Non Parametric and Semi-Parametric Modeling of a Sequence of Graphs for Testing Abnormality: Application to Cybersecurity

Clarisse Boinay[1], Christophe Biernacki[2], Cristian Preda[3], Thomas Anglade[4]

[1] Inria and Seckiot
clarisse.boinay@seckiot.fr
[2] Inria
christophe.biernacki@inria.fr
[3] Inria and University of Lille
cristian.preda@inria.fr
[4] Seckiot
thomas.anglade@seckiot.fr

The growing numerical dependency leads to more and more opportunities for malicious activities across all technologies including the Operational Technology (OT) field. Conventional detection methods (including IOC-based, signature-based methods) require the integration of static discriminant elements based on forensic analysis. Information about previous attacks are needed to partially detect the next mutated attack. It motivates the need for anomaly-based detection system which can detect behaviour deviations without previous knowledge. We model network data from Industrial Control System (ICS) with a graph where the nodes are the IP addresses and the edges are messages between 2 IP addresses. Such modelization is implemented in [3] and graph clustering has been used to detect intrusion in IT in [3, 4, 6] but not in OT as far as we know. We aim at classifying whether an input graph is normal given observations of a sequence of graphs considered as normal. To achieve this goal, we assume independence of the graphs and learn a probabilistic behaviour of the training graphs with different methods and then test new graphs with a likelihood parametric and non parametric boostrapping process.

We are thus looking for flexible models to learn the probabilistic behaviour of the normality. First we use a semi-parametric model of latent classes, the "stochastic block model", and we extend it to a sequence of graphs by adapting the Variation Expectation Maximization algorithm [5]. The choice of the number of classes with the criterion ICL [1] extended too to a sequence of graphs brings flexibility to such a model. Second, we try non parametric approaches leading to a variety of situations. Most of them are based on Poisson and Gaussian kernels. As a first example, we consider either a Poisson kernel of all the values of an edge in the past, or a Poisson kernel of all the values of all the edges where we assume similar behaviour across time and edges. The statistical unit can be the graph or the edge. As an extension such distribution can be mixed for increasing flexibility further and the corresponding weights of the mixtures are computed with an Expectation Maximization algorithm [2].

We compare the different proposed methods on real data from two firms and of different learning time through the empirical level and the power. As a preliminary step, such analysis of pros and cons of each method will able us to do more innovative and efficient modeling in our future works.

References

[1] Christophe Biernacki, Gilles Celeux, and Gérard Govaert. Assessing a Mixture Model for Clustering with the Integrated Classification Likelihood. Technical Report RR-3521, INRIA, October 1998.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[3] Qi Ding, Natallia Katenka, Paul Barford, Eric Kolaczyk, and Mark Crovella. Intrusion as (anti)social communication: Characterization and detection. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD'12, page 886–894, New York, NY, USA, 2012. Association for Computing Machinery.

[4] Laetitia Leichtnam, Eric Totel, Nicolas Prigent, and Ludovic Mé. 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 Assessment*, volume 12223 of *Lecture Notes in Computer Science*, pages 238–258, Lisbon, Portugal, June 2020.

[5] Mahendra Mariadassou, Stéphane Robin, and Corinne Vacher. Uncovering latent structure in valued graphs: A variational approach. *The Annals of Applied Statistics*, 4(2):715 – 742, 2010.

[6] Amine Medad, Baptiste Gregorutti, Edouard Genetay, and Alexandre Peter Nguema. Real-time graph clustering for network intrusion detection. 2021.