ELPaaS: Event Log Privacy as a Service

Martin Bauer[1], Stephan A. Fahrenkrog-Petersen[1], Agnes Koschmider[2], Felix Mannhardt[3], Han van der Aa[1], and Matthias Weidlich[1]

[1] Humboldt-Universität zu Berlin, Berlin, Germany
{martin.bauer,fahrenks,vanderah,weidlima}@hu-berlin.de
[2] Kiel University, Kiel, Germany
ak@informatik.uni-kiel.de
[3] SINTEF Digital, Trondheim, Norway
felix.mannhardt@sintef.no

Abstract. The privacy of an organization's workers represents a crucial concern in process mining settings, where data on an individual's performance is recorded and possibly shared for analysis. To enable users to appropriately deal with privacy concerns in process mining, this paper introduces ELPaaS (Event Log Privacy as a Service), a web application that offers state-of-the-art techniques for event log sanitization and privacy-preserving process mining queries. By employing our techniques, users obtain event logs and process mining results that provide privacy guarantees such as differential privacy and k-anonymity. Hence, the privacy of an organization's workers is protected.

1 Introduction

Process mining represents a family of techniques for the data-driven analysis of business processes [1]. These techniques utilize event data recorded by information systems during the execution of a business process, stored in the form of *event logs*. Event logs are employed for a variety of use cases, such as *process discovery*, in which a process model is constructed on the basis of the recorded event data, *conformance checking*, in which event data is compared to a process model, and *model enhancement*, in which, for example, performance information is added to an obtained process model.

Recognizing the potential of process mining, organizations strive to record event data in an accurate and fine-granular manner. While this enables organizations to ensure the efficient and correct execution of their processes, it can also result in the disclosure of sensitive information regarding an organization's employees. Event logs may breach an individual's privacy [5], violating their ability to control who has access to their personal data [2]. Therefore, disclosure of recorded event data in the form of event logs should be assessed in light of ethical considerations, as well as in the context of privacy regulations, such as the European General Data Protection Regulation (GDPR) [7] and the California Consumer Privacy Act. For instance, the GDPR prohibits the processing of personal data unless explicit consent has been given.

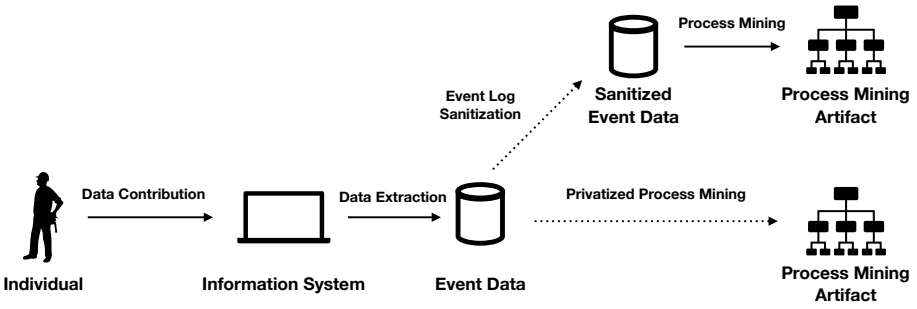


Fig. 1: General approaches for privacy-preserving process mining [3].

To enable the appropriate handling of recorded event data in process mining, we introduce the ELPaaS (Event Log Privacy as a Service) web application. As visualized in Figure 1, the application supports two fundamental ways in which privacy guarantees in process mining can be provided: (1) event log sanitization and (2) privatized process mining. *Event log sanitization* involves the transformation of an extracted event log into one that satisfies established privacy metrics. An event log obtained in this manner can subsequently reduce the disclosure of personally identifiable records in the log. *Privatized process mining*, by contrast, involves process mining techniques that have been specifically designed such that the obtained process mining artifacts, e.g., derived process models or query results, meet desired privacy guarantees. ELPaaS offers state-of-the-art techniques for both directions. For event log sanitization, the application offers the PRETSA [4] (PREfix-Tree based event log SAnitisation for t-closeness) algorithm, which sanitizes event logs to guarantee k-anonymity and t-closeness. For privatized process mining techniques, differential privacy mechanisms for common queries on event logs, developed by Mannhardt et al. [6], are offered.

The remainder of this paper is structured as follows. Section 2 describes and visualizes the functionality of ELPaaS, Section 3 discusses the maturity and availability of the application, before concluding in Section 4.

2 Functionality

This section describes the main input, functions, and output of ELPaaS.

Input. Figure 2 presents a snippet of the opening page of the web application. The first step for a user is to upload an event log. Event logs should be provided as XES (eXtensible Event Stream) or CSV (Comma Separated Value) files. CSV files should at least contain a column representing a *case ID* and an *activity*, and should be sorted according to the execution order of the events.

Event Log Sanitization. Users that want to sanitize an event log can directly do so by selecting the PRETSA algorithm [4] in the upload screen, as shown in

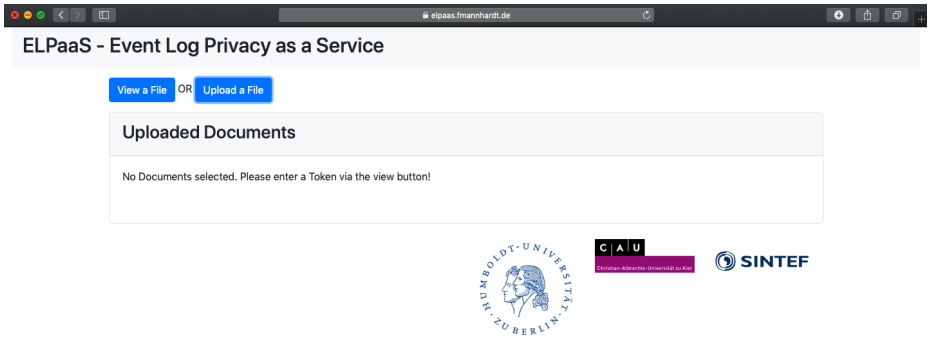


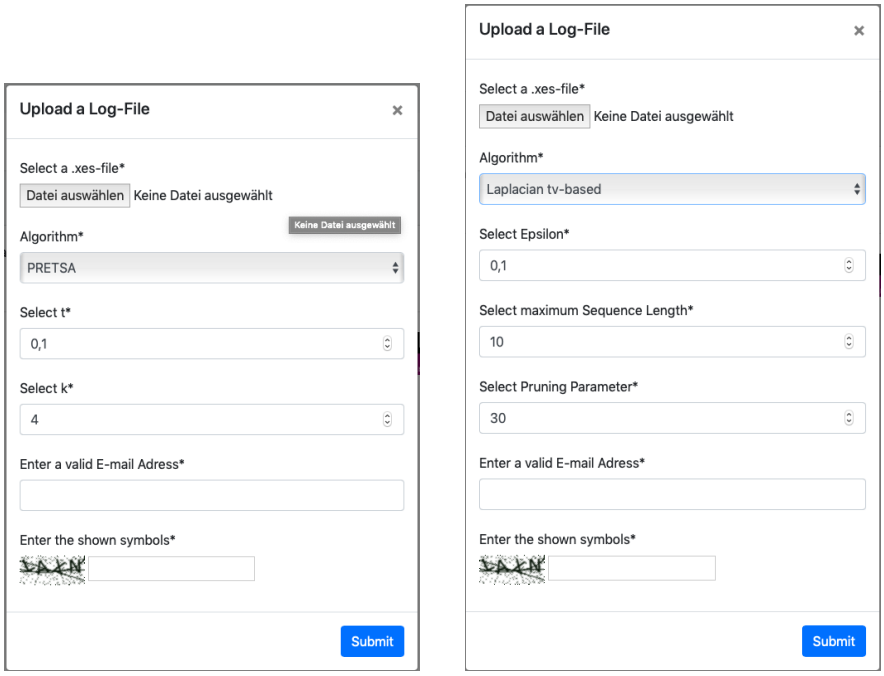
Fig. 2: Opening screen of ELPaaS.

Figure 3a. It then transforms an event log into one that satisfies k-anonymity and t-closeness requirements, while striving to preserve maximum utility for process discovery and process enhancement. When an event log satisfies k-anonymity, any event execution can be related to at least k different actors, mitigating an attacker's ability to confidently associate events to specific workers. An event log that satisfies t-closeness furthermore ensures that the performance of individual workers (e.g., in terms of throughput time) cannot be derived from a sanitized event log. We kindly refer the reader to Fahrenkrog-Petersen et al. [4] for insights on the impact of different k and t values on process mining utility.

Privatized Process Mining Queries. Users are currently able to execute two distinct privacy-preserving process mining queries: the derivation of the frequency of a directly-follows relation (i.e., how often one activity is executed after another one) and the derivation of a trace variants (i.e., how often a particular sequence of activity executions is contained in the log). Both queries follow the privacy-preserving techniques developed by Mannhardt et al. [6]. Here, the general idea is to introduce so-called *Laplace noise* according to a user-defined ϵ value. By doing so, the obtained results will be guaranteed to satisfy ϵ-differential privacy. This means that the query result will not allow attackers to accurately determine the sequencing of activity executions, which could identify a particular worker involved in the execution of the process (e.g., through a particular pattern of activity executions).

To execute the directly-follows query, i.e., the *Laplacian df-based* algorithm, solely the parameter ϵ needs to be specified. For the trace variant query, shown in Figure 3b, a user needs to select ϵ, as well as a maximum sequence length, and a pruning parameter. To avoid exploring a possibly infinite amount of trace variants, the technique only explores trace variants up to a certain maximum length. The pruning parameter is used to further limit the search space by avoiding the consideration of infrequent variants. For a detailed explanation of the queries and their privacy-preserving mechanisms, the reader is referred to [6].

Output. When an event log and algorithm have been selected, the event log will be uploaded and the application of the algorithm will be started as a batch



(a) PRETSA. (b) Differential private query

Fig. 3: Upload an event log and apply the selected algorithm

job. The user will be notified at a provided e-mail address when the execution has been completed. The user can retrieve the obtained process mining artifact (either a sanitized event log or a query result) through the token from the e-mail.

3 Maturity and Availability

ELPaaS, its source code, a tutorial, and all other information is available at github.com/samadeusfp/elpaas and accessible without registration. The source code from ELPaaS is available under the MIT licence. The project was implemented using Python. We used the Django framework[4] as a basis. A screencast of our tool is available under: https://youtu.be/XLq124VpZ6Q

Our application is an online service that has been developed to provide privacy-preserving process mining techniques to other researchers. As such, the web application in its current form is not optimized for industry-scale usage. Nevertheless, the application is suitable to handle real-world event logs, such as those of the BPI challenges[5]. Given its computational complexity, the event log

[4] https://www.djangoproject.com
[5] https://data.4tu.nl/repository/collection:event_logs_real

sanitization algorithm requires up to several hours to complete, whereas privacy-preserving queries can be executed in a matter of seconds. Our web deployment is available through a secure connection and does not store the original event log permanently. With these features we provide security to the users of our application. Alternatively, it is possible for users to host the application themselves. We provide our application as an isolated container, so it can be run on a Docker[6] instance. Given the ongoing developments occurring in the area of privacy-preserving process mining, the ELPaaS architecture is designed for simple integration of novel techniques in the future.

4 Conclusion

In this paper, we introduced ELPaaS, a web application that supports privacy-preserving process mining. The application bundles approaches for both event log sanitization, which transforms an event log into one that meets privacy criteria, as well as privatized process mining techniques, which ensure that obtained process mining results adhere to privacy requirements. As such, the application enables users to choose a technique that best suits their purposes.

As research into privacy-preserving process mining is ongoing, we intend to continuously expand the techniques offered by the application in the future.

Acknowledgments

This work was in part supported by Bane NOR and the Alexander von Humboldt Foundation.

References

1. Van der Aalst, W.M.: Process Mining - Data Science in Action. Springer (2016)
2. Asikis, T., Pournaras, E.: Optimization of privacy-utility trade-offs under informational self-determination. FGCS, in press (2018)
3. Fahrenkrog-Petersen, S.A.: Providing privacy guarantees in process mining. Proceedings of the CAiSE Doctoral Consortium. pp. 23–30 (2019)
4. Fahrenkrog-Petersen, S.A., van der Aa, H., Weidlich, M.: Pretsa: Event log sanitization for privacy-aware process discovery. IEEE ICPM, in press (2019)
5. Mannhardt, F., Petersen, S.A., Oliveira, M.F.: Privacy challenges for process mining in human-centered industrial environments. In: 14th Int'l Conf. on Intelligent Environments. pp. 64–71. IEEE (2018)
6. Mannhardt, F., Koschmider, A., Baracaldo, N., Weidlich, M., Racz, P., Michael, J.: Privacy-preserving Process Mining: Differential Privacy for Event Logs. BISE, accepted (2019)
7. Voss, W.G.: European union data privacy law reform: General data protection regulation, privacy shield, and the right to delisting. Business Lawyer **72**(1), 221–233 (2017)

[6] https://www.docker.com