Process Mining meets Statistical Model Checking to Explain Threat Models: Novel Approach to Model Validation and Enhancement (Extended Abstract)

Roberto Casaluce[1,2]

[1]Department of Computer Science, University of Pisa, Lungarno Pacinotti 43 56126 Pisa, Italy
[2]Sant'Anna School of Adv. Studies, Piazza Martiri della Libertà, 33 - 56127 Pisa, Italy

Abstract

Formal verification of the dynamics of a system can be conducted by employing statistical analysis techniques, such as Statistical Model Checking (SMC). SMC techniques resort to probabilistic simulations to evaluate the system properties to help to circumvent the state space explosion problem, the well-known curse of the classic model checking techniques. Nevertheless, SMC provides only estimations and confidence intervals of the evaluated properties of the system without explaining why the analysis estimated a particular property value. This project aims to present a novel methodology that integrates SMC with process-oriented data-driven techniques known as process mining (PM) applied to threat models. This methodology will empower modelers to see their models' unfolded behavior instead of just numerical aggregated values obtained by SMC analysis. In the present work, there are two research goals. The primary research goal focus on implementing and validating the novel methodology in which we enrich SMC techniques with PM techniques. The secondary research goals focus on implementing an approach to extract an attack pattern from its textual description and another to extract a textual description of the salient information from the process model discovered using PM techniques. The secondary research goals add the necessary means to assist a modeler in using this novel methodology.

Keywords

Process Mining, Statistical Model Checking, Validation, Natural Language Processing, Natural Language Generation

1. Introduction

Statistical Model Checking (SMC) techniques have been applied to various domains to analyze the dynamics of the systems, even when those systems have complex dynamics. Indeed, those complex dynamic systems can only be statistically analyzed by resorting to the simulations of their properties to avoid running into the space explosion problem common to the other classic numerical techniques [1]. When SMC is used to analyze the dynamics of a system without prior knowledge of the overarching behavior of the formal model is called black-box SMC [2]. As discussed in the next section, the state-of-the-art methodology for validating a model through SMC techniques can mainly rely on plots, numerical results, or counterexamples to study the properties of the model. However, when the results of the analyses are not what the modeler

ICPM 2022 Doctoral Consortium and Tool Demonstration Track
roberto.casaluce@santannapisa.it (R. Casaluce)
0000000257869167 (R. Casaluce)

expected, they need to make an informed guess on how to intervene in the model to fix an unexpected mode's behavior. The present project aims to help the modeler rely upon more than just an informed guess to improve the model by integrating SMC with process mining (PM) techniques.

Two main research goals are presented below, to which we give different levels of priority. Our project's primary research goal, with the highest priority, is dedicated to working on the central part of the novel methodology where we enrich SMC with PM techniques. The secondary research goals are directed at completing the primary research goal. One of the secondary research goals is to implement ways to automate the extraction of the attack model and the attacker behavior from a textual description. The other secondary research goal is directly to present through a textual description the most salient information discovered by mining simulations of the formal model using PM techniques. When the salient information extracted from the simulations communicates the presence in the model of some unwanted behaviors, we could use this information to fix the formal model automatically.

2. Research goals

2.1. Primary Research Goal

The first research goal is to propose a novel methodology in which SMC is enriched with PM techniques that assists a modeler in validating and identifying flows in the model or enhancing opportunities. Fig. 1 depicts the life cycle of the state-of-the-art methodology based on SMC that starts with the model creation followed by SMC analysis that returns numeric results, plots, and counterexamples [3]. When the numerical results are not what the modeler expected, e.g., if the probability of success of an attack is too low, the modeler needs to make an informed guess on how to change the model to improve its performance. In our previous work, on the first experimental results of our methodology [4], we called this way to validate a model *SMC-guided black-box model validation* since the modeler evaluates the model and makes changes based only on the numerical results without any other information.

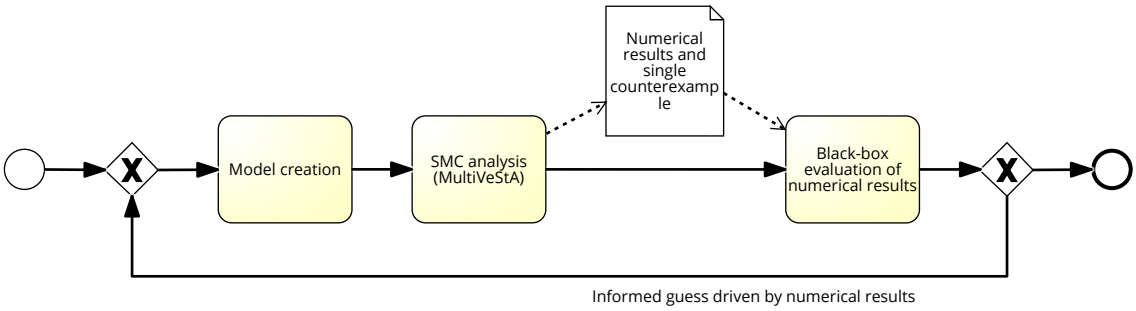

Figure 1: SMC-guided black-box model validation.

The novel methodology proposed here, depicted in Fig. 2, is defined as an *SMC- and PM-guided white-box behavioral model validation* in which the state-of-the-art method, depicted in Fig. 1, is

augmented with PM techniques. The additions of the novel methodology are colored in green in Fig. 2. Now, besides the numerical results, the modeler also has, thanks to PM techniques, a behavioral evaluation of the model to support them in identifying flaws and improving the model. Indeed, this is defined as a white-box behavioral model validation precisely because the graphical results obtained by the PM analysis represent a closer representation of the original model displaying an overview of its overarching behavior.

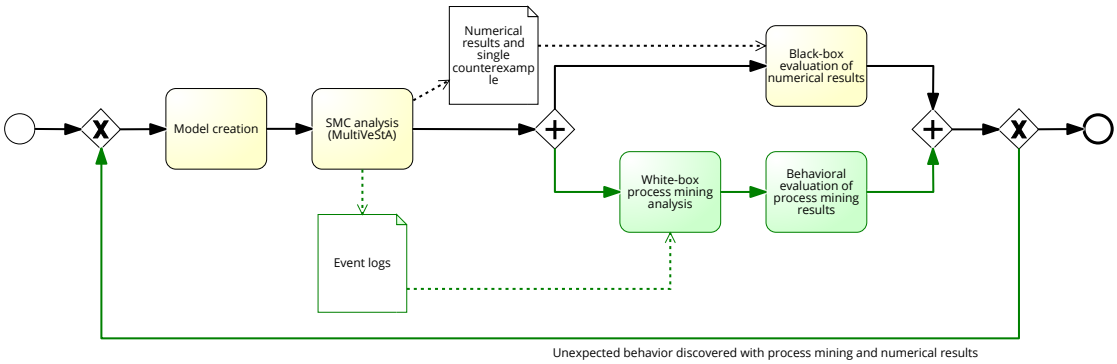

Figure 2: SMC- and PM-guided white-box behavioral model validation.

Methodology. We use RisQFLan, a quantitative graphic-based framework for risk modeling and analysis [5] belonging to the so-called QFLan family [6]. RisQFLan uses graph-based security models called Attack Defense Trees, a variant of the classic Attack trees that can represent an intuitive visual language to describe an attack scenario. In addition, RisQFLan supports SMC analysis using MultiVeStA [7, 8], a black-box statistical model checker that can be integrated with different simulators to add more reliable statistical analysis techniques. We have enriched MultiVeStA with PM-oriented logging capabilities by extending its interface with simulators with new features that allow saving the complete traces of each simulation as event logs. Although we implement the novel methodology proposed using MultiVeStA and RisQFLan, this can be applied to any simulation models and SMC tools since it does not rely on the internal mechanics of either the analyzer or the model but exploits logs of the computed simulations. We implement PM techniques using Fluxicon Disco and PM4PY, a Python library, to experiment with different PM algorithms.

2.2. Secondary Research Goals

Secondary research goals mainly aim to help non-experts apply this novel methodology to validate threat models. On the one hand, we could automate the extraction of the attack models and the attacker behaviors directly using the textual description from different data sources. Currently, the model creation (Fig. 2) is performed by manually encoding the model in RisQFLan; therefore, an automatic creation from a textual description of the attack model could improve the usability of our methodology. Moreover, a model extracted from a description could help the modeler automatically enhance the threat model when new attack strategies are available to the attacker. On the other hand, another way to increase the usability of the

novel methodology is achieved by describing the most salient information of the process model obtained from the event logs. Indeed, a process model extracted from the simulated event logs representing the model's behavior might need to be explained to the non-expert of PM when they visually inspect the process model. Here, we could present the user with a textual description of the unwanted behaviors of the model discovered to help the modeler to identify what needs to be correct in the formal model.

Methodology. To implement the automatic creation of the attack model and attacker behavior, we experiment with NLP techniques using pre-trained models, such as the GPT3 model that [9] used with promising results. In their work [9], they demonstrated how it is possible to extract meaningful information from the text describing a process using Large Language Models (LLM). In our case, we would extract from the textual description of a threat model the activities/strategies an attacker can use to complete their attack. We will investigate techniques to enhance the performance of the GPT-3 model by improving the prompt design [10, 11]. Then, we could fill in templates with the information extracted and feed them into RisQFLan to create the actual attack models. Furthermore, within the secondary research goals, we could implement a method to automatically extract a textual description of the most salient information found in the process model by employing fuzzy quantification techniques and natural language generation (NLG) tools [12]. This approach would help identify unwanted behaviors in the formal model during the phase of evaluating the model's behavior discovered by mining the event logs.

3. Planned Research

In [4], we presented a prototypical instantiation of our approach. We demonstrated how even a trivial threat model could display unexpected behaviors. Indeed, we could find and fix unwanted behaviors in the formal model thanks to PM techniques. Although these experimental results showed the potentiality of the methodology, this needs to be validated with experiments on complete threat models. Therefore, the next step in our project would be to work on real threat models to validate our methodology. At the same time, we will experiment with different discovery PM algorithms or create an ad hoc algorithm to extract the attack models from the event logs. If an ad hoc algorithm is needed, we will also work on a model-to-model metric to measure how much the normative models overlay with the discovered ones. We will experiment with the secondary research goals once we have completed the previous steps for the primary research goal. The last step would be to put together the primary and secondary research goals in a final work.

4. Conclusion

Simulation-based validation approaches run statistical analysis to evaluate the properties of a simulated model returning numerical estimations of those properties. However, without providing behavioral explanations on why the analyses returned those results, the modeler can make only an informed guess based on the numeral results obtained to adjust the model

and fix unwanted behaviors. The present project proposes a novel methodology that can help reason upon the whole behavior of the model and understand why the analysis of the model's properties has returned some estimations of those properties. We achieve that by integrating the simulation-based analysis technique from formal quantitative methods known as SMC with the data- and process-driven techniques known as PM. Thanks to the widespread use of simulation models, this new methodology would be valuable among several other disciplines to help identify issues in the model (validation) or suggest relevant improvements (enhancement) to the modeler.

References

[1] G. Agha, K. Palmskog, A survey of statistical model checking, ACM Trans. Model. Comp. Simul. 28 (2018) 6:1–6:39.

[2] K. Sen, M. Viswanathan, G. Agha, Statistical model checking of black-box probabilistic systems, in: International Conference on Computer Aided Verification, Springer, 2004, pp. 202–215.

[3] P. E. Bulychev, A. David, K. G. Larsen, M. Mikucionis, D. B. Poulsen, A. Legay, Z. Wang, UPPAAL-SMC: statistical model checking for priced timed automata, in: Proc. QAPL'12, volume 85, 2012, pp. 1–16.

[4] R. Casaluce, A. Burattin, F. Chiaromonte, A. Vandin, Process mining meets statistical model checking: Towards a novel approach to model validation and enhancement, in: C. Cabanillas, N. F. Garmann-Johnsen, A. Koschmider (Eds.), Business Process Management Workshops, Springer International Publishing, Cham, 2022.

[5] M. H. ter Beek, A. Legay, A. L. Lafuente, A. Vandin, Quantitative security risk modeling and analysis with RisQFLan, Computers & Security 109 (2021) 102381.

[6] A. Vandin, M. H. ter Beek, A. Legay, A. Lluch-Lafuente, QFLan: A Tool for the Quantitative Analysis of Highly Reconfigurable Systems, in: FM, 2018.

[7] S. Sebastio, A. Vandin, MultiVeStA: statistical model checking for discrete event simulators, in: 7th Int. Conf ValueTools'13, ICST/ACM, 2013, pp. 310–315.

[8] A. Vandin, D. Giachini, F. Lamperti, F. Chiaromonte, Automated and distributed statistical analysis of economic agent-based models, Journal of Economic Dynamics and Control (2022) 104458.

[9] P. Bellan, M. Dragoni, C. Ghidini, Leveraging pre-trained language models for conversational information seeking from text, arXiv (2022).

[10] S. Arora, A. Narayan, M. F. Chen, L. J. Orr, N. Guha, K. Bhatia, I. Chami, F. Sala, C. Ré, Ask me anything: A simple strategy for prompting language models, arXiv preprint arXiv:2210.02441 (2022).

[11] A. Saparov, H. He, Language models are greedy reasoners: A systematic formal analysis of chain-of-thought, arXiv preprint arXiv:2210.01240 (2022).

[12] Y. Fontenla-Seco, M. Lama, V. González-Salvado, C. Peña-Gil, A. Bugarín-Diz, A framework for the automatic description of healthcare processes in natural language: Application in an aortic stenosis integrated care process, Journal of Biomedical Informatics 128 (2022).