Graphs in OT Testing Graph Abnormality Application to a Real OT Data Set Ongoing & Future Works References

C Boinay, C Biernacki, C Preda

Testing Abnormality of a Sequence of Graphs: Application to Cybersecurity

C. Boinay[1], C. Biernacki[2], C. Preda[3]

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

54[e] Journées de Statistique

3-7 juillet 2023, Université libre de Bruxelles (ULB)

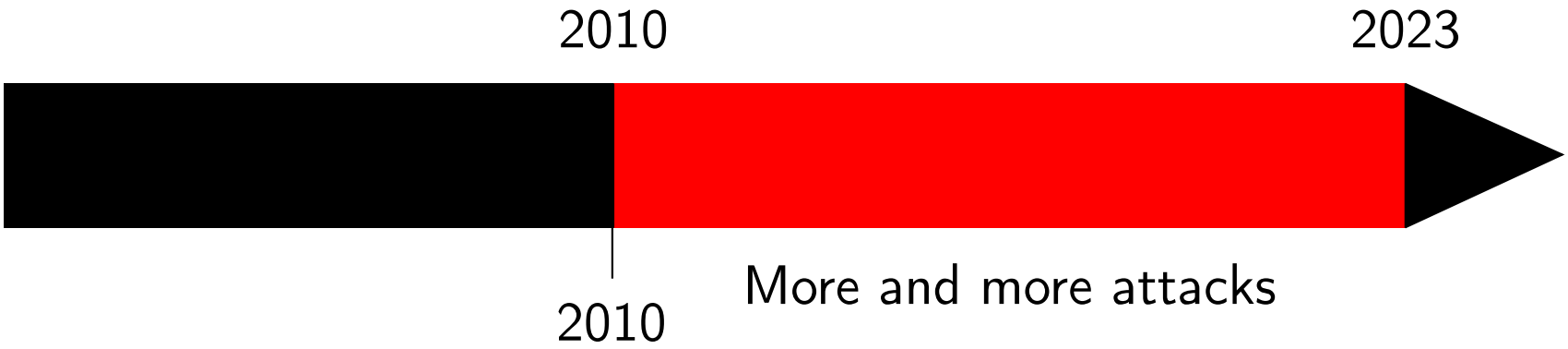
Outline

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)

Cyber attack (the National Cyber Security Centre)

Malicious attempts to damage, disrupt or gain unauthorised access to computer systems, networks or devices, via cyber means

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...

Graph anomaly detection

Graphs as natural structures to detect attack Neil et al. 2013

"Attacks do not happen in isolation on a single endpoint. Instead, they are exhibited across multiple endpoints, and in the communications between these endpoints."

- 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 and scan statistics see Ding et al. 2012 ; Neil et al. 2013 ; Leichtnam et al. 2020)
- But only one statistical work to test if there is an anomaly (Neil et al. 2013), otherwise poor statistical framework. . .

Our data: dynamical graphs of counting

- N IP adresses 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

Outline

Our solution for testing abnormality of a graph

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 ($i \geq n+1$)

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

Use the general bootstrap principle

Compute the distribution of the log-likelihood L_0 of the distribution $\mathbb{P}_0$ with a bootstrap to get a quantile $q_{0.05}$: for $i \geq n+1$, H_0 is said to hold if $L_0(\mathcal{G}_i) \geq q_{0.05}$

Flexible candidate families (competitors) $\mathcal{F}$ for $\mathbb{P}_0$

Given a graph $\mathcal{G}^*$, and considering graphs of the learning set as iid:

- Stochastic bloc model (SBM) with the number of classes K as hyperparameter. A variational EM (VEM) (Mariadassou, Robin et Vacher 2010) is adapted to learn over $\boldsymbol{\mathcal{G}}$. The log-likelihood associated to $\mathcal{G}^*$ $L_0^K(\mathcal{G}^*)$ (intractable) is approximated with the log-likelihood of the complete data.
- Gaussian kernel with the window $h > 0$:

$$L_0^h(\mathcal{G}^*) = \sum_{k \neq l} \log \left(\frac{1}{nh} \sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X_{kl}^* - X_{kl}^k}{h}\right)^2} \right)$$

- Poisson kernel with the window $h > 0$:

$$L_0^h(\mathcal{G}^*) = \sum_{k \neq l} \log \left(\frac{1}{n} \sum_{i=1}^{n} \frac{(X_{kl}^i + h)^{X_{kl}^*}}{X_{kl}^*!} e^{-(X_{kl}^i + h)} \right)$$

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...

Outline

Tuning the hyperparameters for a targeted level

- We use the dataset of a firm in OT (confidential)
- It is split into 3 datasets: a learning dataset to learn $\mathbb{P}_0$, a validation dataset to tune the hyperparameter for a targeted level, a test dataset to estimate the (unbiased) level

Model	Hyperparameter	Empirical level	Time of learning
SBM	$K = 30$	4%	46 hours
Gaussian kernel	$h = 90$	5%	1/2 hours
Poisson kernel	$h = 150$	10% [1]	1/2 hours

1. This model was not flexible enough to reach the targeted level of 5%

Two realistic hypotheses H_1: "star" and "path"

According to Neil et al. 2013, star and directed path in graphs are typical of cyber attacks.

Star construction

Star of size $\theta_s \in [\![1, N-1]\!]$ with β edges on each couple of nodes

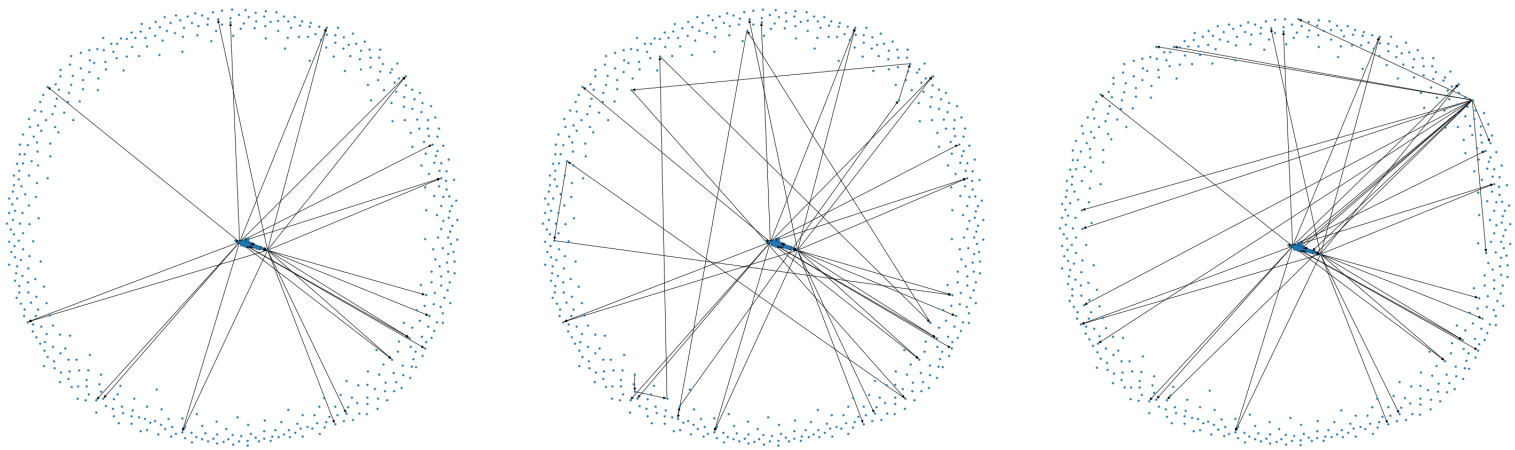
- Take a normal graph from the test set
- Choose uniformly a node k which will be the center of the star
- Choose uniformly θ_s nodes in $\{1, \ldots, N\}_{\{k\}}$
- Add a value β on the edges of the star

Path construction

Directed path of length $\theta_p \in [\![1, N]\!]$ with β edges on each couple of nodes

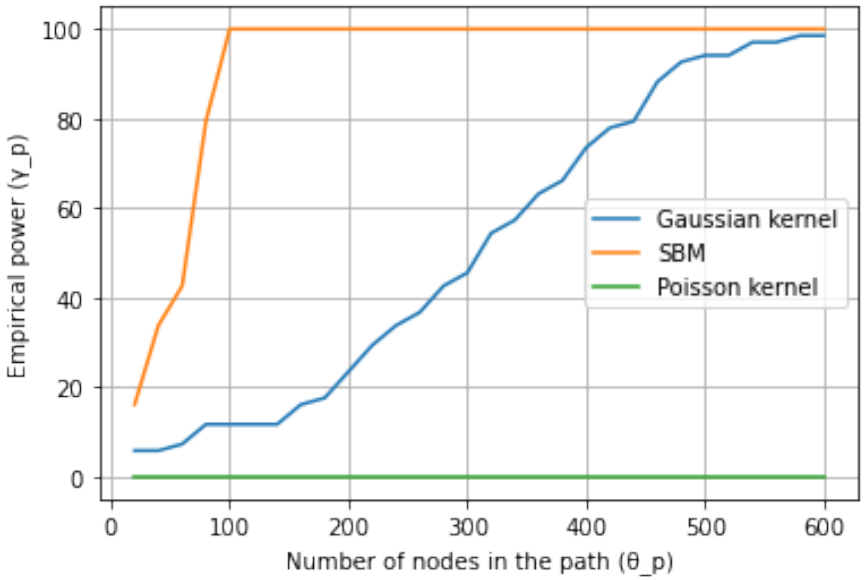
- Take a normal graph form the test dataset
- Choose uniformly θ_p nodes
- Add a value β on the edges of the path

Illustrating star and path

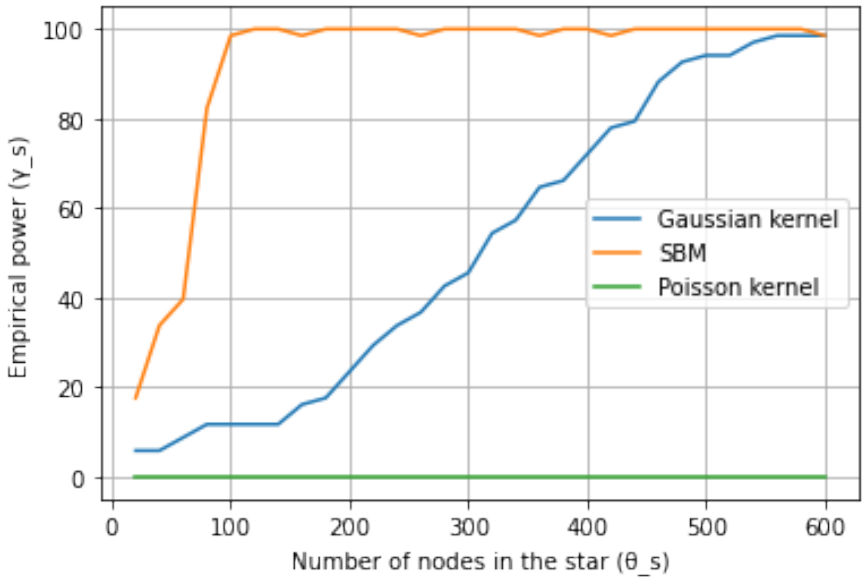


Normal graph (left), graph with a path (middle), graph with a star (right)

Empirical power



Empirical power for the path



Empirical power for the star

The SBM greatly outperforms Gaussian and Poisson kernels

Outline

Ongoing work: a sparse SBM as a new competitor (1/3)

- The SBM showed promising results
- However, 99% of the data in the adjacency matrices equal 0
- SBM doesn't take into account this great sparsity

Definition of the sparse SBM

We take into account sparsity by changing the Poisson distribution with a mixture of a Dirac in 0 and a Poisson truncated in 0 ($\beta \in [0,1]$, $\lambda \in \mathbb{R}_+$):

$$F(x;\beta,\lambda) = \mathbb{1}_{x=0}\beta + \mathbb{1}_{x\neq 0}(1-\beta)\frac{\lambda^x}{(e^\lambda - 1)x!}$$

Ongoing work: a sparse SBM as a new competitor (2/3)

Fix the value of K for a targeted theoretical level

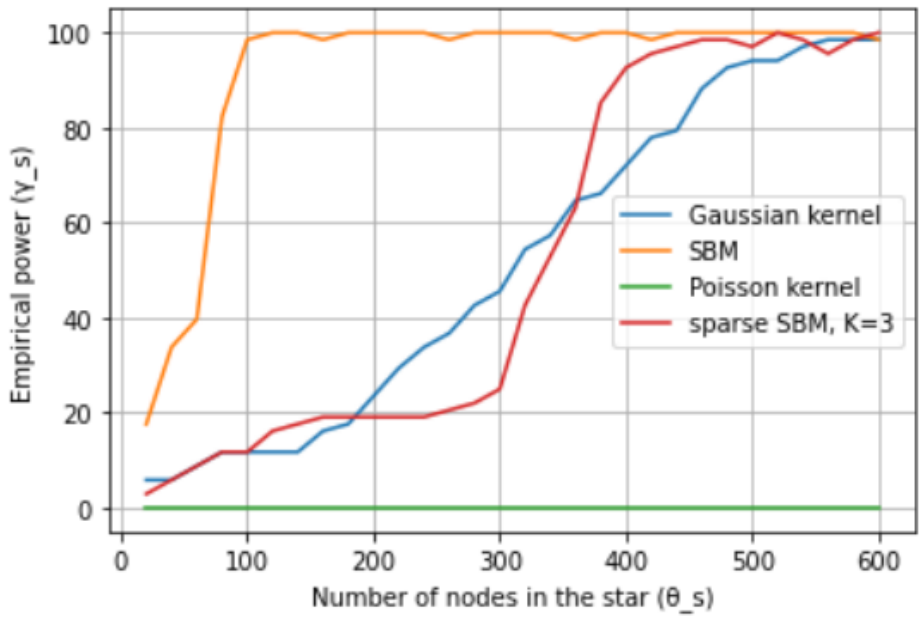
K	Empirical level	Time of learning
1	4.4%	2 hours
2	4.4%	4 hours
3	5.8%	6 hours
5	5.8%	18 hours

Effect of sparsity

We notice that we need much lower K value for reaching a given level than with SBM...

Ongoing work: a sparse SBM as a new competitor (3/3)

Empirical power of the star with the sparse SBM ($K = 3$)



The sparse SBM gives poor result in comparison with the SBM, thus this exploring work should be still continued...

Future works

- Pursue competitor proposals for $\mathbb{P}_0$
- Use scan statistics for relaxing iid assumption?
- Accelerate/avoid SMB-like estimation (hierarchical SBM?)
- Adapt the test to multiple testing (Benferroni correction)
- Investigate other real OT data sets
- Try different values of split Δ_t

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.


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.


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.


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.