Testing Abnormality of a Sequence of Graphs: Application to Cybersecurity

C. Boinay[1], C. Biernacki[2], C. Preda[3], F. Foyer[4]

[1] *Seckiot & Inria,* [2] *Inria & U. Lille,* [3] *U. Lille & Inria* [4] *Seckiot*

Juin 2024

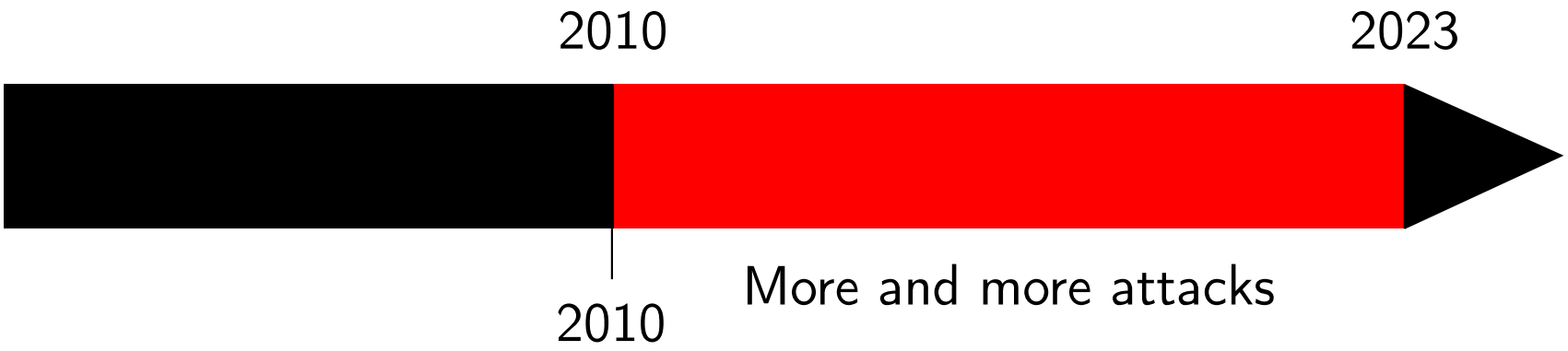
Outline

Operational Technology (OT)

- Part of modern critical infrastructures such as water treatment plants, oil refineries, power grids, and nuclear and thermal power plants
- Composed of heterogeneous and complex components: sensors and actuators, programmable logic controllers, supervisory control and data acquisition and human-machine interface

It is thus essential, but also challenging, to preserve OT from malicious actions (attacks)

Attacks in OT: Stuxnet as a game-changer



Stuxnet: 1st attack of an industrial system (Iranian nuclear power plant)

- Multiplication of the attacks since Stuxnet
- Stuxnet has shown that isolation of the network isn't enough to prevent attack

Standard approaches to detect an attack

- Solutions in IT (Information Technology) not sufficient to stop OT attacks (Raman, Ahmed et Mathur 2021)
- Firms use attacks history signature-based methods (Umer et al. 2022), but
  - What happens with a novel type of attack?
  - What happens if the signature is not well-chosen?
- Anomaly detection is the most efficient to stop a new attack since it can detect deviation of the normal behaviour (Raman, Ahmed et Mathur 2021)

Thus we focus on signature-free anomaly detection...

A way to see anomalies in the network: the graph

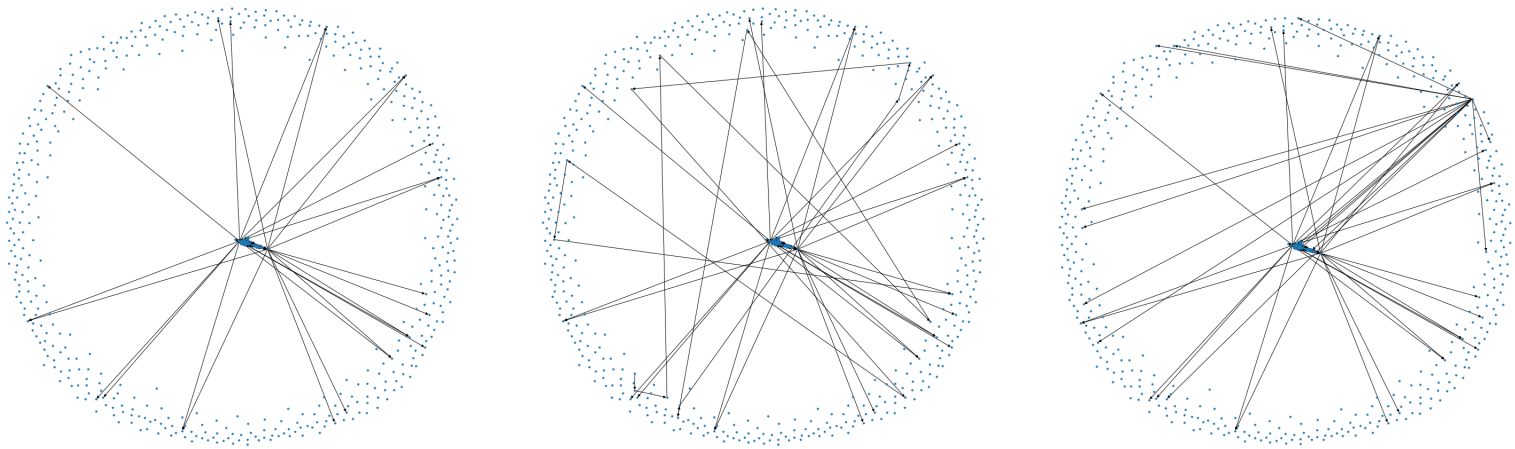
According to Neil et al. 2013, an attack in a network don't happen in isolation, but implies **an increase of communication** between multiple endpoints. Modeling the network with a graph ables to see such anomalies.
Typical behaviour are:

- Exploration of the attacker: a **star** in the graph
- Lateral movement: a **directed path** in the graph

Possible modeling:

- a node = an IP address
- a packet sent = an edge

Illustrating star and path



Normal graph (left), graph with a path (middle), graph with a star (right)

Graph anomaly detection in cybersecurity, a sparkling subject

- OT: up to our knowledge, no graph anomaly detection
- IT: graphs have been already used, for instance:
  - Calls of binary functions (Cohen, Yger et Rossi Nov 2021)
  - Stream of messages sent between IP adresses (in classification see Xiao et al. 2020 ; Abou Rida, Parrend et Amhaz 2021, in unsupervised learning with community detection, auto-encoder, scan statistics and edge streaming based on node embedding through random walk (see Ding et al. 2012 ; Neil et al. 2013 ; Leichtnam et al. 2020, Paudel et Huang 2022)
- But only one statistical work to test if there is an anomaly (Neil et al. 2013), otherwise poor statistical framework. . .

Outline

Our data: dynamical graphs of counting

- N IP addresses communicate over a time $[0, T]$ at different times $t \in [0, T]$ by sending messages
- $[0, T] = \cup_{i=1}^{n} I_i$ divided into n intervals of equal length Δ_t
- Only the number of messages is recorded for each I_i
- The aggregated data is $\boldsymbol{\mathcal{G}} = (\mathcal{G}_i)_{1 \leq i \leq n}$ where $\mathcal{G}_i = (\mathcal{N}, \mathcal{E}_i)$ with the set of nodes $\mathcal{N} = \{1, \ldots, N\}$ and $\mathcal{E}_i$ the list of (possibly duplicated) edges which send messages during I_i
- Equivalently to the $\mathcal{G}_i$s, we can construct the adjacency matrices X^is such that $\forall 1 \leq k, l \leq N, X^i_{k,l}$ is the number of messages sent by the IP address k to the IP address l

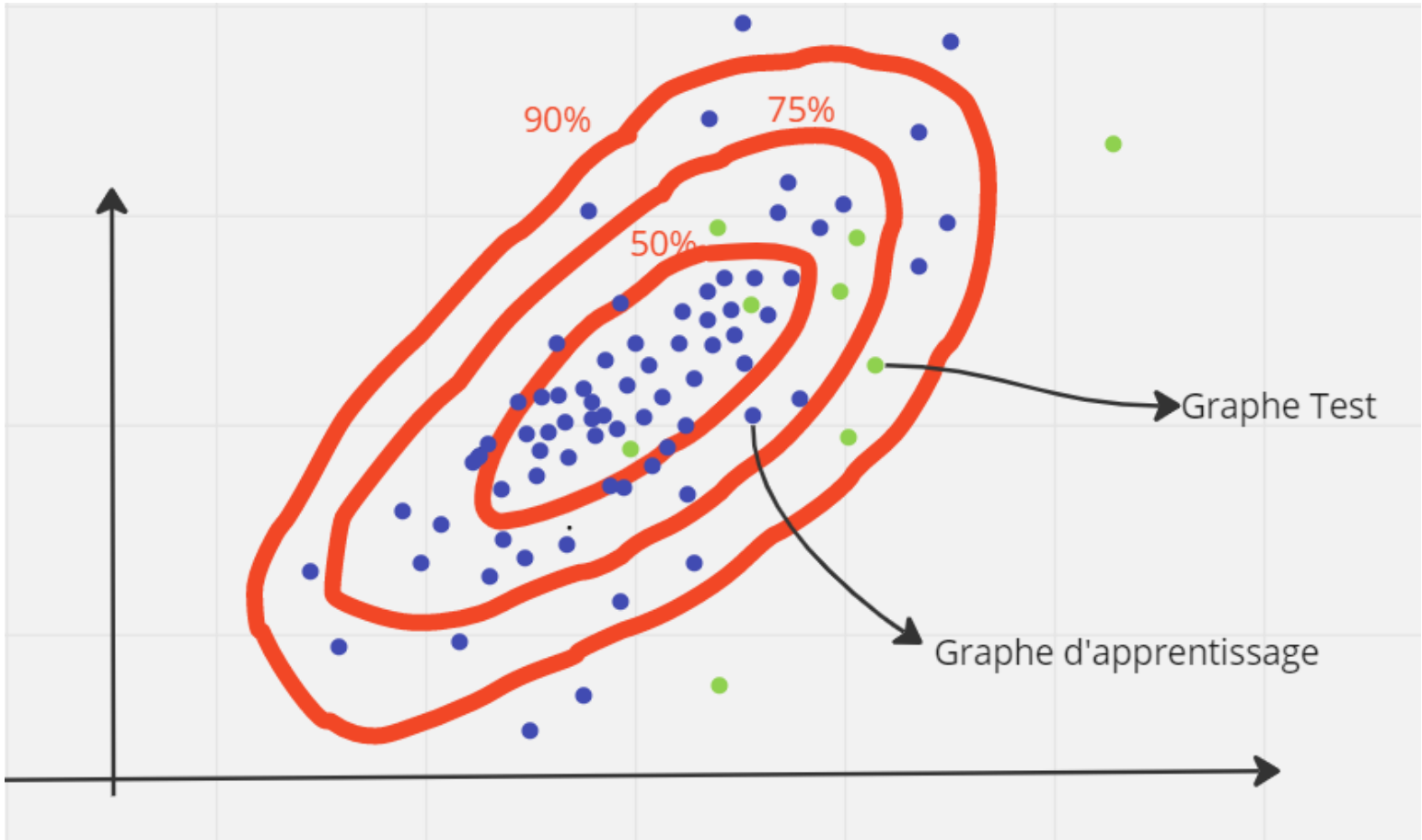
Justifying aggregation

Aggregating a Markov chain implies a quick loss of dependence between the aggregated time series. Independence of the aggregated values implies the independence of the dynamical graphs of counting,simplifying then the analysis.

Aggregating the communication over Δ_T ables to detect the increase of communication, specifically to the dynamicity Δ_T

Aggregating the instantaneous graphs ables to see the paths and the star in the same aggregated graph (agnostic representation).

Big picture



For a time step Δt

Which distribution in the network space?

Our solution for testing abnormality of a graph

We assume that the normal graphs are independent and identically distributed.

1. Learn a normal behaviour (distribution $\mathbb{P}_0$) over a sequence of graphs $\boldsymbol{\mathcal{G}} = (\mathcal{G}_i)_{1 \leq i \leq n}$ with a flexible family $\mathcal{F}$ of probability distributions such that $\mathbb{P}_0 \in \mathcal{F}$
2. Test if a new graph $\mathcal{G}_i$ has the normal behaviour $(i \geq n+1)$

$$\begin{cases} H_0 : \mathcal{G}_i \sim \mathbb{P}_0 \\ H_1 : \mathcal{G}_i \not\sim \mathbb{P}_0 \end{cases}$$

Test Statistics

Compute the distribution of the log-likelihood L_0 of the distribution $\mathbb{P}_0$ to get a pvalue of $L_0(\mathcal{G}_i)$

Choosing between different competitors

Retain the distribution family $\mathcal{F}$ which products the greater power for a given alternative distribution $\mathbb{P}_1$ $(i > n)$

$$H_1 : \mathcal{G}_i \sim \mathbb{P}_1$$

The distribution $\mathbb{P}_1$ represents a kind of attack, thus different scenarios to be tested...

A generic candidate: the Stochastic Bloc Model

What is the Stochastic Bloc Model?

- A mixture of probability of K latent classes
- A clustering model on graphs with K clusters

Why the Stochastic Bloc Model?

- The state space is the graph: a generic approach to detect anomalies
- We think that any probability of graphs can be approached by a Stochastic Bloc Model by increasing the number of classes K as continuous density functions can be approximated by finite mixture of Normal (Nguyen et al. 2020)

How to compute the Stochastic Bloc Model on multiple graphs?

- We assume the independence of the graphs given the partition
- We have done an adaptation of the Variational Expectation Maximization algorithm (VEM, Mariadassou, Robin et Vacher 2010) to multiple graphs

An efficient computation of the Variational Expectation Maximization

Problem:

- The VEM may be inefficient as it is slow and it can find local maxima
- Spectral methods have been used to initialize the VEM in the undirected and unvalued case (Lei et Rinaldo 2015)

Our solution:

- Initialize with a Singular Value Decomposition: as the partition based on the Singular Value Decomposition has been shown to converge to the SBM in the directed and unvalued case (Sussman et al. 2012), we demonstrate through experimentations, that it converges efficiently also to the SBM in the directed and valued case

Outline

Laboratory Data

- A laboratory of Seckiot
- 6 hours aggregated into $\Delta_T = 1$ minute
- 5 hours of benign traffic followed by 9 attacks
- 13 IP addresses

Careful: a test is performed on a controlled system, test on more complex data would be done in the future.

Interpretability of the abnormality

Two other statistics of test:

- Degree-out
- Log-likelihood of the edges

$\rightarrow$ We learn the distribution of such statistics over the learning dataset and compute the pvalues of each degree-out and log-likelihood of edges

Detection of the attacks and interpretability

Attack	pvalue of L_0	Abnormal nodes ($\alpha =$ 0.027%)	Abnormal edges ($\alpha =$ 0.027%)
ping sweep	12.66%	1	12
ARP scan	9.5%	0	6
overwrite registers	0%	2	6
get PLC info	11%	2	2
replay authent	0%	2	2
scan port modbus	9.9 %	1	6
man in the middle	1.8%	1	0
restart PLC	0%	1	2
stop PLC	9.9 %	2	2

Analysis of the result

True positives on the test set

- Some attacks have a low pvalue for the statistic log-likelihood of the graph
- With further analysis, we can say that abnormal nodes or edges are rightfully detected
- For each attack, there is at least one true positive of one of the three statistics

False Positive Rate for each of the statistics on the validation set:

- Log-likelihood of the graph: 0% ($\alpha = 0.027\%$)
- Degree-out: 0% ($\alpha = 0.027\%$)
- Log-likelihood of the edges: 0% ($\alpha = 0.027\%$)

Outline

A changing set of nodes over time

Problem: in various situations, the IP addresses can change, new equipments are installed, internet IP addresses appear or disappear in the network

Solution: the SBM can be adapted to this case through the missing value setting. A node which disappears is said to be missing. A node appearing was said to be missing. The test doesn't change.

Ongoing works

- Search for the rightful time step of aggregation: try different values of split Δ_t. An attack might not be seen for any Δ_t
- Interpretability of the classes: what is the link between the VLAN and the clusters of IP addresses

Thank you for your attention


Abou Rida, Amani, Pierre Parrend et Rabih Amhaz (2021). "Anomaly Detection for CyberSecurity Using Inductive Node Embedding with Convolutional Graph Neural Networks". In : Complex Networks and their Applications 2021, 30 novembre - 2 décembre 2021, Madrid, Madrid, Spain. url : https://hal.archives-ouvertes.fr/hal-03393640.


Cohen, Roxane, Florian Yger et Fabrice Rossi (Nov 2021). "Adding semantic to level-up graph-based Android malware detection". In : Complex Networks and their Applications 2021, 30 novembre - 2 décembre 2021, Madrid, Madrid, Spain. url : https://hal.archives-ouvertes.fr/hal-03393640.


Ding, Qi et al. (2012). "Intrusion as (Anti)Social Communication : Characterization and Detection". In : Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery KDD '12. Beijing, China : Association for Computing Machinery, p. 886-894. isbn : 9781450314626. doi : 10.1145/2339530.2339670. url : https://doi.org/10.1145/2339530.2339670.


Lei, Jing et Alessandro Rinaldo (fév. 2015). "Consistency of spectral clustering in stochastic block models". In : The Annals of Statistics 43.1. issn : 0090-5364. doi : 10.1214/14-aos1274. url : http://dx.doi.org/10.1214/14-AOS1274.


Leichtnam, Laetitia et al. (juin 2020). "Sec2graph : Network Attack Detection Based on Novelty Detection on Graph Structured Data". In : DIMVA 2020 : 17th Conference on Detection of Intrusions and Malware, and Vulnerability T. 12223. Lecture Notes in Computer Science. Lisbon, Portugal, p. 238-258. doi : 10.1007/978-3-030-52683-2_12. url : https://hal.inria.fr/hal-02950489.


Mariadassou, Mahendra, Stéphane Robin et Corinne Vacher (2010). "Uncovering latent structure in valued graphs : A variational approach". In : The Annals of Applied Statistics 4.2, p. 715-742. doi : 10.1214/10-AOAS361. url : https://doi.org/10.1214/10-AOAS361.


Neil, Joshua et al. (2013). "Scan Statistics for the Online Detection of Locally Anomalous Subgraphs". In : Technometrics 55.4, p. 403-414. doi : 10.1080/00401706.2013.822830. eprint : https://doi.org/10.1080/00401706.2013.822830. url : https://doi.org/10.1080/00401706.2013.822830.


Nguyen, T. Tin et al. (jan. 2020). "Approximation by finite mixtures of continuous density functions that vanish at infinity". In : Cogent Mathematics amp ; Statistics 7.1. Sous la dir. de Lishan Liu, p. 1750861. issn : 2574-2558. doi : 10.1080/25742558.2020.1750861. url : http://dx.doi.org/10.1080/25742558.2020.1750861.


Paudel, Ramesh et H. Howie Huang (2022). "Pikachu : Temporal Walk Based Dynamic Graph Embedding for Network Anomaly Detection". In : NOMS 2022-2022 IEEE/IFIP Network Operations and Management Symposium, p. 1-7. doi : 10.1109/NOMS54207.2022.9789921.


Raman, Gauthama, Chuadhry Mujeeb Ahmed et Aditya Mathur (2021). "Machine learning for intrusion detection in industrial control systems : challenges and lessons from experimental evaluation". In : IEEE Transactions on Industrial Informatics. doi : 10.1186/s42400-021-00095-5.


Sussman, Daniel L. et al. (2012).
A consistent adjacency spectral embedding for stochastic blockmodel graphs.
arXiv : 1108.2228 [id='stat.ML' $full_name =' MachineLearning' is_active = True alt_name = None in_archive =' stat' is_general = False description =' Covers machine learning papers (supervised, unsupervised, semi - supervised learning, graphical models, reinforcement learning, bandits, high dimensional inferen


Umer, Muhammad Azmi et al. (2022). "Machine learning for intrusion detection in industrial control systems : Applications, challenges, and recommendations".
In : International Journal of Critical Infrastructure Protection, p. 100516. issn : 1874-5482. doi : https://doi.org/10.1016/j.ijcip.2022.100516. url : https://www.sciencedirect.com/science/article/pii/S1874548222000087.


Xiao, Qingsai et al. (2020). "Towards Network Anomaly Detection Using Graph Embedding". In : Computational Science – ICCS 2020 12140, p. 156-169.