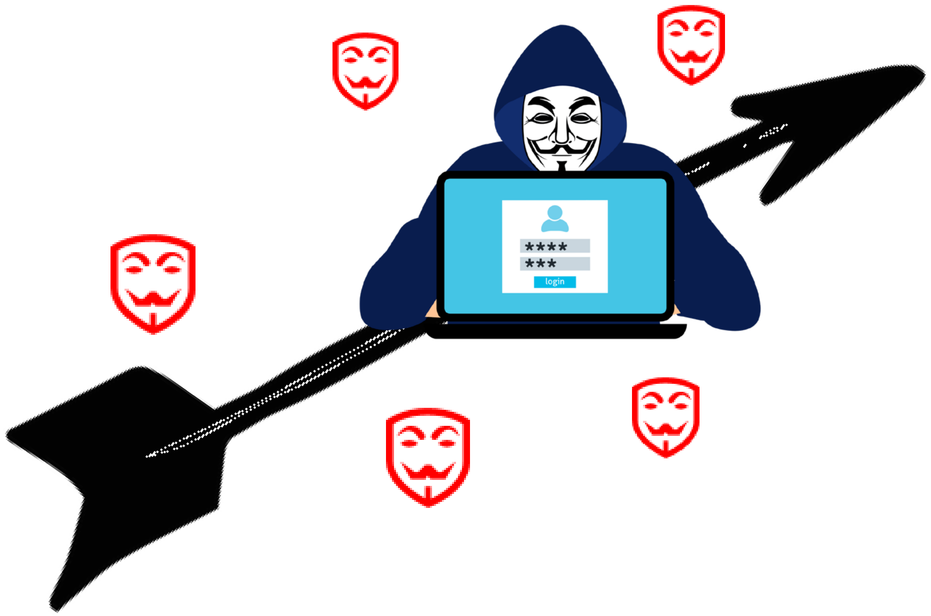
Spectrum Interesting EV - Example

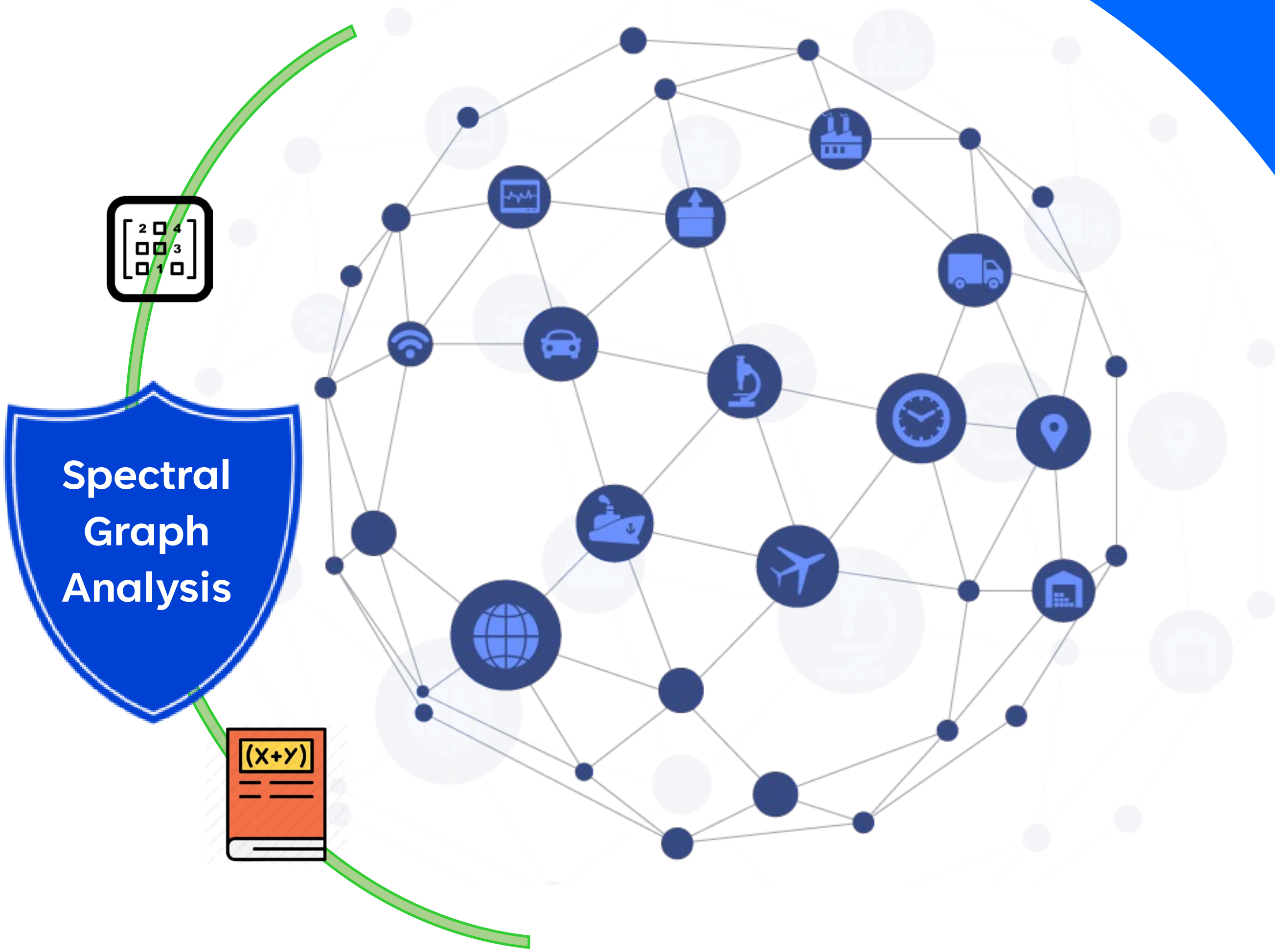




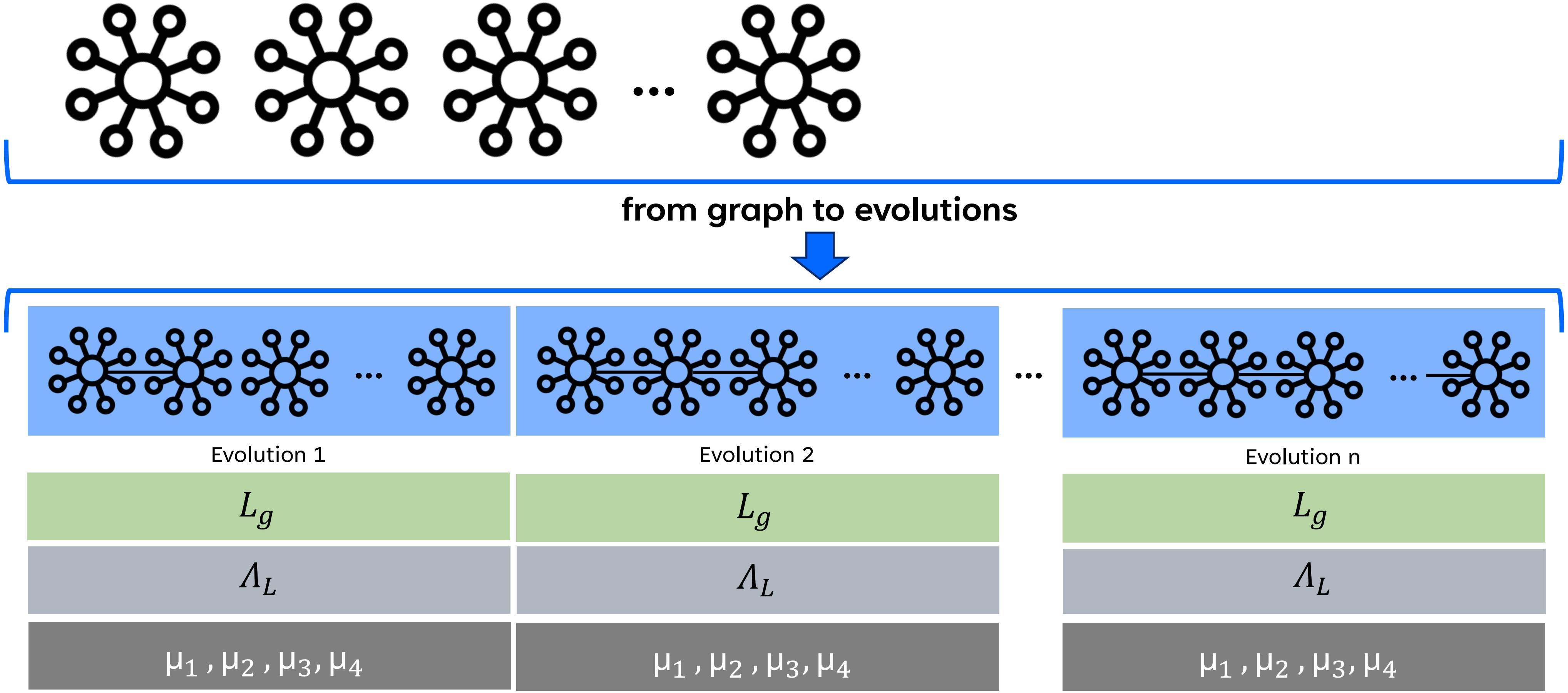
Research Question

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





Dynamicity of graph



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

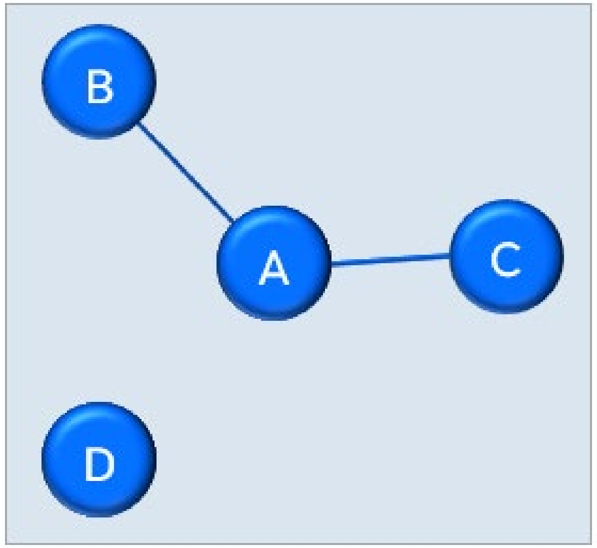
Metric 1 - Connectedness

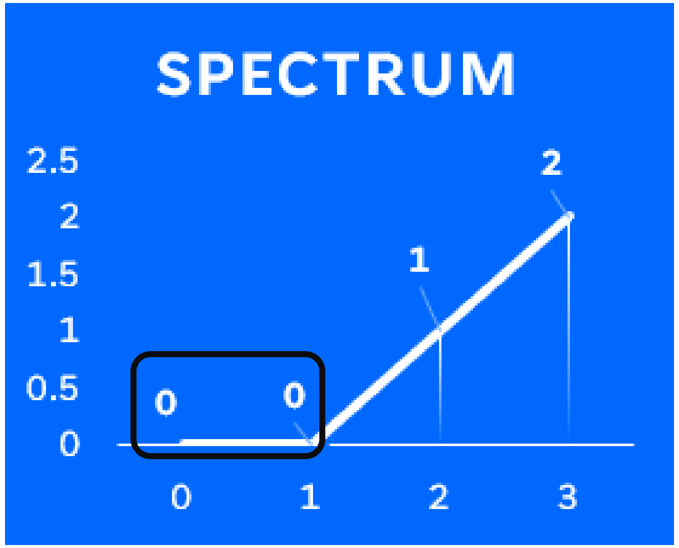
$$\mu_1(t) = \frac{\exp \frac{1}{\mathcal{Z}(t)}}{\exp(1)}$$

$\mathcal{Z}(t)$ 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

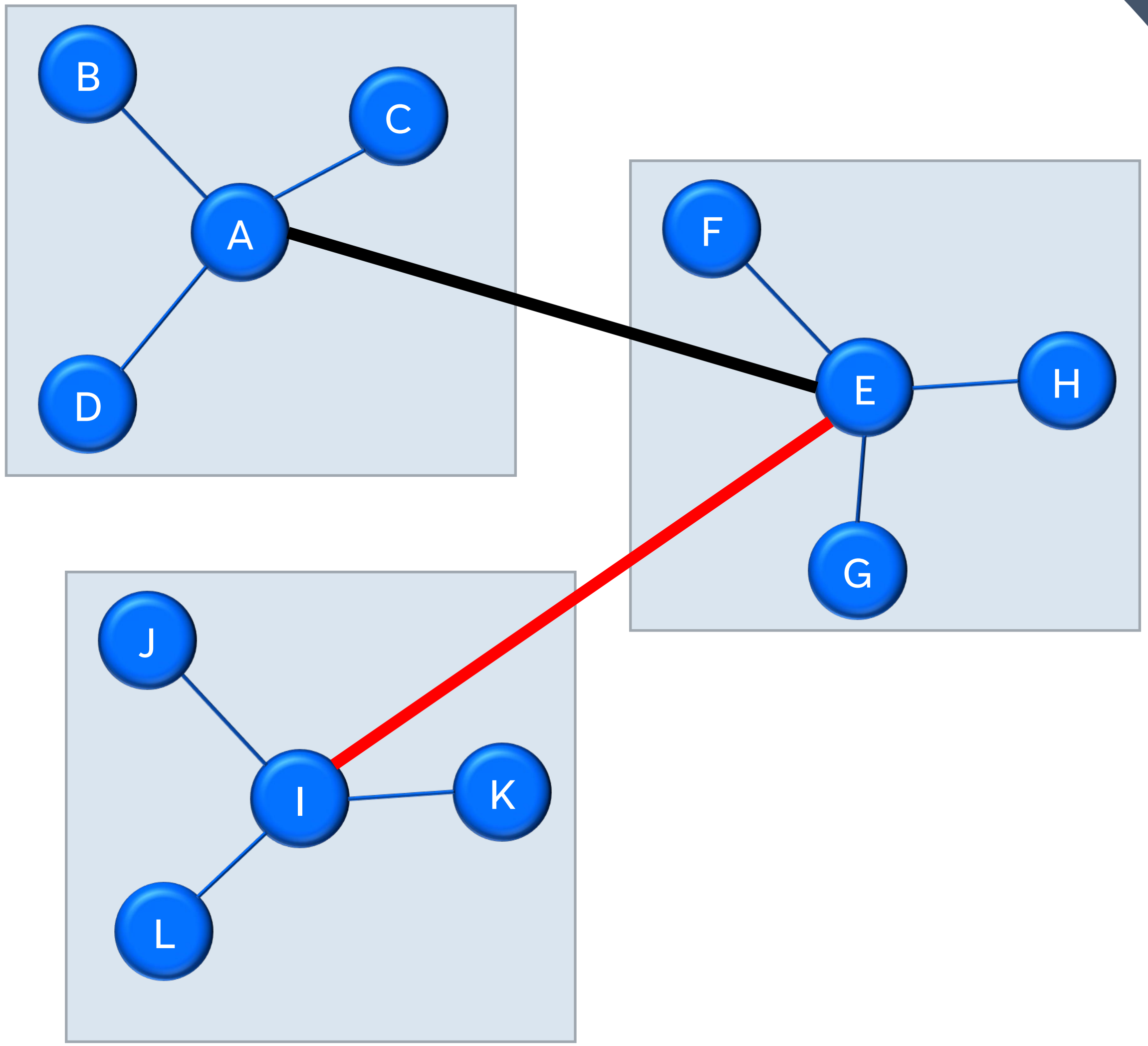
$$\mu_2(t) = \sum_{p=2}^{p=\mathcal{N}} (\exp^{\lambda_p(t)} - 1)$$

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

Connections/Flooding

Results in

Metric 2

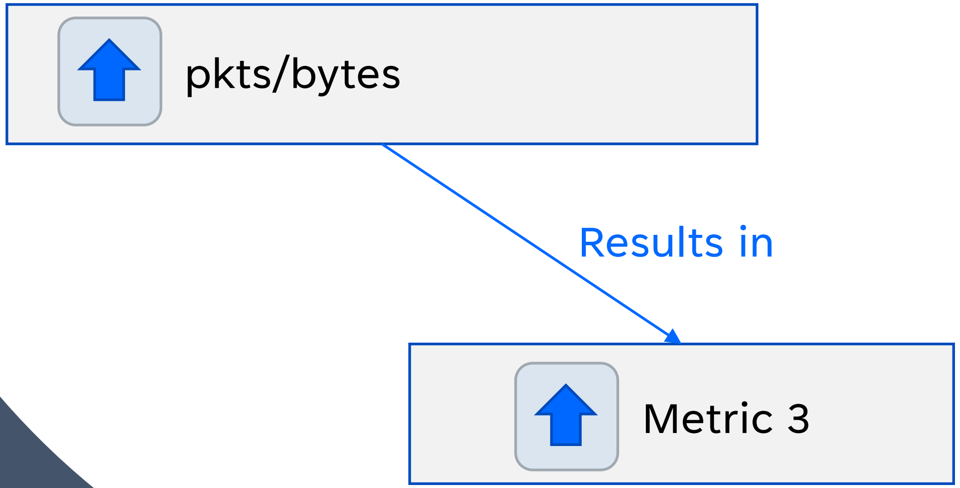


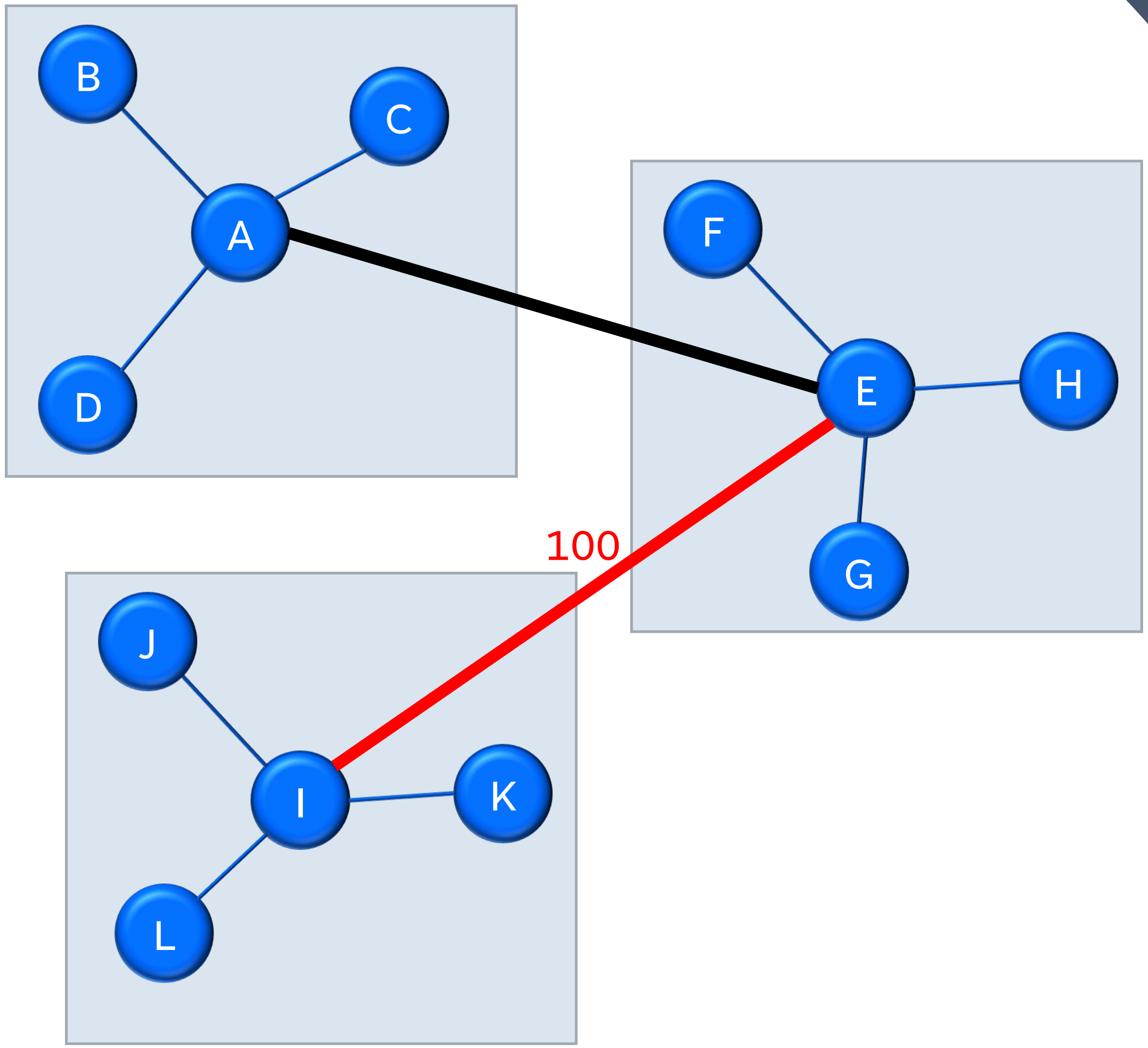
De Abreu, N. M. M. (2007). Old and new results on algebraic connectivity of graphs. Linear algebra and its applications, 423(1), 53-73.

Metric 3 - Wiriness

$$\mu_3(t) = \sum_{p=n-\mathcal{N}+1}^{p=n} \lambda_p(t)$$

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





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

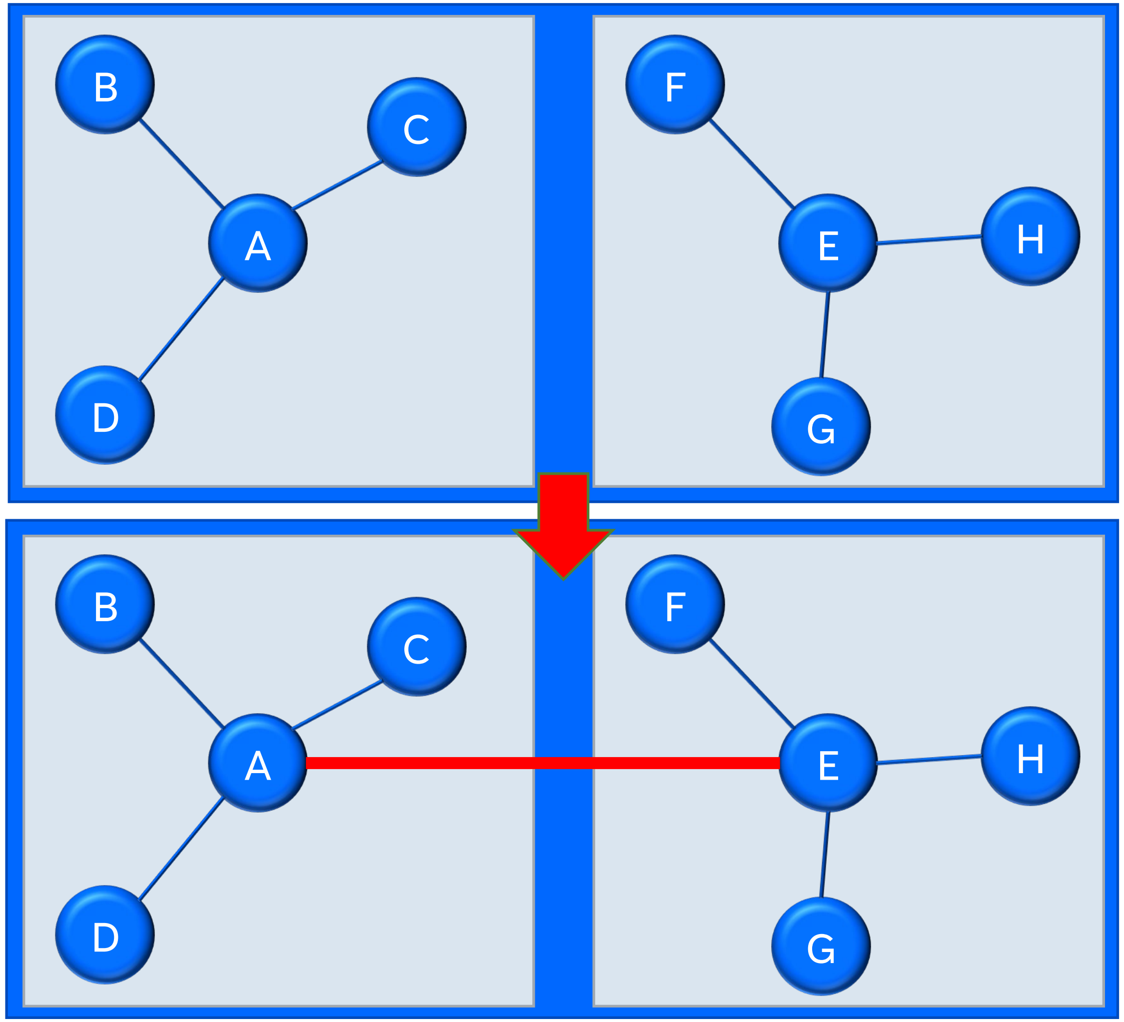
$$\mu_4(t) = \#\{k \in [2, n] \; ; \; \Lambda(t)[k] - \Lambda(t)[k-1] > \varepsilon\}$$

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

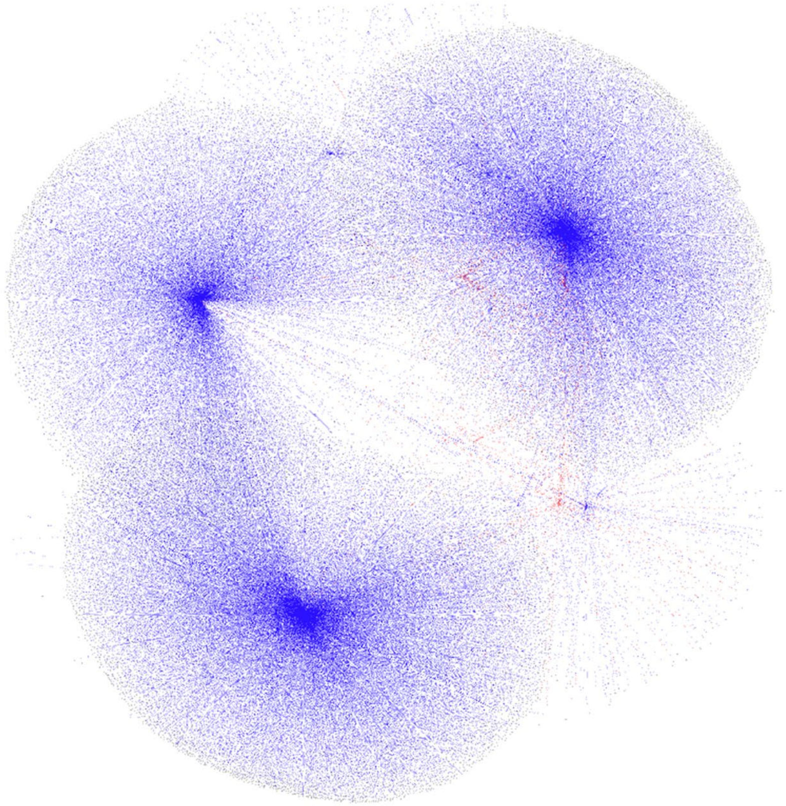
Identical patterns/symmetry

Results in

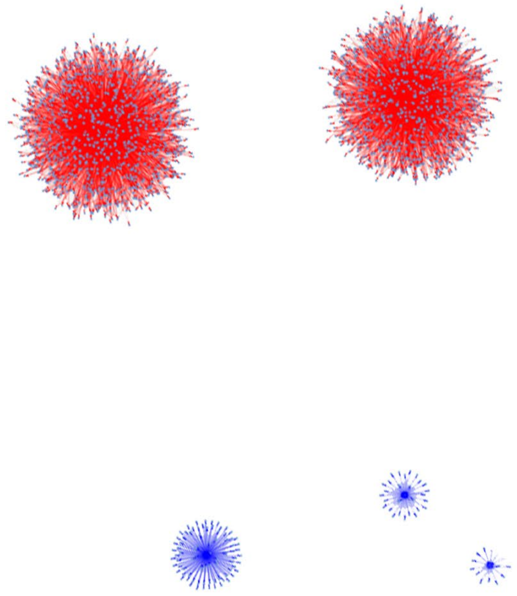
Metric 4



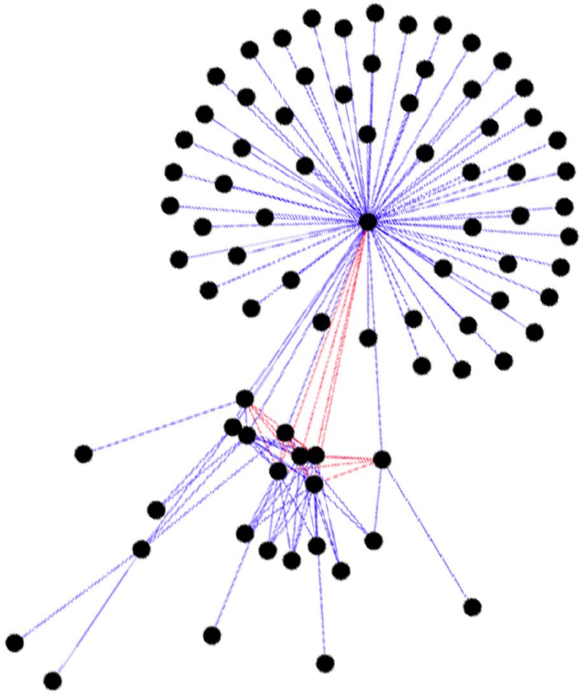
Implementation and datasets



Ton IoT



Healthcare IoT

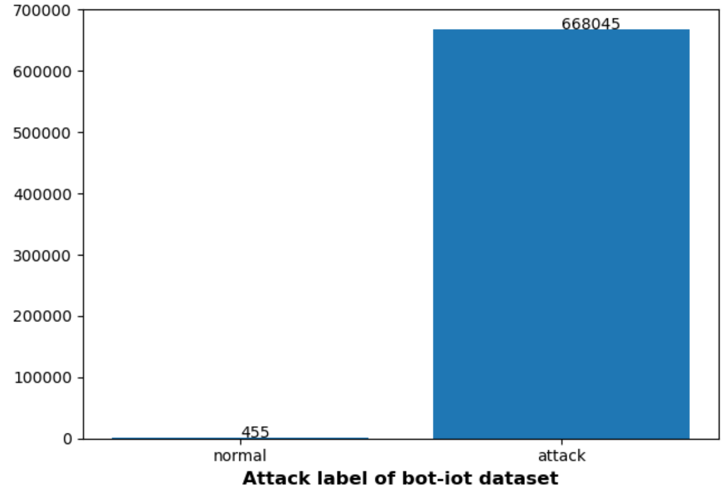


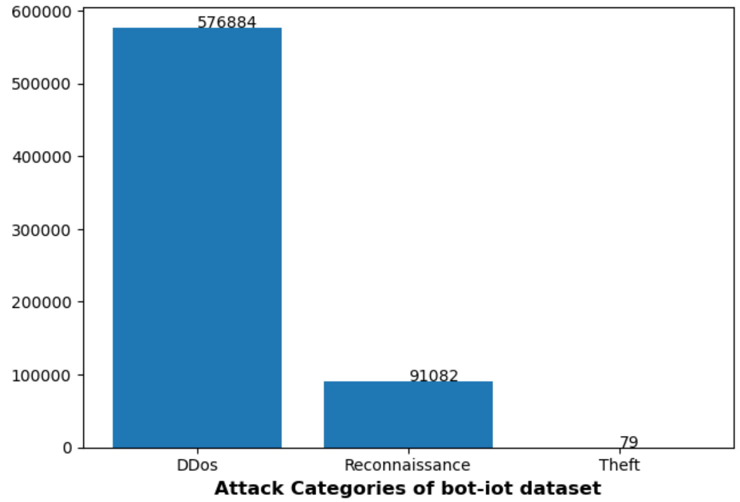
Botnet IoT

[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.

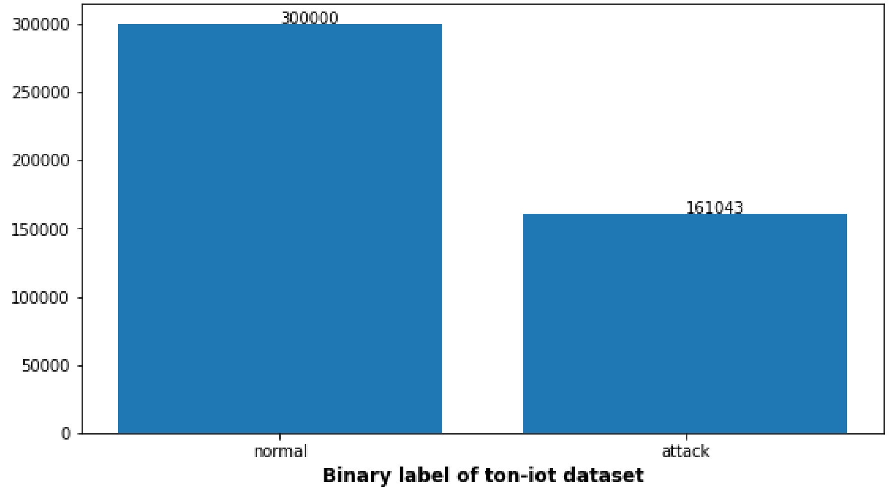
Attack analysis

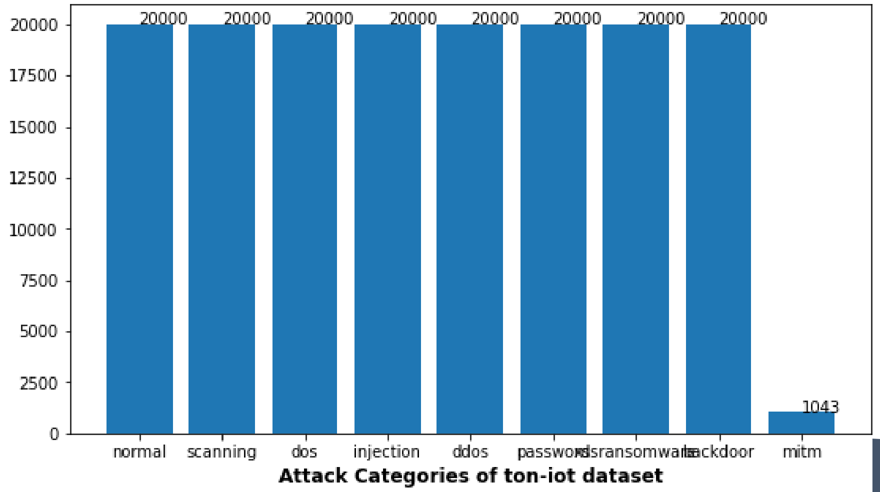
Botnet IoT dataset



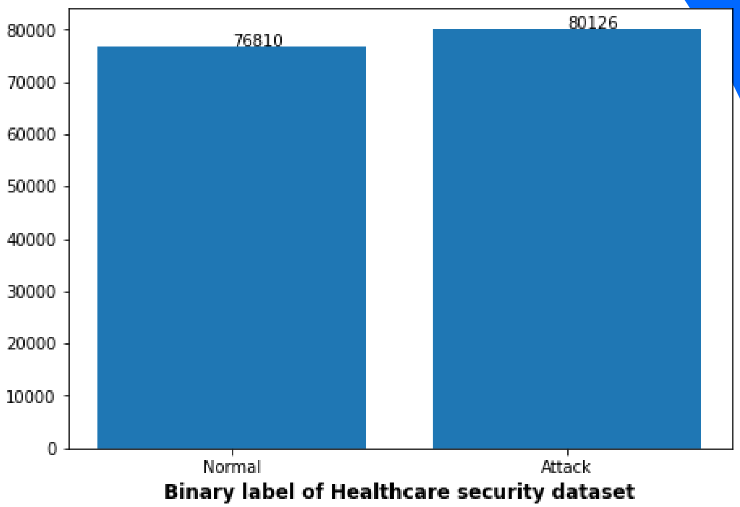


Ton IoT dataset



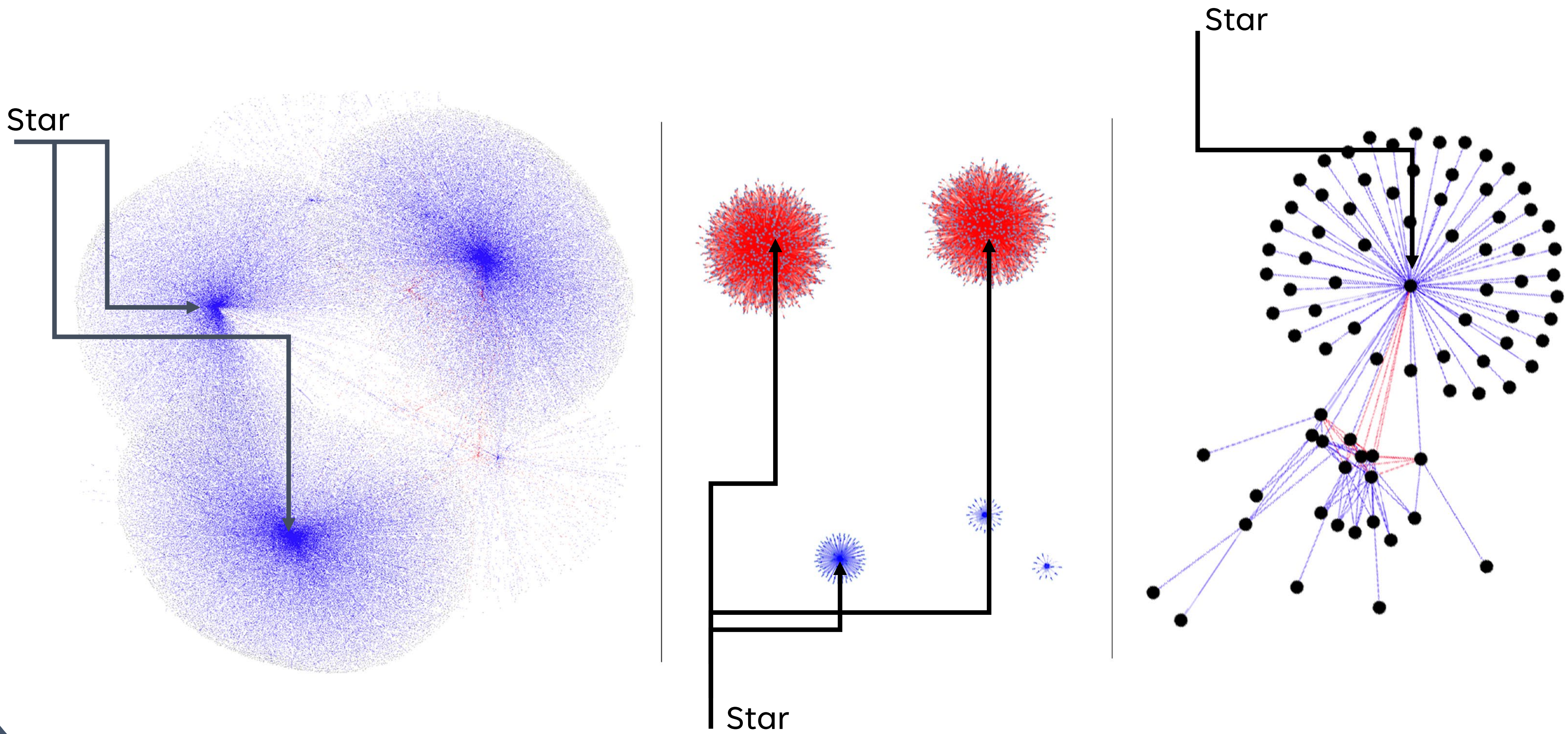


IoT Healthcare Security Dataset



Network patterns

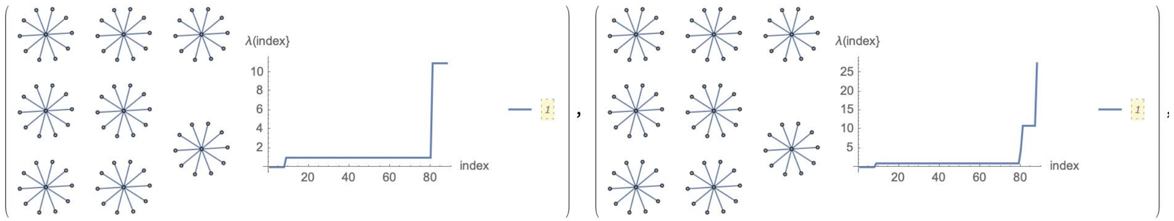
First step for detection



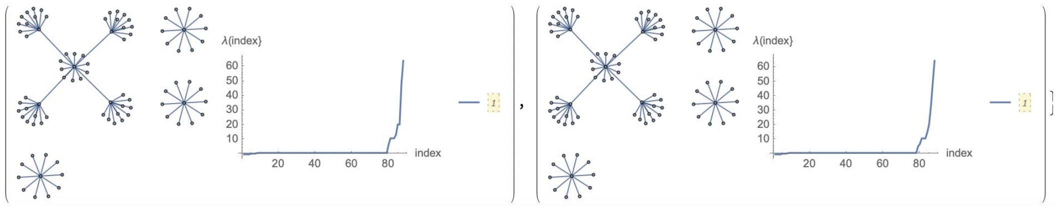
[Boo+21] Tim M Booij et al. “ToN_IoT: The role of heterogeneity and the need for standardization of features and attack types in IoT network intrusion data sets”. In: IEEE Internet of Things Journal 9.1 (2021), pp. 485–496.
[Kor+19] Nickolaos Koroniotis et al. “Towards the development of realistic botnet 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.

First Methodology

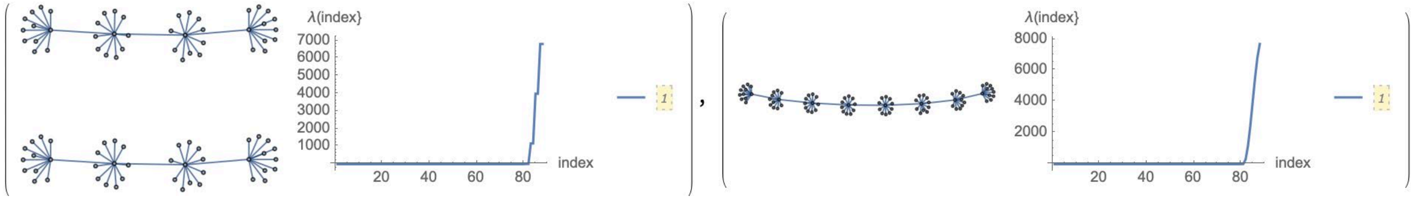
Starting with
Star graph topologies



Normal case scenario

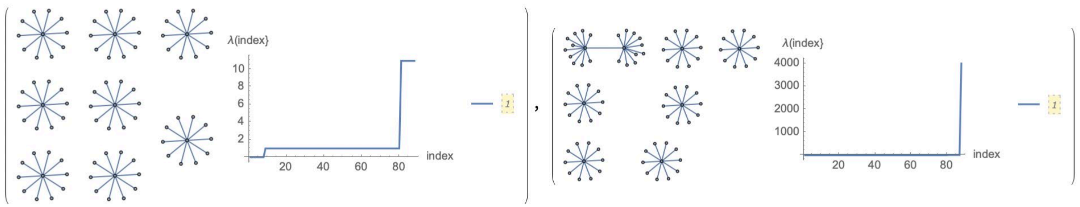


Suspicious case scenario
DoS/DDoS attack behavior

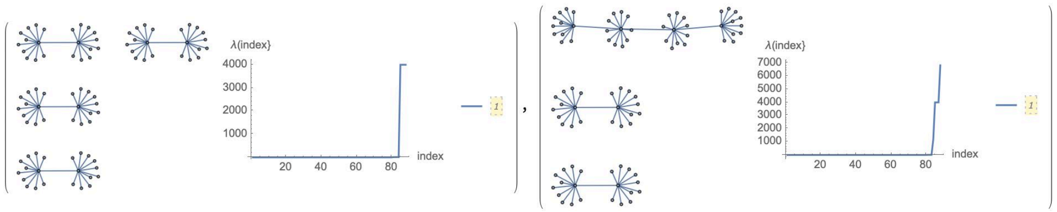


Experiments – Scenario 1 – Attack behavior

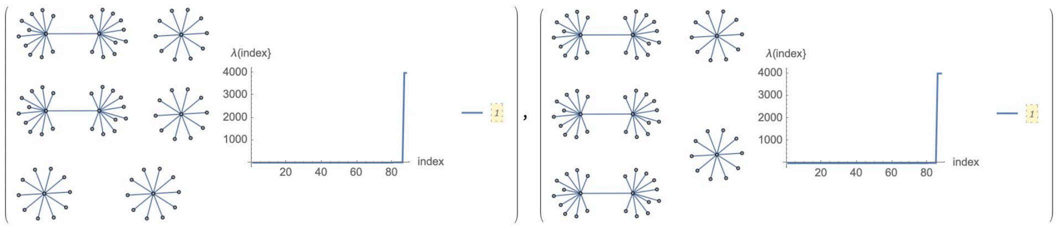
1



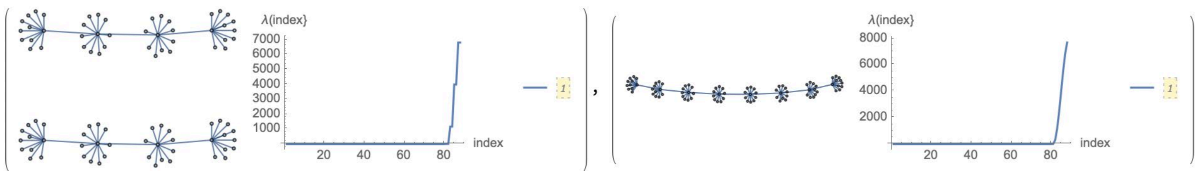
3

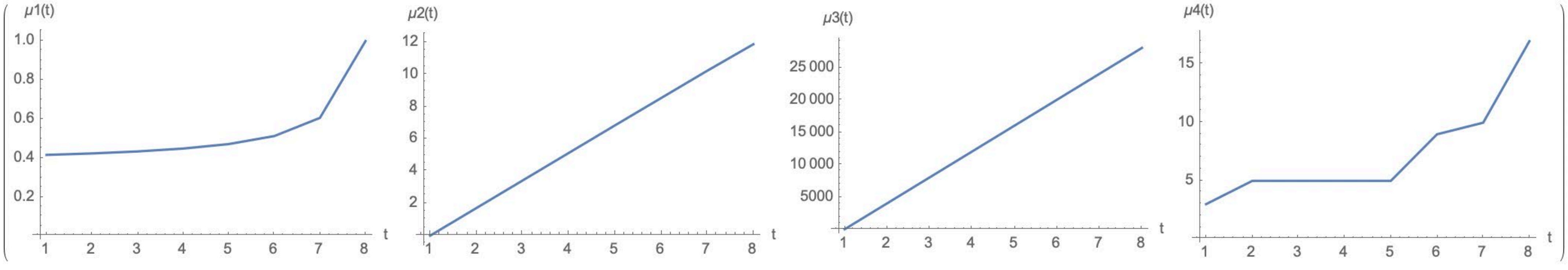


2



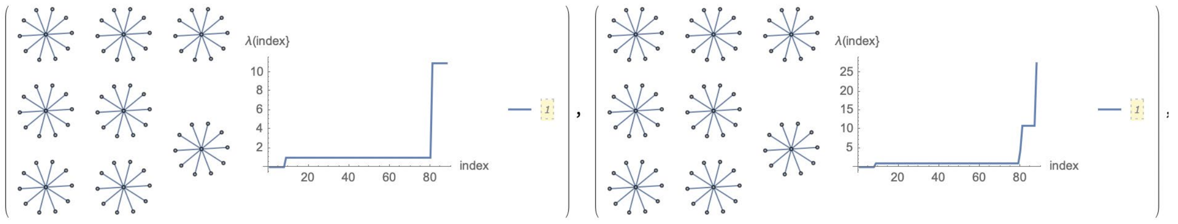
4



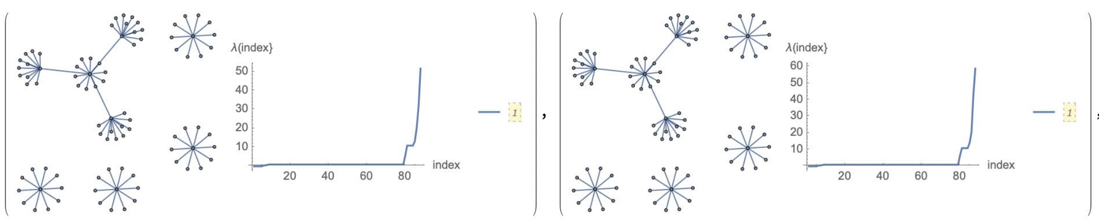


Experiments – Scenario 2 – Normal behavior

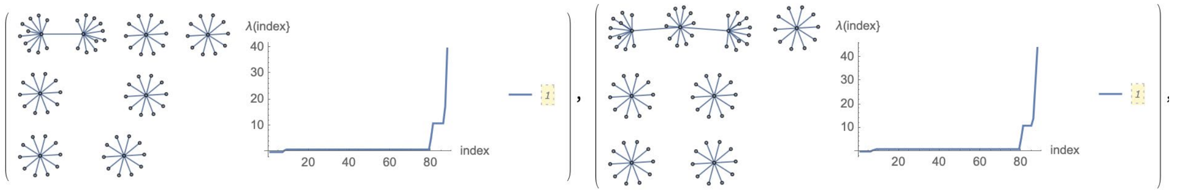
1



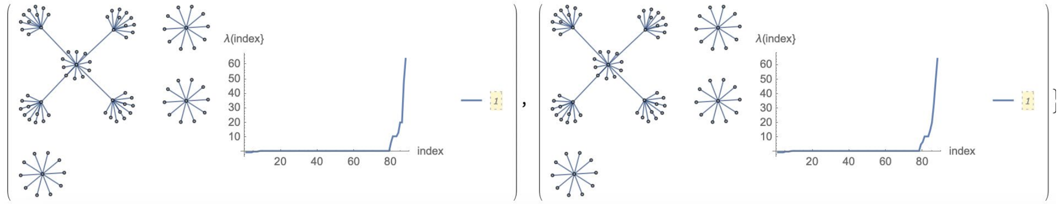
3

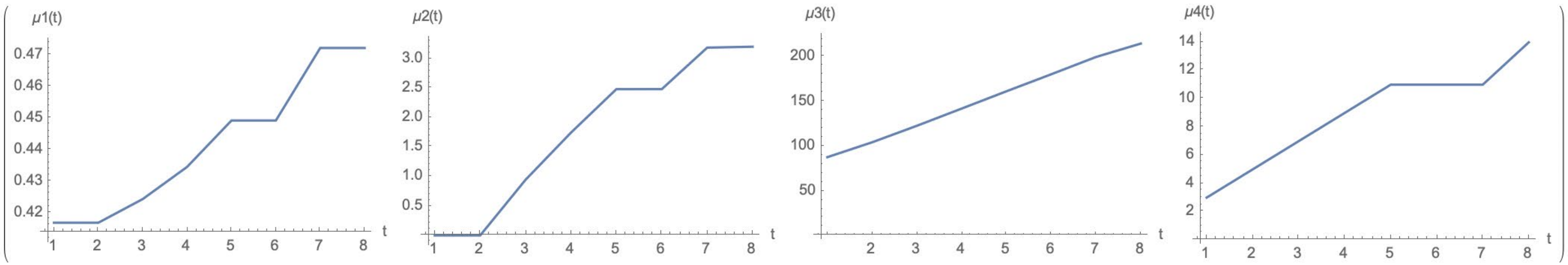


2

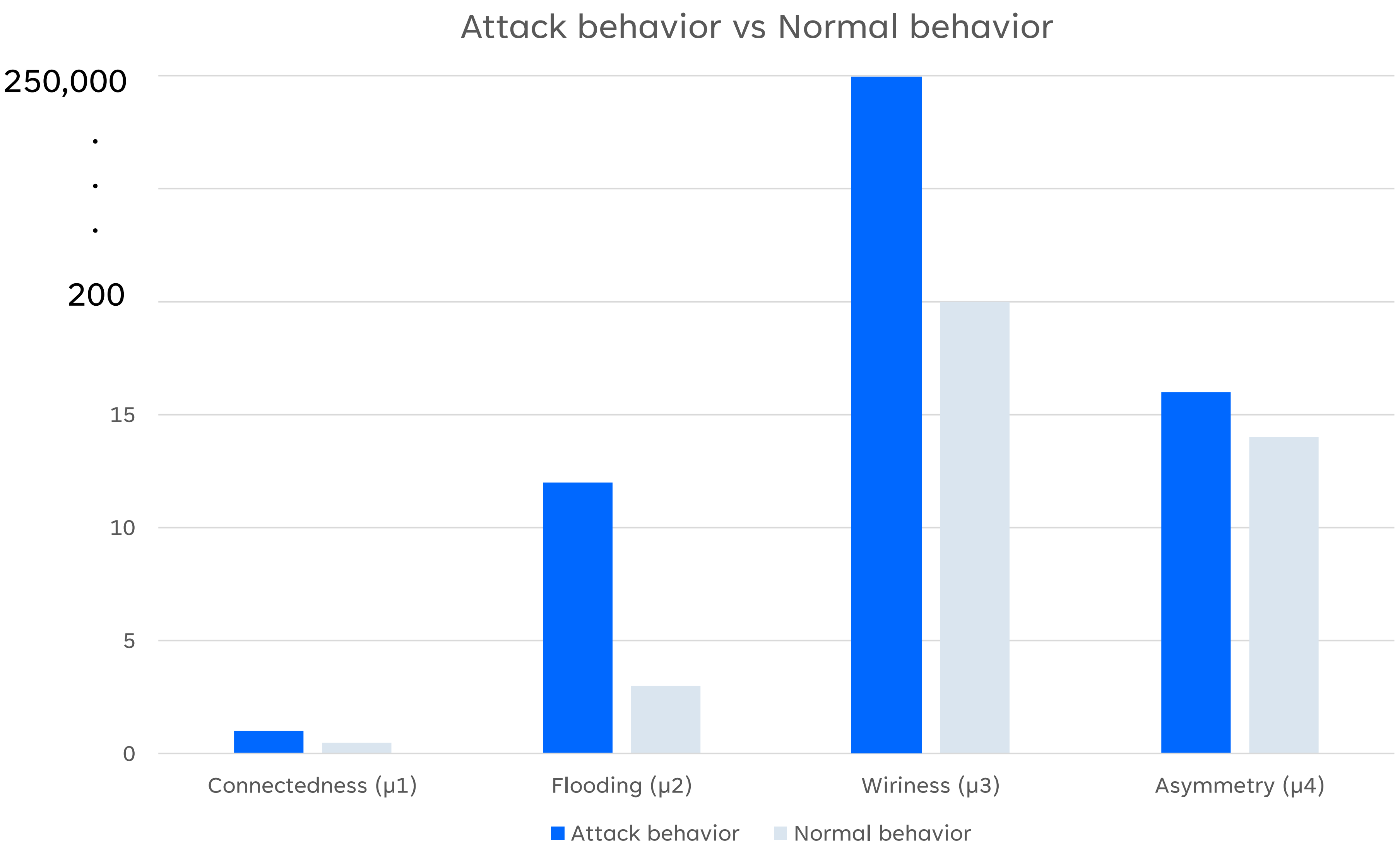


4





Experiments Evaluation



Metrics over real dataset

datasets

Apply metrics over datasets
Detect Suspicious behaviors

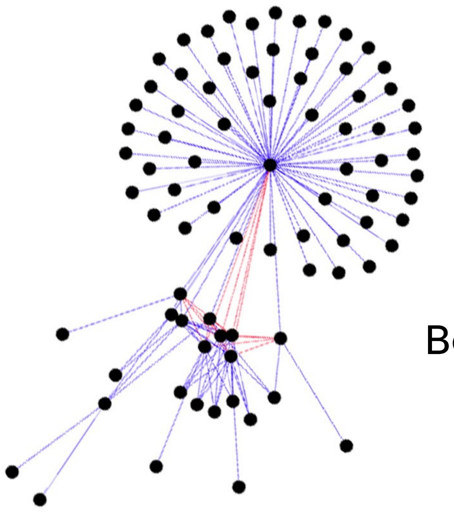
$\mu_1 , \mu_2 , \mu_3 , \mu_4$

Connectedness

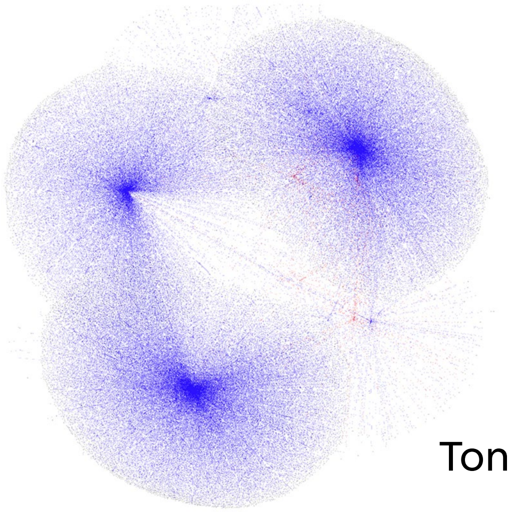
Flooding

Wiriness

Asymmetry



BotIoT graph representation



TonIoT graph representation

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	0
576916	1526344032	192.168.100.46	192.168.100.5	29726	0
576921	1526344032	192.168.100.3	13.55.154.73	3018	1
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

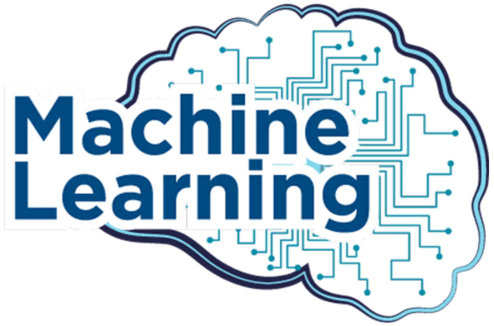
Methodology

Evolution 1
Evolution 2
...
Evolution n

Evolutions

$\mu_1, \mu_2, \mu_3, \mu_4$
$\mu_1, \mu_2, \mu_3, \mu_4$
...
$\mu_1, \mu_2, \mu_3, \mu_4$

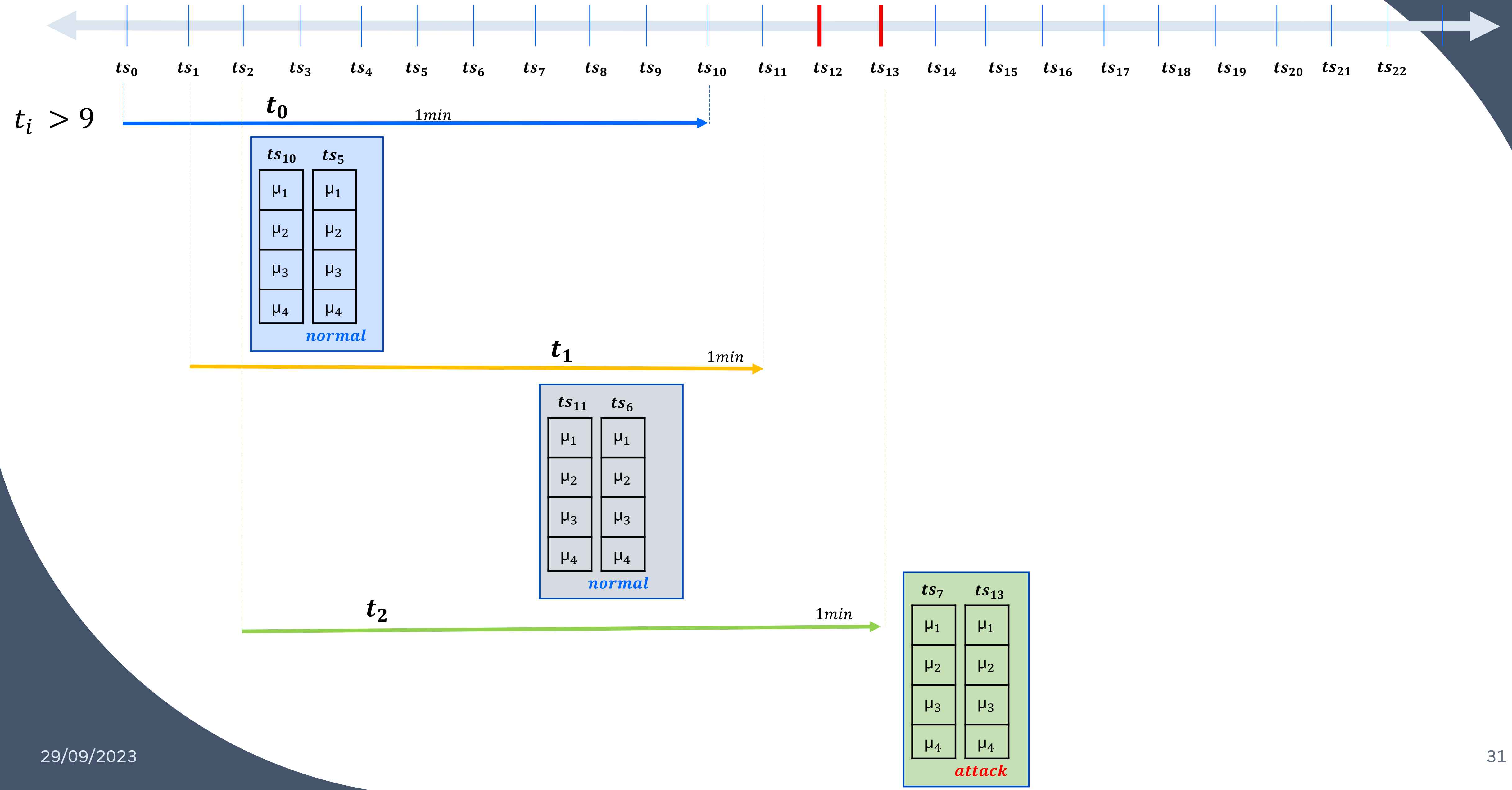
Calculate Metrics



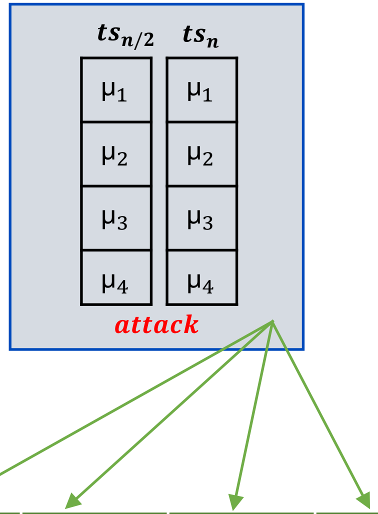
Apply Classification Methods

- Random Forest
- Decision Tree
- MLP
- XGBoost
- SVM

Timeseries to Evolutions



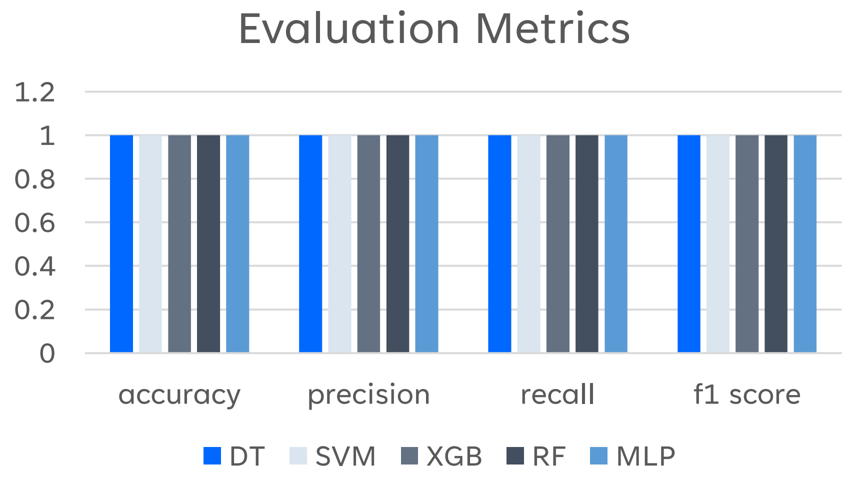
Train-Test data



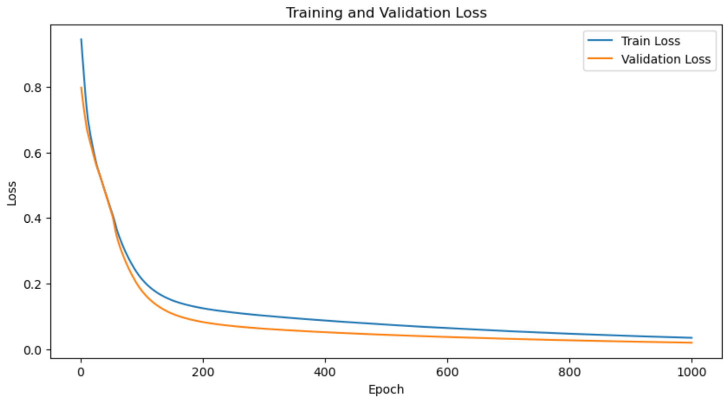
t1_m1	t1_m2	t1_m3	t1_m4	t2_m1	t2_m2	t2_m3	t2_m4	label
…	…	…	…	…	…	…	…	1

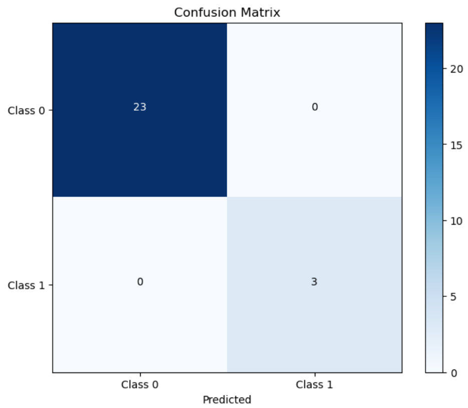
Machine Learning

Evaluation – BotIoT dataset

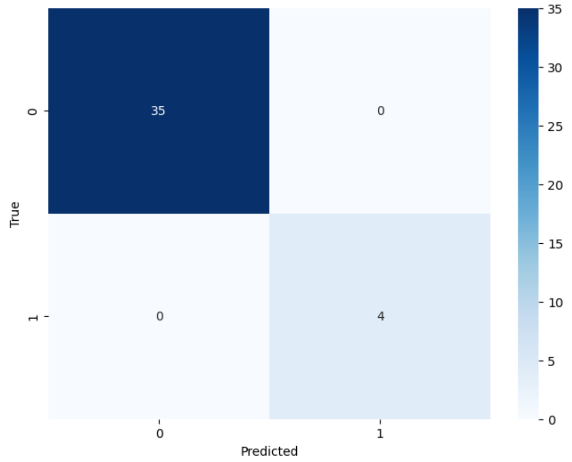


MLP

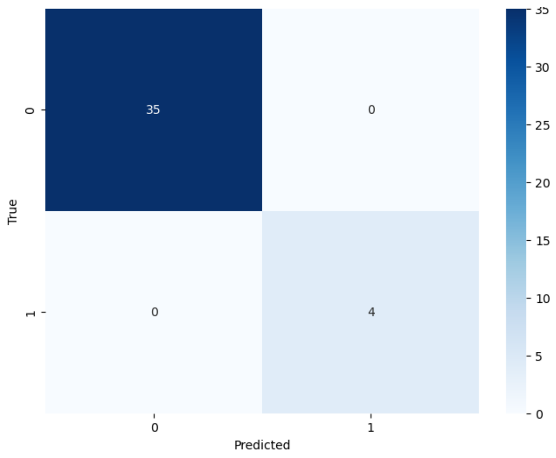




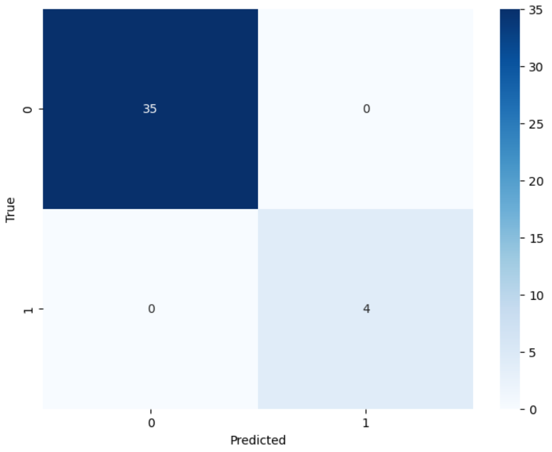
Random Forest



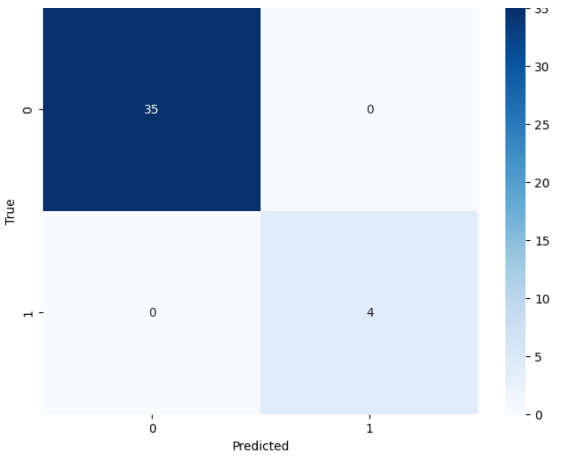
Decision Tree



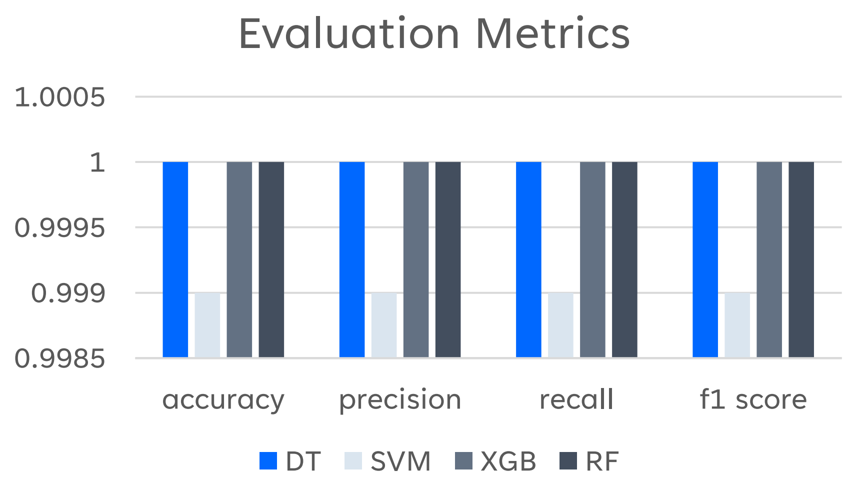
SVM



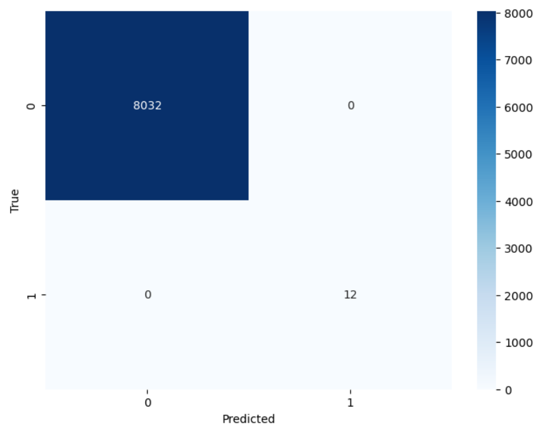
XGBoost



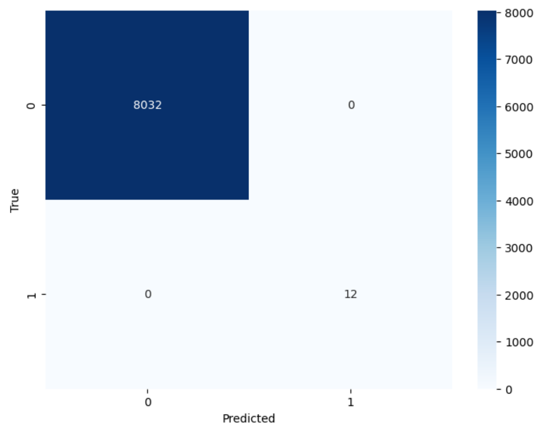
Evaluation – TonIoT dataset



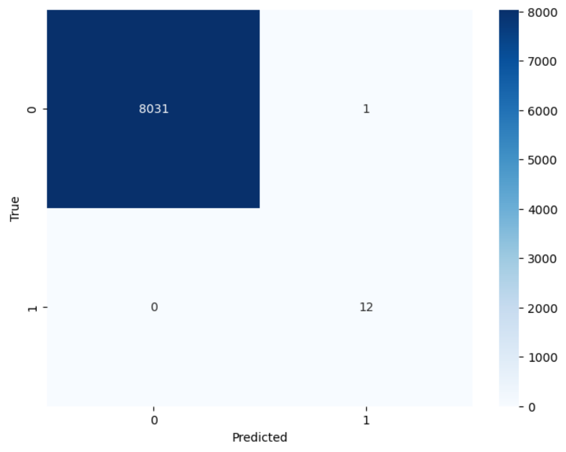
Random Forest



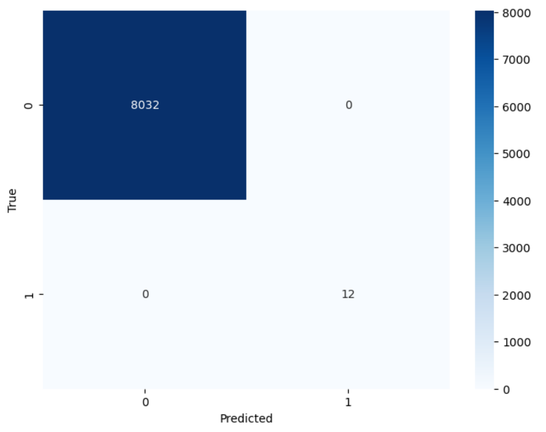
Decision Tree



SVM



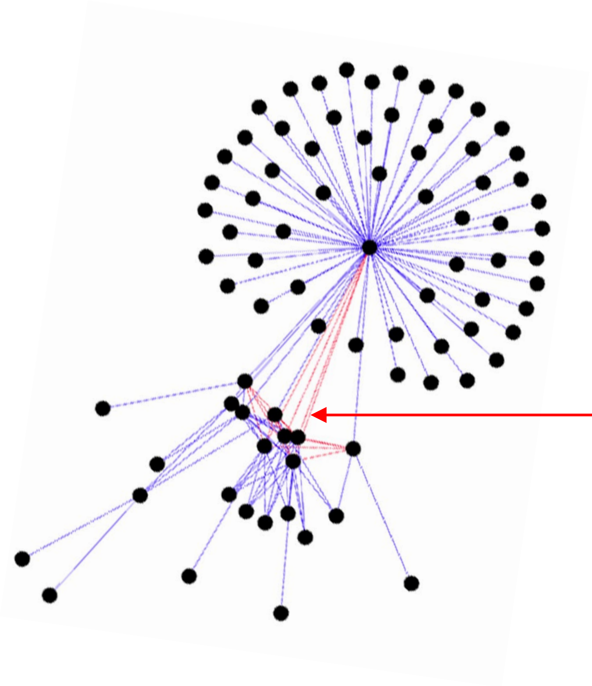
XGBoost



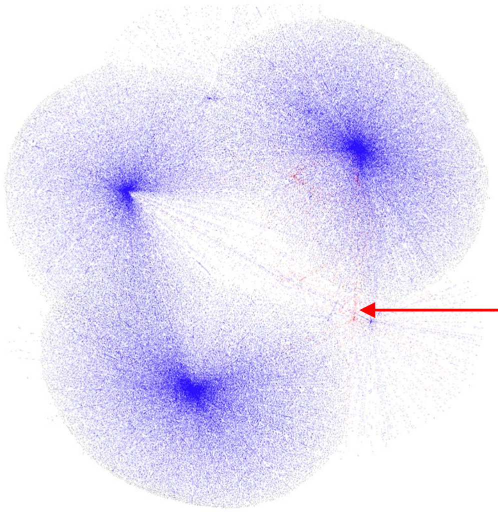
Achieved results

Metrics

Were able to detect attacks
using our supposed metrics

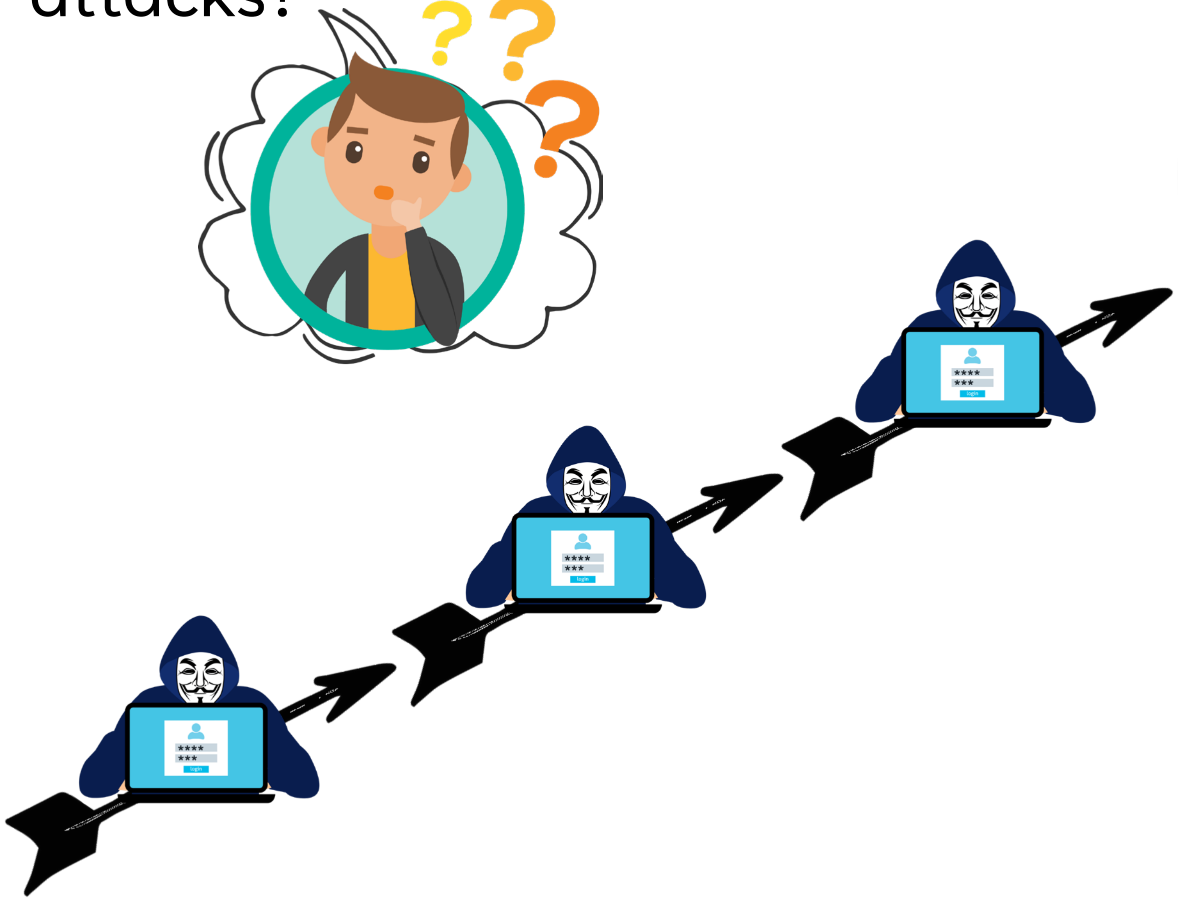


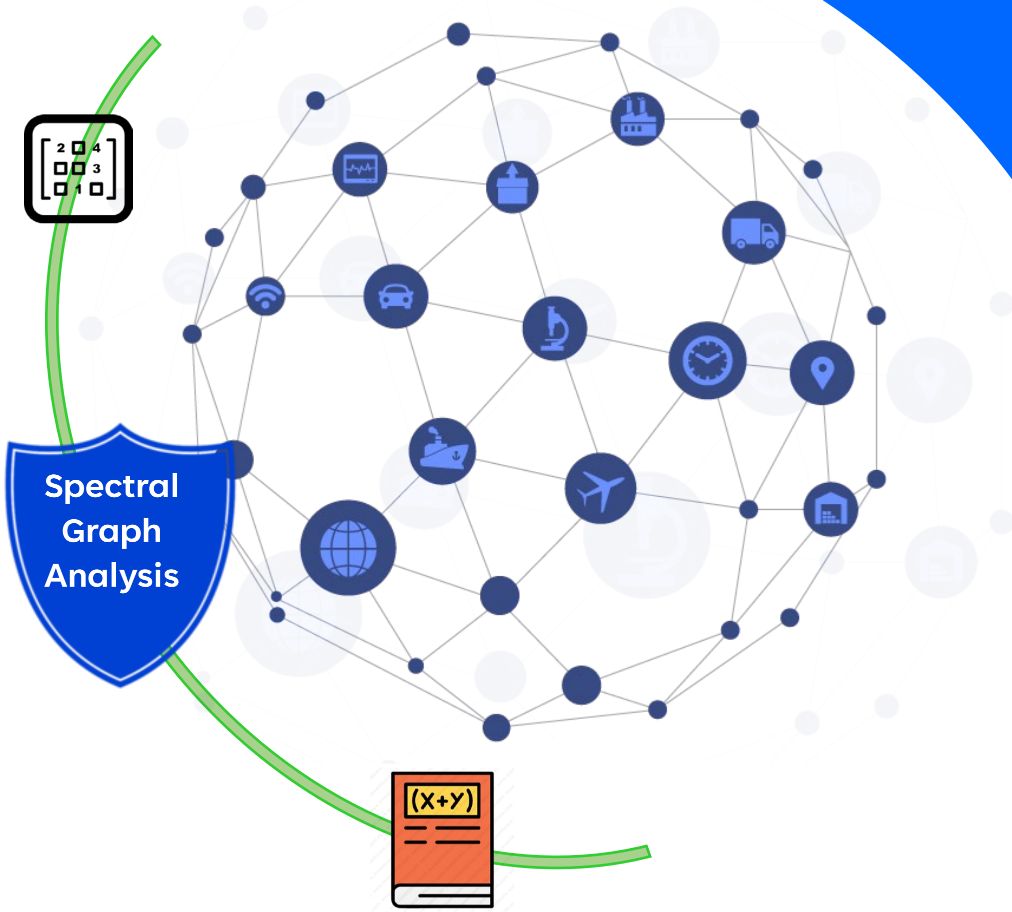
DOS
DDOS



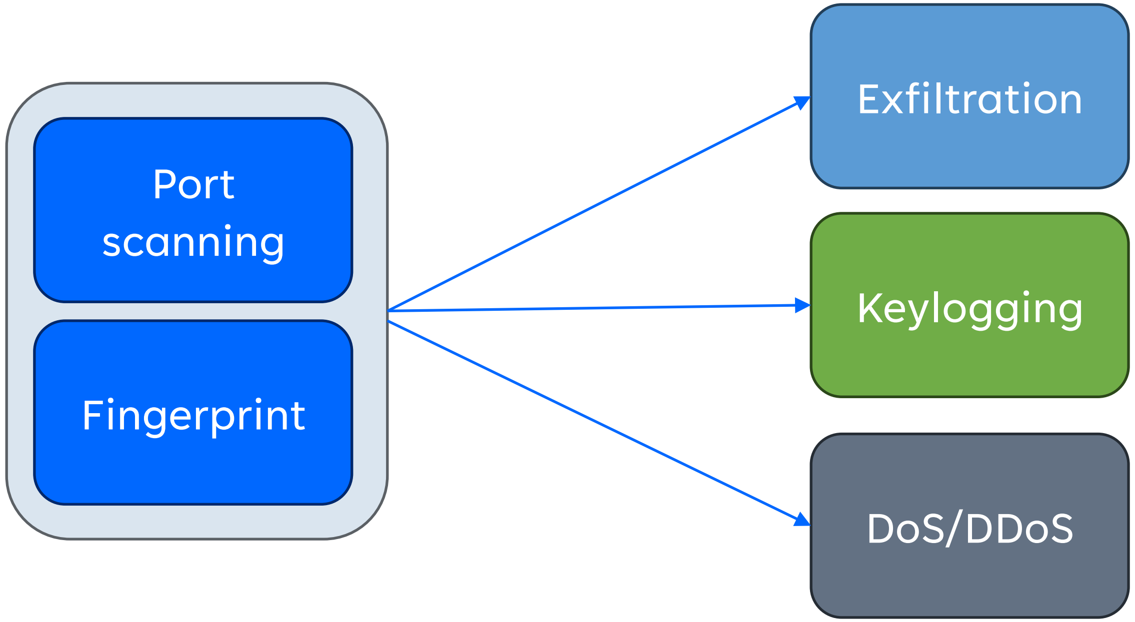
Second Contribution

Can spectral analysis detect advanced attacks, a multistep attacks?

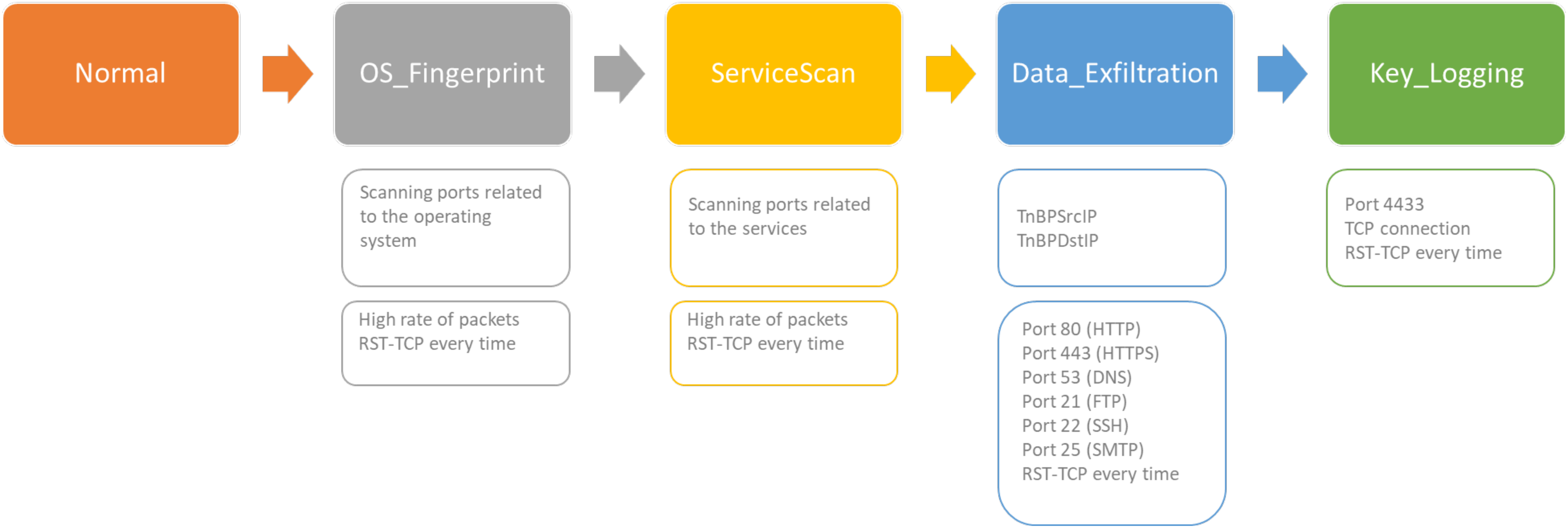




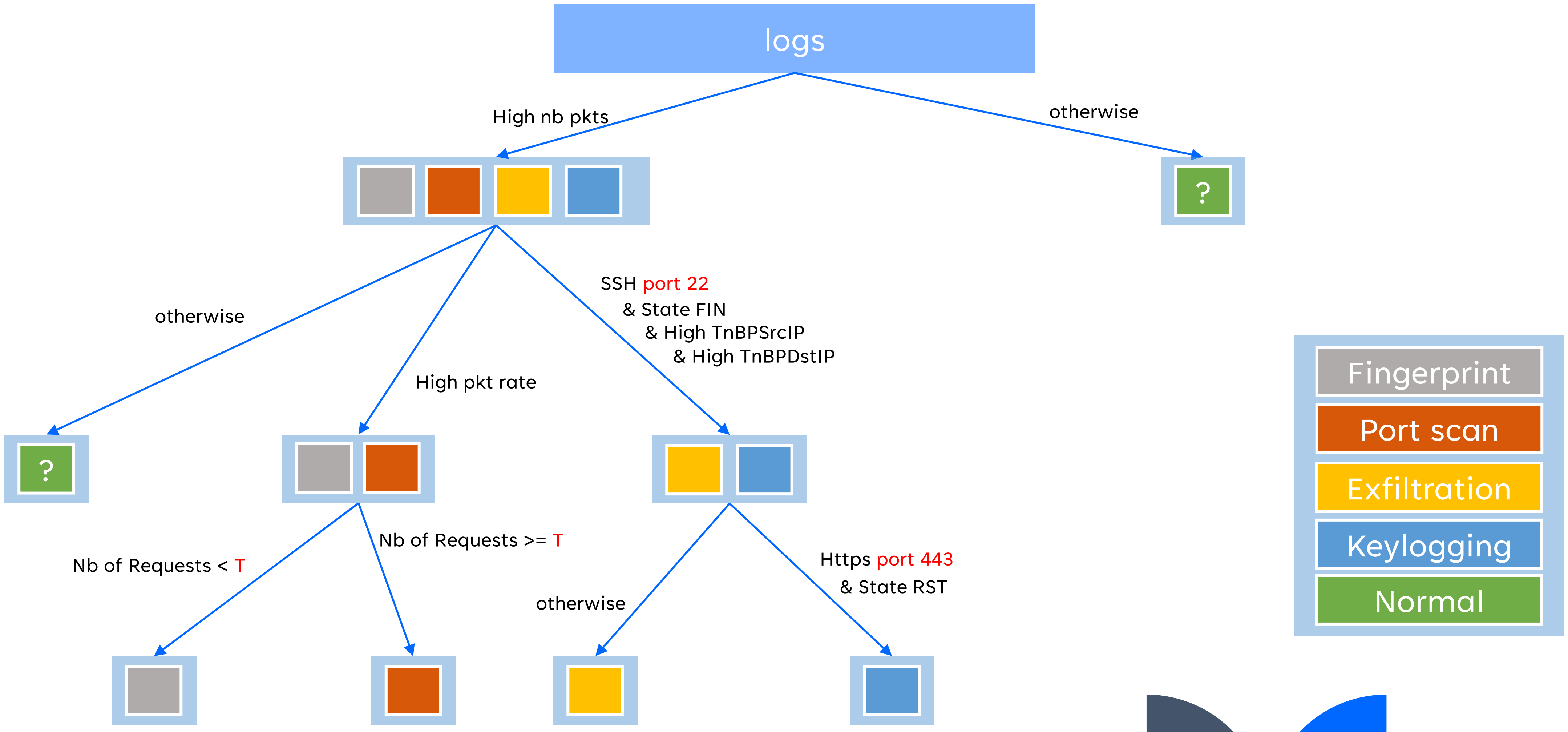
Multistep attack usecase



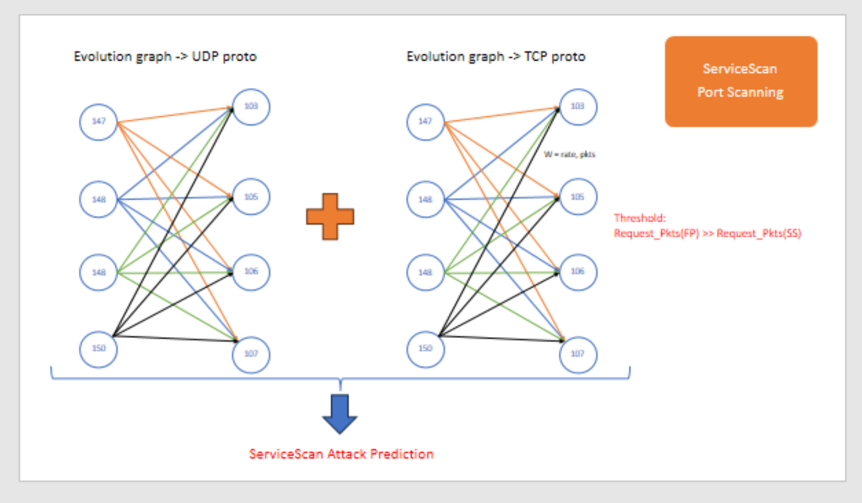
Sequence of multistep attack in BotIoT dataset



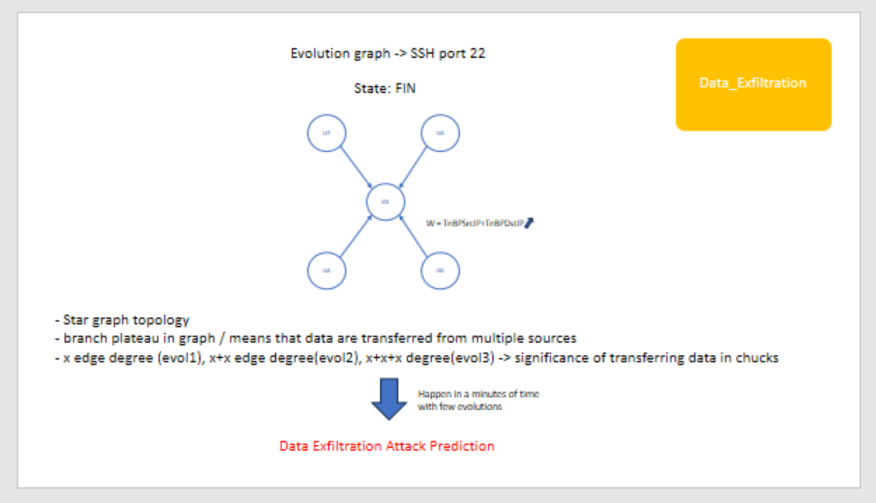
Multistep attack criteria



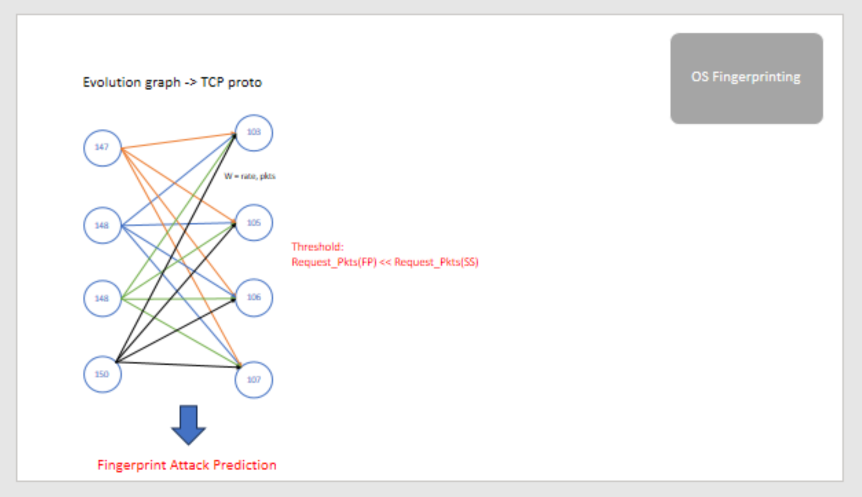
Coming work



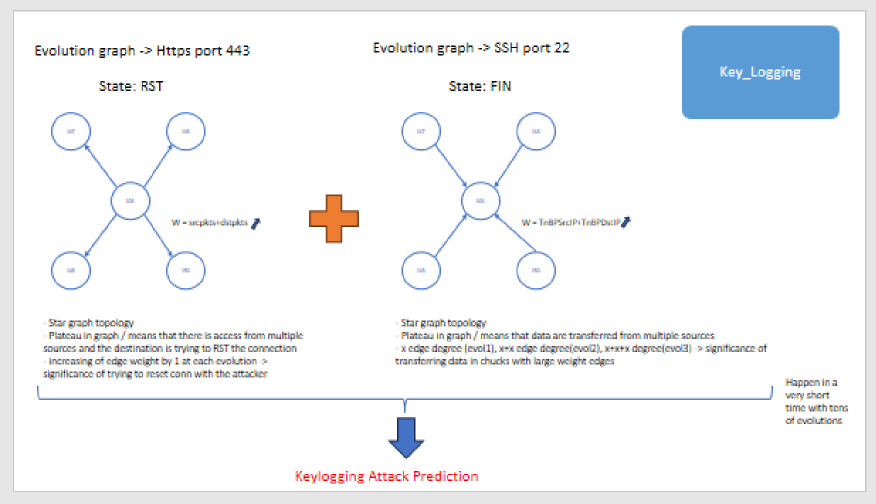
Evolution graph -> UDP proto
Evolution graph -> TCP proto
ServiceScan
Port Scanning
W = rate, pkts
Threshold:
Request_Pkts(FP) >> Request_Pkts(SS)
ServiceScan Attack Prediction



Evolution graph -> SSH port 22
State: FIN
Data_Exfiltration
- Star graph topology
- branch plateau in graph / means that data are transferred from multiple sources
- x edge degree (evol1), x+x edge degree(evol2), x+x+x degree(evol3) -> significance of transferring data in chucks
Happen in a minutes of time
with few evolutions
Data Exfiltration Attack Prediction



OS Fingerprinting
Evolution graph -> TCP proto
W = rate, pkts
Threshold:
Request_Pkts(FP) << Request_Pkts(SS)
Fingerprint Attack Prediction



Evolution graph -> Https port 443
Evolution graph -> SSH port 22
Key_Logging
State: RST
State: FIN
- Star graph topology
- Plateau in graph / means that there is access from multiple sources and the destination is trying to RST the connection
- increasing of edge weight by 1 at each evolution -> significance of trying to reset conn with the attacker
- Star graph topology
- Plateau in graph / means that data are transferred from multiple sources
- x edge degree (evol1), x+x edge degree(evol2), x+x+x degree(evol3) -> significance of transferring data in chucks with large weight edges
Happen in a very short time with tens of evolutions
Keylogging Attack Prediction

Thank you

Majed Jaber majed.jaber@epita.fr

Nicolas Boutry nicolas.Boutry@epita.fr

Pierre Parrend pierre.parrend@epita.fr

Any Questions