GRIFIN Research Internship: Robust and adaptive data-driven anomaly detection in IoT

Sébastien Tixeuil, Gregory Blanc, Thomas Silverston

Application deadline: January 4th, 2021

Keywords — anomaly detection, classification errors, robustness, adaptation, IoT

Context

According to Cisco[1], the Internet of Things (IoT) is expected to be the fastest growing trend in terms of connections on the 2018-2023 period. By 2023, IoT will represent half of the total devices and connections on the Internet. While it will contribute to connecting more people, homes and sites to the Internet and to offering a plethora of services, it is also expected to represent a major threat vector. As a matter of fact, a general lack of security awareness from both the IoT manufacturers and the consumers has led to an overall vulnerability of legacy IoT devices, and probably more recent ones, to a number of attacks, potentially leading to large-scale hijacking and massive attacks, as it has been witnessed for the last 5 years.

IoT devices are peculiar beasts as they are characterized by their high diversity in terms of types and communications, mobility, and pervasiveness. Some may run on constrained resources, others may be deployed in critical environments or be completely autonomous. Most times, they manipulate sensitive data. Nonetheless, they are often running proprietary software, which make them hard to protect, let alone update. For these reasons, resorting to network monitoring is reasonable in terms of computation, privacy and security. In particular, the state-of-the-art has produced complementary tools that allow to identify devices [1, 2] and detect deviant ones [3], based machine/deep learning techniques leveraging their distinct network behaviours. On small-scale networks and with very diverse devices, distinguishing proves to be easy, as IoT devices are also characterized by the fact that they perform a small number of specific tasks. At a greater scale, more features are needed to better separate them, and prevent misidentification.

Once the monitoring framework has identified the IoT devices, it can apply to the network a number of anomaly detection models to detect deviant behaviours from these devices. More generally, a number of behaviours may be acceptable from the set of IoT devices in a given network, regardless of whether the IoT devices are identified or not, as IoT is also characterized by its increased mobility, leading devices to come and go, and be attached to countless networks consecutively. But due to the high heterogeneity of IoT devices, the set of legitimate behaviours may be large and some may be expressed intermittently, or even rarely. All of these peculiarities

[1]Cisco Annual Internet Report (2018-2023) White Paper, available at: `https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html`

contribute to a possibly high number of false positives, that is events that are wrongly considered anomalies in the monitored network. Typically, software/firmware updates may alter the functions of a device and change its behaviour. While this is not malicious, the new behaviour will certainly trigger an anomaly.

In fact, anomaly detection struggles with the changing nature of anomalies itself [4], as traffic changes over time, deviating from the once learned representation. This requires the anomaly detection models to be regularly retrained to cope with the new behaviours, and sometimes forget the extinct ones. *Adaptive anomaly detection* has been studied in the early days of intrusion detection [5], but only recently regained interest thanks to the emergence of deep-learning techniques. *Reinforcement learning* allows to tune parameters automatically, reducing the operator's workload and improving the detection rate of anomalies [6].

Additionally, retraining needs to collect new data, and deep-learning techniques often require an excessive amount of data to be successful, which may not be amenable when deploying new IoT devices in a network. A possible solution is to use *transfer learning*, which enables the transfer of a previously learned model from one input space (the data collected from a first device) to another input space (characterizing another device). For exampe, deep transfer learning, which relaxes the assumption that training and testing data must be in the same domain (feature space and distribution), has shown promising results on images and time series anomaly detection [7].

Finally, learning can be distributed to overcome the constraints of resource-scarce devices. Beyond legacy distributed learning, federated architectures enable to collaboratively learn anomaly detection models efficiently, provided the models are uniform to the different federated networks. Federated learning [8] typically relies on a 2-tier architecture with a centralised model that aggregates local models scattered across the edge networks. In return, the local models are then updated. These communications repeat until convergence. This will allow to anticipate the occurrence of devices moving to networks where they were not previously seen.

This internship aims at surveying diverse training techniques to cope with data-driven anomalies and avoid the issues related to retraining (data scarcity, computation overhead). It will likely lead to the proposal of a monitoring architecture tailored to the chosen learning approach, as well as a protocol for adaptive learning in IoT networks.

Activities

- survey of retraining approaches and parameters
- survey of adaptive anomaly detection approaches
- focus on a specific IoT use-case
- design of a tailored learning architecture and communication protocol
- evaluation of the approach on some available datasets

Practical information

The internship will take place at LIP6, a laboratory of Sorbonne Université (Paris). It will be 5 months long.

Applicants are about to complete their Master 2 level degree (or equivalent engineering school degree) and should have the following skills:

- intermediate to strong knowledge and practice of machine/deep learning

- fundamentals in networking, and basic practice of traffic analysis
- concepts in cybersecurity, in particular intrusion detection
- practice in code development

The internship topic is linked to a Ph.D offer in the context of the GRIFIN project (funded by ANR), a research collaboration between Télécom SudParis, Sorbonne Université and LORIA.

Applications (resume, motivation letter, academic transcripts, recommendation letters) must be sent to `sebastien.tixeuil[at]lip6.fr` and `gregory.blanc[at]telecom-sudparis.eu`.

References

[1] Mustafizur R Shahid, Gregory Blanc, Zonghua Zhang, and Hervé Debar. Iot devices recognition through network traffic analysis. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 5187–5192. IEEE, 2018.

[2] Nesrine Ammar, Ludovic Noirie, and Sébastien Tixeuil. Autonomous identification of iot device types based on a supervised classification. In *ICC 2020-2020 IEEE International Conference on Communications (ICC)*, pages 1–6. IEEE, 2020.

[3] Mustafizur R Shahid, Gregory Blanc, Zonghua Zhang, and Hervé Debar. Anomalous communications detection in iot networks using sparse autoencoders. In *2019 IEEE 18th International Symposium on Network Computing and Applications (NCA)*, pages 1–5. IEEE, 2019.

[4] Robin Sommer and Vern Paxson. Outside the closed world: On using machine learning for network intrusion detection. In *2010 IEEE symposium on security and privacy*, pages 305–316. IEEE, 2010.

[5] Wenke Lee, Salvatore J Stolfo, and Kui W Mok. Adaptive intrusion detection: A data mining approach. *Artificial Intelligence Review*, 14(6):533–567, 2000.

[6] Tong Wu and Jorge Ortiz. Towards adaptive anomaly detection in buildings with deep reinforcement learning. In *Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, pages 380–382, 2019.

[7] Vincent Vercruyssen, Wannes Meert, and Jesse Davis. Transfer learning for time series anomaly detection. In *CEUR Workshop Proceedings*, volume 1924, pages 27–37, 2017.

[8] Jakub Konečnỳ, H Brendan McMahan, Daniel Ramage, and Peter Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*, 2016.