Detecting Privacy, Data and Control-Flow Deviations in Business Processes

Detecting Privacy, Data and Control-flow Deviations in Business Processes

Azadeh S. Mozafari Mehr[0000−0001−8156−6492], Renata M. de Carvalho[0000−0001−6129−9278], and Boudewijn van Dongen[0000−0002−3978−6464]

Department of Mathematics and Computer Science
Eindhoven University of Technology, Eindhoven, The Netherlands
{a.s.mozafari.mehr, r.carvalho, b.f.v.dongen}@tue.nl

Abstract. Existing access control mechanisms are not sufficient for data protection. They are only preventive and cannot guarantee that data is accessed for the intended purpose. This paper proposes a novel approach for multi-perspective conformance checking which considers the control-flow, data and privacy perspectives of a business process simultaneously to find the context in which data is processed. In addition to detecting deviations in each perspective, the approach is able to detect hidden deviations where non-conformity relates to either a combination of two or all three aspects of a business process. The approach has been implemented in the open source ProM framework and was evaluated through controlled experiments using synthetic logs of a simulated real-life process.

Keywords: Process Mining · Multi-layer Alignment · Data privacy · Conformance Checking · Multi-perspective Analysis

1 Introduction

In recent years, data privacy issues are of increasing concern to organisations and governments. Organisations often define sets of rules as privacy policies for protecting sensitive data of their processes. However, regulations like GDPR (https://gdpr-info.eu) impose more strict privacy requirements. New privacy rules which denotes"who can access data for which purpose" relate to multiple perspectives of a business process, as they are closely related to the tasks being executed (control-flow perspective), the flow and processing of information (data perspective) and legitimate role allocation (resource or privacy perspective). Employees should follow these policies while performing activities within business processes. However, it is well documented in the literature that real process behavior often deviates from the expected process which often opens the way to the fraudulent behaviour or performance issues [5, 13]. Unfortunately, standard preventative access control, which regulates who may carry out which data operations in a system is not sufficient for data protection as access is independent of context since it is not checked for which purpose data are processed after access to data has been granted [10]. In this paper, we address this issue by proposing a novel approach for multi-perspective conformance checking. By considering all control-flow, data, and privacy perspectives of a business process simultaneously, our approach brings two main contributions: a) we detect spurious data access and identify privacy infringements where data have been

Table 1: Data model of treatment process. R:Read,C:Create

Activity	Data Operations
Identify patient (ip)	R(ID)
Admission (ad)	R(ID,PatientID,Name) C(AdmissionID)
Visit (vi)	R(AdmissionID,PatientID,MedicalHistoryID) C(VisitID,PrescriptionID)
Lab appointment (la)	R(AdmissionID,PatientID C(LabAppointment)
Basic lab test (bt)	R(AdmissionID,PatientID,PrescriptionID) C(BLabPID)
Advanced tests (at)	R(AdmissionID,PatientID,PrescriptionID) C(ALabPID)
Evaluate (ev)	R(AdmissionID,PrescriptionID, BLabPID,ALabPID) C(TestResultID)
Consult request (co)	R(AdmissionID,PatientID) C(ConsAppointment)
Inter-colleague consultation (in)	R(AdmissionID,PatientID,VisitID, PrescriptionID,TestResultID,MedicalHistoryID) C(VisitID,CPrescriptionID)
Treatment prescription (tr)	R(AdmissionID,VisitID,MedicalHistoryID) C(TreatmentPlan)
Discharge (di)	R(AdmissionID,PatientID) C(Confirmation)
Billing (bi)	R(AdmissionID,PatientID,PaymentID) C(PaymentReceipt)

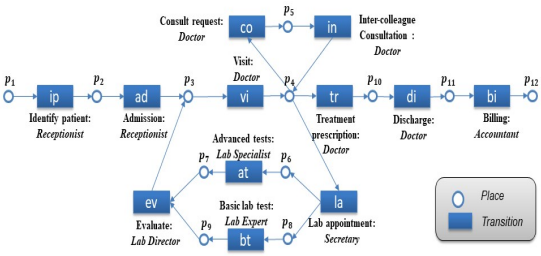

Fig. 1: An example of healthcare treatment process in Petri net notation (adapted from [3])

processed for unclear or secondary purposes by an authorised role; and b) we detect important deviations in each perspective such as unexpected activities, missing data operations or illegitimate role allocations. As a proof of concept, we implemented and tested our approach over synthetic logs generated from simulation of a real-life process.

This paper is organized as follows. Section 2 introduces a running example along with some scenarios as the motivation of this work. Section 3 illustrates our approach. Section 4 presents experimental results. Section 5 discusses related work. Section 6 concludes the paper and provides directions for future works.

2 Motivating Example

As a running example, consider a healthcare treatment process derived from Alizadeh *et.al.* [3]. Fig. 1 shows the process as a Petri net. The process starts with the patient identification and admission by the receptionist. Next, the patient is visited by a doctor. The doctor might request a basic lab test and advanced tests such as MRI scans, for which the secretary makes an appointment. After a lab expert and a lab specialist perform the tests, a lab director evaluates the results. Based on the evaluation, the doctor may request inter-colleague consultation ((co) followed by (in)), request more lab tests, or prescribe a treatment plan. Finally, the patient is discharged and a bill is sent to the patient's insurance company by an accountant. In this process, certain data operations on specific data fields are required to be performed during each activity. Table 1, presents these data operations.

An example of an execution of this process is depicted in Fig. 2. This figure shows observed behavior from three perspectives which can be extracted from the recorded behavior in the process and data logs. For each activity, a start event and a complete event are expected. Whenever they both occur and are performed by the same resource, they are linked as a yellow rectangle as shown in Figure 2(a). The sequence of yellow triangles in Figure 2(b) shows a data trace consisting of twenty data events. The events

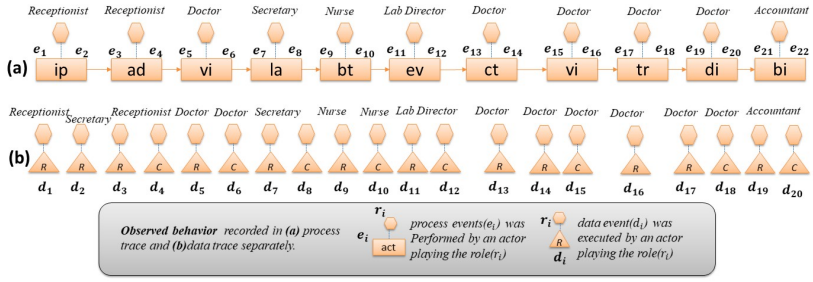

Fig. 2: An executed process instance of the healthcare process depicted in Figure 1.

in the process trace and data trace record information regarding the process instance or case, the corresponding activity and data operation, the time of the execution, and the actor who executed the activity or data operation, separately. Each hexagon presents the role of the actor under whose name the event is registered in the system.

Below, we present some scenarios to motivate the need for investigating the data, privacy and control-flow conformance to detect the hidden deviations:

Scenario 1: According to the presented process and data models, several roles like doctors and lab experts are allowed to access sensitive data of patients. A curious actor may exploit this privilege to access patient information for personal or financial gain. For instance, a doctor who has access to patient information for providing medical treatment, can use this information to conduct a clinical trial (ct) which does not contribute to the fulfilment of the treatment process.

Scenario 2: A nurse, instead of a lab expert, takes a blood sample from the patient. Based on both data and control-flow perspectives, the occurrence of this activity and related data operations are allowed but from the privacy perspective they are not. This is a case where all three layers of process, data and privacy should be considered together to detect the hidden deviation.

Scenario 3 [3]: During each visit, doctors are expected to add a prescription or treatment plan to the patient's medical history. A doctor may negligently forget to update it. This missing data operation may cause other doctors to prescribe an incompatible drug to the patient. In this case, from a control-flow perspective there is no violation while from the data perspective there is a missing data operation.

3 Proposed Multi-Layer Alignment Approach

In this section, we propose our approach for multi-perspective conformance checking. The main goal of this approach is to align process, data and privacy policy layers to find hidden deviations between these three perspectives of a business process in addition to detecting the deviations in each layer.

Fig. 3 shows an overview of our approach together with its inputs and outputs. A Process log (❶) records process executions and a data log (❷) contains data operations showing which user accessed which data. These two inputs indicate observed behaviors.

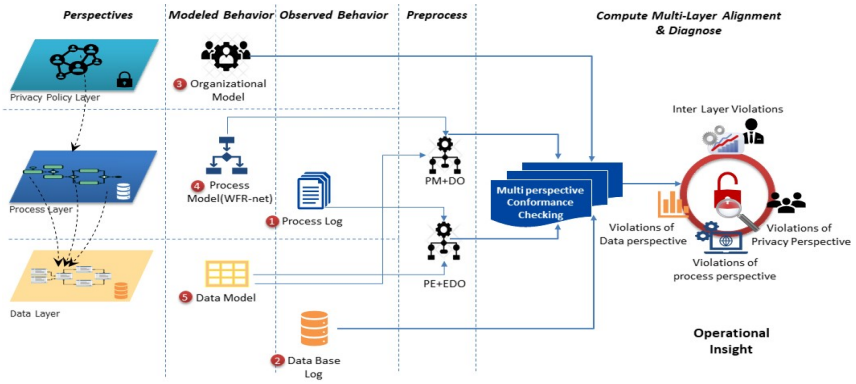

Fig. 3: An overview of the proposed approach

To represent the modeled behaviors the approach considers a process model (❹), a data model (❺) and an organisational model (❸). A process model describes the activities to be performed in a specific order to reach a certain business goal. The data model relates the process logic to the data layer by indicating which data operations must be executed in order to complete a given activity. The organisational model links users to their roles. The role of actors in the process log and data log can be retrieved from this model. As discussed before, using only an organisational model for access control is not sufficient to check data privacy. Therefore, in order to find the context of data access, first we integrate the activities with their corresponding roles in the process model to unify the two perspectives of process and privacy into a single model. Second, using the data model, we enrich the aforementioned process model with expected data operations in a pre-processing step, shown as "PM+DO" in Fig. 3. In another pre-processing step, we enrich the events of the process log with the expected data operations using the data model ("PE+EDO" in Fig. 3).

The combination of the process model with role information, the process event log showing the start and complete of activities performed by specific resources and the data log showing who accessed what at which time is translated into a large *synchronous product model*. Such synchronous product is the foundational model for conformance checking and techniques exist [1] to find the optimal execution given a cost function that penalizes specific deviations.

In this synchronous product, totally synchronous moves represent expected behavior. We further distinguish six kinds of deviations:

- A *move on data log* happens when a not-allowed data operation was executed.
- A *move on process log* happens when an unexpected activity was performed.
- A *move on model* happens when there is a missing activity in the process log.
- A *partially synchronous move with correct role* happens when there is a missing data operation in the data log. In this case, the expected activity was performed by a legitimate role.
- A *partially synchronous move with wrong role*, as the previous, but performed by a not-allowed role.

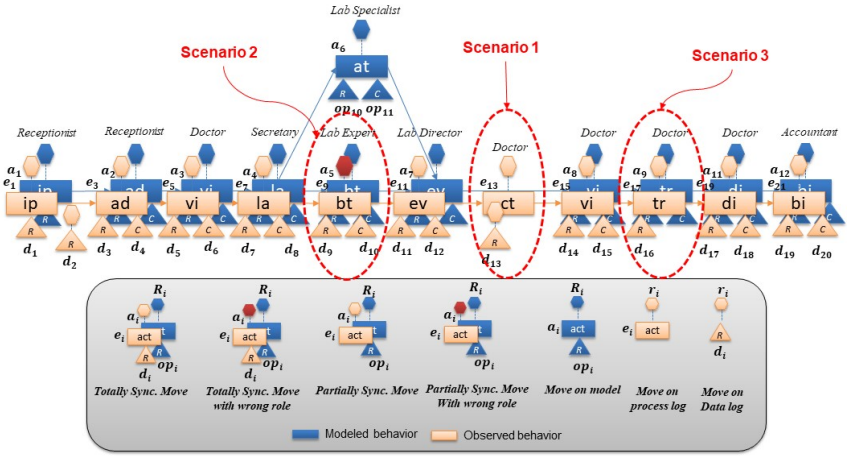

Fig. 4: Multi-perspective alignment between modeled and observed behavior

– A *totally synchronous move with wrong role* happens when an expected activity and data operation were done by a not-allowed role.

A cost function in our approach assigns a cost equal to 4 for move on data log, move on process log and move on model. It assigns a cost equal to 2 to partially synchronous moves and cost 0 to totally synchronous moves. Finally, it adds the penalty cost 1 if the actor plays a not-allowed role. This cost function is a parameter of the approach and can be changed per use-case, but it is an essential parameter to compute the optimal alignment.

The optimal alignment we get from the synchronous product model is translated back into a multi-perspective alignment as shown in Figure 4. Returning to the running example in section 2, by using our technique for multi-perspective alignment we can identify the three scenarios of section 2 as shown in Fig 4.

4 Evaluation

We implemented the approach illustrated in Fig. 3 as a package named MultiLayer-Alignment in the ProM framework (https://www.promtools.org/). The output is a csv file including the alignment results that can be used by other applications for visualization or further analysis. In order to conduct controlled experiments, we simulated the process model depicted in Fig. 1 using CPN tools (http://cpntools.org) to generate process and data logs with real-life complexity(e.g. loops or considerable trace length). Table 2 summarizes the differences of conducted experiments in terms of the type of deviations and the perspectives in which the deviations happened. The numbers in parentheses show the percentage of inserted noise and the filled cells in each row represent the type of deviations that were included in the experiment. For instance, E0 is the fully fitting base line and experiments E1 to E3 are the simulation of the three scenarios described in section 2. We inserted all kinds of deviations at the level of traces in the E7 and at the level of the entire log in E8.

Table 2: The Result of Experiments

deviation happened in	All Three layers	Process layer	data layer	data layer	data layer & privacy layer	privacy layer	-
Legal Move	Move on Model	Move on Process Log	Move on Data Log	Partially Sync. Move	Partially Sync. Move with Penalty cost	Totally Sync. Move with Penalty cost	Totally Sync. Move
	P-R-F1	**P- R- F1**	**P- R- F1**	**P- R- F1**	**P- R- F1**	**P- R- F1**	**P- R- F1**
E0 (0 %)							1.00-1.00-1.00
E1 (5 %)		1.00-1.00-1:00	1:00-1:00-1:00				
E2 (5 %)						1.00-1.00-1.00	
E3 (5 %)				1.00-1:00-1:00			
E4 (5 %)	1.00-1.00-1.00						
E5 (5 %)			1.00-1.00-1.00				
E6 (5 %)					1.00-1.00-1.00		
E7 (5 %)	1.00-1.00-1.00	1.00-1.00-1.00	1.00-1.00-1.00	1.00-1.00-1.00	1.00-1.00-1.00	1.00-1.00-1.00	
E8 (26 %)	1.00-1.00-1.00	1.00-1.00-1.00	1.00-1.00-1.00	1.00-1.00-1.00	1.00-1.00-1.00	1.00-1.00-1.00	

To assess the approach's capability of detecting different kinds of deviations and the accuracy of obtained results, we computed the precision, recall, and F_1-measure [9]. Precision is computed as the fraction of detected deviations that are actual deviations, whereas recall is the fraction of the inserted deviations that are detected. The F_1-measure is the harmonic mean of precision and recall. In each experiment, the ground truth was known since deviations were introduced artificially.

As shown in Table 2, overall, our results show high precision and recall. Considering all experiments, we conclude that the approach is able to detect all deviations that happened in one, two, or all three combinations of process perspectives (control-flow, data and privacy policy).

5 Related Work

Process mining is a set of techniques that aim at analyzing business process execution data recorded in event logs. We limit related work to the research approaches most related to our contribution to the field of conformance checking.

Besides the control-flow, there are also other perspectives like data or resources that are often crucial for conformance analysis. Few approaches have investigated how to include these perspectives in the conformance analysis: De Leoni *et.al.* [4] extend the alignment approach to bring other perspectives' impact in the identification of non-conformity. This approach considers data, resource, and time as data attributes of process events. Thus, control-flow is aligned first, and then data are considered. Mannhardt *et.al.* [6] extend the work in [4] to propose a more balanced approach using data-aware Petri net as the prescribed model and check executed behaviors in the process log with respect to the values of the variables in the guards in addition to control-flow conformance. Both approaches are unable to consider the three perspectives separately since these methods give priority to the control-flow. Accordingly, some important violations such as missing data operations or not allowed data access can be missed in the alignment results.

Alizadeh *et.al.* [3] proposed an approach for linking data and process perspectives for conformance analysis. Similarly to [4] and [6] they extend the alignment approach

to handle the data perspective in which control-flow is aligned first and then data are considered. In contrast to the proposed approaches in [4] and [6], Alizadeh *et.al.* [3] aligned data and process perspectives independently. They applied a CRUD matrix that relates process activities to data operations and defined two criteria functions to link data operations in the data traces and events in the process traces.

We have extended the work in [3] and added privacy perspective in addition to process and data perspectives. To this end, we integrate the activities with corresponding roles in the process model in addition to using organisational model. Therefore, our approach can provide more comprehensive diagnostics than [3]. Similar to [3], we use a data model that relates process activities to data operations. However, we employed the data model in a completely different way to bring data perspective into conformance analysis. In [3], the approach applies a data model along with two criteria functions to link data operations in data traces with events in process traces. They performed this step in post-processing (after alignment computation) locally for each event in the alignment trace to find the deviations related to the data layer. This is the reason why their approach is not able to identify all the deviations correctly. For instance, in the presence of concurrent process events, a data operation can be linked to different process events with the same activity name. We solved this problem globally by allowing the alignment algorithm to find the best match. In contrast to [3], we use the data model in the pre-processing step to enrich the process model with related data operations in order to model prescribed behavior from all three perspectives. By constructing it, our approach is able to link data and process layer in a more robust way.

A large body of literature is related to privacy-preserving process/data mining i.e. [2, 7, 8, 11, 12] . They are not compared here since they consider privacy issue at design time to minimise privacy risks while maximising data utility for analysis. However, they do not consider the run-time perspective of business process management.

To the best of our knowledge, the work in this paper is the first work that proposes a novel technique for computing alignment by considering all control-flow, data, and privacy perspectives of a business process at the same time without giving priority to one perspective.

6 Conclusion

In this work, we presented a new method for multi-perspective conformance checking. We discussed that by considering more perspectives, our approach is able to find the context of data accesses in addition to detect hidden deviations between control-flow, data, and privacy perspectives of business processes.

As proof of concept, we implemented the approach in the ProM framework. An evaluation of the proposed approach has been carried out using synthetic logs generated from the simulation of a real-life process. The evaluation shows the applicability of our implementation to real-life complexity. The experiments confirm that our approach is able to provide more accurate diagnostics of deviations than control-flow based conformance checking approaches. The results also implied that the proposed approach allows the user to identify violations that cannot be detected by taking into consideration only one or two aspects.

In future work, we plan to improve the visual representation of the results to guide users towards an in depth identification of problems in the business processes execution. Extending the application of the approach and making it suitable for online process mining would be another direction of future work.

Reproducibility. The source code and inputs required to reproduce the experiments can be found at https://github.com/AzadehMozafariMehr/Multi-Layer-Alignment

Acknowledgement. The author has received funding within the BPR4GDPR project from the European Union's Horizon 2020 research and innovation programme under grant agreement No 787149.

References

1. Adriansyah, A., van Dongen, B.F., van der Aalst, W.M.P.: Towards robust conformance checking. In: zur Muehlen, M., Su, J. (eds.) Business Process Management Workshops. pp. 122–133. Springer Berlin Heidelberg, Berlin, Heidelberg (2011)
2. Aggarwal, C.C.: Data Mining: The Textbook. Springer, Cham (2015)
3. Alizadeh, M., Lu, X., Fahland, D., Zannone, N., van der Aalst, W.: Linking data and process perspectives for conformance analysis. Computers and Security **73**, 172–193 (Mar 2018)
4. de Leoni, M., van der Aalst, W.M.P.: Aligning event logs and process models for multi-perspective conformance checking: An approach based on integer linear programming. In: Daniel, F., Wang, J., Weber, B. (eds.) Business Process Management. pp. 113–129. Springer Berlin Heidelberg, Berlin, Heidelberg (2013)
5. de Leoni, M., van der Aalst, W.M.P., van Dongen, B.F.: Data- and resource-aware conformance checking of business processes. In: Abramowicz, W., Kriksciuniene, D., Sakalauskas, V. (eds.) Business Information Systems. pp. 48–59. Springer Berlin Heidelberg, Berlin, Heidelberg (2012)
6. Mannhardt, F., Leoni, de, M., Reijers, H., Aalst, van der, W.: Balanced multi-perspective checking of process conformance. BPMcenter. org (2014)
7. Mannhardt, F., Koschmider, A., Baracaldo, N., Weidlich, M., Michael, J.: Privacy-preserving process mining differential privacy for event logs. Business Information Systems Engineering **61**, 1–20 (2019)
8. Michael, J., Koschmider, A., Mannhardt, F., Baracaldo, N., Rumpe, B.: User-centered and privacy-driven process mining system design for iot. In: Cappiello, C., Ruiz, M. (eds.) Information Systems Engineering in Responsible Information Systems. pp. 194–206. Springer International Publishing, Cham (2019)
9. Perry, J.W., Kent, A., Berry, M.M.: Machine literature searching x. machine language; factors underlying its design and development. American Documentation **6**(4), 242–254 (1955)
10. Petković, M., Prandi, D., Zannone, N.: Purpose control: Did you process the data for the intended purpose? In: Jonker, W., Petković, M. (eds.) Secure Data Management. pp. 145–168. Springer Berlin Heidelberg, Berlin, Heidelberg (2011)
11. Pika, A., Wynn, M.T., udiono, S., Ter Hofstede, A.H.M., van der Aalst, W.M.P., Reijers, H.A.: Privacy-preserving process mining in healthcare. International journal of environmental research and public health **17** (2020)
12. Rafiei, M., van der Aalst, W.M.P.: Privacy-preserving data publishing in process mining. CoRR **abs/2101.02627** (2021)
13. Zhang, S., Genga, L., Yan, H., Lu, X., Kaymak, U.: Towards multi-perspective conformance checking with fuzzy sets. In: Workshop on Data Fusion for Artificial Intelligence (DAFUSAI 2020) (2020)