PC4PM: A Tool for Privacy/Confidentiality Preservation in Process Mining

Majid Rafiei, Alexander Schnitzler and Wil M.P. van der Aalst

Chair of Process and Data Science, RWTH Aachen University, Aachen, Germany

Abstract

Process mining enables business owners to discover and analyze their actual processes using event data that are widely available in information systems. Event data contain detailed information which is incredibly valuable for providing insights. However, such detailed data often include highly confidential and private information. Thus, concerns of privacy and confidentiality in process mining are becoming increasingly relevant and new techniques are being introduced. To make the techniques easily accessible, new tools need to be developed to integrate the introduced techniques and direct users to appropriate solutions based on their needs. In this paper, we present a Python-based infrastructure implementing and integrating state-of-the-art privacy/confidentiality preservation techniques in process mining. Our tool provides an easy-to-use web-based user interface for privacy-preserving data publishing, risk analysis, and data utility analysis. The tool also provides a set of anonymization operations that can be utilized to support privacy/confidentiality preservation. The tool manages both standard XES event logs and non-standard event data. We also store and manage privacy metadata to track the changes made by privacy/confidentiality preservation techniques.

Keywords

process mining, privacy preservation, confidentiality, event data

1. Introduction

Process mining techniques employ event logs to provide insights into actual processes [1]. Event logs contain detailed information about operational processes and can be extracted from various types of information systems, e.g., EPR systems. Events are considered as the smallest units of process execution which are distinguished by their attributes. The main attributes of events are as follows: *activity*, *case id*, *timestamp*, and *resource*. For instance, a heart surgery (*activity*) performed by Dr. John (*resource*) for a patient with id=10 (*case id*) at timestamp 2021.06.10-10:00:00 is an event recorded by an information system in a hospital. The attributes that directly or indirectly refer to individuals raise privacy concerns. For example, in the healthcare context, the *case id* attribute may refer to the patients whose data are processed, and the *resource* attribute may refer to the employees who perform activities for the patients, e.g., nurses. Furthermore, other attributes can also be considered as confidential information, e.g.,

Proceedings of the Demonstration & Resources Track, Best BPM Dissertation Award, and Doctoral Consortium at BPM 2021 co-located with the 19th International Conference on Business Process Management, BPM 2021, Rome, Italy, September 6-10, 2021

✉ majid.rafiei@pads.rwth-aachen.de (M. Rafiei); Alexander.Schnitzler@outlook.com (A. Schnitzler); wvdaalst@pads.rwth-aachen.de (W. M.P. v. d. Aalst)

0000-0001-7161-6927 (M. Rafiei); 0000-0002-8223-6097 (A. Schnitzler); 0000-0002-0955-6940 (W. M.P. v. d. Aalst)

CEUR Workshop Proceedings (CEUR-WS.org)

the *activity* attribute may contain a confidential activity name that must not be exposed. Respect for privacy when analyzing personal data is also dictated by regulations, e.g., the European General Data Protection Regulation (GDPR)[1]. Such legitimate and ethical requirements have recently resulted in more attention to privacy and confidentiality issues in process mining [2, 3, 4, 5]. Some tools have also been introduced to provide specific privacy/confidentiality requirements [6, 7, 8].

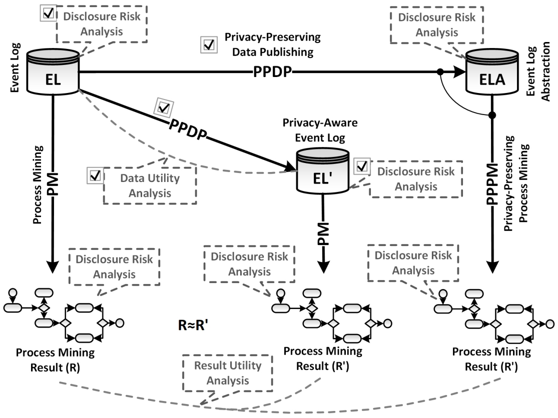


Figure 1: The general overview of privacy-related activities in process mining.

Figure 1 shows the general overview of privacy-related activities in process mining including Privacy-Preserving Data Publishing (PPDP), Privacy-Preserving Process Mining (PPPM), and Privacy Analysis (PrAn). PPDP tries to obscure the identity and/or sensitive data of individuals to preserve their privacy. PPDP techniques often apply one or more *anonymization operations*, e.g., *suppression*, *generalization*, etc., to provide the desired privacy requirements. PPPM intends to expand existing process mining algorithms to cope with intermediate results, so-called *abstractions* [9], generated by some PPDP techniques. Note that PPPM algorithms are closely linked with the corresponding PPDP approaches, and PPPM may refer to the entire privatization process, starting with an event log and finishing with process mining findings. PrAn, indicated with dashed lines in Figure 1, includes two types of activities: *risk analysis* and *utility analysis*. Both PrAn activities could be done for data and results. In this paper, we introduce a tool, named PC4PM, mainly focusing on the activities indicated by the check-boxes in Figure 1. PC4PM is the successor of the privacy tool introduced in [7], and it offers new privacy preservation techniques, privacy analysis, a set of anonymization operations, and user guidance that directs users to the right techniques based on their requirements. In the rest of the paper, we demonstrate the functionality and characteristics of PC4PM. We also describe the maturity and availability of the tool.

2. Functionality and Characteristics

PC4PM is implemented in Python using Django framework[2]. Figure 2 shows a high-level view of the architecture. PC4PM includes eight main Django applications and each application provides at least one main privacy-related activity implemented as a Python package. The Django templates, accessible from any web browser, provide a web interface for the applications. Implementing each technique as an independent Django application enables users to simultaneously run different techniques on event logs. Such architecture makes the process of maintenance and

[1] http://data.europa.eu/eli/reg/2016/679/oj

[2] https://www.djangoproject.com/

integration simple. To integrate new techniques, one can create a Python package and integrate it as an independent application. Moreover, Python packages can independently be imported into other Python-based tools.

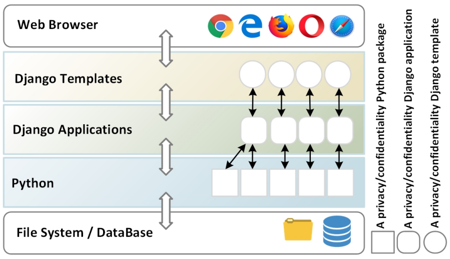
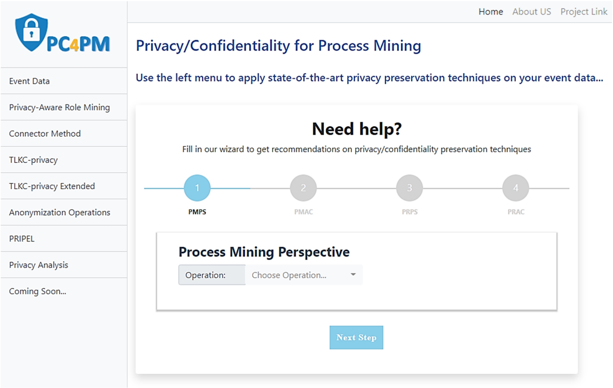

Web Browser
Django Templates
Django Applications
Python
File System / DataBase
A privacy/confidentiality Python package
A privacy/confidentiality Django application
A privacy/confidentiality Django template

Figure 2: The architecture of PC4PM.

Figure 3 shows the home page of PC4PM. The left menu shows the main Django applications including *event data management*, *privacy-aware role mining*, *connector method*, *TLKC-privacy*, *TLKC-privacy extended*, *anonymization operations*, *PRIPEL*, and *privacy analysis*. The *event data management* application manages both standard XES event logs and non-standard event data, called *event log abstraction* [9]. The *privacy-aware role mining* application implements the decomposition method, proposed in [10], to discover roles from event logs while preserving privacy. This method perturbs the frequency of activities in an event log to eliminate frequency-based attacks. The *connector method* is an encryption-based method for securely discovering directly follows graphs from event logs. This method breaks down traces into a collection of directly follows relations to prevent linkage attacks.



Privacy/Confidentiality for Process Mining

Use the left menu to apply state-of-the-art privacy preservation techniques on your event data...

Need help?

Fill in our wizard to get recommendations on privacy/confidentiality preservation techniques

Process Mining Perspective

Operation: Choose Operation...

Figure 3: The home page of PC4PM.

The *TLKC-privacy* application implements the TLKC-privacy model providing group-based privacy guarantees for process discovery and performance analysis. The *TLKC-privacy extended* application extends the TLKC-privacy model and considers all the main perspectives of process mining [3]. The *anonymization operation* application, implements all the main anonymization operations proposed in [9] including *suppression*, *addition*, *substitution*, *condensation*, *swapping*, *generalization*, and *cryptography*. The *PRIPEL* application presents the PRIPEL method [2] which applies the notion of *differential privacy* to provide privacy guarantees for event logs. The *privacy analysis* application includes three components for analyzing *disclosure risks*, *data utility* [11], and *FCB-anonymity* [12].

PC4PM supports users with a four-step user guide to help them choose the right technique(s) based on their needs. The user guidance works based on a four-dimension *signature* assigned to each technique. The signature reflects the following aspects: *process mining perspective* (PMPS), *process mining activity* (PMAC), *privacy perspective* (PRPS), and *privacy activity* (PRAC). PMPS indicates the process mining perspective that a privacy technique focuses on, e.g., control-flow. PMAC shows the process mining activity, e.g, process discovery, for which the utility of event data is preserved. PRPS shows the privacy perspective of a privacy technique, i.e., resource or case. PRAC indicates the main privacy-related activity of a privacy preservation technique, i.e., PPDP, PPPM, or PrAn. Moreover, PC4PM helps users with *help tooltips* provided for the parameters used by techniques. PC4PM also inherits all the characteristics of its predecessor [7]. Some of those are as follows: (1) Each Django application provides the results in an independent output section, (2) It enables a cycle of privacy/confidentiality preservation techniques such that the results from one technique can be added to the event data repository and used as an input for other techniques, (3) The *privacy metadata* [9] which specify the order and type of the main anonymization operations are added to anonymized event logs.

3. Availability and Maturity

The source code, a screencast, a user manual, and all other resources are available in our GitHub repository: https://github.com/m4jidRafiei/PC4PM. Each privacy/confidentiality Python package is linked to a separate GitHub project. The main GitHub project contains links to all those projects. In the corresponding GitHub project of each privacy/confidentiality Python package, one can find the name of the Python package, the link to the main paper, and a sample source code that shows the usage. In terms of performance and time complexity, each privacy preservation technique which is linked to a Django application behaves differently w.r.t. the size of the input event log. Based on our experiments, the applications are able to handle real-world event logs, e.g., BPI challenge datasets: https://data.4tu.nl/. Moreover, all the complicated and time-consuming functions, developed in the Python packages, have a parameter to be run using multi-processing which is enabled by default. In this case, the input event log is divided into smaller pieces w.r.t. the cores of the processor hosting PC4PM. PC4PM is provided as a Docker container that can simply be hosted by users: https://hub.docker.com/r/m4jid/pc4pm. The Docker usage is also explained in the GitHub repository.

4. Conclusion

In this paper, we introduced a tool for publishing event data w.r.t. privacy concerns. Our web-based tool is mainly focused on privacy-/confidentiality-preserving data publishing and privacy analysis considering both data utility and disclosure risk analyses. PC4PM can be considered as a sanitizer that provides sanitized event logs that can be used by any process mining tool. The architecture has been designed in such a way that other privacy preservation techniques can easily be integrated, e.g., we integrated *PRIPEL* as an external library. The goal of PC4PM is to provide a comprehensive set of techniques that can cover all the aspects of privacy-related activities for different perspectives of process mining. We invite other researchers to integrate their solutions as independent applications into the provided framework.

Acknowledgments

Funded under the Excellence Strategy of the Federal Government and the Länder. We also thank the Alexander von Humboldt (AvH) Stiftung for supporting our research.

References

[1] W. M. P. van der Aalst, Process Mining - Data Science in Action, Second Edition, Springer, 2016. doi:10.1007/978-3-662-49851-4.

[2] S. A. Fahrenkrog-Petersen, H. van der Aa, M. Weidlich, PRIPEL: privacy-preserving event log publishing including contextual information, in: Business Process Management - 18th International Conference, BPM, volume 12168 of *Lecture Notes in Computer Science*, 2020, pp. 111–128.

[3] M. Rafiei, W. M. P. van der Aalst, Group-based privacy preservation techniques for process mining, Data & Knowledge Engineering 134 (2021) 101908. doi:https://doi.org/10.1016/j.datak.2021.101908.

[4] G. Elkoumy, S. A. Fahrenkrog-Petersen, M. F. Sani, A. Koschmider, F. Mannhardt, S. N. von Voigt, M. Rafiei, L. von Waldthausen, Privacy and confidentiality in process mining - threats and research challenges, CoRR abs/2106.00388 (2021). URL: https://arxiv.org/abs/2106.00388. arXiv:2106.00388.

[5] G. Elkoumy, S. A. Fahrenkrog-Petersen, M. Dumas, P. Laud, A. Pankova, M. Weidlich, Secure multi-party computation for inter-organizational process mining, in: Enterprise, Business-Process and Information Systems Modeling - 21st International Conference, BPMDS, Springer, 2020.

[6] G. Elkoumy, S. A. Fahrenkrog-Petersen, M. Dumas, P. Laud, A. Pankova, M. Weidlich, Shareprom: A tool for privacy-preserving inter-organizational process mining, in: Proceedings of the Best Dissertation Award, Doctoral Consortium, and Demonstration & Resources Track at BPM 2020, volume 2673 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020, pp. 72–76.

[7] M. Rafiei, W. M. P. van der Aalst, Practical aspect of privacy-preserving data publishing in process mining, in: Proceedings of the Best Dissertation Award, Doctoral Consortium, and Demonstration & Resources Track at BPM 2020 co-located with the 18th International Conference on Business Process Management (BPM 2020), CEUR-WS.org, 2020.

[8] M. Bauer, S. A. Fahrenkrog-Petersen, A. Koschmider, F. Mannhardt, H. van der Aa, M. Weidlich, Elpaas: Event log privacy as a service, in: Proceedings of the Dissertation Award, Doctoral Consortium, and Demonstration Track at BPM 2019, 2019.

[9] M. Rafiei, W. M. P. van der Aalst, Privacy-preserving data publishing in process mining, in: Business Process Management Forum - BPM Forum 2020, Seville, Spain, September 13-18, 2020, Proceedings, volume 392 of *Lecture Notes in Business Information Processing*, Springer, 2020, pp. 122–138. doi:10.1007/978-3-030-58638-6_8.

[10] M. Rafiei, W. M. P. van der Aalst, Mining roles from event logs while preserving privacy, in: Business Process Management Workshops - BPM 2019 International Workshops, Vienna, Austria, 2019, pp. 676–689.

[11] M. Rafiei, W. M. P. van der Aalst, Towards quantifying privacy in process mining, in: International Conference on Process Mining - ICPM 2020 International Workshops, Padua, Italy, October 4-9, 2020, 2020.

[12] M. Rafiei, W. M. P. van der Aalst, Privacy-preserving continuous event data publishing, CoRR abs/2105.11991 (2021). URL: https://arxiv.org/abs/2105.11991. arXiv:2105.11991.