Unsupervised and Incremental Learning Orchestration for Cyber-Physical Security

Lúcio Henrik Amorim Reis, Andrés Felipe Murillo Piedrahita, Sandra Julieta Rueda Rodríguez, Natália Castro Fernandes, Dianne Scherly Varela de Medeiros, Marcelo Dias de Amorim, Diogo Menezes Ferrazani Mattos

Unsupervised and Incremental Learning Orchestration for Cyber-Physical Security

Lúcio Henrik A. Reis, Andrés F. Murillo, Sandra Rueda, Natália C. Fernandes, Dianne S. V. Medeiros, Marcelo Dias de Amorim, Diogo M. F. Mattos

Abstract—Attacks on cyber-physical systems, such as nuclear and water treatment plants, have physical consequences that impact the lives of thousands of citizens. In such systems, it is mandatory to monitor the field network and detect potential threats before a problem occurs. This work proposes a hybrid approach that combines unsupervised and incremental learning methods to detect threats that impact the control loops in a plant. We use the idea of *online processing of honeypot data* to identify new attack vectors and to train the online incremental learning method as new attacks arrive. We also apply a one-class support vector machine to each monitored sensor or actuator to retrieve abnormal behaviors of their closed control loop. The proposed solution orchestrates the outputs from the two machine learning methods and alerts the system operators when it detects a threat. We evaluate the proposal on the Secure Water Treatment (SWaT) testbed dataset, and the results reveal that the proposed machine learning orchestration detects threats at more than 90% precision and with accuracy higher than 95%.

Keywords—Cyber-physical security, Honeypot, Machine learning, Incremental learning.

1 Introduction

Cyber-physical systems (CPSs) are widely distributed, large-scale heterogeneous systems in which sensors, actuators, Programmable Logic Controllers (PLCs), and other elements form a network to measure and control a physical system [1]. The critical characteristic of CPS requires integration of ubiquitous computing, efficient network communication, and control in physical processes. As any critical networked system, CPSs are prone to security vulnerabilities that may imply severe economic losses or social disorder.

Two key challenges to assure the operational and security goals of a cyber-physical system are to *anticipate the threats* and to *identify anomalous behaviors* in both field and supervisory networks of the physical system. To address these issues, a possible solution is to rely on machine learning techniques. Nevertheless, although attacks against CPS are a reality nowadays, there are only a few available datasets describing physical or cyber attacks – and such datasets are required to train the learning system. As a consequence, the development of security solutions is hindered. Moreover, each CPS has its particularities, which hardens even more the design of effective security countermeasures.

In this work, to circumvent the lack of data, we propose the use of a high-interaction honeypot to gather attack data targeting the CPS infrastructure. With such data in hands, we propose to apply unsupervised and incremental machine learning techniques to detect threats. Our proposal orchestrates both zero-day threat detection through honeypot data analysis and anomaly detection through pattern discovery on regular CPS traffic. To this end, we first deploy an unsupervised one-class classification mechanism to infer anomalies on the monitored data of each sensor or actuator in the field network. Then, we deploy an incremental learning classifier, fed by the honeypot traffic, as a labeled dataset of threats. The trained incremental classifier discriminates incoming

- *L. H. A. Reis, N. C. Fernandes, D. S. V. Medeiros, and D. M. F. Mattos are with the Department of Telecommunications Engineering, Universidade Federal Fluminense, Niterói, Brazil. E-mail: lucioreis@id.uff.br*
- *A. M. Piedrahita, and S. Rueda are with Universidad de los Andes, Bogotá, Colombia.*
- *M. D. Amorim is with LIP6/CNRS, Sorbonne Université, Paris, France.*

Manuscript received xxxx; revised xxxx.

traffic from the private network of the protected CPS, as well as it enhances the training with each new sample. The proposed orchestration of learning methods generates alerts whenever either the incremental learning or the one-class classifier emits a threat classification. We evaluate our proposal using the Secure Water Treatment (SWaT) testbed dataset, and the results show that the proposed machine learning orchestration achieves an accuracy of at least 95% and detects threats with a precision of more than 90%.

2 Unsupervised and Incremental Learning for Cyber-Physical Threat Detection

Detecting threats in cyber-physical systems is challenging, as it is necessary to analyze specific features in data coming from both the communication network and sensors/actuators. Modeling such a system with a finite automaton is impractical because the number of states tends to explode. Thus, it is impractical to run a monolithic model that encompasses both the communication network and sensors/actuators. Machine learning-based models are serious candidates as these models learn patterns from the data [2].

In this work, we identify threats in a CPS through a dual analysis of the communication network and physical sensors' monitored data. The communication network data represent the cyber features, bringing information about protocols and applications running on the system. The physical data consists of the sensor measurements and actuator commands, extracted from the application data in the packets. We orchestrate two machine learning strategies, *incremental learning* and *unsupervised one-class classification*, as shown in Figure 1. The incremental learning component implements an online mechanism to learn from previously unknown attacks. The unsupervised one-class classification, in turn, is an offline solution that focuses on detecting anomalies in the data. In a nutshell, the incremental classifier seeks for zero-day threats on the real CPS, while the one-class device-based classifier seeks for control loop deviation for each physical process. Thus, our proposal monitors both the cyber-attacks based on zero-day threats and the sabotage on physical processes due to a deviation of the control loop.

2.1 Honeypots and Incremental Learning

One of the challenges we face is that there are very few datasets available to train the machine learning models. Furthermore, the ones available are either synthetic or contain just a few samples. To overcome this limitation, we propose to use a honeypot that emulates the real topology to gather abnormal data. As the honeypot is not intended to offer any legitimate public service, it is safe to assume that all interactions come from malicious or curious users [3]. Either way, we consider that the honeypot attracts unwanted activities exclusively. The honeypot runs on public, vulnerable IP addresses and we assume as ground-truth that all traffic on the honeypot is considered illegitimate. Thus, we analyze it as threat behavior.

The honeypot traffic feeds an incremental learning classifier as a labeled dataset of threats. The incremental learning mechanism updates its learning parameters to identify new cyber-attacks. We adopt the incremental version of the Support Vector Machine (SVM) with the incremental step based on the Stochastic Gradient Descent (SGD). The SVM is a binary classifier based on a decision hyperplane that defines the class boundaries. A hyperplane built in a multidimensional space divides the data. The incremental SVM algorithm uses a smooth margin approach with the hinge loss, a convex function that is not differentiable, but it has sub-gradients. Thus, the basic SGD algorithm searches for the optimal maximal margin of separability between classes. The SVM is a particularly useful linear predictor for high dimensional feature spaces, which represents a computationally complex learning problem. Moreover, the SGD algorithm is useful when the data are sparse and takes less than linear time and space per iteration to optimize a system. Therefore, the SVM predictor with incremental SGD algorithm for optimizing the maximal separability margin well fits the need for an incremental learning method that does not introduce heavy computation resource requirements.

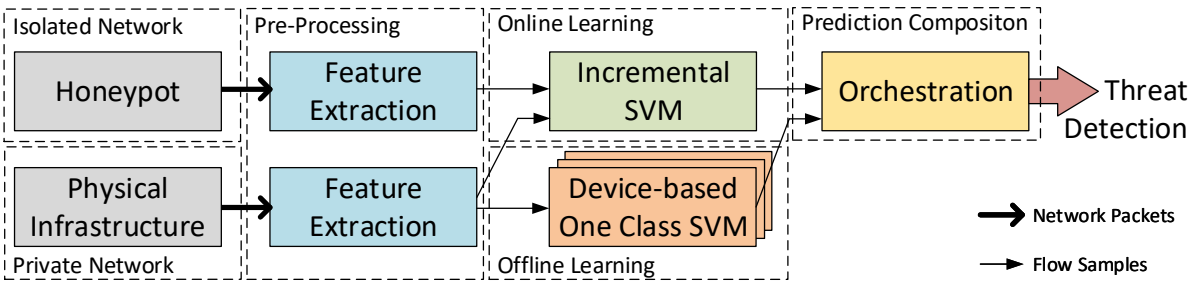


Fig. 1. Orchestration of two learning methods, incremental and one-class learning. A honeypot provides malicious activity data for the incremental SVM which updates the classification parameters. Monitoring of the physical infrastructures provides normal-behavior samples for both one-class and incremental SVM. The orchestration generates alarms when a new threat is detected by any of the learning method.

2.2 Data, Anomalies, and Unsupervised Learning

Besides the lack of data, some CPS protocols used within the EtherNet/IP stack rely on device-specific I/O messages that follow non-standard formats of fixed lengths defined by the control system vendors. As these messages are non-standardized, it becomes a challenge to develop a monitoring system based on machine learning that extracts knowledge from the network communication. In our proposal, the unsupervised one-class classification mechanism infers the probability distribution of the monitored data on each sensor or actuator in the field network. The key idea of inferring the normal behavior of the control data in the field devices derives from the predictability and closed-loop behavior of the control functions. As the field network is always generating data, from time to time, the one-class strategy updates learn normal behavior of the physical process. Any sample that does not fit inside the surface of the normal class is reported as an anomaly.

The one-class learning strategy relies on the one-class SVM method. One-class SVM is an unsupervised learning algorithm that derives a decision hyperplane for anomaly detection. New data are classified as similar to or different from the training set. In contrast with typical SVM implementations, the one-class takes into consideration a training set of samples from a single class. Any new sample that does not fit into the decision surface defined by the training set is considered an instance of a new class and, thus, an anomaly. As the one-class is more susceptible to variations on the distribution of probabilities of the features, we model a classifier for each monitored sensor/actuator.

2.3 Specific Design Choices

An essential step before data analysis is preprocessing, which refers to the organization of data in windows, the extraction of useful features, and the enrichment of the data. Both implemented learning strategies use fixed-duration time windows to organize data. The key idea of organizing data in windows is to evaluate instantaneous values as well as recent variations of features values, e.g., the variation on the measure of a sensor is available when considering a set of recent measures of the sensor. Thus, means, variations, and counts of values on a time window are added to a sample from the network to enrich data.

Considering the incremental learning strategy, we deploy a single classifier for the entire network, primarily focused on retrieving features that discriminate cyber-attacks on the honeypot and the normal behavior of the infrastructure network. Thus, the incoming samples of the incremental learning component include network features, such as device identity, network protocol, type of messages, transport protocol, count of packets, and also physical-related features, such as Modbus message type, read/write information, register, and value. On the other

hand, the one-class strategy applies a classifier per monitored device and considers just features that express the physical state of the device, such as the value of each device register and recent variations of these values. This dual strategy of having different features in each classifier enables monitoring of both physical and cyber threats.

3 Case Study: Water Treatment System

3.1 Dataset

To evaluate our proposal, we use the Secure Water Treatment (SWaT) dataset [4]. This dataset comes from a realistic testbed representing a scaled-down version of a real-world industrial water treatment plant. The testbed has a six-stage filtration process corresponding to the physical and control components of the water treatment facility. Each stage has a dedicated PLC that communicates directly with the sensors and actuators of the specific stage, and with the SCADA server. In the tanks related to the first (raw water), third (filtering) and fourth (dechlorination) processes, there are Level Indicator Transmitters (LITs), which are sensors to indicate the level of the water.

The data in the SWaT testbed [4] was collected for 11 days. The resulting dataset contains both the physical properties of the plant and the water treatment process, as well as the network traffic, comprising a total of 946,722 samples. As the authors assume that no severe attacks can be launched within less than one second, data collected from sensors and actuators are recorded once every second. The network traffic concerns the communication between the SCADA system and each PLC. A total of 36 types of attacks were launched against the testbed. Unfortunately, in the SWaT dataset, the only attack that involves both the network and the physical devices is the replay attack – for this reason, in our analyses, we focus on this specific attack.

The SWaT testbed relies on the EtherNet/IP protocol stack in the supervisory system networks and on the Fieldbus [5]. However, at each layer of the network, the encapsulated protocol is different. The Common Industrial Protocol (CIP) communicates devices in the control and supervisory network as well as in the Fieldbus network. Communication between controllers and sensors/actuators relies on device-specific I/O messages. The CIP stack meets three primary needs of cyber-physical systems: control, configuration, and data collection. CIP defines the application layer as an encapsulated and object-oriented protocol. CIP messages have rich semantics about information exchanged in the network, which helps understand the process and define the process-based machine learning models. CIP messages follow the object-oriented paradigm and allow the transmission of a variable number of data registers with distinct types. In the Fieldbus network of the SWaT testbed, device-specific I/O EtherNet/IP messages follow non-standard formats of fixed lengths defined by the vendor.

3.2 Evaluation and Results

We implemented a prototype of the proposed machine learning approaches. Our prototype is written in Python and relies on the *sklearn*[1] library as the substrate for the learning algorithms. Because we use the SWaT dataset, we focus on the deployment and analysis of the machine learning methods, instead of implementation details of the field network or the honeypot. To achieve a honeypot-like dataset, we fragmented the attacks on CPS on small batches that simulate attacks on a honeypot that mimics the real CPS network. We establish as ground-truth that all traffic on the honeypot is malicious. In the SWaT dataset, attacks are discriminated by the hour of the day and while an attack happens, attack flows and normal flows coexist.

We analyze the Modbus/TCP protocol to help us design the machine learning strategies. We engineer the following features: source and destination IP addresses, Modbus function code, eight registers of Modbus values, number of changes of the values of the register within a sampling window, and number of packets involved in the Modbus communication. We highlight that these features allow the detection of all attacks that influence these features in

1. Available at https://scikit-learn.org/.

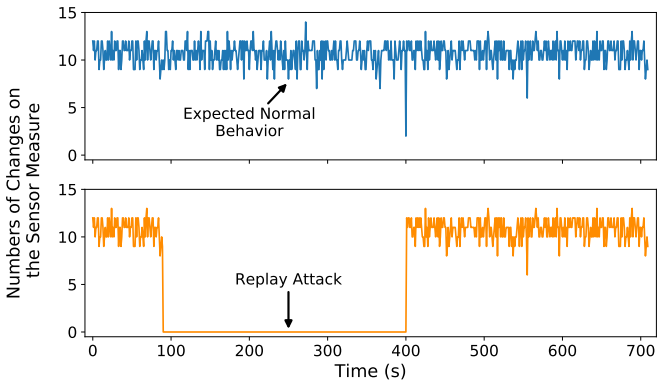


Fig. 2. The key features on the identification of a replay attack are the value of the level sensor and the number of changes of this value within a window. The number of changes of the level sensor within a sampling window is constant while the replay attack happens.

an unusual fashion. As stated previously, we evaluate our proposal against a replay attack. In the SWaT dataset, the replay attacks consist on setting the value of a tank level measurement to a constant. The immediate effect is the underflow or overflow of a tank, damaging the system and causing financial loss. To cause de overflow, for instance, the attacker captures a Modbus message indicating the tank volume when the tank starts to fill and continuously replays this message, indicating that the tank has a low volume, when it is indeed being fulfilled. Hence, while the attack is happening, the number number of changes on the sensor measurement is constant, as shown in Figure 2. We use the number of changes of the Modbus Register values as a primary feature for our attack detection proposal. Using these values as a feature of the machine learning model is the main reason for the incremental learning strategy to behave with high accuracy during the early stages. This happens because the attack differs from the remaining of the normal data previously collected.

We evaluate the performance of the two machine learning strategies using well-known metrics such as precision, accuracy, sensitivity and, specificity. Precision measures how many of the predicted true elements are indeed true. The accuracy is the fraction of results correctly identified. The sensitivity is the fraction of true positives correctly identified. The specificity is the proportion of true negatives correctly identified.

Our first evaluation considers the performance of the one-class SVM on a device-based classification. We train a predictor of each sensor ("LIT x01" where x represents the tank of the corresponding process) in the field network. This assures that the SVM decision surface represents the expected behavior for each sensor/actuator. Figure 3 shows the performance metrics for the attack class, for each sensor, and the average for the three sensors, represented by "System". The precision metric is always around 0.95, while the accuracy of the classifiers is upper-bounded at 0.80. This result is due to the low specificity of the classifiers. The primary reason for obtaining high precision while having low specificity is because the attacks introduce small physical changes for the time they last, but the dataset labels the entire period of the attack as belonging to the attack class. Therefore, as the period of normal behavior is labeled as an attack, the classifiers produce a false negative flow. Nevertheless, we highlight that the precision is high, almost 1.0 for the system as a whole, which means that the identified threats are, in fact, attacks for almost all alarms. The result reveals that one-class SVM is precise at finding anomalies and, when it fails to detect the anomaly, there is no physical change on the monitored process.

The following experiment evaluates the incremental learning strategy. Figures 4(a) and 4(b) show the accuracy and the precision of the system as a function of time using three-time windows. From 0 to 10 s, there was no attack traffic on the honeypot; as a consequence, incremental learning performs poorly. This is the bootstrap phase, in which it is learning the attacks, causing the accuracy and the precision of the classifier to be low (the classifier has not been trained with an attack sample yet). After the first batch of samples from the honeypot, at 10 s, the behavior of the predictor changes, and it reaches more than 0.95 of precision and accuracy. It is noteworthy that accuracy is higher for the incremental SVM than for the one-class SVM, because the incremental considers features from the network, transport protocol, and monitored

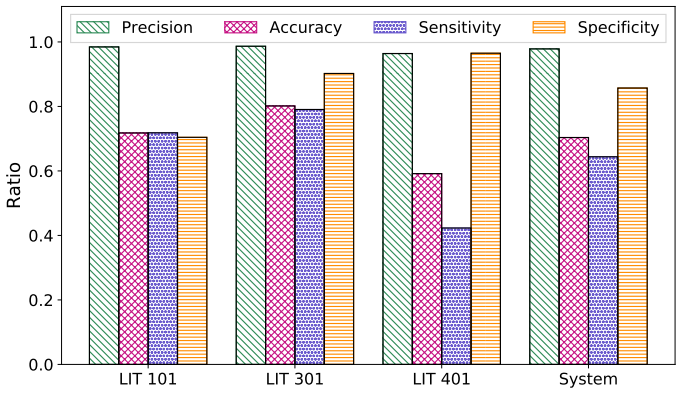


Fig. 3. Performance metrics for the one-class SVM. We show the results for the three sensors and the system as a whole. The low sensitivity associated with high accuracy and precision indicates that some attacks perform very close to normal behavior.

values on physical sensors/actuators, while the one-class SVM considers only the monitored values in the physical devices.

Moreover, the incremental strategy learns both normal and abnormal behaviors from samples, while one-class only derives the normal behavior. Figure 4 also compares the time window size used to gather the samples. We highlight that a 1 s time window performs better than 2 s or 3 s windows because it reduces the latency on reacting to attacks while keeping enough information to promptly train the classifier and improve performance of the next step.

Considering both strategies on detecting intrusions on a CPS, we argue that the orchestration between incremental and one-class SVM approaches is essential to assure a high level of safety on the operation of a CPS. As there is no warranty that a new attack targets a honeypot, it is crucial to apply the one-class approach, which has a high precision rate detecting deviations from the normal behavior of the physical process. Also, the incremental SVM fed by the honeypot is mandatory to learn zero-day attack patterns and leads to high accuracy of attack detection even when the attacks do not imply changes on physical parameters.

4 Conclusion and Future Work

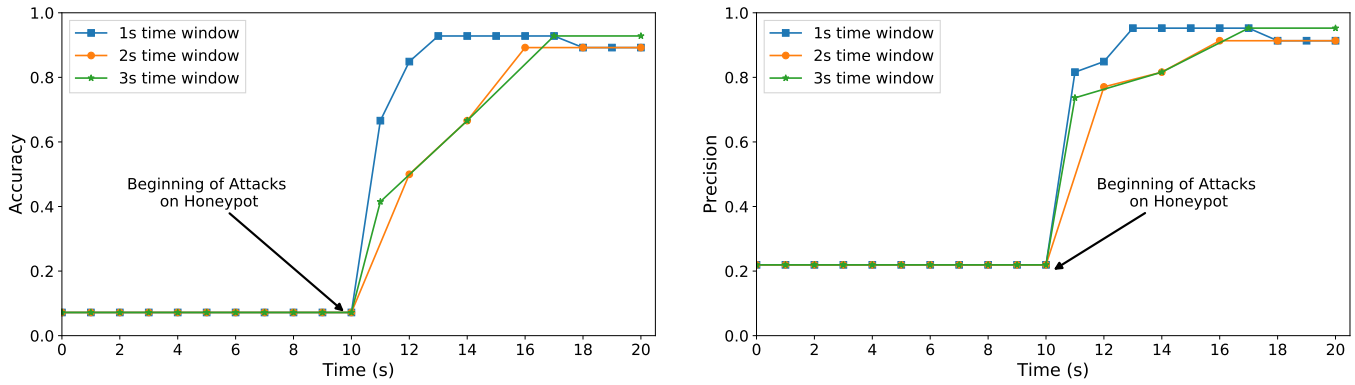
Applying machine learning algorithms to perform intrusion detection on real CPS is a challenging task. Firstly, data used to train the classifiers must be curated appropriately, verifying that it includes not only normal operating conditions but also some conditions such as downtime caused by maintenance and even emergency stops on the CPS. Furthermore, during the data acquisition process, it is crucial to have the expertise of the CPS operators to correctly label the events in the dataset. Second, CPS systems tend to operate for several years, and the aging of the devices must be taken into consideration by the classifiers to avoid false positives. Finally, the latency of the classification system must be considered to ensure that the attack countermeasure is responsive enough.

Cyber-physical systems (CPSs) integrate ubiquitous computing, efficient network communication, and control in physical processes. Since CPSs are part of critical infrastructures, it is mandatory to assure operational and safety goals of any given CPS. In this work, we investigated the deployment of a machine-learning approach for detecting malicious behaviors on the CPS network. Our approach orchestrates a detection of attacks learned from a honeypot through an incremental learning Support Vector Machine and anomaly detection on the measurements of the physical processes through a one-class Support Vector Machine. Our results show that, for the SWaT testbed dataset, this approach reaches more than 0.95 of accuracy on detecting learned attacks and detects physical anomalies at more than 0.90 precision.

Further work includes improving the false-positive rate of our classifier. Also, we intend to apply our detection system to other types of plants and industrial protocols. We also plan to address privacy-related questions about knowledge extracted from the CPS analytics and about leakage of the CPS topology information as an outcome for honeypot exploitation.

5 Acknowledgements

We would like to acknowledge CNPq, CAPES, FAPERJ, and RNP for the partial funding of this research. We would also like to warmly thank Álvaro Cárdenas from the University of California, Santa Cruz, USA, for his valuable contributions to this work.



(a) Accuracy for the incremental learning strategy.

(b) Precision for the incremental learning strategy.

Fig. 4. The performance metrics for the incremental SVM are closely related to the amount of attack traffic that is used as test samples, which is determined by the size of the time window. Before experiencing attacks on the honeypot, the system presents low accuracy (a) and low precision (b). As the attacks reach the honeypot, the system learns their behavior, shown at 20 s.

REFERENCES

[1] Derui Ding and Qing-Long Han and Yang Xiang and Xiaohua Ge and Xian-Ming Zhang, "A Survey on Security Control and Attack Detection for Industrial Cyber-Physical Systems," *Neurocomputing*, vol. 275, pp. 1674 – 1683, 2018.

[2] R. Boutaba, M. A. Salahuddin, N. Limam, S. Ayoubi, N. Shahriar, F. Estrada-Solano, and O. M. Caicedo, "A Comprehensive Survey on Machine Learning for Networking: Evolution, Applications and Research Opportunities," *Journal of Internet Services and Applications*, vol. 9, no. 1, p. 16, Jun 2018.

[3] A. G. P. Lobato, M. A. Lopez, I. J. Sanz, A. A. Cardenas, O. C. M. B. Duarte, and G. Pujolle, "An Adaptive Real-Time Architecture for Zero-Day Threat Detection," in *2018 IEEE International Conference on Communications (ICC)*, May 2018, pp. 1–6.

[4] J. Goh, S. Adepu, K. N. Junejo, and A. Mathur, "A Dataset to Support Research in the Design of Secure Water Treatment Systems," in *Critical Information Infrastructures Security*, G. Havarneanu, R. Setola, H. Nassopoulos, and S. Wolthusen, Eds., 2017, pp. 88–99.

[5] D. I. Urbina, J. A. Giraldo, N. O. Tippenhauer, and A. A. Cárdenas, "Attacking Fieldbus Communications in ICS: Applications to the SWaT Testbed." in *Proceedings of the Singapore Cyber-Security Conference (SG-CRC) 2016*, ser. Cryptology and Information Security Series, vol. 14, 2016, pp. 75–89.

Lúcio Henrik A. Reis is an undergraduate student at Universidade Federal Fluminense. Currently, Lúcio is researching the application of incremental learning on cyber physical systems security.

Andrés F. Murillo is currently a doctorate candidate at the Universidad de los Andes. He received a Master's degree in Electric Engineering from the Univerisdade Federal do Rio de Janeiro, Brazil, in 2014. He received the Electronic Engineer degree from Universidad Autónoma de Occidente of Santiago de Cali (UAO), Colombia, in 2009.

Sandra Rueda is associate professor at Universidad de los Andes. She received her Ph.D. degree in Computer Science and Engineering from the Pennsylvania State University, PA, USA. She received her M.Sc. degree in Computer and Systems Engineering from the Universidad de los Andes, Bogotá, Colombia.

Natalia C. Fernandes is currently a Professor at the Universidade Federal Fluminense (Niterói, Brazil). She received her degrees of D.Sc. and M.Sc. in Electrical Engineering from Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil, in 2011 and 2008. She received a Electronic and Computer Engineer degree from the same university, in 2006, awarded with *Cum Laude*.

Dianne Scherly Varela de Medeiros is a professor at the Universidade Federal Fluminense (UFF). Dianne received her Master's degree on Telecommunications Engineering from UFF in 2013, and her D.Sc. degree on Electric Engineering from the Universidade Federal do Rio de Janeiro in 2017.

Marcelo Dias de Amorim is a Research Director at the French National Center for Scientific Research (CNRS) and a member of the LIP6 laboratory of Université Pierre et Marie Curie, France, where he leads the Networks and Performance Analysis team. His scientific interests are in the area of networked systems, with an emphasis on mobile wireless systems.

Diogo Menezes Ferrazani Mattos is currently a Professor at the Universidade Federal Fluminense (Niterói, Brazil). He received his degrees of D.Sc. and M.Sc. in Electrical Engineering from Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil, in 2017 and 2012. He received a Computer and Information Engineer degree from the same university, in 2010, awarded with *Magna Cum Laude*.